Prefacio

El nivel de abstracción conceptual que distingue a un magíster/doctor de un ingeniero se verifica principalmente en que el ingeniero tiene la habilidad de seleccionar y utilizar el modelo matemático apropiado para un problema dado (diseño), mientras el magíster/doctor es capaz de elaborar el modelo matemático adecuado para un problema novedoso (investigación). Por eso el magíster/doctor necesita una fundamentación matemática mucho más sólida, pues él debe abstraer de una realidad compleja los aspectos más relevantes asociados con un objetivo particular de estudio, y formularlos en el contexto formal y riguroso de las matemáticas para obtener conclusiones que permitan comprender o controlar el sistema bajo estudio. En redes de comunicaciones, estos modelos matemáticos deben relacionar la capacidad de la red y la demanda de los usuarios con las medidas de desempeño que la red ofrece, de manera que resulten útiles en el diseño, el control y el análisis de las redes. En particular, los modelos matemáticos asociados con las redes de comunicaciones deben representar, fundamentalmente, la incertidumbre de los ingenieros respecto a su objeto de estudio. En efecto, si bien podría pensarse que los componentes de la red (equipos y protocolos) tienen un comportamiento determinístico, ellos existen para satisfacer las demandas de los usuarios, las cuales se presentan en cantidades aleatorias y en instantes de tiempo aleatorios[1].

Considere, por ejemplo, una sesión *ftp* (file transfer protocol). La transferencia de un archivo de cierto tamaño (en bytes) toma cierto tiempo (en segundos), de donde se puede calcular la velocidad de transmisión o tasa de transferencia (en bits por segundo, bps). *ftp* ofrece mediciones de estas tres cantidades al presentar resultados como los siguientes (datos consecutivos reales):

Tamaño del archivo (bytes)	Tiempo de trans-ferencia (segundos)	Tasa de trans-ferencia (bps)
1’352.663	5.2	2’081.020
768.023	3.3	1’861.874
3’114.271	11.9	2’093.627
569344	1.8	2’530.418
8’682.209	34.1	2’036.882
2’118.124	8.0	2’118.124
935.936	3.2	2’339.840

Se puede observar cómo no sólo el tamaño de los archivos y el tiempo de transferencia varían de manera imprevisible, sino que el caudal mismo también cambia entre un archivo y otro a pesar de tratarse de la misma conexión usada de manera consecutiva. Por supuesto, esto no debe sorprendernos: diferentes conexiones se establecen con el servidor *ftp* en instantes de tiempo que no

[1] Algunos protocolos (de acceso múltiple, de control de congestión, de seguridad, etc.) hacen uso de números pseudo-aleatorios, esto es, tampoco son determinísticos.

podemos predeterminar; el número de archivos que se transferirán en cada conexión tampoco es predecible con anterioridad; no se puede conocer de antemano la duración de cada archivo; *ftp* ofrece un servicio confiable, de manera que debe detectar y corregir errores, los cuales se presentan de manera impredecible en los medios de transmisión debido al ruido y a la imperfección de los enlaces; etc. Por supuesto, el magíster/doctor en redes de comunicaciones debe ser capaz de incluir dicha incertidumbre dentro de sus modelos matemáticos, para lo cual se usa el modelado probabilístico de los tiempos entre llegadas y las intensidades de las demandas de los usuarios de una red, así como las pérdidas por errores de transmisión, entre otros eventos inciertos.

En estas notas estudiaremos los fundamentos de la teoría de la probabilidad, las variables aleatorias y los procesos estocásticos, los cuales proporcionan herramientas útiles en el análisis de redes de comunicaciones. El enfoque hacia las redes de comunicaciones se debe a que estas notas constituyen los apuntes de clase de los cursos que he dictado en algunos programas de maestría y doctorado de la Universidad Distrital, que suelen incluir muchos ingenieros electrónicos y de sistemas que inician sus estudios de posgrado con énfasis en teleinformática y temas afines. Sin embargo, los conceptos fundamentales que se estudian constituyen abstracciones matemáticas adecuadas para una gran cantidad de realidades en todas las áreas de las ciencias y las ingenierías. Por eso, éste es un libro de matemática y no de redes de comunicaciones o de ingeniería de teletráfico. Pretende ser una introducción general pero formal a la teoría de las probabilidades, variables aleatorias y procesos estocásticos para estudiantes que inician su formación posgradual, ya sea en teleinformática o en cualquier otra disciplina de la ingeniería. La única particularidad es que, en los ejemplos de aplicación, los dados, las monedas y las bolas de billar han sido cambiados por bits, paquetes de datos y enrutadores. De hecho, el principal objetivo de estas notas no es aprender a usar y desarrollar modelos probabilísticos en redes de comunicaciones; el principal objetivo es migrar desde el enfoque pragmático con que se estudian la probabilidad, las variables aleatorias y los procesos estocásticos en el pregrado hacia la formalidad y la rigurosidad que se requiere en un programa de posgrado en ingeniería. El uso de un tipo específico de ejemplos no es tan significativo pues, después de todo, un estudiante de maestría/doctorado debe tener la capacidad de abstracción necesaria para reconocer las posibles interpretaciones de un ejemplo particular: ¿Cuánto se parece lanzar una moneda para ver si cae cara o sello a recibir un bit para ver si es uno o cero? ¿Cuán diferente es el problema de un juez que tiene que determinar la culpabilidad o inocencia de un acusado a partir de las evidencias, y el problema de un módem que debe determinar si se transmitió un uno o un cero a partir de la señal recibida? Es ese nivel de abstracción y de conceptualización el que se espera de un magister/doctor en ingeniería y, en ese sentido, la escogencia particular de los ejemplos de aplicación no debe ser relevante.

Para facilitar el estudio de los temas del libro, se han enumerado los conceptos, ejemplos y ejercicios de manera consecutiva. En efecto, he notado que, ante el amplio contenido de los cursos que podrían basarse en este libro, resulta conveniente considerar pequeñas unidades conceptuales que puedan irse interconectando, de una en una, hasta formar la intrincada malla conceptual de las probabilidades, las variables aleatorias y los procesos estocásticos. En mi experiencia docente, el uso de conceptos breves y precisos ha permitido a los estudiantes ir construyendo la estructura general del conocimiento gracias a que les facilita recordar conceptos precisos de manera rápida en los momentos en que se

necesitan. Espero que este mismo efecto se logre al subdividir de la misma manera el contenido de estas notas. Por supuesto, la construcción del conocimiento siempre será incompleta si no se hacen muchos ejercicios de aplicación de cada uno de los conceptos enumerados. Aunque iremos añadiendo más ejemplos y ejercicios a estas notas, se recomienda al estudiante complementarlos con muchos ejercicios adicionales de la literatura sugerida.

Por último, quisiera agradecer sinceramente a mis estudiantes de los últimos 26 años, a quienes considero coautores de estas notas.

Marco Alzate, Bogotá, junio de 2016

I. Introducción

1. Modelado Matemático en Ingeniería

*El modelado matemático consiste en **abstraer** de una realidad compleja los aspectos más relevantes asociados con un **objetivo particular** de estudio, y formularlos en el contexto formal y riguroso de las matemáticas. Para ello, se asocia la realidad bajo estudio con un **sistema**, se identifican los **componentes** que intervienen en dicho sistema, se determinan las **variables descriptivas** de esos componentes, se identifican las **reglas de interacción** entre los componentes y se determina los **parámetros descriptivos** de esas interacciones. Ese sistema así formulado constituye un **modelo matemático** de la realidad compleja que se quiere estudiar. Al operar con dicho modelo matemático, se obtendrán conclusiones que podrían remitirse directamente a la realidad bajo estudio. La utilidad de estas conclusiones para comprender o controlar dicha realidad determinará la **validez** del proceso de modelado matemático. Para evaluar el modelo matemático se pueden usar **técnicas analíticas, experimentales o de simulación**. Estas técnicas no son excluyentes sino que, al contrario, se complementan entre ellas.*

La **ingeniería** es la aplicación creativa de **principios científicos** para diseñar, construir y operar estructuras, máquinas o procesos, de manera que se pueda conocer su comportamiento bajo condiciones específicas de operación, todo con respecto a una función precisa y con criterios estrictos de desempeño, economía y seguridad[2].

Por principios científicos nos referimos al proceso de generación de conocimiento verificable mostrado en la Figura 1, según el cual observamos el mundo natural (realizamos **experimentos**) de manera que, a partir del **análisis** de las observaciones, obtengamos alguna generalización del comportamiento de la naturaleza. Por supuesto, dicha observación debe obedecer a una pregunta significativa que queremos responder empíricamente, de manera que el análisis de los resultados obedezca a un proceso de razonamiento lógico y coherente. Después de haber acumulado suficiente evidencia empírica, se usa un proceso de **inducción** para construir una **teoría** sobre dicho comportamiento, la cual debe ser **verificable**. Esto es, a partir de la teoría podremos **deducir** algunas **hipótesis** que nos permitirían predecir el comportamiento de la naturaleza ante condiciones particulares. Si estas predicciones resultan satisfechas en experimentos de verificación adecuados, podemos ir ganando confianza en la teoría propuesta. Por supuesto, estos experimentos de verificación deben conducirse con suficiente rigor como para que cualquier persona pueda **replicarlos** independientemente y obtener conclusiones semejantes.

[2] En particular, la ingeniería electrónica aplica los principios físicos del electromagnetismo y la mecánica cuántica para capturar, almacenar, transmitir y procesar información. Las diferentes áreas de actuación de la ingeniería electrónica (dispositivos de estado sólido, circuitos electrónicos, telecomunicaciones, sistemas de radio, sistemas de control, automática, procesamiento de señales, ingeniería de sistemas, ingeniería de computadores, instrumentación, etc.), algunas de las cuales se han constituido en disciplinas ingenieriles independientes, están todas ellas relacionadas íntimamente con las redes de comunicaciones.

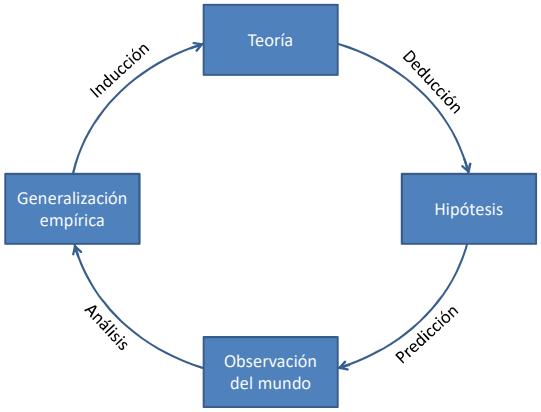


Figura 1. Método científico

Nótese que con este procedimiento se hace imposible demostrar la verdad de una teoría científica: **La ciencia no tiene verdades**, sólo teorías. Si construimos una teoría con una abrumadora evidencia empírica, bastará un solo resultado en contra para desbaratar la teoría. Las teorías científicas son sólo ideas que están evolucionando permanentemente.

En todos los procesos del método científico, desde el planteamiento de la pregunta a resolver y el diseño de experimentos hasta los procesos lógicos de análisis, deducción e inducción, se usa un **modelo de la realidad**. En el caso de los principios científicos que usa la ingeniería, estos modelos son, invariantemente, **modelos matemáticos**. Sidney Harris ha capturado apropiadamente muchas de las características del método científico y de su modelamiento matemático, como muestra la Figura 2.

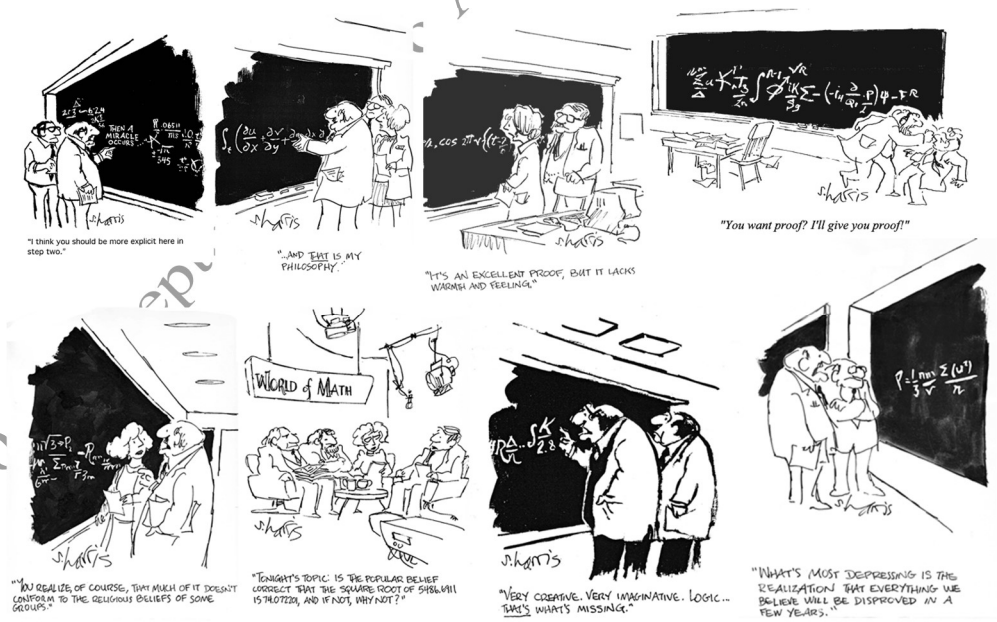


Figura 2. Sobre el método científico (© Sidney Harris, http://www.sciencecartoonsplus.com)

Así pues, mientras la ciencia usa el método científico para analizar las observaciones hechas de la naturaleza, proponiendo teorías para comprender y explicar su comportamiento, la ingeniería usa sus resultados para predecir dicho comportamiento de manera que se pueda hacer diseños que permitan

actuar sobre ella misma para facilitar y mejorar las condiciones de vida de la humanidad. Esta consideración suele conducir a comparaciones como las de la Figura **3** aunque, a medida que la ciencia y la ingeniería se enfocan más hacia el estudio de los sistemas complejos (sistemas que se auto-organizan con base en procesos locales de percepción, aprendizaje, evolución y adaptación), se empieza a difuminar esa diferencia entre ciencia e ingeniería, pues el proceso de comprensión del mundo ya no se distingue del proceso de actuación sobre el mundo, como veremos en la definición 2.

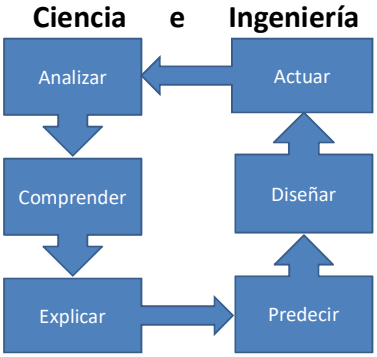


Figura 3. Una comparación simplista pero reveladora entre Ciencia e Ingeniería

Claro, el ingeniero desarrolla sus diseños sobre un modelo matemático antes de implementarlo para actuar en la realidad. **Si no se usa un modelo matemático, no se hace ingeniería**.

Como se mencionó en el resumen de este apartado, el modelado consiste en **abstraer** de una realidad compleja los aspectos más relevantes asociados con un **objetivo particular** de estudio, y formularlos en el contexto formal y riguroso de las matemáticas. Al operar con el modelo así formulado, se obtendrán conclusiones que deberían poderse aplicar a la realidad bajo estudio. La utilidad de estas conclusiones para comprender o controlar el sistema bajo estudio determinará la **validez** del proceso de modelado matemático.

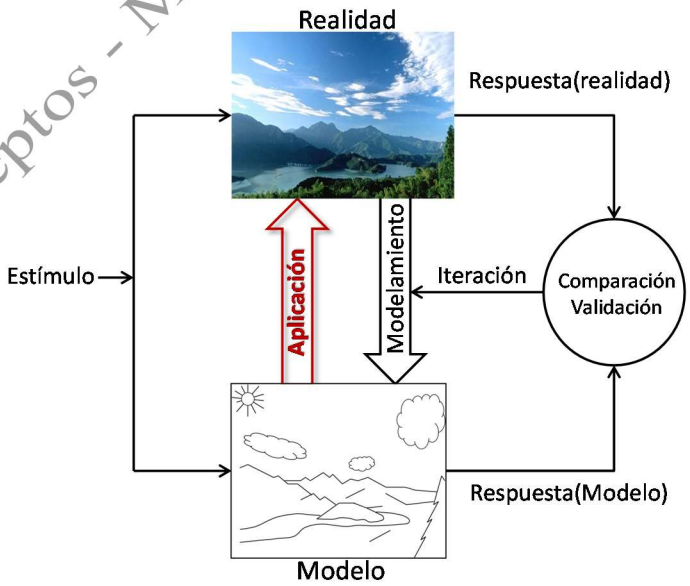


Figura 4. Proceso de modelado matemático

Así pues, el modelado matemático comprende algunas etapas

- Identificación de **componentes**
- Determinación de **variables descriptivas** de esos componentes
- Identificación de **reglas de interacción**
- Determinación de **parámetros** de esas interacciones

Por ejemplo, para describir la traslación de la luna alrededor de la tierra podemos proceder así:

- Identificación de componentes: La luna y la tierra (¡suponemos que ningún otro objeto del universo las perturba!).
- Determinación de variables descriptivas de esos componentes: Posición y velocidad de la tierra y de la luna (¡suponemos que son puntos infinitesimales o, al menos, esferas homogéneas!)
- Identificación de reglas de interacción: Dos objetos se atraen con una fuerza directamente proporcional al producto de sus masas e inversamente proporcional al cuadrado de la distancia entre ellos (ley universal de la gravedad –Newton–).
- Determinación de parámetros descriptivos de esas interacciones: La constante de proporcionalidad ($G\approx 6.674\times10^{-11}$ $m^3kg^{-1}s^{-2}$), la masa de la tierra ($M_T\approx 5.9724\times10^{24}$ kg), la masa de la luna ($M_L\approx 7.346\times10^{22}$ kg), la distancia entre la tierra y la luna ($d\approx 3.84403\times10^5$ km en promedio, 3.63026×10^5 km en el perigeo, 4.05610×10^5 km en el apogeo) y la velocidad de la luna ($|v|=1.022\times10^3$ ms^{-1} en promedio, 1.082×10^3 ms^{-1} en el perigeo, 9.7×10^2 ms^{-1} en el apogeo).

La siguiente figura muestra otro aspecto de la realidad: una pila, un interruptor, una resistencia y un condensador en serie; también se muestra un modelo matemático asociado con esa realidad. Por supuesto, hay muchos aspectos de la realidad que el modelo matemático no captura: La inductancia presente en la resistencia, los efectos de las soldaduras potencialmente defectuosas, la polaridad del condensador electrolítico, los rebotes mecánicos del interruptor, etc. Pero, como se mencionó hace un momento, se trata de construir una representación conceptual de la realidad abstrayendo los aspectos más relevantes asociados con un **objetivo particular**. Si deseamos un modelo para estudiar la carga del condensador, la abstracción alcanzada con el modelo matemático de la Figura 5 es muy apropiada, aunque el mismo modelo no podrá ser útil para un objetivo diferente, como estudiar los efectos térmicos sobre el condensador, por ejemplo.

Otro aspecto fundamental del modelamiento matemático es la generalidad de los modelos. Como muestra la Figura 6, la aplicación de las leyes de Ohm y Kirchhoff al circuito y la aplicación de las leyes de Newton al automóvil (nuevamente, despreciando muchos aspectos de la realidad) conducen a una misma ecuación diferencial lineal ordinaria de primer orden que puede representar un sistema genérico abstracto.

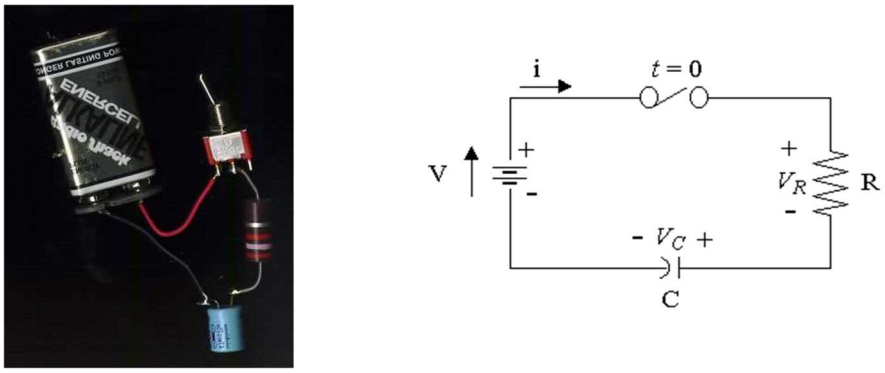


Figura 5. Realidad y modelo matemático

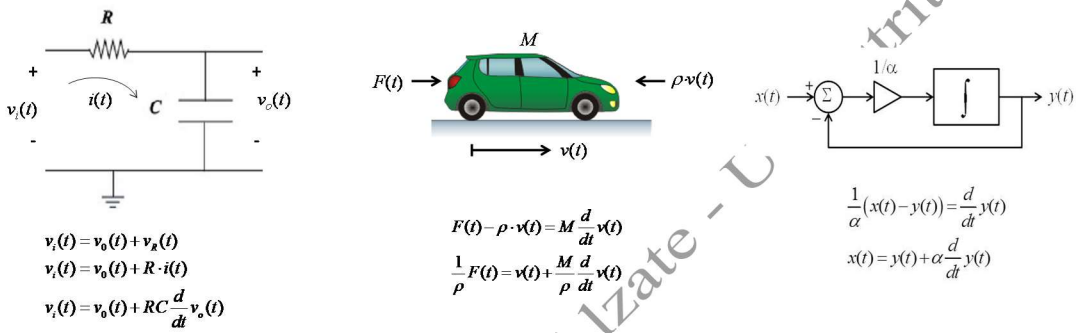


Figura 6. Un modelo matemático es más útil entre más general sea

Lo importante es tener en cuenta que todo modelo matemático en ingeniería es una abstracción del problema de la realidad que queremos modelar, lo cual implica una **simplificación** que limita la validez de los resultados del modelo a las suposiciones que se hayan hecho. Por eso, siempre un modelo matemático particular está asociado necesariamente con un **objetivo** de estudio particular, como muestra la Figura 7. Por ejemplo, mientras mi hijo de cuatro años modelaba un computador como una pantalla, un CD-drive y un joystick, un ingeniero de hardware ve en él unidades de procesamiento, buses, memoria y dispositivos de entrada/salida y un ingeniero de software ve una máquina física, un sistema operativo y un conjunto de aplicaciones. Cada uno de ellos encontrará variables y reglas de interacción para completar su modelo, el cual será válido únicamente para el propósito con que se construyó. Para nosotros, interesados en redes de comunicaciones, un computador es una fuente de tráfico y sus variables descriptivas estarán asociadas con las estadísticas de la generación de paquetes: cuántos paquetes genera en una hora, cuál es la distribución de los tiempos entre llegada, cómo están correlacionados, cuál es el tamaño de los paquetes, etc.

Como se mencionó en el prefacio, en este libro veremos modelos matemáticos que intentan representar redes de comunicaciones. Claro, cada modelo estará asociado con un objetivo particular, lo cual limitará fundamentalmente su rango de aplicabilidad, pero dará luces sobre la dinámica de los fenómenos que ocurren en una red de comunicaciones. Es muy importante que tengamos en cuenta que el modelo no es la realidad y, por lo mismo, las predicciones del modelo no son resultados absolutos del comportamiento que podemos esperar del sistema bajo estudio.

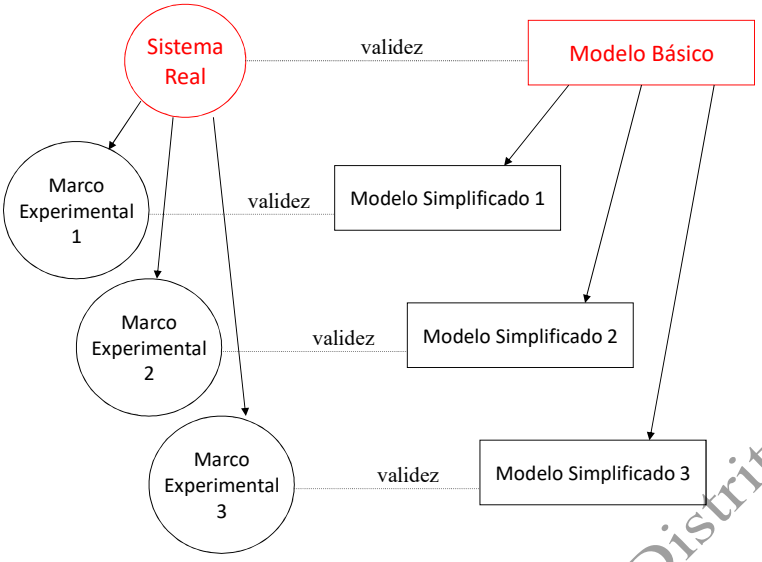


Figura 7. Cualquier modelo matemático es una simplificación ajustada a un marco experimental particular

Lo anterior nos conduce a un punto final igualmente importante: La **evaluación de un modelo matemático**. Por alguna razón que desconozco, siempre que se habla de un modelo matemático las personas (inclusive algunos de los ingenieros, los magísteres y los doctores que estudiarían este libro) piensan en sistemas de ecuaciones. Pero, insistamos, el modelo matemático es apenas una descripción de los componentes, las reglas de interacción y las variables descriptivas de dichos componentes y dichas reglas. Cuando estos elementos se especifican claramente, ya se ha formulado un modelo matemático. Otra cosa es la manera como queramos evaluar dicho modelo. Si la formulación lo permite, el **análisis matemático** (la solución de un sistema de ecuaciones, por ejemplo), es una alternativa deseable –**modelo analítico**–. Pero, generalmente, un modelo que sea capaz de admitir este tipo de soluciones habrá sufrido muchas simplificaciones que lo alejan de la realidad que se quiere modelar. Otra posibilidad es la realización de **experimentos en el laboratorio** con versiones parciales o reducidas del sistema bajo estudio –**modelo experimental**–. Por supuesto, si el ingeniero no tiene un modelo matemático que represente el sistema bajo estudio, le será imposible diseñar experimentos significativos para resolver las preguntas pertinentes del estudio. De todas maneras, los resultados de la experimentación en el laboratorio pueden ser mucho más significativos porque incluirían los efectos de otros componentes que ni siquiera se consideraron explícitamente en el modelo. Una alternativa intermedia es usar la **simulación por computador**, esto es, elaborar un programa que reproduzca las interacciones entre los componentes a medida que transcurre un tiempo virtual en el computador –**modelo de simulación**–. Otra alternativa poco común es experimentar con el mismo sistema que se estudia, pero aún en este caso el ingeniero trabaja con un modelo matemático, así este modelo lo tenga en su mente sólo de manera implícita. Por supuesto, entre más explícito y específico sea el modelo matemático que usa el ingeniero, más significativos serán los resultados de su estudio.

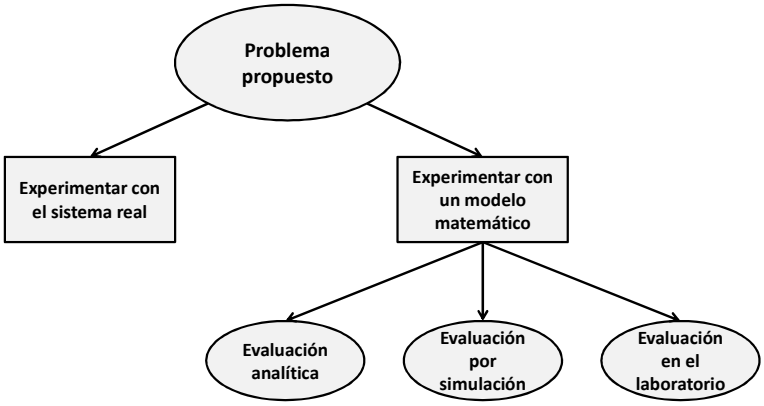


Figura 8. Técnicas de evaluación de un modelo matemático

Por ejemplo, considere el modelo del circuito eléctrico de la Figura 6 cuando RC = 5 ms, el voltaje de entrada es un escalón de 1 voltio y el condensador estaba inicialmente descargado. El modelo se puede evaluar analíticamente, pues la ecuación diferencial se reduce a $dv_0(t)/dt=200(1-v_0(t))$ para $t\geq 0$. Haciendo $u(t) = 1 - v_0(t)$ podemos separar las variables para obtener $\int \frac{du}{u} = -200\int dt$, de donde $u(t)=Ce^{-200t}$ o, lo que es lo mismo, $v_0(t) = 1 - Ce^{-200t}$. Como $v_0(0) = 0$, entonces $C=1$, por lo que la respuesta deseada es

$$v_0(t) = 1 - e^{-200t}, \quad t \geq 0$$

La segunda alternativa es la simulación: Si consideramos un incremento de tiempo Δt en vez del diferencial dt, y consideramos sólo instantes de tiempo múltiplos de Δt, podríamos aproximar la ecuación diferencial mediante

$$v_i(n\Delta t) \approx v_0(n\Delta t) + RC\frac{v_0(n\Delta t) - v_0((n-1)\Delta t)}{\Delta t}$$

Este es un sistema en tiempo discreto que podemos re-escribir así

$$v_i[n] = v_0[n](1+\beta) - \beta v_0[n-1], \quad \text{donde } \beta = \frac{RC}{\Delta t}$$

Por ejemplo, con Δt=0.2 ms, tendríamos $v_0[n]=(v_i[n]+25v_0[n-1])/26$ y la simulación se podría realizar, por ejemplo, con el siguiente programa en matlab©:

```
n  = -100:400;
vi = (n>=0);
vo = zeros(size(vi));
for m=-100:400
    k = m+101;
    if m<0
        vo(k) = 0;
    else
        vo(k) = (vi(k)+25*vo(k-1))/26;
    end
end
plot(n/5000,vo)
```

Listado 1. Simulación de un circuito RC en matlab®

Una tercera alternativa es implementar el circuito de la Figura 5 en el laboratorio y ver la señal del condensador en el osciloscopio. La Figura 9 muestra el resultado del modelo desde los tres tipos de evaluación (tres técnicas de evaluación distintas para un mismo modelo matemático).

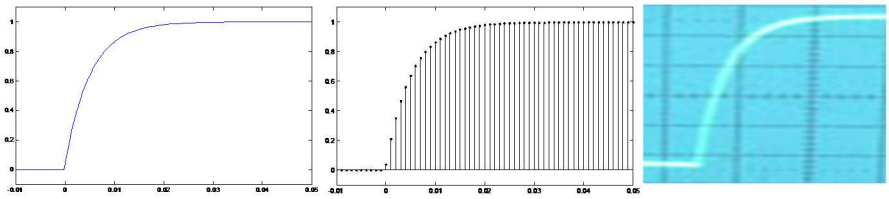


Figura 9. Soluciones analítica, por simulación y experimental de un modelo matemático

En el área de la ingeniería de teletráfico, generalmente se requiere demasiadas simplificaciones para lograr que un modelo matemático se mantenga analíticamente tratable. Por otro lado, resulta demasiado costoso experimentar con el sistema real. En consecuencia, en ingeniería de redes típicamente se usa el método de la simulación por computador para estudiar problemas de diseño y operación de sistemas, especialmente como apoyo a la toma de decisiones. Es importante notar que los dos métodos son complementarios, pues para validar un programa de simulación se suele verificar que, ante escenarios que tengan una solución analítica, ambos métodos de evaluación obtengan los mismos resultados. Esta condición incrementa la confianza de que los resultados del escenario propuesto sean correctos. La validación funciona también en el sentido contrario: Después de haber obtenido un resultado analítico, se puede usar la simulación para verificar los rangos de validez del resultado. En ambos casos, la mejor validación se obtiene al comparar los resultados analíticos o de simulación con resultados experimentales.

Por último, vale la pena aclarar que la simulación es tanto una herramienta como una ciencia. Existe una formalidad muy rigurosa y estricta (la estadística) para asegurar que los estudios de simulación sean significativos y válidos. Sin el cuidado necesario, el uso de la simulación puede conducir a grandes equivocaciones. La simulación de eventos discretos apropiada para la evaluación de modelos matemáticos en ingeniería de teletráfico se revisará con cuidado a partir de la definición 134.

2. Modelos Matemáticos de Sistemas Simples y de Sistemas Complejos

*En el modelo matemático de un **sistema simple** existen pocos componentes, pocas variables descriptivas y pocas reglas de interacción, de manera que el comportamiento global del sistema se puede explicar como el resultado directo de las reglas de interacción entre los componentes. En efecto, tanto los componentes como el sistema entero actúan en una misma escala de tiempo y espacio y resulta fácil encontrar relaciones de causalidad entre las propiedades particulares de los componentes y las propiedades globales del sistema. En el modelo matemático de un **sistema complejo** existen muchos componentes y muchas variables descriptivas cuyas relaciones son no lineales, de manera que, aunque las reglas de interacción sean muy simples, surgen propiedades globales del sistema que no se pueden atribuir directamente a sus componentes sino que son el resultado emergente de las múltiples interacciones no lineales entre ellos. En efecto, los componentes del sistema actúan en escalas de tiempo y espacio diferentes a aquellas en las que se observa el*

comportamiento global del sistema, lo que da lugar a fenómenos de emergencia y auto-organización (formación de patrones, múltiples lazos de realimentación, sincronización, cooperación/competencia, etc.). En ambos casos, los procedimientos de modelado matemático mencionados en el numeral anterior siguen siendo los mismos (componentes, interacciones y variables relacionadas con el objetivo de estudio), pero con sistemas simples se busca una descomposición reduccionista/mecanicista del sistema que permita diseñar el comportamiento deseado, y con sistemas complejos se busca guiar la auto-organización para obtener los comportamientos emergentes que permitan la adaptación autónoma del sistema a los cambios dinámicos del ambiente en que se desenvuelve, mediante aprendizaje o evolución.

Una forma de construir los modelos matemáticos del apartado anterior es considerando que las propiedades y los comportamientos globales de un sistema se pueden deducir o explicar a partir las propiedades y los comportamientos de las partes que lo componen. En este contexto se hace muy apropiado el principio de superposición propio de los sistemas lineales. En la misma línea se encuentra el principio del determinismo, según el cual la incertidumbre se limita a la presencia de errores en las mediciones, en cuyo caso la distribución Gaussiana resulta muy apropiada. Sin embargo, la linealidad y la gaussianidad son, en realidad, poco comunes en la naturaleza.

Históricamente, en ausencia de computadores electrónicos, hasta hace algunas décadas sólo se contaba con métodos analíticos de evaluación de modelos matemáticos, o métodos numéricos calculados manualmente. Los primeros, en general, casi siempre existen para sistemas lineales y gaussianos pero casi nunca existen para sistema no-lineales y no-gaussianos. Los segundos pueden requerir un número tan exagerado de cálculos que, si no se cuenta con computadores electrónicos, resulta inútil aplicarlos a sistemas con un gran número de componentes cuyas variables descriptivas se relacionan de manera no lineal y no gaussiana. Fue la introducción del computador la que permitió evaluar modelos matemáticos no-lineales y no-gaussianos con muchos componentes, y logró redescubrir fenómenos fascinantes que ya grandes pensadores habían intuido, como el caos, la fractalidad y la emergencia. Con semejante herramienta, la ciencia y la ingeniería pudieron abordar problemas mucho más complejos al permitirse incluir la no-linealidad y la no-gaussianidad en sus modelos matemáticos evaluados numéricamente o mediante simulación. El computador explica en gran medida el vertiginoso desarrollo científico y tecnológico de las últimas décadas.

Entre las herramientas que el computador ayudó a desarrollar para evaluar modelos matemáticos no-lineales y no-gaussianos están los métodos bio-inspirados (redes neuronales artificiales, algoritmos genéticos, inteligencia de enjambre, sistemas inmunes artificiales, sistemas difusos, etc.). Todos ellos han sido muy útiles en la solución de complejos problemas de optimización, regresión o clasificación, fundamentales en el diseño de nuevas soluciones a problemas complejos. La inclusión de estos métodos, entonces, ha permitido extender la aplicabilidad de los modelos matemáticos en ingeniería, especialmente al facilitar la inclusión de requerimientos de robustez (el sistema debe seguir operando satisfactoriamente en un amplio rango de condiciones) y su forma extendida, la adaptabilidad (el sistema debe poder ajustar automáticamente sus parámetros de acuerdo a los cambios en el ambiente). Los grandes avances que estos modelos han permitido alcanzar en materia de comunicaciones,

también han permitido incluir una característica formidable en los diseños de ingeniería: Los sistemas distribuidos, en los que la información local se distribuye para tomar decisiones colectivas en busca de un objetivo global. También estos modelos matemáticos de sistemas robustos, adaptables y distribuidos han contribuido al fascinante mundo tecnológico de hoy.

Así pues, a medida que crece la complejidad de los fenómenos que la ciencia estudia y de los problemas que la ingeniería aborda, crece también la versatilidad de los modelos matemáticos que ellas usan. Hoy, la experiencia ingenieril con sistemas bio-inspirados y la comprensión científica sobre la organización de muchos sistemas naturales, han permitido explorar también otros fenómenos comunes en la naturaleza y otras formas de plantear soluciones de ingeniería a mayores problemas de la humanidad: La auto-organización y la emergencia. Efectivamente, en física, biología, ciencias sociales y ciencias de la computación, entre muchas otras áreas, se ha venido comprendiendo que los comportamientos complejos observados a cierta escala surgen como fenómenos emergentes de procesos de auto-organización a escalas menores, casi siempre mediados por interacciones simples pero no-lineales. Esta comprensión indica que la ingeniería puede enfrentar nuevos problemas básicos de la humanidad, en los ámbitos ambiental, tecnológico, biológico, económico y político, si extiende los nuevos modelos matemáticos que capturan estos fenómenos para lograr diseñar, construir, operar y controlar aparatos, procesos y estructuras que resuelvan problemas de la humanidad, que es el propósito de la ingeniería. Pero estos aparatos, procesos y estructuras serán fundamentalmente diferentes en un sentido básico: Sus partes deberán percibir el mundo, aprender de él, decidir autónoma e inteligentemente comportamientos apropiados, comunicarse entre ellas, al menos en un ámbito local, y, de manera colectiva, permitir la emergencia de patrones funcionales, temporales o espaciales que resuelvan el problema que se trata. A estos nuevos modelos matemáticos en ingeniería se les ha llamado “modelos de inteligencia colectiva”, “modelos de auto-organización guiada”, “modelos de sistemas dinámicos cognitivos”, "modelos de control cooperativo", "modelos morfogenéticos", etc. Estos modelos matemáticos ya no buscan que el ingeniero diseñe la solución del problema sino que diseñe el sistema complejo adaptivo que sea capaz de emerger mediante auto-organización la solución del problema, pues el problema mismo cambia dinámicamente en un ambiente complejo. De hecho, el desarrollo de estos modelos matemáticos va difuminando la frontera entre ciencia e ingeniería, pues además de permitir el diseño del sistema de ingeniería, permite proponer explicaciones a los fenómenos emergentes de la naturaleza.

La Figura 10 muestra esquemáticamente el concepto más general de un sistema complejo. Las interacciones no lineales entre sus muchos componentes se presentan en múltiples escalas, de manera que de una escala a otra ocurren fenómenos de emergencia de estructuras y patrones que aparecen por auto-organización, todo como producto de una evolución en el tiempo caracterizada por una alta sensibilidad a las condiciones iniciales. Estas propiedades aparecen en los mercados de valores, las organizaciones políticas, la Internet, las economías nacionales y globales, el clima global, los sistemas ecológicos, etc. En todos estos sistemas existe una gran cantidad de microcomponentes que interactúan de manera tal que la información no se encuentra en los componentes sino en sus interacciones. Por eso, los modelos matemáticos de sistemas complejos involucran herramientas de modelado adicionales a las típicas de los sistemas simples, como se aprecia en la Figura 11. Áreas tan fascinantes como las dinámicas no-lineales y el caos, la realimentación, la evolución, la adaptación,

la interacción entre agentes autónomos mediante teoría de juegos, transiciones de fase, redes, etc. se unen a las técnicas de modelado de sistemas simples (ecuaciones diferenciales y de diferencia lineales, fundamentalmente) para abstraer y conceptualizar los fenómenos propios de los sistemas complejos.

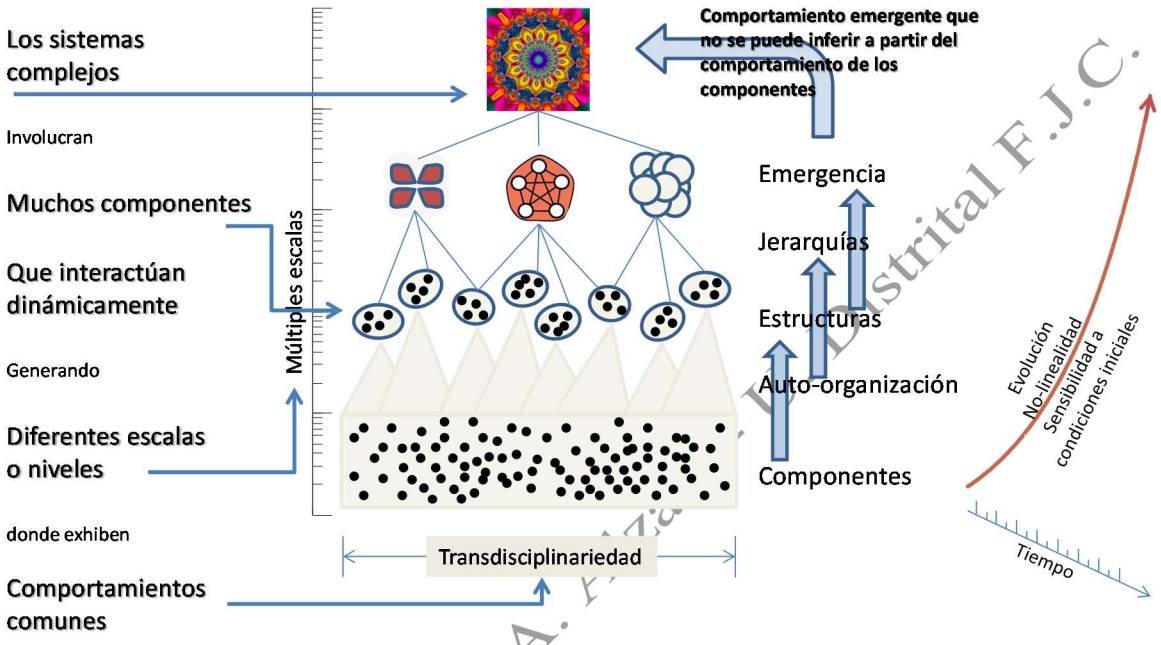


Figura 10. Sistemas Complejos (adaptado de www.art-sciencefactory.com/complexitymap_feb09.html)

Figura 11. Herramientas para el modelado matemático de sistemas complejos (adaptado de Hiroki Sayama, D.Sc., State University of New York, Binghamton)

En lo que respecta a la ingeniería de redes de comunicaciones, el modelado matemático de sistemas complejos busca diseñar sistemas compuestos por agentes cognitivos, esto es, capaces de percibir información del medio ambiente y actuar sobre él de manera que pueda comunicarse directa o indirectamente con otros agentes del sistema, aprender de las señales percibidas incluyendo los efectos de sus propias acciones, y adaptarse así a los cambios ambientales. De esta manera, cada agente cognitivo construye su propio modelo predictivo sobre los aspectos de interés del ambiente y desarrolla sus propias reglas de comportamiento que le permitan actuar sobre el ambiente para llevar a cabo algunas tareas prescritas de manera efectiva y eficiente, ya sea en cooperación o en competencia con otros agentes. En particular, hablando de redes de comunicaciones, ellas están compuestas por un gran número de componentes que interactúan entre ellos simultáneamente con una alta no-linealidad en las dinámicas de los protocolos que lo controlan (potencialmente caótica), fractalidad y multi-fractalidad en las trazas de tráfico que cursa por sus enlaces, topologías físicas y lógicas libres de escala, auto-organización de la demanda de los usuarios al borde de la congestión, y leyes de potencia en la distribución del tamaño de los archivos y la duración de las sesiones que circulan por la red. Todas estas condiciones actuales de las redes de comunicaciones, típicas en todos los sistemas complejos, contrastan con los principios filosóficos de diseño que les dieron origen, los cuales se resumen en una estructura modular jerárquica en la que las diferentes capas de la jerarquía sólo deberían interactuar mediante interfaces precisas y limpias, con protocolos diseñados exclusivamente para la función particular de cada capa. La insuficiencia de estos principios originales, especialmente debida al descubrimiento de interacciones imprevistas entre las capas de la jerarquía funcional, ha conducido a técnicas de diseño "cross-layer" cuyos problemas de optimización distribuida multiobjetivo se resuelven desde la teoría de control de sistemas dinámicos, la teoría de la información y la teoría de juegos, con base en el aprendizaje, la adaptabilidad y la robustez a las condiciones variantes del entorno, utilizando de manera ubicua técnicas de inteligencia computacional como las redes neuronales, la lógica difusa, los algoritmos genéticos y la inteligencia de enjambre, por ejemplo. En la siguiente definición veremos cómo, en ingeniería de teletráfico, se deben usar modelos simples y complejos, siempre considerando la aleatoriedad dentro de ellos.

3. Ingeniería de Teletráfico

> *La **ingeniería de teletráfico** consiste en el uso de **modelos matemáticos** formales para representar las interacciones que existen entre la **demanda**, la **capacidad** y el **desempeño** de una red de comunicaciones, como muestra la Figura 12. Dichos modelos deben permitir **comprender**, **analizar**, **diseñar** y **controlar** redes de comunicaciones.*

Cada uno de estos tres aspectos puede requerir una gran cantidad de variables para ser descrito adecuadamente, de manera que los símbolos a los que se refiere la figura para representar la demanda (λ), la capacidad (μ) y el desempeño (D) pueden ser arreglos de variables (vectores, posiblemente de dimensión infinita).

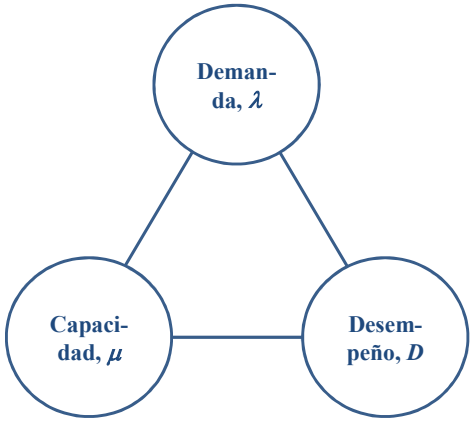


Figura 12. La ingeniería de tráfico estudia las relaciones entre demanda, capacidad y desempeño en una red de comunicaciones

En un ejemplo de extrema sencillez, λ puede ser un escalar que represente la **tasa promedio de llegadas** de paquetes a un enlace de comunicaciones (demanda, en paquetes por segundo), μ puede ser un escalar que represente la **tasa promedio de transmisión** de paquetes que dicho enlace puede alcanzar (capacidad, en paquetes por segundo) y D puede ser un escalar que represente el **retardo promedio** que experimentan los paquetes en dicho enlace, ya sea siendo transmitidos o esperando a ser transmitidos (desempeño, en segundos). Bajo condiciones muy específicas de variabilidad alrededor de esos valores promedios[3], las tres variables se relacionan de la siguiente manera:

$$D = \frac{1}{\mu - \lambda} \tag{1}$$

Un resultado así es muy útil porque nos permite determinar el desempeño dadas la capacidad y la demanda (**análisis de redes**, $D=1/(\mu-\lambda)$), o dimensionar la capacidad dada la demanda y el desempeño deseado (**diseño de redes**, $\mu=\lambda+1/D$), o regular la demanda dada la capacidad y el desempeño deseado (**control de redes**, $\lambda=\mu-1/D$). Además, resulta muy significativo el hecho de que, para que dicha expresión tenga sentido, la capacidad μ deba ser superior a la demanda λ o, lo que es lo mismo, que se satisfaga la relación $\rho \equiv \lambda/\mu < 1$, donde ρ, que se conoce como **intensidad de tráfico**, indica la relación entre la demanda (lo que los usuarios exigen, λ) y la capacidad (lo que la red puede ofrecer, μ). En efecto, si la demanda se acerca a la capacidad por la izquierda, el retardo se acerca a infinito; pero si la demanda supera la capacidad, el retardo se haría negativo, indicando que ese caso no está en el rango de validez del análisis que condujo a dicha expresión. La ingeniería de teletráfico busca exactamente este tipo de relaciones que proporcionan comprensión de la dinámica

[3] Como veremos más adelante, las llegadas deben formar un proceso de Poisson y los tiempos de transmisión deben ser independientes y estar exponencialmente distribuidos

de las redes de comunicaciones y herramientas matemáticas para análisis, diseño y control de las mismas.

Para determinar relaciones útiles entre la demanda, la capacidad y el desempeño de una red de comunicaciones es necesario estudiar las interacciones dinámicas que se forman cuando los usuarios compiten o colaboran para utilizar los recursos de la red. En efecto, estas redes están conformadas por un conjunto de recursos de capacidad limitada que deben atender las demandas impuestas por los usuarios de la red. La Figura 13 muestra un modelo matemático de una red compuesta por seis enrutadores, cada uno de ellos con dos enlaces de entrada y dos enlaces de salida. Suponemos que cada paquete que llega a un enrutador debe pasar por un proceso de enrutamiento (círculo azul) en el que se decide a cuál de los dos enlaces de salida debe dirigirse el paquete. Como el proceso de decisión toma un tiempo diferente de cero, es posible que un paquete que llegue a un enrutador encuentre el procesador de enrutamiento ocupado con algún otro paquete que llegó con anterioridad, por lo que se dispone también de una memoria (o *buffer*) donde almacenar los paquetes que deben hacer cola para esperar a ser enrutados (rectángulos azules). Los paquetes que salen del proceso de decisión se dirigen a su respectivo enlace de salida (círculos amarillos) pero, nuevamente, como el tiempo de transmisión de un paquete es diferente de cero, es posible que un paquete encuentre el enlace ocupado, por lo que se dispone también de un buffer donde almacenar los paquetes que deben hacer cola para esperar a ser transmitidos (rectángulos amarillos). En este proceso, los paquetes sufren **retardos** (tanto por los tiempos de servicio como por los tiempos de espera en cola) y potenciales **pérdidas** (si, por ejemplo, a su llegada no hay espacio suficiente en el buffer), lo que conduce a una tasa efectiva de paquetes por unidad de tiempo o **caudal**, que puede ser inferior a la demanda impuesta por los usuarios.

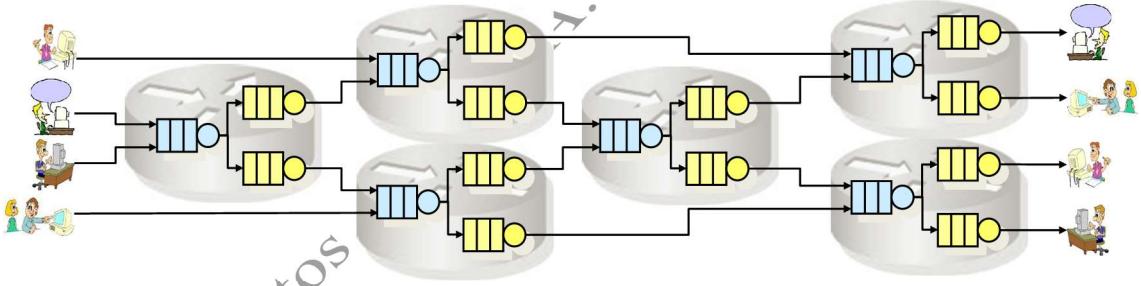


Figura 13. La red de comunicaciones es un conjunto de recursos de capacidad finita que deben ser compartidos por los usuarios

El anterior ejemplo explica por qué los modelos matemáticos más ampliamente utilizados en ingeniería de teletráfico corresponden a la **Teoría de Colas**, que es el área de la investigación operacional que estudia los procesos estocásticos asociados con líneas de espera, donde unos "clientes" hacen fila en espera de un "servicio". De hecho, fue **Agner Erlang** quien fundó la teoría de colas al modelar matemáticamente la red telefónica de Copenhague a comienzos del siglo XX. Posteriormente la teoría de colas se ha aplicado a una gran cantidad de problemas prácticos en flujo de tráfico vehicular, procesos industriales, sistemas operativos, servicio a clientes en diferentes tipos de oficinas, planeación de sistemas de transporte de carga, atención a pacientes en hospitales y centros médicos, etc. Aun así, sigue siendo en el área de las redes de comunicaciones donde mayores avances se han alcanzado en esta disciplina matemática.

En efecto, ha sido en las redes de comunicaciones donde mayor necesidad ha habido de análisis novedosos de sistemas de colas. Con la introducción de los conmutadores automáticos a finales del siglo XIX y su posterior digitalización en la década de 1970, las actividades de análisis de tráfico y optimización de la red para satisfacer unos criterios dados de calidad de servicio (QoS) correspondían al diseño de **redes conmutadas por circuitos** con una sola clase de tráfico. Entonces se desarrolló un método tradicional basado en la **fórmula B de Erlang** que relacionaba la intensidad de tráfico (demanda, A, igual al producto entre el número promedio de llamadas por minuto y la duración promedio de una llamada en minutos), el número de circuitos disponibles (capacidad, N) y la probabilidad de que una llamada entrante no encontrara un circuito libre (desempeño, P_B, igual a la fracción de llamadas perdidas), cuando las llegadas forman un proceso de Poisson y sus duraciones son independientes unas de otras:

$$P_B = \frac{A^N / N!}{\sum_{n=0}^{N} A^n / n!} \tag{2}$$

Esta es otra relación que obedece a la Figura 12, donde la demanda es A (en vez de λ), la capacidad es N (en vez de μ) y el desempeño es P_B (en vez de D), y que también permite hacer análisis (encontrar P_B dados A y N), diseño (encontrar N dados P_B y A) y control (encontrar A dados P_B y N) de redes.

Mientras se desarrollaba esta infraestructura para las redes telefónicas, en el mundo de la informática se desarrollaban paralelamente las técnicas de transmisión de datos entre computadores, lo cual condujo a las **redes conmutadas por paquetes** con múltiples arquitecturas jerárquicas de protocolos, todas orientadas filosóficamente por el **modelo de referencia OSI**. El propósito fundamental de estas redes era compartir recursos físicos y prestar servicios de transferencia de archivos, correo electrónico y log-in remoto, especialmente. La Figura 13 y la Ecuación (1) se refieren más a este tipo de redes. En la década de 1980, con el desarrollo de las redes públicas de datos tipo **X.25**, las redes telefónicas digitales con **señalización por canal común**, y el desarrollo de la **Internet mediante TCP/IP**, se reconoció la necesidad de integrar la comunicación telefónica con la transmisión de datos por lo que se desarrollaron técnicas híbridas en la llamada **Red Digital de Servicios Integrados** (ISDN), de donde surgieron tecnologías aún usadas hoy, como **Frame Relay** (relevo de tramas). En la década de 1990 se tuvo la visión de una sola red que integrara no sólo voz y datos sino todo tipo de medios de comunicación (texto, imágenes, video, teleconferencia, etc.), para lo cual se decidió que la conmutación de paquetes era la más apropiada y se desarrolló la red digital de servicios integrados de banda ancha (**B-ISDN**) y su técnica de conmutación, el **Modo de Transferencia Asincrónico (ATM)**. Todos estos desarrollos estuvieron asociados con avances correspondientes en la teoría de colas, incluyendo los conceptos teóricos para ofrecer diferente **calidad de servicio** (QoS –Quality-of-Service-) a diferentes **clases de tráfico** en una misma red conmutada por paquetes. Estos desarrollos incluyen caracterización de tráfico con tiempos entre llegadas y tiempos de servicio correlacionados a corto y a largo plazo, técnicas de control de admisión y control de acceso, tecnologías de conmutación por hardware, conceptos de calidad de servicio con cotas de desempeño entre extremos, el concepto de capacidad equivalente, mecanismos de asignación de recursos de transmisión, gestión activa de colas y de control de congestión mediante realimentación, etc. Sin embargo, mientras se planeaba y se empezaba a implementar la B-ISDN y la tecnología ATM,

apareció la **world-wide-web (WWW)**, una aplicación que disparó a la Internet y su arquitectura TCP/IP como la red donde finalmente se dio la integración fundamental que se buscaba con ATM.

Efectivamente, es al navegar por la web que se evidencian las bondades de tener video, voz, datos, imágenes y audio integrados en una misma aplicación y en una misma red. El fenómeno de la WWW sobre Internet hizo que los conceptos desarrollados para QoS en ATM se aplicaran de alguna manera en redes TCP/IP, con lo que se desarrollaron arquitecturas orientadas a QoS sobre IP tales como servicios integrados (**IntServ**) y servicios diferenciados (**DiffServ**), y protocolos adicionales que buscan acercar conceptualmente IP a ATM, tales como **MPLS**, **RSVP**, **RTP**, **IPv6**, **QoS-OSPF**, etc. Todos estos mecanismos integran la teoría de colas con **teoría de control** de sistemas dinámicos, en sus formas más avanzadas de control no-lineal, control estocástico, control óptimo, control robusto y control distribuido, simultáneamente, abriendo ampliamente el espectro de los modelos matemáticos aplicables en las redes de comunicaciones y en la ingeniería de teletráfico. Por ejemplo, los usuarios que usan TCP como protocolo de transporte ajustan la tasa de transmisión de acuerdo con la información retroalimentada por los mecanismos de gestión de las colas en los enrutadores (**AQM, Active Queue Management**). Esta información realimentada puede ser en forma de paquetes perdidos, variaciones en el retardo de los paquetes o notificaciones explícitas de congestión. De esta manera, el conjunto usuario/enrutador forma un sistema de control realimentado, como se muestra en la Figura 14. El algoritmo TCP modula el flujo de datos de usuario de manera que, en el instante t, genera $r(t)$ paquetes por segundo. La longitud de la cola en el enrutador, $q(t)$, varía con el flujo $r(t)$ de manera que, en el instante t, se produce una pérdida de paquetes $p(t)$. Debido a los retardos en la red, los usuarios se enteran de estas pérdidas τ segundos tarde y, de acuerdo con ellas, reajusta la tasa $r(t)$ tratando de maximizar alguna función de utilidad.

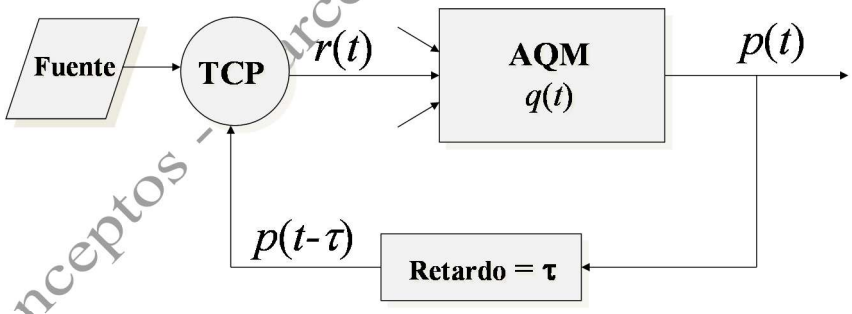


Figura 14. Modelo de control realimetnado para el flujo TCP

Cada uno de los mecanismos mostrados en la Figura 14 (TCP, AQM y la dinámica de la cola) determina las tasas de cambio del flujo, la longitud de la cola y la tasa de pérdidas mediante funciones no lineales correspondientes (R, P y Q, resepctivamente):

$$r'(t) = R\big(p(t-\tau)\big), \quad q'(t) = Q\big(q(t), r(t)\big), \quad p'(t) = P\big(q(t), r(t)\big) \tag{3}$$

Claramente, ajustar los parámetros de estos mecanismos para maximizar una función de utilidad adecuada es un problema de control óptimo, predictivo, no lineal, distribuido y estocástico, lo cual exige modelos matemáticos adecuados de ingeniería de teletráfico.

Otro desarrollo vertiginoso de las últimas tres décadas se ha dado en el contexto de las **redes inalámbricas**. Junto con la integración de medios de comunicación y la satisfacción de requerimientos de QoS, los usuarios quieren disponibilidad de los servicios de comunicación en todo instante y en todo lugar. Esto sólo es posible si se usan medios inalámbricos de acceso que permitan la movilidad de los usuarios, para lo cual se desarrolló la **telefonía celular**: Una estación base coordina la comunicación en una "célula" o "celda" (una región de cobertura o área de servicio en la que se usa un conjuntos particular de canales de radio), permitiendo el acceso inalámbrico de múltiples usuarios móviles a la infraestructura de la red conmutada. A lo largo de cuatro generaciones de tecnología celular se pasó de la transmisión análoga de canales de voz conmutados por circuitos (1G) a una conexión 100% IP entre extremos, sin conmutación de circuitos, donde los paquetes que se intercambian pueden llevar datos, acceso móvil a la www, telefonía IP, televisión (de media o alta resolución), videoconferencia, etc. (4G). Ya se hacen pruebas de una quinta generación con velocidades de acceso de varios Gbps con millones de conexiones por kilómetro cuadrado, lo que traerá a nuestros dispositivos de bolsillo la realidad virtual y la realidad aumentada, además de habilitar la Internet de las cosas (IoT). Con tantas alternativas de acceso, los dispositivos móviles actuales pueden escoger libremente entre acceso mediante **WiFi**, **Bluetooth**, **WiMax**, **LTE** o cualquier tecnología celular, de manera oportunista, para acceder a cualquiera de los servicios mencionados o para conectarse directamente entre ellos sin ninguna infraestructura subyacente, incluyendo redes móviles ad hoc (**MANET**), las cuales se pueden usar para acceso a redes fijas (**redes mesh**), y redes de sensores inalámbricos (**WSN**), incluyendo tecnologías tan útiles como las VANET (Vehicle Ad hoc Network) que ya conducen a la Internet de Vehículos (IoV) para gestión de tráfico urbano.

En este contexto los principios básicos de las arquitecturas funcionales no aplican transparentemente, por lo que se requiere usar técnicas de **diseño *cross-layer***. Adicionalmente, la capacidades de adaptabilidad de los sistemas de radio (**SDR** –software defined radio–) hacen posible la implementación de redes de **radio cognitiva**, en donde los equipos cognitivos aprovechan los períodos de tiempo en que las bandas licenciadas no son utilizadas por sus usuarios legítimos para transmitir su propia información. Los terminales cognitivos utilizan propiedades como la **percepción**, el **aprendizaje** y la **adaptación** para llevar a cabo sus funciones, propiedades asociadas con el cerebro humano, lo cual conecta los modelos matemáticos de las redes de comunicaciones con la **inteligencia computacional** y con la teoría de **sistemas dinámicos complejos**. Estos mecanismos permiten resolver cooperativamente los problemas de detección y estimación de señales mediante sistemas **MIMO** (MultiInput-MultiOutput), y permiten la virtualización de los servicios de comunicación móvil sobre los esquemas de computación distribuida "en la nube" (**cloud-computing**). Más aún, con técnicas de comunicación por campo cercano **NFC** (Near Field Communication) como **RFID** (Radio Frequency Identification), que permiten comunicaciones a muy corta distancia para intercambio de datos entre dispositivos, empieza a hacerse realidad la **IoT** (Internet of Things), la cual permite percibir y controlar objetos remotamente a través de la infraestructura de redes, en un paso más de virtualización del mundo real que conducirá rápidamente a sistemas cyber-físicos de objetos inteligentes (smart grids, smart homes, smart cities, etc.).

Todo este desarrollo en las tecnologías de las redes de comunicaciones ha traído grandes retos al modelado matemático para análisis, diseño y control de las mismas, por lo que la ingeniería de teletráfico vive un momento de grandes desafíos en su propósito de encontrar relaciones formales entre la capacidad, la demanda y el desempeño de las redes modernas de comunicaciones. En efecto, con los avances tecnológicos en redes de comunicaciones, avanzan también las técnicas de modelado matemático para ingeniería de teletráfico, partiendo desde la teoría de colas y avanzando hacia conceptos como el ancho de banda equivalente (el cual permite diseñar mecanismos de QoS con base en el estudio de las cotas propuestas por la teoría de grandes desviaciones), la aplicación formal de la teoría de control de sistemas dinámicos (incluyendo conceptos básicos de estabilidad y robustez en control distribuido, robusto, óptimo, no-lineal y estocástico), desarrollos basados en inteligencia computacional (redes neuronales, algoritmos genéticos, lógica difusa, algoritmos de enjambre, sistemas inmunes, etc.) y, en últimas, aplicación formal de la teoría de los sistemas complejos adaptivos (comportamientos emergentes, auto-organización, algoritmos de consenso, fractales, etc.). Algunas propuestas para ingeniería de sistemas complejos con aplicaciones en ingeniería de redes incluyen los sistemas dinámicos cognitivos, la auto-organización guiada y la ingeniería morfogenética, entre otros. Considere por ejemplo un enjambre de robots cuyo movimiento debe ser tal que mantengan un mínimo de conectividad en la red de comunicaciones ad-hoc formada entre ellos, a pesar de que la estructura de la red cambie con el movimiento. Cualquier solución centralizada o jerárquica podría imponer una excesiva sobrecarga a la red en términos de información de control, por lo que se necesitan algoritmos de control descentralizados que relacionen en un modelo matemático distribuido la conectividad inalámbrica local y la movilidad de los robots, de manera que cada robot, haciendo uso únicamente de la información local, tome decisiones individuales y autónomas tales que el conjunto de robot mantengan la conectividad necesaria.

En todos estos efuerzos recientes sigue siendo necesario modelar la incertidumbre, aun más que en los modelos matemáticos de comienzos del siglo XX. El modelamiento probabilístico que se estudia en este libro es un método muy poderoso para representar la incertidumbre cuantitativamente que ha demostrado su gran utilidad desde hace más de cuatrocientos años y que se complementa maravillosamente con desarrollos más recientes como los sistemas dinámicos caóticos y la lógica difusa, entre otras posibilidades.

II. Conceptos Básicos de Teoría de Probabilidad

4. Experimento Aleatorio

Un experimento es un proceso de observación mediante el cual se selecciona un elemento de un conjunto de posibles resultados. Un experimento aleatorio es aquel en el que el resultado no se puede predecir con anterioridad a la realización misma del experimento.

El modelado matemático consiste en lograr un nivel de abstracción tal que podamos agrupar una gran cantidad de problemas en un solo concepto. Todos sabemos que lanzar una moneda para ver qué lado queda hacia arriba es un experimento aleatorio, cómo también lo fue observar las señales del detector Atlas del LHC (el colisionador de hadrones de 27 km, y 8 mil millones de euros) después de una colisión y detectar si hay en ellas evidencia o no de la existencia del bosson de Higgs[4]. ¿Qué tienen en común estos dos experimentos para poder extraer los elementos de una definición, que aplique también para lanzar un dado y contar los puntos de la cara que queda hacia arriba, para implementar un testbed en el laboratorio y medir el desempeño de un nuevo protocolo de comunicaciones, o para construir y correr un programa de simulación que evalúe el mismo protocolo? Pues bien, en todos ellos hemos llevado a cabo un proceso de observación y, como consecuencia del mismo, hemos seleccionado uno de un conjunto de posibles resultados. Eso es lo que hace el Banco Emisor cuando ajusta las tasas de interés para ver sus efectos en la inflación y lo que hace el protagonista de un libro de probabilidades cuando selecciona una bola de una bolsa para observar su color ¡Así es la capacidad de abstracción del modelado matemático!

Los ejemplos típicos de los cursos de probabilidad incluyen, como acabamos de ver, lanzar una moneda, que equivale a seleccionar un elemento del conjunto $\{cara, sello\}$, lanzar un dado, que equivale a seleccionar un elemento del conjunto $\{1,2,3,4,5,6\}$, o escoger una carta de la baraja de naipes, que equivale a seleccionar un elemento del conjunto $\{(f,n) : f \in \{picas, tréboles, corazones, diamantes\}, n \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K\}\}$. Otros ejemplos más interesantes para nosotros incluyen medir la tasa de pérdida de paquetes en una conversación VoIP, que equivale a seleccionar un elemento del conjunto $\{x \in \mathbb{Q} \mid 0 \leq x \leq 1\}$, medir el retardo de un paquete de voz en esa misma conversación, que equivale a seleccionar un elemento del conjunto $\mathbb{R}^+$ (los reales positivos), o verificar el estado de ocupación de un canal de comunicaciones, que equivale a seleccionar un elemento del conjunto $\{libre, ocupado\}$.

En un experimento aleatorio, aunque se mantengan constantes las condiciones bajo las cuales se realizan diferentes instancias del mismo, el resultado no se puede predecir con anterioridad a la realización del experimento. Por ejemplo, como vimos en el prefacio, generalmente no es posible

[4] En 2012 los experimentos del LHC cruzaron el umbral de los "7 sigmas": Hay una certeza superior al 99.9999999999% de que exista el bosón de Higgs. Si el bosón de Higgs no existe, los resultados de los experimentos se dieron al azar con una probabilidad de 0.0000000001%. Note que lanzar una moneda es un experimento que podemos repetir. Detectar la exitencia del bosón de Higgs es un experimento que se hizo una vez y no se podrá volver a hacer, porque ya no sería aleatorio.

predecir el caudal, el tamaño del archivo ni el tiempo de transferencia en una transacción *ftp*, lo que indica que transferir un archivo de un servidor a un cliente mediante *ftp* constituye un experimento aleatorio. Igualmente, si desde la ventana de comandos de nuestro PC ejecutamos la instrucción

```
C:>netstat –e 10 > estadisticas.txt
```

y navegamos por Internet durante una hora, generaremos un archivo con algunas estadísticas de la red, incluyendo el número de bytes que se han recibido en períodos de diez segundos. La Figura 15 presenta una gráfica del número de bytes recibidos durante varios períodos en una instancia del experimento. Evidentemente, no estamos en condiciones de predecir cuántos bytes llegarán en el siguiente período, aun cuando podemos afirmar que, por ejemplo, sería muy extraño si llegaran más de 20 Mbytes y, en cambio, sí sería de esperar que llegaran más de 2 Mbytes. De cualquier manera, queda claro que observar el número de bytes recibidos en 10 segundos mientras se navega por Internet constituye un experimento aleatorio.

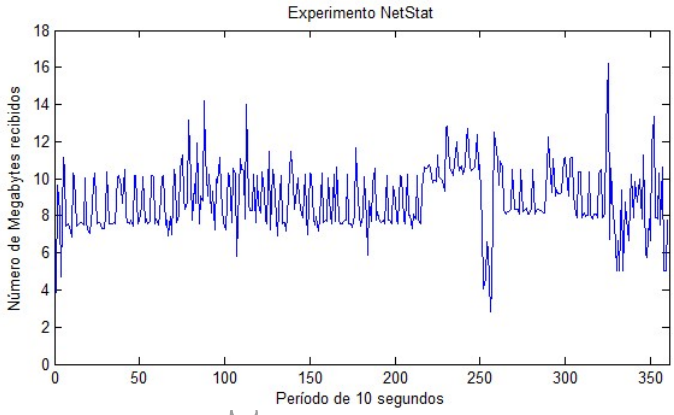


Figura 15. Observar el número de bytes que llegan de la red en un período de 10 segundos constituye un buen ejemplo de lo que es un experimento aleatorio.

¿Por qué no podemos predecir el resultado de un experimento aleatorio? En principio, esta pregunta ha desvelado a muchos científicos de muchas maneras distintas. Las siguientes son tres posibles razones: (1) Desconocemos las leyes naturales que rigen el experimento, (2) conocemos dichas leyes pero son tan complejas que nos es imposible –o resulta indeseable- evaluarlas, (3) existe una indeterminación básica en el universo. La tercera razón es propia de la mecánica cuántica, en la que cada partícula se describe mediante una función de onda cuya magnitud al cuadrado representa la incertidumbre sobre su posición en cada instante. Las primeras dos razones, en cambio, hablan de nuestra ignorancia o nuestra incapacidad, lo que haría de la aleatoriedad un concepto subjetivo, que podría desaparecer con el desarrollo del conocimiento o de la tecnología. Una cuarta fuente de incertidumbre, aunque no de aleatoriedad, es la impredecibilidad de los sistemas dinámicos caóticos. A pesar de ser completamente determinísticos, presentan una sensibilidad a las condiciones iniciales tal que, si queremos predecir algo sobre ellos en un futuro mediano, deberíamos conocer las condiciones iniciales con una precisión absurdamente alta. A pesar del determinismo de estos sistemas, a veces conviene usar modelos probabilísticos para cuantificar la incertidumbre en sus estados futuros. Lo cierto es que, como muestran los experimentos `netstat` o `ftp`, ni siquiera el

más experto ingeniero conocedor de los más íntimos detalles de la implementación de cada protocolo de una red de comunicaciones a todos los niveles de su jerarquía funcional podría predecir los instantes en que cada usuario de la red generará demandas o la magnitud de esas demandas. En consecuencia, aunque una mente privilegiada con infinitos poderes divinos pudiera considerar una red de comunicaciones como un sistema determinístico, a nosotros, pobres mortales, nos toca aceptar nuestra incertidumbre sobre el comportamiento de la red y conformarnos con el hecho de que, al observar la red, estamos llevando a cabo un experimento aleatorio.

5. Espacio Muestral

El espacio muestral de un experimento aleatorio es el conjunto de todos los posibles resultados que podrían observarse en una realización del experimento,

$$\Omega = \{\omega : \omega \text{ es un posible resultado del experimento aleatorio}\}$$

Cuando definimos un experimento como un proceso de observación mediante el cual se selecciona *un elemento de un conjunto de posibles resultados*, queda claro que, si queremos especificar adecuadamente un experimento, lo primero que debemos describir con precisión es ese conjunto de posibles resultados. En este libro, como es costumbre en la mayoría de textos sobre probabilidades, denotaremos el espacio muestral mediante la letra griega mayúscula Ω (ómega) y sus elementos, de manera genérica, se denotarán mediante la correspondiente letra minúscula ω. Algunos ejemplos que ya se mencionaron en la definición 4 son:

1. Lanzar una moneda y ver qué lado queda hacia arriba: $\Omega = \{cara, sello\}$.
2. Lanzar un dado y contar los puntos en la cara que queda hacia arriba: $\Omega = \{1,2,3,4,5,6\}$.
3. Escoger una carta de la baraja de naipes: Ω = {*picas*, *tréboles*, *corazones*, *diamantes*}×{1,2,3,4,5,6,7,8,9,10,*J,Q,K*}, donde × representa el producto cartesiano entre los dos conjuntos.
4. Verificar el estado de ocupación de un canal de comunicaciones: $\Omega = \{libre, ocupado\}$.
5. Medir la fracción de paquetes perdidos durante una hora en una red IP: $\Omega = \{x \in \mathbb{R} : 0 \le x \le 1\}$.
6. Medir el retardo experimentado por un paquete de datos mientras transita por una red IP: $\Omega = \mathbb{R}^+ = \{x \in \mathbb{R} : x > 0\}$.
7. Contar el número de canales PCM libres en un enlace E1: $\Omega = \{0,1,2,\ldots,32\}$.
8. Mirar el estado de ocupación de cada uno de los canales en un enlace E1 : $\Omega = \{libre, ocupado\}^{32}$. La potencia indica que se deben ejecutar 32 productos cartesianos del conjunto {*libre*, *ocupado*} con sigo mismo, con lo que se construye el conjunto de todas las cadenas de 32 símbolos en las que cada símbolo puede tomar uno de los valores *libre* u *ocupado*. Nótese que, aunque el experimento parece sencillo, la *cardinalidad* del espacio muestral (su número de elementos) es mayor a cuatro mil millones ($|\Omega|$ = 4.294'967.296).
9. Determinar si un bit, transmitido sobre un canal de comunicaciones, llega correctamente al receptor en el otro extremo del canal: $\Omega = \{si, no\}$

10. Contar el número de transmisiones que requiere un paquete de datos hasta llegar correctamente a su destino: $\Omega = \{1,2,3,\ldots\}$
11. Contar el número de bits recibidos con error en una trama de L bits que llega a través de un canal ruidoso: $\Omega = \{0,1,2,\ldots,L\}$
12. Medir durante una hora la fracción de tiempo que un enlace de comunicaciones permanece ocupado: $\Omega = \{x \in \mathbb{R} : 0 \leq x \leq 1\}$.
13. Contar el número de paquetes que llegan a un enrutador de una red de comunicaciones durante un período de una hora: $\Omega = \{0,1,2,\ldots\}$.
14. Medir el tiempo que transcurre entre la llegada de dos paquetes consecutivos a un enrutador de una red de comunicaciones: $\Omega = \mathbb{R}^+ = \{ x \in \mathbb{R} : x > 0\}$.
15. etc.

Nótese que, aunque cada experimento aleatorio puede tener solamente un espacio muestral, la ignorancia del modelador respecto a detalles particulares del experimento le puede llevar a considerar espacios muestrales más grandes, lo cual no está mal mientras el espacio muestral supuesto por el modelador, Ω_m, contenga al espacio muestral verdadero, Ω_v. En efecto, en ese caso, simplemente los "posibles resultados" pertenecientes a $\Omega_m \cap \Omega_v^C$ nunca ocurrirán. Por ejemplo, en el experimento 7 supusimos $\Omega = \{0,1,2,\ldots,32\}$ porque un enlace E1 tiene 32 canales TDM; sin embargo, un modelador con mayor *conocimiento a priori* podría saber que ese canal E1 hace parte de una red PCM para telefonía, en cuyo caso reduciría su espacio muestral a $\Omega = \{0,1,\ldots,30\}$ al considerar que el canal 0 siempre está ocupado con bits de sincronización y el canal 16 siempre está ocupado con bits de señalización. Igualmente, en el experimento 6, algún modelador podría saber cuál es la mínima longitud de los paquetes, L bits, y la máxima capacidad de la ruta, C bps, con lo que podría reducir el espacio muestral a $\Omega = \{x \in \mathbb{R} : x \geq L/C\}$. Cabe anotar que, en muchas ocasiones, un modelador con una gran cantidad de conocimiento a priori que le permita encontrar un conjunto Ω_m muy cercano a Ω_v, puede decidir escoger un espacio muestral aún mayor a Ω_m con el único propósito de simplificar el tratamiento analítico posterior. Por ejemplo, como el número de paquetes que llegan en una hora es un número entero, un modelador podría saber que en el experimento 5 un espacio muestral más cercano al verdadero está contenido en $\{m/n \in \mathbb{Q}^+ : n>0, m \leq n\}$, donde $\mathbb{Q}^+$ son los números racionales no negativos. Sin embargo, parece intuitivamente claro que podría ser más fácil considerar el espacio muestral constituido por el intervalo real [0,1].

La Figura 16 muestra un diagrama de Venn que incluye el conjunto Ω compuesto por todos los posibles resultados de todos los posibles experimentos (¿cuán grande es este conjunto?) y, en él, algunos espacios muestrales asignados a un experimento particular, $\Omega_v \subset \Omega_1 \subset \Omega_2 \subset \Omega_3$ y Ω_4. El verdadero espacio muestral, Ω_v, puede ser un conjunto muy complejo. Ω_1 es el espacio muestral que podría seleccionar un modelador juicioso con una gran cantidad de conocimiento a priori. Ω_2 es el espacio muestral que decidiría seleccionar este mismo modelador para facilitar el análisis posterior. Ω_3 es el espacio muestral seleccionado por otro modelador igualmente juicioso pero que tiene muy poco conocimiento a priori. Por último, Ω_4 es el espacio muestral que seleccionaría un modelador poco juicioso y muy desafortunado, pues no podrá llegar a ningún destino útil por haber empezado parándose sobre arenas movedizas.

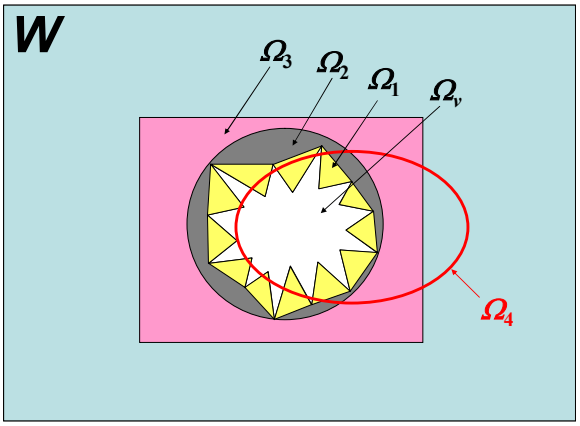


Figura 16. Algunos espacios muestrales asignados a un mismo experimento

6. Evento

Por lo pronto, diremos que un evento es un subconjunto del espacio muestral de un experimento aleatorio. Más adelante (después de la definición 14) restringiremos la definición de evento a un subconjunto "medible", esto es, un subconjunto al que se le pueda asignar una medida de probabilidad.

El evento $A \subset \Omega$ ocurre al realizar una instancia del experimento si el resultado obtenido, ω, pertenece a A, $\omega \in A$.

Supongamos, por ejemplo, que medimos la fracción de paquetes perdidos en una videoconferencia (experimento 5), de manera que los posibles resultados son $\Omega = \{x \in \mathbb{Q} : 0 \leq x \leq 1\}$. Si nos interesa satisfacer un requerimiento de calidad de servicio según el cual no se pueden perder más del 0.1% de los paquetes, deberíamos buscar que una fracción importante de las instancias del experimento correspondieran a elementos del subconjunto $A = \{ x \in \Omega : x \leq 0.001\}$. De la misma manera, en cada uno de los experimentos propuestos en la definición 5 podemos definir algunos eventos apropiados:

1. Lanzar una moneda y ver qué lado queda hacia arriba: Los posibles eventos de interés incluyen a los subconjuntos unitarios de Ω, A={*cara*} y B={*sello*}. No sobra recordar que, mientras *cara* es un posible resultado del experimento, esto es, un elemento de Ω, {*cara*} es un subconjunto unitario de Ω. ¡Y es muy importante reconocer la diferencia! De otro lado, además de los eventos A y B mencionados antes hay otros dos posibles eventos : Ω (*el evento cierto*) y Φ, (vacío o *el evento nulo*), pues ellos dos siempre son subconjuntos de Ω.
2. Lanzar un dado y contar los puntos en la cara que queda hacia arriba: En el espacio muestral Ω= {1,2,3,4,5,6} están incluidos conjuntos como A = {*hay más de tres puntos*} = {4,5,6} y B = {*hay un número par de puntos*} = {2,4,6}, así como $A \cap B$ = {4,6}, por ejemplo.
3. Escoger una carta de la baraja de naipes: En el espacio muestral descrito antes están contenidos, por ejemplo, los eventos A = {*Una figura de pinta roja*} = {*corazones*, *diamantes*}×{*J*,*Q*,*K*} y B = {*Un as negro*} = {(*picas*, 1), (*tréboles*, 1)}.

4. Verificar el estado de ocupación de un canal de comunicaciones: Como en el ejemplo 1, los posibles eventos de interés son los subconjuntos unitarios {*libre*} y {*ocupado*}, que son diferentes a los elementos de Ω, *libre* y *ocupado*.
5. Medir la fracción de paquetes perdidos durante una hora en una red IP: El espacio muestral es el conjunto de racionales en el intervalo [0,1] de la recta real, donde podemos definir un evento A que dispararía una alarma en el centro de gestión de la red, $A = \{x \in \Omega : 0.1 \leq x\}$: ¡En la última hora se perdió más del 10% de los paquetes!
6. Medir el retardo experimentado por un paquete de voz mientras transita por una red IP dotada con mecanismos VoIP: En este caso, como un paquete que llegue con más de 100 ms (p.ej.) de retardo es descartado en el receptor, un evento de gran interés sería $A = \{x \in \Omega : x > 0.1\} = \{$*El paquete no alcanza a ser reproducido en el receptor*$\}$.
7. Contar el número de canales libres en un enlace E1: Si una videoconferencia requiere 384 kbps, un evento de interés podría ser $A = \{$*Se puede establecer una videoconferencia*$\} = \{6,7,8,\ldots,32\}$.
8. Mirar el estado de ocupación de cada uno de los canales en un enlace E1 : $\Omega = \{$*Libre*, *Ocupado*$\}^{32}$. Si definimos 33 eventos diferentes [$X_i = \{$*Hay* i *canales libres*$\}$, i=0,1,2,…,32], estaríamos "reconstruyendo" el experimento 7. Sin embargo, mientras "16 *canales libres*" es un elemento del espacio muestral del experimento 7, en el experimento 8 se trata de un evento (un subconjunto del espacio muestral) compuesto por ¡601'080.390 elementos!
9. Determinar si un bit, transmitido sobre un canal de comunicaciones, llega correctamente al otro lado: $\Omega = \{$*si*, *no*$\}$. Como en el experimento 1, no tenemos muchos más eventos que los unitarios {*si*} y {*no*}, aunque siempre podemos escoger también el evento cierto y el evento Nulo.
10. Contar el número de transmisiones (a través de un canal ruidoso) que requiere un paquete de datos hasta llegar correctamente a su destino:$\Omega = \mathbb{N} = \{1,2,3,\ldots\}$. El evento {*No hay errores de transmisión*} corresponde al subconjunto unitario {1}.
11. Contar el número de bits con errores en una trama de L bits que se recibe de un canal ruidoso: $\Omega = \{0,1,2,\ldots,L\}$. El evento {*Es necesario retransmitir el paquete*} corresponde al subconjunto $\{1,2,3,\ldots,L\} = \{0\}^C$, donde el superíndice C indica complemento en Ω : Será necesario retransmitir el paquete si se daña al menos un bit.
12. Medir durante una hora la fracción de tiempo que un enlace de comunicaciones permanece ocupado: $\Omega = \{x \in \mathbb{R} : 0 \leq x \leq 1\}$. Un posible criterio de gestión de la red podría interpretar el evento $\{x \in \Omega : x \geq 0.8\}$ como {*Es necesario dispersar el tráfico que cursa sobre el enlace*}.
13. Contar el número de paquetes que llegan a un enrutador de una red de comunicaciones durante una hora: $\Omega = \{0,1,2,\ldots\}$. Si el enrutador es capaz de atender hasta μ paquetes por hora, un evento de sumo interés para el administrador de la red será $A = \{n \in \Omega : n > \mu\}$, pues la ocurrencia del evento A indica que el enrutador está experimentando congestión.
14. Medir el tiempo que transcurre entre la llegada de dos paquetes consecutivos a un enlace de una red de comunicaciones: $\Omega = \mathbb{R}^+ = \{ x \in \mathbb{R} : x > 0\}$. ¿Cuál sería el evento A={*Cuando llega el segundo paquete, el primer paquete ya ha sido transmitido*}? La respuesta puede no ser fácil porque depende del estado del enlace (cuántos paquetes había en espera de servicio) cuando llegó el primer paquete, cuya observación constituye otro experimento aleatorio. Sin embargo, si la longitud mínima de los paquetes es L bits y la capacidad del enlace es C bps, sabemos con

seguridad que el evento que nos preguntan está contenido en otro evento mayor, $A \subseteq B = \{x \in \Omega : x > L/C\}$.

En esta definición hemos considerado "evento" como sinónimo de "sub-conjunto de Ω". Más adelante, cuando definamos la medida de probabilidad (definición 14), distinguiremos entre subconjuntos medibles y subconjuntos no medibles según se les pueda asignar o no una medida de probabilidad. Entonces restringiremos el significado de evento a "subconjunto medible".

7. Frecuencia Relativa

Sea A *un subconjunto del espacio muestral de un experimento aleatorio. Si repetimos* N *veces el experimento y observamos que en* N_A *de esas repeticiones se obtuvo un elemento de* A*, decimos que* $f_N(A) = N_A/N$ *es la frecuencia relativa del subconjunto A en* esas *N repeticiones del experimento.*

Nótese que la notación es muy imprecisa pues $f_N(A)$ no es una función de A subindicada por N. En efecto, en una secuencia diferente de N repeticiones del mismo experimento podríamos obtener un valor distinto de $f_N(A)$. Por ejemplo, considérense las siguientes dos secuencias de 10 lanzadas de un dado:

Secuencia 1 : 2 4 2 6 1 5 3 6 3 3
Secuencia 2 : 6 4 2 5 3 1 5 3 6 4

Si observamos la frecuencia relativa del subconjunto A = {el resultado es menor que cuatro} = {1, 2, 3} obtenemos que $f_{10}(A) = 0.6$ en la primera secuencia y $f_{10}(A) = 0.4$ en la segunda secuencia, mientras que, considerando las dos secuencias conjuntamente, obtenemos $f_{20}(A) = 0.5$. Con respecto al subconjunto B = {el resultado es un número par} = {2, 4, 6}, en cada secuencia individual se obtiene $f_{10}(B) = 0.5$ al igual que en la secuencia conjunta, $f_{20}(B) = 0.5$.

Así pues, calcular la frecuencia relativa de un subconjunto de posibles resultados, A, en N repeticiones de un experimento aleatorio resulta ser **otro experimento aleatorio** (un proceso de observación mediante el cual se selecciona un elemento del conjunto $\{0/N, 1/N, 2/N, ..., N/N, \}$). Afortunadamente, en muchas ocasiones las frecuencias relativas observadas en diferentes secuencias de experimentos parecen converger a un número muy preciso a medida que el número de repeticiones aumenta en cada secuencia, como se menciona a continuación.

8. Regularidad Estadística

La regularidad estadística es la propiedad que tienen muchos experimentos aleatorios según la cual, al repetir el experimento un gran número de veces bajo condiciones constantes, la frecuencia relativa de los posibles eventos tiende a un valor preciso a medida que aumenta el número de repeticiones.

Aceptar con humildad nuestra incapacidad de predecir el comportamiento de una red de comunicaciones no quiere decir que debamos considerar imposible el diseño de dichas redes con estrictos requerimientos de desempeño. Al contrario, lo que debemos hacer (y lo que han hecho los ingenieros de redes de comunicaciones en los últimos 150 años) es tratar de cuantificar la incertidumbre para así poder usarla como una herramienta a nuestro favor. Afortunadamente, muchos experimentos aleatorios presentan cierta regularidad estadística que facilitan la cuantificación de nuestra incertidumbre.

Considérese, por ejemplo, el experimento netstat de la Figura 15. Supongamos que después de haber observado el número de bytes recibidos durante n períodos de 10 segundos medimos la fracción de períodos en los que llegaron más de 8 Mbytes. Esta fracción es la *frecuencia relativa* del *evento* $E = \{x \in \mathbb{N} : x < 8000000\}$,

$$f_n(E) = \frac{1}{n}\sum_{i=1}^{n} 1\left(x_i \in E\right)$$

donde x_i es el número de bytes recibidos en el i-ésimo período de 10 s y $1(p)$ es la *función indicadora* de la proposición p, igual a 1 si la proposición p es cierta e igual a 0 si la proposición p es falsa. La Figura 17 muestra una gráfica de $f_n(E)$ vs n, en la que se puede apreciar cómo $f_n(E)$ parece tender a algún valor específico a medida que aumentamos el número de experimentos, n.

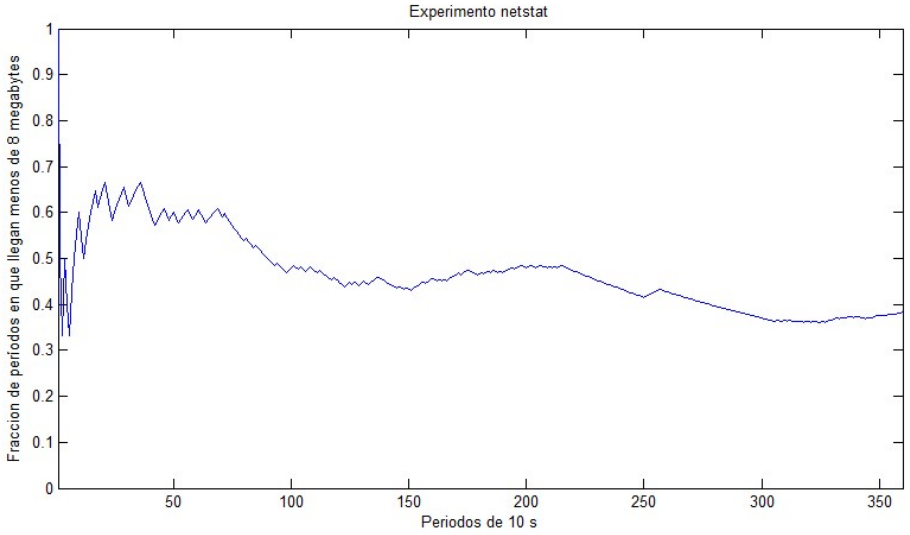


Figura 17. Aunque no podemos predecir el número de bytes que llegarán en el próximo período de 10 segundos, podemos afirmar que cerca del 38% del tiempo se reciben más de 8 Mbytes.

Es precisamente la regularidad estadística la que nos permite estudiar con rigurosidad los experimentos aleatorios que a diario tiene que realizar un ingeniero a cargo de una red de comunicaciones, pues ella nos permite saber que, *a la larga*, se pueden esperar comportamientos claramente predecibles. Por ejemplo, si la persona que estuvo navegando por la web durante los 60 minutos que duró el experimento netstat de la Figura 15 y de la Figura 17, sigue haciendo el mismo

tipo de consultas durante los siguientes 60 minutos, podríamos afirmar con un alto grado de certeza que “en cerca del 38% de los períodos de 10 s, se espera que lleguen más de 8 Mbytes”.

La teoría de la probabilidades pretende estudiar estas tendencias observadas en las estadísticas que se pueden asociar con un gran número de repeticiones de un experimento aleatorio, pero librándonos de términos imprecisos como “a la larga”, “se espera que”, “cerca de”, etc. Por ejemplo, la teoría de probabilidades querría que dijéramos que “con una confianza del 38%, en el próximo período de 10 s llegarán más de 8 Mbytes”. Así pues, es la regularidad estadística de muchos experimentos aleatorios la que le permite a la teoría de la probabilidad convertirse en una herramienta para cuantificar nuestra incertidumbre.

Es interesante notar que, cuando calculamos estadísticas como la frecuencia relativa de un evento, nos referimos a instancias ya realizadas del experimento aleatorio, sobre las cuales no hay ninguna incertidumbre. La parte más interesante es cuando queremos hacer "inferencia estadística", esto es, cuando queremos usar esas estadísticas para estudiar la incertidumbre que existe sobre instancias por realizar del experimento aleatorio. Es entonces cuando hablamos de la teoría de probabilidades.

9. Conceptos básicos sobre Conjuntos

Un conjunto es una colección de elementos. El conjunto de todos los elementos de interés en un contexto dado es el ***conjunto universal, Ω****. Si* ***a*** *es un elemento del conjunto* ***A****, decimos que* $a \in A$ *(a "pertenece a" A). En otro caso, decimos que* $a \notin A$ *(a "no pertenece a" A). Un conjunto está completamente determinado por sus elementos, al punto que dos conjuntos A y B son iguales (*$A=B$*) si tienen la propiedad que* $a \in A \Leftrightarrow a \in B$ *(a pertenece a A "si y sólo si" a pertenece a B). La propiedad* $a \in A \Rightarrow a \in B$ *("Si" a pertenece a A "entonces" a pertenece a B) define la relación* $A \subset B$ *(A "es un subconjunto de" B o A "está contenido en" B). En consecuencia, la igualdad entre los conjuntos A y B corresponde a las propiedades* $A \subset B$ *y* $B \subset A$*. Siempre es cierto que* $A \subset A$*. Si* $A \subset B$ *y* $B \subset C$*, entonces* $A \subset C$*. Podemos describir un conjunto especificando la propiedad que determina la pertenencia de un elemento:* $A = \{a \in \Omega: P(a)\}$ *("el conjunto A está compuesto por los elementos de Ω que satisfacen la propiedad P"). Como un conjunto está determinado exclusivamente por sus elementos, existe un único conjunto que no tiene elementos, el* ***conjunto vacío,*** $\phi = \{a \in \Omega: a \neq a\}$*. Para cualquier conjunto A,* $\phi \subset A$*.Si a es un elemento de A* ($a \in A$)*,* $\{a\}$ *es un* ***subconjunto unitario*** *de A* ($\{a\} \subset A$)*. Siempre es cierto que* $a \in \{a\}$ *aunque sería muy raro que* $a \in a$ *(Aquí evitaremos conjuntos que pertenezcan a sí mismos para no perdernos en paradojas como la* ***paradoja de Russell****: Si* $A = \{X : X \notin X\}$*,* $¿A \in A?$).

La ***unión*** *de dos conjuntos A y B, que se representa mediante* $A \cup B$*, es el conjunto de los elementos de Ω que pertenecen a A o que pertenecen a B (o que pertenecen a ambos):* $A \cup B = \{x \in \Omega : x \in A \vee x \in B\}$*. La* ***intersección*** *de dos conjuntos A y B, que se representa mediante* $A \cap B$*, es el conjunto de los elementos de Ω que pertenecen a A y que pertenecen a B:* $A \cap B = \{x \in \Omega : x \in A \wedge x \in B\}$*. Ambas operaciones son conmutativas y asociativas*

$(A\cup B = B\cup A$, $A\cap B = B\cap A$, $(A\cup B)\cup C = A\cup(B\cup C)$, $(A\cap B)\cap C = A\cap(B\cap C))$, además de ser distributivas entre ellas $(A\cap(B\cup C) = (A\cap B)\cup(A\cap C)$, $A\cup(B\cap C) = (A\cup B)\cap(A\cup C))$. Estas propiedades se extienden a cualquier secuencia de subconjuntos de Ω: $B\cap(\cup_n A_n)=\cup_n(A_n\cap B)$, $B\cup(\cap_n A_n)=\cap_n(A_n\cup B)$, donde $\cup_n A_n = A_1\cup A_2\cup A_3\cup\ldots$ y $\cap_n A_n = A_1\cap A_2\cap A_3\cap\ldots$. Claramente, $A\subset B \Rightarrow A\cap B=A$ y $A\cup B=B$. Dos conjuntos A y B son ***mutuamente excluyentes*** *si $A\cap B = \phi$.*

El ***complemento de A, A^C****, está compuesto por los elementos de Ω que no pertenencen a A: $A^C = \{b\in\Omega : b\notin A\}$. Evidentemente, $(A^C)^C = A$, $\phi^C=\Omega$, $\Omega^C=\phi$, $A\cap A^C=\phi$, $A\cup A^C=\Omega$. Las* ***leyes de DeMorgan*** *establecen que $(A\cup B)^C = A^C\cap B^C$ y que $(A\cap B)^C = A^C\cup B^C$. Estas leyes se extienden a cualquier secuencia de subconjuntos de Ω: $(\cup_n A_n)^C=\cap_n A_n{}^C$, $(\cap_n A_n)^C=\cup_n A_n{}^C$. Con uniones, intersecciones y complementos podemos definir la* ***diferencia $A\backslash B$*** *$= A\cap B^C = \{a\in A : a\notin B\}$ (el conjunto de los elementos de A que no pertenecen a B) y la* ***diferencia simétrica $A\Delta B$*** *$= (A\cap B^C)\cup(A^C\cap B) = \{a\in A\cup B : a\notin(A\cap B)\}$ (el conjunto de los elementos que pertenecen a A ó a B, pero no a ambos).*

Otra operación importante entre conjuntos es el ***producto cartesiano, $A\times B$*** *$= \{(a,b) : a\in A, b\in B\}$. Cada elemento del conjunto $A\times B$ es una* ***pareja ordenada*** *en la que el primer componente es un elemento de A y el segundo es un elemento de B, y $A\times B$ es el conjunto de todas esas parejas. Una* ***función $f:A\rightarrow B$*** *se construye como un subconjunto del producto cartesiano entre el dominio A y el rango B, donde a cada elemento de A le corresponde uno y sólo uno de los elementos de B, $(f:A\rightarrow B)\subset(A\times B)$.*

La ***cardinalidad*** *$|A|$ de un conjunto A es el número de elementos que pertenecen a A. Sólo hay un conjunto con cardinalidad cero, el conjunto vacío ($|\phi|=0$). Algunos conjuntos pueden tener una cardinalidad infinita contable (como el conjunto de los números racionales) o no contable (como el conjunto de los números reales). A veces existen resultados intuitivos con conjuntos finitos que no se pueden generalizar a conjuntos infinitos.*

Un conjunto finito con n elementos distintos se puede ordenar de n! maneras diferentes. Si sólo queremos k de ellos, hay n!/(n-k)! maneras de escogerlos en diferentes órdenes, esto es, hay n!/(n-k)! permutaciones de k elementos entre n. Si cualquier orden en que los escojamos da igual, obtenemos que k! de esas permutaciones son equivalente, de manera de hay n!/[(n-k)!k!] combinaciones de k elementes entre un conjunto de n elementos.

Una secuencia de subconjuntos de Ω, $\{A_n, n=1,2,\ldots\}$ es una ***partición*** *de Ω si $A_i\cap A_j = \phi\ \forall i\neq j$ y $\cup_n A_n = \Omega$. Una secuencia $\{B_n, n=1,2,\ldots\}$ es* ***decreciente*** *si, $\forall n$, $B_{n+1}\subseteq B_n$. Una secuencia $\{C_n, n=1,2,\ldots\}$ es* ***creciente*** *si, $\forall n$, $C_n\subseteq C_{n+1}$. El* ***límite de una secuencia decreciente*** *$\{B_n, n=1,2,\ldots\}$ es $\lim_{n\rightarrow\infty}B_n = \cap_{i=1\ldots\infty}B_i$. El* ***límite de una secuencia creciente*** *$\{C_n, n=1,2,\ldots\}$ es $\lim_{n\rightarrow\infty}C_n = \cup_{i=1\ldots\infty}C_i$. Dada cualquier secuencia de conjuntos $\{A_n\}$,*

podemos construir una secuencia decreciente $D_n = \cap_{i=1\ldots n} A_i$ *y una secuencia creciente* $C_n = \cup_{i=1\ldots n} A_i$. D_n *es el conjunto más grande que está contenido en todos los* $\{A_i, i=1\ldots n\}$ *mientras* C_n *es el conjunto más pequeño que contiene a todos los* $\{A_i, i=1\ldots n\}$. *Podemos tomar los respectivos límites* $D = \lim_{n\to\infty} D_n = \cap_{i=1\ldots\infty} D_i$ *y* $C = \lim_{n\to\infty} C_n = \cup_{i=1\ldots\infty} C_i$, *donde D es el conjunto más grande contenido en todos los conjuntos de la secuencia y C es el conjunto más pequeño que contiene a todos los conjuntos de la secuencia. Cuando consideramos una secuencia infinita de eventos* $A_1, A_2, A_3, \ldots \subset \Omega$, *podemos construir otras dos secuencias, una decreciente y una creciente así:*

$$\left\{ S_m = \sup A_{n\geq m} = \bigcup_{n=m}^{\infty} A_n \,, \quad m = 1,2,3... \right\} \qquad \left\{ I_m = \inf A_{n\geq m} = \bigcap_{n=m}^{\infty} A_n \,, \quad m = 1,2,3,... \right\}$$

S_m *es el* ***supremo*** *de* $\{A_m, A_{m+1}, A_{m+2}, \ldots\}$ (*el conjunto más pequeño que contiene a todos los* $A_{n\geq m}$) *e* I_m *es el* ***ínfimo*** *de* $\{A_m, A_{m+1}, A_{m+2}, \ldots\}$ (*el conjunto más grande contenido en todos los* $A_{n\geq m}$), *de manera que* $\{S_m, m=1,2,3\ldots\}$ *es una secuencia decreciente y* $\{I_m, m=1,2,3\ldots\}$ *es una secuencia creciente. A medida que m tiende a infinito, podemos construir el límite de cada secuencia,*

$$A = \limsup_{n\to\infty} A_n = \bigcap_{m=1}^{\infty} \bigcup_{k=m}^{\infty} A_k \qquad\qquad B = \liminf_{n\to\infty} A_n = \bigcup_{m=1}^{\infty} \bigcap_{k=m}^{\infty} A_k$$

Cuando ocurre el evento A en un experimento aleatorio con espacio muestral Ω, *decimos que ocurrió un número infinito de eventos de la secuencia* $\{A_n, n=1,2,\ldots\}$. *Cuando ocurre el evento B decimos que ocurrieron todos los eventos* $\{A_n, n=1,2,\ldots\}$, *con la posible excepción de un número finito de ellos,*

$$\limsup_{n\to\infty} A_n = \bigcap_{m=1}^{\infty} \bigcup_{k=m}^{\infty} A_k = \left\{ \omega \in \Omega : \forall n \exists k > n : \omega \in A_K \right\}$$

$$\liminf_{n\to\infty} A_n = \bigcup_{m=1}^{\infty} \bigcap_{k=m}^{\infty} A_k = \left\{ \omega \in \Omega : \exists n : \forall k > n : \omega \in A_K \right\}$$

Claramente, $\lim_{n\to\infty}\inf A_n \subseteq \lim_{n\to\infty}\sup A_n$. *Si también ocurre que* $\lim_{n\to\infty}\sup A_n \subseteq \lim_{n\to\infty}\inf A_n$, *entonces la secuencia de eventos* $A_1, A_2, A_3, \ldots \subset \Omega$ *tiene un límite, el evento L, dado por*

$$L = \lim_{n\to\infty}\sup A_n = \lim_{n\to\infty}\inf A_n$$

Una manera muy útil de visualizar las relaciones lógicas entre una colección finita de conjuntos es el ***diagrama de Venn***. *El espacio muestral se representa como un rectángulo y los eventos se presentan como áreas sombreadas dentro de él. Cuando hay demasiados conjuntos por representar en un diagrama de Venn se puede recurrir al mapa de Karnaugh, que representa en una tabla la partición constituida por todos los eventos de interés.*

Como podemos inferir de las definiciones 4 a 8, el lenguaje de la teoría de conjuntos es fundamental para la teoría de probabilidades. El concepto de conjunto es una "noción primitiva", es decir, un conjunto no puede definirse en términos de objetos definidos previamente. Algo semejante ocurre

con el concepto de punto, definido en la geometría Euclidiana como "lo que no tiene partes": La definición intuitiva de punto como una ubicación única del espacio que no tiene longitud, área, volumen ni ningún otro atributo dimensional, es el concepto primitivo sobre el cual se construye toda la geometría Euclidiana. Igualmente, la definición intuitiva de conjunto como una "colección de objetos" es el concepto primitivo sobre el cual se construye toda la teoría de conjuntos y, sobre ella, casi toda la matemática. Un niño en pre-escolar entiende fácilmente qué es un conjunto de lápices, así como un niño en primaria entiende qué es el conjunto de los mamíferos y un ingeniero entiende qué es el conjunto de los números reales (o al menos eso creemos). Cantor, Frege y otros matemáticos de finales del siglo XIX supusieron que estos conceptos ingenuos de la teoría de conjuntos eran suficientemente básicos, elementales e intuitivos como para construir toda la matemática sobre ella. Y, hasta cierto punto, Cantor y Frege no se equivocaron: La mayoría de las matemáticas de hoy se basan en la teoría de conjuntos, aunque dicha teoría no es tan elemental, tan básica ni tan intuitiva como se creyó en un principio, pues un uso poco juicioso del concepto de conjunto como "colección de objetos" puede conducir a contradicciones.

Básicamente, Cantor proponía que: (1) Un conjunto es una colección de objetos (los elementos del conjunto) que cumplen con cierta propiedad y, por consiguiente, el conjunto queda definido por la propiedad que determina a sus elementos. (2) Se pueden definir clases de conjuntos: conjuntos cuyos elementos son, a su vez, conjuntos que satisfacen cierta propiedad. (3) Dos conjuntos son iguales si tienen los mismos elementos. Sobre estas tres suposiciones, Frege construyó una rigurosa formalidad basada en sus propios desarrollos de lógica matemática y de lenguajes formales. Cuando, en 1902, Frege había culminado su monumental obra "Las Leyes Fundamentales de la Aritmética", en la que re-escribía toda la artitmética con base en la lógica y la teoría de conjuntos, Bertrand Russell le indicó una grave inconsistencia en su sistema lógico, la paradoja de Russell: ¿El conjunto de los conjuntos que no pertenecen a si mismos pertenece a si mismo? Si no, pertenece al tipo de conjuntos que no pertenecen a si mismos y, por lo tanto, pertenece a si mismo. Si sí, pertenece al tipo de conjuntos que pertenecen a si mismos y, por lo tanto, no pertenece a si mismo. Es decir, este conjunto pertenece a si mismo sólo si no pertenece a si mismo. Aunque son raros, hay conjuntos que pertenecen a si mismos. Por ejemplo, los principios de cantor permiten definir el conjunto $U = \{a : a=a\}$ o "el conjunto de todo", "el universo", para el cual $U \in U$, obviamente. El conjunto $C = \{a : a$ es un concepto matemático$\}$ es, en sí mismo, un concepto matemático, de manera que $C \in C$. El mismo Russell explica su paradoja más claramente: Sólo hay un barbero en la ciudad, quien dice que afeitará sólo a todos aquellos que no se afeiten a sí mismos. ¿Quién afeitará al barbero? Posteriormente Russell reformuló la aritmética sobre una nueva teoría de conjuntos en la que no se permiten los conjuntos que sean elementos de sí mismos pero, igualmente, Gödel encontró una inconsistencia insalvable en la formulación de la aritmética de Russell (el teorema de la incompletitud, que demuestra que en la formulación de Russell de la aritmética –y en cualquier otra formulación que sea recursiva y consistente– hay problemas indecidibles dentro del sistema, como el teorema que dice "este teorema no es demostrable").

En fin, la teoría de conjuntos debió desarrollarse cuidadosamente para evitar el tipo de contradicciones que aparecen fundamentalmente cuando se consideran conjuntos "demasiado grandes" como, por ejemplo, el conjunto de todas las cosas ($U = \{a : a = a\}$). Los matemáticos

Zermelo y Fraenkel redujeron la teoría de conjuntos a un sistema axiomático restringido que no permite paradojas como la de Russell. Existen otras formalizaciones axiomáticas exitosas como la propuesta posteriormente por John von Neumann, Paul Bernays y Kurt Gödel (NBG, una extensión de la de ZF). Ambas teorías se basan en unos principios básicos (axiomas) y unas reglas precisas de obtención de teoremas a partir de axiomas o teoremas previamente demostrados, de manera que no se presente ninguna contradicción (Al menos hasta el día de hoy no se ha encontrado ninguna contradicción… ¡pero aún no se sabe!). Sin embargo, a diferencia de Zermelo y Fraenkel, que siguen usando el conjunto como noción primitiva, Neumann, Bernays y Gödel no usan el concepto de conjunto como noción primitiva sino el concepto de clase: La clase es una "colección de objetos" y "un objeto pertenece a una clase". Un conjunto, en cambio, es un tipo especial de clase que está contenida en alguna otra clase. Así $U=\{a: a=a\}$ no es un conjunto sino una clase y a ella aplican sólo algunos axiomas NBG, pues otros axiomas aplican sólo a las clases que son conjuntos. De esta forma se evitan las paradojas conocidas hasta ahora.

Si evitamos los experimentos aleatorios en los que el espacio muestral mismo sea un posible resultado del experimento (¿el lector puede imaginar un experimento así?), no necesitaremos estudiar los formalismos axiomáticos de teorías de conjuntos. En cambio, nos basaremos en los conceptos intuitivos mencionados al comienzo como resumen de esta definición 9, y de los cuales no daremos más detalles que el ejemplo de operaciones básicas en diagramas de Venn de la Figura 18, que se pueden asociar con las operaciones lógicas entre predicados que involucran la ocurrencia de eventos, como muestra la Figura 19. De hecho, como se mencionó, una diagrama de Venn y un mapa de Karnaugh son equivalentes, como muestra la Figura 20. Las secuencias infinitas y sus límites los tratamos en las definiciones 23, 24 y 25.

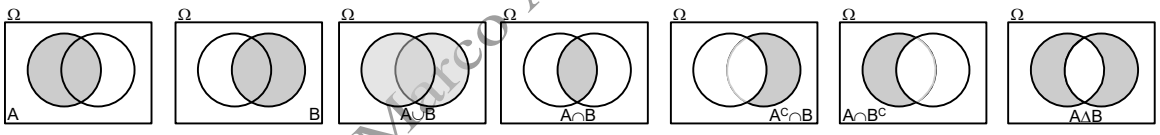


Figura 18. Diagram de Venn de algunas operaciones básicas entre dos conjuntos

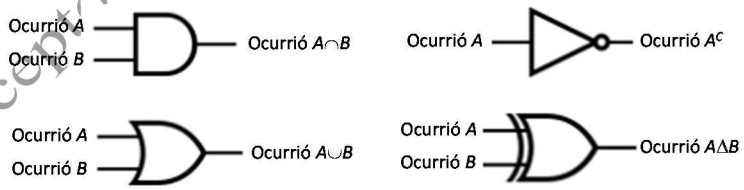
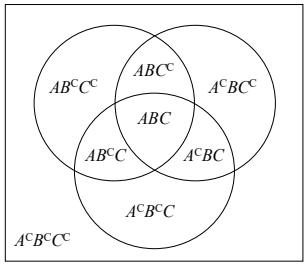


Figura 19. Operaciones entre conjuntos y sentencias lógicas sobre la ocurrencia de eventos



C \ AB	00	01	11	10
0	$A^CB^CC^C$	A^CBC^C	ABC^C	AB^CC^C
1	A^CB^CC	A^CBC	ABC	AB^CC

Figura 20. Comparación entre el diagram de Venn de tres subconjuntos y el mapa de Karnaugh de tres proposiciones lógicas (la ocurrencia o no ocurrencia de cada uno de los tres eventos asociados)

10. Conjunto Potencia de Ω, $\{0,1\}^{\Omega}$

El Conjunto Potencia de Ω es el conjunto de todos los posibles eventos, esto es, la clase de conjuntos conformada por todos los subconjuntos contenidos en Ω,

$$\{0,1\}^{\Omega} = \{A : A \subseteq \Omega\}.$$

En aquellos experimentos aleatorios en los que el espacio muestral tiene una cardinalidad finita, es legítimo pensar en enumerar todos los posibles eventos que pueden ocurrir, esto es, todos los posibles subconjuntos de Ω. Para construir esta clase de conjuntos basta con considerar todas las secuencias binarias de $|\Omega|$ bits, donde $|\Omega|$ es la cardinalidad de Ω, de manera que a cada posición en la secuencia le corresponde un elemento de Ω. Así, con cada secuencia construimos un subconjunto conformado por los elementos asociados con un uno en la posición correspondiente de la secuencia. Por ejemplo a la secuencia 0, compuesta por $|\Omega|$ ceros, le corresponde el conjunto vacío, que siempre es un subconjunto de cualquier conjunto; a la secuencia 2^{i-1}, con $i \in \{1,2,\ldots, |\Omega|\}$, compuesta por $|\Omega|$-1 ceros y un uno en la posición i, le corresponde el evento unitario $\{\omega_i\}$; a la secuencia $2^{i-1} + 2^{j-1}$, con $i,j \in \{1,2,\ldots, |\Omega|\}$, $i \neq j$, compuesta por $|\Omega|$-2 ceros y dos unos en las posiciones i y j, le corresponde el evento binario $\{\omega_i, \omega_j\}$; a la secuencia $2^{|\Omega|}$-1, compuesta por $|\Omega|$ unos, le corresponde el espacio muestral mismo que, por definición, es un subconjunto de sí mismo. Debido a esta metodología de construcción, es razonable que al conjunto potencia del espacio muestral Ω se le denote como $\{0,1\}^{\Omega}$. Más aún, como en $\{0,1\}^{\Omega}$ hay un conjunto vacío, $|\Omega|$ conjuntos unitarios, $\binom{|\Omega|}{2}$ conjuntos binarios –donde $\binom{m}{k}$ es el número de combinaciones de k elementos escogidos entre m posibles–, $\binom{|\Omega|}{3}$ conjuntos ternarios, etc., la cardinalidad de $\{0,1\}^{\Omega}$ es $\left|\{0,1\}^{\Omega}\right| = \sum_{n=0}^{|\Omega|}\binom{|\Omega|}{n} = 2^{|\Omega|}$.

En el experimento 4 de la definición 5, por ejemplo, en el que verificamos el estado de ocupación de un canal de comunicaciones, solamente hay dos posibles resultados, por lo que tenemos solamente cuatro posibles eventos:

Libre	*Ocupado*	*Evento*
0	0	Φ
0	1	{*Ocupado*}
1	0	{*Libre*}
1	1	Ω

Pero si viéramos el estado de ocupación de dos canales, considerando cada uno individualmente, tendríamos 16 posibles eventos:

(libre, libre)	*(libre, ocupado)*	*(ocupado, libre)*	*(ocupado, ocupado)*	*Evento*
0	0	0	0	Φ
0	0	0	1	{*No hay canales libres*}
0	0	1	0	{(*ocupado*, *libre*)}

0	0	1	1	{*El primer canal está ocupado*}
0	1	0	0	{(*libre*, *ocupado*)}
0	1	0	1	{*El segundo canal está ocupado*}
0	1	1	0	{*Sólo hay un canal libre*}
0	1	1	1	{*Al menos un canal está ocupado*}
1	0	0	0	{*No hay canales ocupados*}
1	0	0	1	{*Ambos canales están libres o ambos están ocupados*}
1	0	1	0	{*El segundo canal está libre*}
1	0	1	1	{*El primer canal está ocupado o el segundo está libre*}
1	1	0	0	{*El primer canal está libre*}
1	1	0	1	{*El primer canal está libre o el segundo está ocupado*}
1	1	1	0	{*Al menos un canal está libre*}
1	1	1	1	Ω

Con tres canales tendríamos 256 posibles eventos y con cuatro canales deberíamos considerar 65536 eventos… En el experimento 8, por ejemplo, en el que inocentemente queremos ver el estado de ocupación de cada canal en un enlace E1, tendríamos $2^{4.294'967.296}$ posibles eventos, ¡más de $10^{1.292'913.986}$ eventos (un uno seguido por mil trescientos millones de ceros)! Para hacernos a una idea de la cardinalidad de este conjunto, consideremos que en el universo hay del orden de 2^{265} átomos. Si nos regalan un átomo por cada subconjunto de este espacio muestral, necesitaríamos $2^{4.294'967.031}$ universos como éste. En ese número de universos es altamente probable encontrar uno idéntico al nuestro, excepto porque nuestro amable lector tendría otro color de ojos ¿Cómo nos pudimos meter en un problema tan grande si sólo queríamos monitorear un simple enlace E1?

De hecho, aunque observar los 32 canales de un inofensivo enlace E1 puede generar un número mucho más que astronómicamente grande de posibles eventos, todavía se trata de un conjunto describible (tanto que podemos contar cada uno de los eventos). Pero, ¿podría el lector imaginarse el conjunto potencia del experimento 10? Los más de $10^{1.000'000.000}$ posibles eventos del experimento 8 siguen siendo un número infinitesimalmente pequeño de eventos en comparación con la cardinalidad del espacio muestral del número de transmisiones que se deben hacer para que una trama llegue bien a su destino, pues ese espacio muestral ya tiene un número infinito de elementos, aunque sea "el infinito más pequeño", $\aleph_0$, que es la cantidad de números naturales (un conjunto tiene cardinalidad infinita si se puede poner en correspondencia biunívoca con un subconjunto propio de si mismo. Por ejemplo, los naturales son infinitos porque se pueden poner en correspondecia biunívoca con los pares y los reales son infinitos porque se pueden poner en correspondencia biunívoca con el intervalo unitario). Como nos demuestra la paradoja del "Hotel de Hilbert"[5], $\aleph_0 = \aleph_0 + n = n\aleph_0 = (\aleph_0)^n \ \forall n \in \mathbb{N}$.

[5] El único hotel de un pueblo pequeño tiene infinitas habitaciones ($|\mathbb{N}| = \aleph_0$ habitaciones). En el pueblo se juega un campeonato de fútbol en el que cada equipo tiene un número infinito de jugadores. Cuando llega el bus con el primer equipo, se ubica al jugador n en la habitación $n \ \forall n \in \mathbb{N}$, de manera que todas las habitaciones quedan

Pero, como demostró Georg Cantor, el número de subconjuntos que se pueden construir con los números naturales es mayor que el número de números naturales, $\aleph_0 < \aleph_1 = 2^{\aleph_0}$, donde $\aleph_1$ es la cantidad de números reales que existen[6]. El conjunto potencia del espacio muestral del experimento 10 tiene tantos eventos como números reales existen!

Más aún, ¿podría el lector imaginar el conjunto potencia del experimento 14? ¡Es el conjunto de todos los subconjuntos que se pueden formar con los números reales no negativos! Si los famélicos 32 canales de un enlace E1 lograrán atemorizarnos de hoy en adelante cada vez que pasemos cerca de la pequeña PBX de la oficina, y temblaremos de terror al tener que contar el número de transmisiones que requiere una trama ¿qué podría hacernos un experimento cuyo espacio muestral sea el conjunto de los números reales? El número de subconjuntos que se pueden formar con los números reales es $\aleph_2 = 2^{\aleph_1} > \aleph_1$. Grandes matemáticos como Bolzano, Cauchy, Weierstrass, Dedekind y Cantor han estudiado estos "monstruos matemáticos", los "trans-finitos", algunos de ellos con apreciables consecuencias en su salud mental. Como nos preocupa la salud mental de nuestros lectores, resulta conveniente definir el siguiente concepto, campo-σ de eventos. Esto es, en vez de pretender que se atemorice la próxima vez que vaya a revisar el PBX de la oficina, sólo queremos motivar al lector a seleccionar un conjunto razonablemente pequeño de eventos de interés cada vez que decida modelar un experimento aleatorio (donde "pequeño" puede ser $\aleph_0$ ó $\aleph_1$, pero no $\aleph_2$ ni ningún otro trans-finito mayor a $\aleph_1$).

11. Campo-σ de Eventos

Un Campo de Eventos, $\mathcal{F}$, es una clase de subconjuntos de Ω que satisface los siguientes axiomas: (1) *$\mathcal{F}$ es no vacío*, (2) *si $A \subset \Omega$ es tal que $A \in \mathcal{F}$, $A^C \in \mathcal{F}$,* (3) *si $A,B \subset \Omega$ son tales que $A,B \in \mathcal{F}$, $A \cup B \in \mathcal{F}$. Un campo-σ de eventos es un campo contablemente aditivo, esto es, que satisface la condición adicional* (3ª) *si $\{A_n \in \mathcal{F}, n=1,2,...\}$, $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.*

ocupadas. Al final se baja el entrenador… ¿dónde ubicar al entrenador? Se le solicita al jugador que está en la habitación n que, por favor, se pase a la habitación $n+1$ $\forall n \in \mathbb{N}$, con eso todos los jugadores siguen con una habitación pero se desocupa la habitación 1 para el entrenador: $\aleph_0 = \aleph_0 + 1 = \aleph_0 + k$ $\forall k \in \mathbb{N}$. Pero resulta que cada jugador trajo a su mamá… ¿dónde ubicamos a las mamás? Se le solicita al jugador que está en la habitación n que, por favor, se pase a la habitación $2n$ $\forall n \in \mathbb{N}$, con eso todos siguen con una habitación pero se desocupan las habitaciones impares para las mamás: $\aleph_0 = 2\aleph_0 = k\aleph_0$ $\forall k \in \mathbb{N}$. Pero resulta que el campeonato se juega entre un número infinito de equipos y empiezan a llegar los buses con los demás equipos. ¿dónde ponemos a todos esos jugadores? Le pedimos al jugador que está en la habitación n que se pase a la habitación 2^n y ubicamos al m-ésimo jugador del r-ésimo equipo en la habitación $p(r)^m$, donde $p(r)$ es el r-ésimo número primo mayor a 2: $\aleph_0 = \aleph_0^2 = \aleph_0^k$ $\forall k \in \mathbb{N}$. Nótese que ahora todos los jugadores de todos los equipos tienen su habitación pero, a diferencia de cuando acomodamos sólo al primer equipo y llenamos el hotel, ¡ahora tenemos un número infinito de habitaciones desocupadas (1, 6, 10, 12, 14, 15, 18, 20, 21, 22, 24…)!

[6] Si podemos contar los puntos en el intervalo unitario [0,1], los tendremos a todos en una correspondencia biunívoca con los naturales en una lista $\{s_1, s_2, s_3, \ldots\}$, donde el i-ésimo número es de la forma $s_i = 0 \cdot x_{i1}x_{i2}x_{i3}x_{i4}\ldots$ siendo x_{ij} el j-ésimo dígito de la expansión binaria de s_i. Pero esa lista es imposible porque en ella hará falta el número $0 \cdot y_1y_2y_3y_4\ldots$ con $y_1 \neq x_{11}$, $y_2 \neq x_{22}$, $y_3 \neq x_{33}$,… En consecuencia, el número de puntos en el intervalo unitario es mayor que el número de naturales, $\aleph_1 > \aleph_0$. $\aleph_1$ es el número de números reales, que es igual al número de secuencias inifnitas de unos y ceros, que es el mismo número de subconjuntos de los naturales, $\aleph_1 = 2^{\aleph_0}$.

La idea es que más adelante vamos a definir la probabilidad como una función que le asigna una medida real a cada evento de interés. Pero una función no queda bien definida si no especificamos claramente su rango y su dominio. Y, como vimos en la definición 10, no podemos especificar como dominio el conjunto de todos los posibles eventos, pues en muchos casos ese conjunto puede ser monstruoso. Sólo cuando el espacio muestral tiene cardinalidad finita (¡y pequeña!), es posible considerar el conjunto de todos los eventos, el cual es un campo-σ, evidentemente. Pero tampoco podemos seleccionar algunos pocos eventos de interés e ignorar el resto si no le damos una estructura al dominio correspondiente, con el que evitemos llegar rápidamente a inconsistencias. Si nos interesa el evento $A \subseteq \Omega$, ¿cómo no nos podría interesar el evento $A^C = \{No\ sucede\ A\}$? O si nos interesan los eventos A y $B \subseteq \Omega$, ¿cómo no nos podría interesar el evento $A \cup B = \{sucede\ por\ lo\ menos\ uno\ de\ los\ dos\ eventos\}$? Al cerrar el campo de eventos sobre las uniones y los complementos, estamos incluyendo en él todos los eventos asociados con los eventos de interés definidos originalmente, con lo cual podemos asignar medidas de probabilidad a cada evento sin preocuparnos por inconsistencias.

Podemos deducir algunas propiedades adicionales de un campo-σ de eventos a partir de los axiomas que lo definen. Por ejemplo,

1. $\Omega \in \mathcal{F}$.
   En efecto, como $\mathcal{F}$ es no-vacío, debe contener al menos un evento A y, por el segundo axioma, también debe contener a A^C. El tercer axioma requiere que la unión de cualquier par de miembros de $\mathcal{F}$ pertenezca también a $\mathcal{F}$, por lo que $A \cup A^C = \Omega \in \mathcal{F}$.
2. $\Phi \in \mathcal{F}$.
   Esta propiedad surge de aplicar el segundo axioma a la propiedad anterior.
3. Si $A \in \mathcal{F}$ y $B \in \mathcal{F}$, $A \cap B \in \mathcal{F}$.
   En efecto, por el segundo axioma $A^C \in \mathcal{F}$ y $B^C \in \mathcal{F}$, por lo que el segundo axioma asegura que $A^C \cup B^C \in \mathcal{F}$ y, aplicando nuevamente el segundo axioma, $(A^C \cup B^C)^C = A \cap B \in \mathcal{F}$ (ley de DeMorgan).
4. Similarmente, usando los axiomas 2 y 3ª, podemos decir que si $\{A_n \in \mathcal{F},\ n=1,2,\ldots\}$, $\bigcap_{n=1}^{\infty} A_n \in \mathcal{F}$.

Así pues, el campo-σ contiene todos los complementos, intersecciones numerables y uniones numerables de cada uno de los conjuntos que lo componen. La virtud de esta construcción es que, con ella, podemos desarrollar todas las funciones lógicas Booleanas sobre los eventos de interés, lo cual nos da la coherencia que necesitamos para expresarnos de manera lógica respecto a los eventos sin salirnos de nuestro campo de eventos, pues estos constituyen una estructura lógica cerrada para la negación, la conjunción y la disyunción. Si nos interesan los eventos $A,B,C \subseteq \Omega$, ¿cómo no nos podría interesar el evento $A \cap B = \{suceden\ ambos\ eventos\}$? ¿O el evento $A \cap B^C = \{sucede\ A\ pero\ no\ sucede\ B\}$? ¿O el evento $(A \cap C^C) \cup (B \cap C \cap A^C) = \{Sucede\ A\ pero\ no\ sucede\ C\ o\ suceden\ B\ y\ C\ pero\ no\ sucede\ A\}$? En un campo sigma ocurren todas las operaciones lógicas posibles sobre la ocurrencia o no de cualquier conjunto contable de eventos de interés.

El campo-σ trivial, el más pequeño posible, es $\{\Phi,\Omega\}$. Otro campo sigma trivial, el más grande posible, es el conjunto potencia $\{0,1\}^{\Omega}$. Si sólo nos interesa saber si ocurrió o no un único evento $A \subset \Omega$, un campo sigma adecuado es $\{\Phi,A,A^C,\Omega\}$. El conjunto de todas las uniones de conjuntos en una partición contable de Ω es un campo sigma. Si Ω es un conjunto infinito no contable, el conjunto

de subconjuntos de Ω que son contables o cuyo complemento es contable forma un campo-σ que contiene a todos los conjuntos unitarios de Ω, pero que es infinitas veces más pequeño que el conjunto potencia de Ω.

Antes de mostrar algunos ejemplos adicionales, es conveniente incluir una definición adicional.

12. Mínimo Campo-σ de Eventos

Dada una clase de eventos $\mathcal{C} \subseteq \{0,1\}^{\Omega}$, el mínimo campo-σ de eventos que contiene a $\mathcal{C}$, $\sigma(\mathcal{C})$, es el campo-σ de menor cardinalidad entre todos los campos-σ que lo contienen.

Como sugeríamos al concluir la definición 10, una vez escogidos el experimento aleatorio y su espacio muestral Ω, lo siguiente por hacer es seleccionar una clase de eventos de interés, $\mathcal{C}$, y, con esta clase, construir el mínimo campo-σ que contiene a todos los eventos en $\mathcal{C}$. Este mínimo campo-σ se representa mediante $\sigma(\mathcal{C})$ y se "puede construir" así: Si denotamos $\mathcal{H}$ como el conjunto de campos-σ que contienen a $\mathcal{C}$, podemos saber que $\mathcal{H}$ no es vacío pues por lo menos $\{0,1\}^{\Omega}$ es un campo-σ de subconjuntos de Ω que contiene a $\mathcal{C}$. Definiendo $\sigma(\mathcal{C})$ como la intersección de todos los campos-σ en $\mathcal{H}$, sabremos que $\sigma(\mathcal{C})$ es el mínimo campo-σ que contiene a $\mathcal{C}$. En efecto, si $\mathcal{F}_1$ y $\mathcal{F}_2$ son dos campos-σ que contienen a $\mathcal{C}$, $\mathcal{F}_1 \cap \mathcal{F}_2 = \{A \in \mathcal{F}_1 : A \in \mathcal{F}_2\}$ es una clase de eventos que también contiene a $\mathcal{C}$ (pues cada evento en $\mathcal{C}$ está tanto en $\mathcal{F}_1$ como en $\mathcal{F}_2$) y que forma un campo-σ porque no es vacío (al menos Ω y Φ pertenecen a ambos), si el evento A pertenece a ambos campos, el evento A^C también pertenece a ambos campos, y si los eventos A y B pertenecen a ambos campos, el evento $A \cup B$ también pertenece a ambos campos. Por supuesto, $|\mathcal{F}_1 \cap \mathcal{F}_2| \leq \min(|\mathcal{F}_1|, |\mathcal{F}_2|)$, por lo que la intersección de todos los campos en $\mathcal{H}$ nos da el mínimo campo-σ, $\sigma(\mathcal{C})$. Siendo así, si $\{\mathcal{F}_n , n=1,2,\ldots\} \in \mathcal{H}$ es el conjunto de todos los campos-σ que contienen a $\mathcal{C}$, $\sigma(\mathcal{C})$ se puede definir como $\sigma(\mathcal{C}) = \bigcap_n \mathcal{F}_n$ … Sí, debemos admitir que éste no es un procedimiento muy práctico para determinar el mínimo campo-σ que contiene a $\mathcal{C}$ (puede ser que la intersección ni siquiera sea contable), pero al menos así podemos estar seguros que dicho mínimo campo-σ existe.

Como ejemplo, consideremos algunos posibles campos-σ definidos en los 10 primeros experimentos propuestos en las definiciones 5 y 6.

1. La sencillez del espacio muestral del experimento que consiste en lanzar una moneda y ver qué lado queda hacia arriba sugiere que un campo-σ apropiado es el conjunto potencia del espacio muestral. Después de todo, la cardinalidad de dicho campo-σ es solamente 4.
2. En el experimento de lanzar un dado y contar los puntos en la cara que queda hacia arriba incluimos los conjuntos A = {*hay más de tres puntos*} = {4,5,6} y B = {*hay un número par de puntos*} = {2,4,6}, con los cuales se puede construir el siguiente campo-σ en Ω = {1,2,3,4,5,6}: $\mathcal{F}$ = {Φ, {5}, {4,6}, {4,5,6}, {2}, {2,5}, {2,4,6}, {2,4,5,6}, {1,3}, {1,3,5}, {1,3,4,6}, {1,3,4,5,6}, {1,2,3}, {1,2,3,5}, {1,2,3,4,6}, Ω}

La cardinalidad de este campo-σ es 16, menor a los 64 eventos del conjunto potencia. Obsérvese que otro campo-σ que también contiene a la clase de eventos $\mathcal{C}=\{A,B\}$ es el siguiente:
$\mathcal{E}$ = {Φ, {6}, {5}, {5,6}, {4}, {4,6}, {4,5}, {4,5,6}, {2}, {2,6}, {2,5}, {2,5,6}, {2,4}, {2,4,6}, {2,4,5}, {2,4,5,6}, {1,3}, {1,3,6}, {1,3,5}, {1,3,5,6}, {1,3,4}, {1,3,4,6}, {1,3,4,5}, {1,3,4,5,6}, {1,2,3}, {1,2,3,6}, {1,2,3,5}, {1,2,3,5,6}, {1,2,3,4}, {1,2,3,4,6}, {1,2,3,4,5}, Ω}
cuya cardinalidad, 32, sigue siendo menor a $|\{0,1\}^{\Omega}|$=64. Sin embargo $\mathcal{F}=\sigma(\mathcal{C})$ es el mínimo campo-σ que incluye a los eventos de $\mathcal{C}$. De hecho, nótese que $\mathcal{F}=\mathcal{F}\cap\mathcal{E}$. Para construir un modelo probabilístico de este experimento en donde sólo interesen los eventos en $\mathcal{C}$, es suficiente con asignar medidas de probabilidad a cada uno de los eventos de $\mathcal{F}$, y no hace falta asignarle probabilidades a cada evento en $\mathcal{E}$ ni mucho menos a cada evento en $\{0,1\}^{\Omega}$.

3. En el experimento de escoger una carta de la baraja de naipes definimos los eventos A = {*Una figura de pinta roja*} = {*corazones, diamantes*}×{*J,Q,K*} y B = {*Un as negro*} = {(*picas*,1), (*tréboles*,1)}. En este caso, como los eventos son excluyentes (no pueden suceder simultáneamente), el mínimo campo-σ de eventos que incluye a A y B es bastante pequeño: $\mathcal{F}$={Φ, A, B, $A\cup B$, A^C, B^C, $(A\cup B)^C$, Ω}. Por supuesto, el conjunto potencia tiene 2^{52} eventos, ¡más de cuatro mil billones (cuatro mil millones de millones)!
4. En el experimento de medir la fracción de paquetes perdidos durante una hora en una red IP teníamos como espacio muestral el intervalo [0,1] de la recta real, donde resulta imposible definir el conjunto potencia. Si definimos una familia de eventos $\mathcal{E}$ compuesto por los intervalos cerrados {[0, x], $x\leq 1$}, podríamos considerar el mínimo campo-σ que contiene a $\mathcal{E}$, $\sigma(\mathcal{E})$. Este conjunto se llama el campo de Borel del intervalo [0,1], $\mathcal{B}([0,1])$ –ver enseguida la definición 13– y, aunque es difícil de describir, sabemos que también contiene todos los intervalos abiertos, semiabiertos, cerrados, puntos aislados y uniones contables de dichos eventos… ¡Todo lo que nos pueda interesar! Claro, hay muchos subconjuntos de [0,1] que no están en $\mathcal{B}([0,1])$ –¡la mayoría!- pero son tan "raros" para nuestros propósitos de modelar la fracción de paquetes perdidos, que no nos interesa incluirlos en nuestro campo-σ de eventos (¡afortunadamente, porque ni siquiera nos queda fácil imaginarlos!).
5. En el experimento de medir el retardo experimentado por un paquete de voz mientras transita por una red VoIP, puede que sólo nos interese el evento $A = \{x\in\Omega : x > 0.1\}$ = {*El paquete no alcanza a ser reproducido en el receptor*}, en cuyo caso el campo-σ de eventos sería elemental: $\mathcal{F}$={Φ, A, A^C, Ω}, a pesar de que el espacio muestral está compuesto por los reales no negativos.
6. Al verificar el estado de ocupación de un canal de comunicaciones los posibles eventos de interés son los subconjuntos unitarios {*libre*} y {*ocupado*}. En este caso, el conjunto potencia resulta un campo-σ perfecto para trabajar: $\{0,1\}^{\Omega}$ = {Φ, {*libre*}, {*ocupado*}, Ω}
7. En el experimento de contar el número de canales libres en un enlace E1 podríamos estar interesados en los siguientes dos eventos: A = {*Se puede establecer una videoconferencia*} = {6,7,8,…,32} y B = {*Se puede transmitir video MPEG-4 a por lo menos* 768 *kbps*} = {12,13,14,…, 32}. En este caso, como el evento A incluye al evento B, $\sigma(\{A,B\})$={Φ, A, B, A^C, B^C, $A^C\cup B$, $A\cap B^C$, Ω}.
8. En el experimento de mirar el estado de ocupación de cada uno de los canales en un enlace E1 podemos definir los siguientes 33 eventos [X_i = {*Hay* i *canales libres*}, i=0,1,2,…,32]. Como se trata de eventos mutuamente excluyentes, el mínimo campo-σ tendría sólo 2^{33} eventos de interés.

Los eventos de este mínimo campo-σ se podrían asociar, en una correspondencia uno-a-uno, con los eventos del conjunto potencia del experimento 7.

9. Como en los experimentos 1 y 4, el conjunto potencia es un campo-σ apropiado al determinar si un bit transmitido sobre un canal de comunicaciones llega correctamente al lado receptor del canal.
10. Al contar el número de transmisiones (a través de un canal ruidoso) que requiere un paquete de datos hasta llegar correctamente a su destino podemos estar interesados sólo en los eventos A={*No hay errores de transmisión*} = {1} y B={*Mejor desistir de seguir intentándolo*} = {16,17,18,…}, con los que se puede construir un pequeño campo-σ con sólo ocho eventos: $\mathcal{F}$ = {Φ, A, B, $A\cup B$, A^C, B^C, $(A\cup B)^C$, Ω}.
11. Al contar el número de bits con errores en una trama de L bits que se recibe de un canal ruidoso podríamos estar interesado en el evento en que no se dañó ningún bit, A = {0}. Un campo-σ apropiado es {Φ, A, A^C, Ω}, donde A^C indica la necesidad de retransmitir la trama.
12. Al medir durante una hora la fracción de tiempo que un enlace de comunicaciones permanece ocupado, dos eventos de interés serían A = {$x\in\Omega$: $x \geq 0.8$}, que indica congestión, y B={$x\in\Omega$: $x \leq 0.2$}, que indica subutilización. Como A y B son mutuamente excluyentes, el mínimo campo-σ tiene 8 posibles eventos: $\mathcal{F}$ = {Φ, [0.8, 1], [0, 0.2], (0.2, 0.8), (0.2, 1], [0, 0.8), [0.8, 1]∪[0, 0.2], Ω}.
13. Contar el número de paquetes que llegan a un enrutador de una red de comunicaciones durante una hora. Como mencionamos, el espacio muestral es un subconjunto de los enteros no negativos, Ω={0,1,2,3,…M}, aunque el valor preciso de M puede ser difícil de establecer ya que depende del número de enlaces de entrada, el número de canales en cada enlace, la longitud de los paquetes, etc. Para no correr el riesgo de establecer un valor de M y que lleguen M+1 paquetes, una medida segura es suponer que pueden llegar infinitos paquetes, $\Omega=\mathbb{Z}^+$. En este caso, la cardinalidad de Ω es el infinito contable, $\aleph_0$. En este caso, el conjunto potencia tiene $\aleph_1 = 2^{\aleph_0}$ eventos y ¡sigue siendo un campo sigma manejable!
14. Medir el tiempo que transcurre entre la llegada de dos paquetes consecutivos a un enlace de una red de comunicaciones. Ahora el espacio muestral es el conjunto de reales no negativos y no vale la pena intentar considerar su conjunto potencia. Nuevamente, si sólo nos interesa saber si transcurrieron menos de 100 ms o no, podemos considerar el campo sigma de eventos {Φ, A, A^C, Ω}, donde A = {el tiempo transcurrido es menor a 100 ms}. Sin embargo, es posible que nos interesen, por ejemplo, todos los eventos de la forma $A(x)$ = {$\omega\in\mathbb{R}$: $-\infty < \omega \leq x$}, $x\in\mathbb{R}$. Si es así, necesitamos encontrar el mínimo campo sigma que incluye a los eventos {$A(x)$, $x\in\mathbb{R}$}, al que definiremos en el siguiente apartado.

En los 12 primeros ejemplos anteriores nos interesaba un número finito o contable de eventos y resultaba fácil construir con ellos una partición del espacio muestral y considerar como campo-σ todas las uniones de los eventos de la partición[7]. Cuando los eventos de interés son infinitos contables, como en el ejemplo 13, podríamos imaginar un procedimiento parecido. Pero cuando los eventos de

[7] Si {C_i, $i\in I$} es una partición de Ω, $\mathcal{F}$ = {$A_J=\cup_{i\in J}C_i$, $J\subset I$} es un campo-σ, como se puede demostrar fácilmente.

interés son infinitos no contables, como en el último ejemplo, debemos considerar campos más abstractos, como muestra la siguiente definición.

13. Campo-σ de Borel de los Reales, $\mathcal{B}(\mathbb{R})$

El campo-σ de Borel de los números reales, $\mathcal{B}(\mathbb{R})$, es el mínimo campo-σ que contiene a todos los intervalos semi-infinitos de la forma $A_x = \{\omega \in \mathbb{R} : -\infty < \omega \leq x\}$, $x \in \mathbb{R}$. Los subconjuntos de $\mathbb{R}$ que pertenecen a $\mathcal{B}(\mathbb{R})$ se denominan "conjuntos de Borel".

En muchas ocasiones el espacio muestral de nuestros experimentos será el conjunto de los números reales, por lo que se hace muy importante definir un campo sigma de eventos "sencillo" que involucre todos los eventos "razonables" que nos puedan interesar. Como veremos enseguida, el campo de Borel de los números reales incluye a todos los intervalos cerrados, abiertos, semiabiertos, finitos ó semi-infinitos, incluyendo todos los puntos aislados. Cuando nos limitamos a uniones numerables de este tipo de eventos en $\mathbb{R}$, podemos construir un espacio de probabilidad coherente sobre el cual podremos aplicar toda la lógica booleana sin llegar a inconsistencias.

Veamos qué tipos de eventos se incluyen en $\mathcal{B}(\mathbb{R})$:

1. $(-\infty,x] \in \mathcal{B}(\mathbb{R})\ \forall x \in \mathbb{R}$, por definición
2. Aplicando el segundo axioma a los eventos anteriores,
   $(-\infty,x]^C = (x,\infty) \in \mathcal{B}(\mathbb{R})\ \forall x \in \mathbb{R}$
3. Como $(-\infty,b]$ y (a,∞) pertenecen a $\mathcal{B}(\mathbb{R})$,
   $(-\infty,b] \cap (a,\infty) = (a,b] \in \mathcal{B}(\mathbb{R})\ \forall a \in \mathbb{R},\ b \in \mathbb{R},\ a<b$.
4. De acuerdo con el punto anterior, $(a-1/n,\ a] \in \mathcal{B}(\mathbb{R})$ y, como un campo-σ es cerrado para las intersecciones contables,
   $$\bigcap_{n=1}^{\infty}\left(a-\frac{1}{n},a\right]=[a] \in \mathcal{B}(\mathbb{R})\ \forall a \in \mathbb{R}$$
5. De los puntos (3) y (4), $[a] \cup (a,b] = [a,b] \in \mathcal{B}(\mathbb{R})\ \forall a \in \mathbb{R},\ b \in \mathbb{R},\ a<b$.
6. $[b]^C \cap (a,b] = (a,b) \in \mathcal{B}(\mathbb{R})\ \forall a \in \mathbb{R},\ b \in \mathbb{R},\ a<b$.
7. $[a] \cup (a,b) = [a,b) \in \mathcal{B}(\mathbb{R})\ \forall a \in \mathbb{R},\ b \in \mathbb{R},\ a<b$.

Si todos los intervalos (abiertos, cerrados, semiabiertos, semi-infinitos) y todos los puntos aislados son conjuntos de Borel, al igual que las uniones e intersecciones numerables de dichos subconjuntos, ¿Puede haber algún subconjunto de $\mathbb{R}$ que no sea un conjunto de Borel? Los hay, y son muchos más los subconjuntos de $\mathbb{R}$ que no son de Borel que aquellos que sí son de Borel ¿Qué tal, por ejemplo, el conjunto de los racionales, $\mathbb{Q}$? Existe un conjunto infinito de números reales entre cada par de números racionales, por lo que $\mathbb{Q}$ forma un conjunto de puntos aislados, cada uno de los cuales es un conjunto unitario de Borel por el punto 4. Sin embargo, sólo la unión contable de conjuntos de Borel forma un conjunto de Borel… ¡Pero $\mathbb{Q}$ es un conjunto de cardinalidad contable porque, como vimos en la definición 10, $\aleph_0 = \aleph_0^2$. ¿Y los irracionales? Ellos forman el complemento de $\mathbb{Q}$ en $\mathbb{R}$ así que, a pesar de ser la unión incontable de puntos aislados (hay infinitos racionales entre cada par de irracionales), los irracionales siguen siendo un conjunto de Borel debido a la segunda propiedad de

los campos sigma ($\mathcal{B}(\mathbb{R})$ es cerrado para el complemento). Veamos otro subconjuto de $\mathbb{R}$ aún más extraño, el conjunto ternario de Cantor:

Comenzando con el intervalo cerrado [0,1], extraemos de él el segmento central correspondiente al intervalo abierto (1/3, 2/3), dejando los dos intervalos cerrados [0,1/3] y [2/3,1]. A cada uno de estos intervalos le extraemos los respectivos segmentos centrales (1/9, 2/9) y (7/9, 8/9), dejando cuatro intervalos cerrados [0,1/9], [2/9, 3/9], [6/9, 7/9] y [8/9,1]. A cada uno de estos intervalos le extraemos los respectivos segmentos centrales (1/27, 2/27), (7/27, 8/27), (19/27, 20/27) y (25/27, 26/27), dejando ocho intervalos cerrados. Así seguimos repitiendo el proceso de extracción del tercio de la mitad de cada intervalo cerrado que nos vaya quedando, *Ad Infinitum*, como sugiere la Figura 21. El conjunto que nos queda cuando repetimos la iteración un número infinito de veces es el conjunto de Cantor, $\mathcal{C}$.

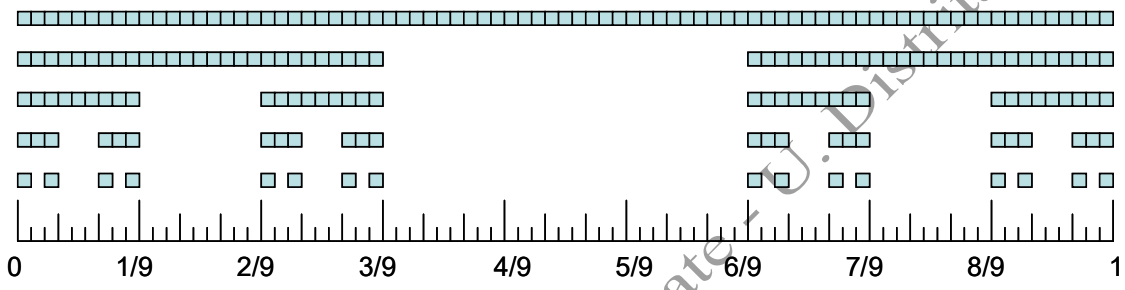


Figura 21. Primeras cuatro iteraciones en la construcción del conjunto de Cantor

¿Es $\mathcal{C}$ un conjunto de Borel? Partiendo de un intervalo de longitud 1, obtenemos 2^1 intervalos de longitud $1/3^1$ en la primera iteración, 2^2 intervalos de longitud $1/3^2$ en la segunda iteración, 2^3 intervalos de longitud $1/3^3$ en la tercera iteración y, en general, 2^n intervalos de longitud $1/3^n$ en la n-ésima iteración. Esto es, después de un número finito de iteraciones seguimos teniendo un número finito de intervalos, que es un conjunto de Borel, pues es la unión de un número finito de intervalos cerrados; pero ¿qué pasa después de un número infinito de iteraciones? ¡Que ya no existe ningún intervalo en $\mathcal{C}$! En efecto, en la iteración n, para n = 1,2,3,…, retiramos 2^{n-1} intervalos, cada uno de longitud 3^{-n}, de manera que la longitud total de los intervalos que extraemos es $\frac{1}{3}\sum_{n=0}^{\infty}\left(\frac{2}{3}\right)^n = \frac{1}{3}\left(\frac{1}{1-\frac{2}{3}}\right) = 1$. Si de un intervalo de longitud 1 quitamos subintervalos cuya longitud total es 1, nos queda sólo una "nube de polvo" que no contiene ningún intervalo, sólo puntos aislados. Pero cada punto individual de $\mathbb{R}$ es un conjunto unitario de Borel y, si los puntos que quedan en el conjunto de Borel son los extremos de los intervalos que van quedando en cada iteración, tendríamos que $\mathcal{C}$ es la unión contable de puntos aislados… ¡pero no es posible numerar los puntos en el conjunto de Cantor! Existe otra cantidad no numerable de puntos que, sin ser el extremo de ninguno de esos intervalos, jamás se eliminan del conjunto de Cantor, como se puede apreciar para el punto ¼ en la Figura 22 : Si no se retiró en la tercera iteración ya jamás será retirado de $\mathcal{C}$. De hecho, Cantor mostró que hay tantos puntos en $\mathcal{C}$ como en $\mathbb{R}$, $|\mathcal{C}|=|\mathbb{R}|=\aleph_1$. Basta con notar que el algoritmo de construcción equivale a quitar del intervalo unitario todos los puntos cuya expansión en base 3 incluya algún 1, de manera que los puntos de Cantor son aquellos cuya expansión en base 3 sólo contiene los dígitos 0 y 2 (se entiende que cualquier número que termine en 1 seguido por un número infinito de ceros se escribe terminando

en 0 seguido por un número infinito de 2). En la primera iteración retiramos todos los puntos que tienen un uno en la posición de 3^{-1}; En la segunda iteración retiramos todos los puntos que tienen un uno en la posición 3^{-2}; En general, en la n-ésima iteración quitamos todos los puntos que tienen un uno en la posición 3^{-n}. Si nunca retiramos el punto ¼ de $\mathcal{C}$ es porque $(¼)_3 = 0.020202\ldots$. Ahora, si expresamos cada punto del intervalo unitario en binario, podemos cambiar cada dígito 1 por 2 e interpretarlo en base 3, con lo cual hemos establecido una relación biunívoca entre los puntos del intervalo unitario y los puntos del conjunto de Cantor:

$$\mathcal{C} \ni x = \sum_{n=1}^{\infty} a_n \cdot 3^{-n} \quad \leftrightarrow \quad y = \sum_{n=1}^{\infty} \frac{a_n}{2} \cdot 2^{-n} \in [0,1], \qquad a_n \in \{0,2\}$$

Como Cantor nos enseñó, dos conjuntos tienen la misma cardinalidad si podemos establecer una relación biunívoca entre ellos: ¡hay tantos puntos en $\mathcal{C}$ como en [0,1] y, por consiguiente, como en $\mathbb{R}$!

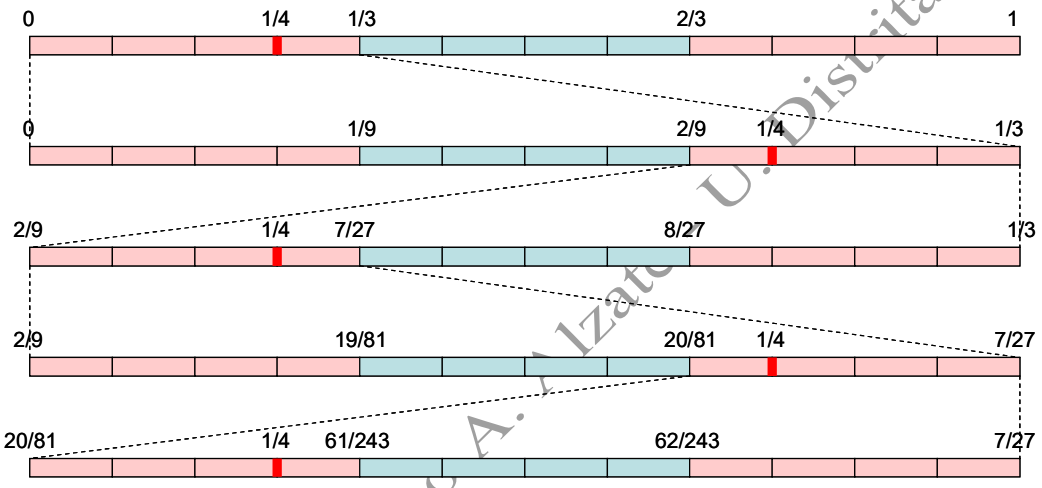


Figura 22. El punto [1/4] pertenece al conjunto de Cantor

En efecto, nótese que si miramos $\mathcal{C}$ en el intervalo $[0, 3^{-n}]$ a través de una lupa que aumente 3^n veces, reconstruiremos el conjunto de Cantor completo. Por eso el conjunto $\mathcal{C}$ constituye un objeto autosemejante (es idénticos a sí mismo en un número infinito de escalas) con dimensión fractal $\ln(2)/\ln(3) \approx 0.6$, a pesar de tener una dimensión topológica 0… $\mathcal{C}$ parece un buen candidato para un subconjunto de $\mathbb{R}$ que no pertence a $\mathcal{B}(\mathbb{R})$. Después de todo, $\mathcal{C}$ es un conjunto no numerable de puntos aislados cuyo complemento es un conjunto no numerable de intervalos abiertos ¡Pero hasta el conjunto de Cantor es un conjunto de Borel! Basta notar que $\mathcal{C}$ surge de una secuencia contable de intersecciones y uniones de intervalos cerrados:

$$\mathcal{C} = \bigcap_{n=0}^{\infty} \bigcup_{m=0}^{\frac{3^n-1}{2}} \left[\frac{2m}{3^n}, \frac{2m+1}{3^n} \right]$$

Sin embargo, se pueden encontrar subconjuntos del conjunto de cantor que no son Borel. Por ejemplo, evaluando la función de Cantor[8] en un conjunto de Vitali (ver enseguida) obtenemos un subconjunto

[8] La función de Cantor f_c:[0,1]→[0,1] se contruye para cada $x \in [0,1]$ así: (1) expresamos x en base 3, (2) si hay un dígito 1, ponemos cero de él en adelante, (3) cambiamos cada dígito 2 por un dígito 1, (4) interpretamos el resultado como un numero binario en [0,1]. A la gráfica de f_c se le llama grandilocuentemente "la escalera del diablo".

del conjunto de Cantor que no es Borel. Vale la pena mencionar que, aunque el conjunto de Cantor parece ser una curiosidad matemática diseñada para mostrar extraños subconjuntos de $\mathbb{R}$, mediante procedimientos de construcción generalizados semejantes al algoritmo de Cantor, se han desarrollado importantísimos modelos de tráfico en redes de comunicaciones, tales como el modelo wavelet multifractal (MWM).

Bueno, pero si hay menos conjuntos de Borel en $\mathbb{R}$ que subconjuntos de $\mathbb{R}$, ¿porqué no hemos podido encontrar uno? Mostraremos enseguida una familia infinita contable de conjuntos infinitos no contables de los reales que no son conjuntos de Borel, los conjuntos de Vitali.

Decimos que dos puntos de la recta real, x y y, son equivalentes si su diferencia $x-y$ es un número racional. No hace falta que x y y sean racionales, aunque pueden serlo; es suficiente con que su diferencia sea racional. Por ejemplo, 1/5 y 1/7 son equivalentes porque $1/5 - 1/7 = 2/35 \in \mathbb{Q}$, π y $\pi - 1/3$ son equivalentes porque su diferencia es $1/3 \in \mathbb{Q}$, pero π y $\pi/2$ no son equivalentes porque su diferencia es $\pi/2 \notin \mathbb{Q}$. Todos los racionales forman una única clase equivalente y existe otra clase equivalente por cada número irracional en los reales, esto es, existe un número infinito no contable de clases equivalentes (una por cada irracional que no esté separado de otro irracional por una cantidad fraccional), donde cada clase equivalente tiene un número infinito contable de elementos (uno por cada racional). De cada una de esas incontables clases escogemos un elemento en el intervalo [0, 1] y al conjunto resultante le llamamos un "conjunto de Vitali", V. Ahora, por cada número racional q en el intervalo [0, 1], hacemos una traslación de V sumando q a cada uno de sus elementos. Los conjuntos resultantes son los conjuntos V_q, de manera que ahora tenemos un conjunto contable de conjuntos incontables cuya unión, $\cup_q V_q$, forma el intervalo [0,2]. ¿Cuál es la longitud de cada V_q? No puede ser cero, porque la longitud del intervalo [0, 2] sería cero. No puede ser diferente de cero, porque la longitud del intervalo [0, 2] sería infinita… Los conjuntos de Vitali no son conjuntos de Borel (ver definición 27).

14. Medida de Probabilidad

Una medida de probabilidad $\mathbf{P}$ *asociada a un experimento aleatorio* $(\Omega,\mathcal{F})$ *es una función* $\mathbf{P}:\mathcal{F}\to\mathbb{R}$ *que asigna a cada evento en* $\mathcal{F}$ *un número real que satisface los siguientes axiomas:* (1) $\mathbf{P}(\Omega) = 1$, (2) *Si* $A\in\mathcal{F}$, $\mathbf{P}(A) \geq 0$, (3) *Si* $A,B\in\mathcal{F}$ *son mutuamente excluyentes* $(A\cap B=\Phi)$, $\mathbf{P}(A\cup B) = \mathbf{P}(A) + \mathbf{P}(B)$. *Si* $\mathcal{F}$ *es un campo-*σ *infinitamente aditivo, también debe satisfacerse el siguiente axioma adicional:* (3^a) *Si* $\{A_n\in\mathcal{F}, n=1,2,3,\ldots\}$ *es una colección de eventos tal que* $A_i\cap A_j = \Phi$ *para* $i\neq j$, *entonces* $\mathbf{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbf{P}(A_n)$.

Esta definición axiomática, tan útil como es, deja de lado el problema de darle un significado al número que se asigna a cada evento. Lo cierto es que si Kolmogorov estableció esta definición en 1933, en respuesta al reto lanzado por Hilbert en 1900 sobre determinar unas bases formales para la

teoría de la probabilidad, fue inspirado en propiedades fundamentales de los conceptos típicos de probabilidad: (1) Que si repito un experimento un gran número de veces y mido la fracción de experimentos en que sucede el evento A (ver definición 7), la fracción obtenida tiende a $\mathbf{P}(A)$ a medida que hago más y más repeticiones. (2) Que si logro describir el espacio muestral como un conjunto de cardinalidad finita en el que ninguno de los eventos unitarios ocurre preferencialmente sobre los otros, la probabilidad de un evento está dada por la cardinalidad del evento sobre la cardinalidad del espacio muestral. (3) Que si consulto a un experto sobre la ocurrencia de un evento en una hipotética realización de un experimento, la probabilidad del evento es el grado de certeza que el experto tiene en que dicho evento ocurra. (4) Que si he acumulado cierta evidencia a favor o en contra de una hipótesis, la probabilidad del evento en el que dicha hipótesis es cierta está dada por el grado de implicación lógica que existe de la evidencia a la hipótesis. (5) etc. Resulta muy afortunado saber que siempre es posible asociar los axiomas propuestos por Kolmogorov a propiedades particulares de la probabilidad en cada una de sus interpretaciones, como describiremos en la definición 15.

En efecto, desde el punto de vista de la formalidad, la matemática no es más que un lenguaje en el que se pueden mezclar unos símbolos de forma adecuada para satisfacer ciertas reglas gramaticales dictadas, por ejemplo, por la lógica, la teoría de conjuntos, etc., de manera que, a partir de unos axiomas, podemos deducir unos teoremas. Pero un modelo matemático es un isomorfismo entre un lenguaje formal de las matemáticas y una simplificación de alguna realidad particular: El isomorfismo induce un significado a las matemáticas. El hecho es que, para cada uno de los diferentes conceptos de probabilidad mencionados en el párrafo anterior, se puede establecer un isomorfismo con la formalidad de Kolmogorov, así como la formalidad de las ecuaciones diferenciales de primer orden con coeficientes constantes pueden asociarse con un circuito RC o con automóvil en movimiento (Figura 6). En un servidor de correo electrónico es posible medir la frecuencia relativa de cada tipo de mensaje spam y aplicar sobre esas medidas la formalidad de Kolmogorov, indicando que la definición formal de probabilidad aplica a la definición (1) de probabilidad. En un juego de cartas es posible contar las cartas para obtener información que permita decidir las acciones que maximicen la probabilidad de ganar, en cuyo caso la formalidad de Kolmogorov aplica a la definión (2) de probabilidad. Si tengo unas evidencias experimentales de las que induzco un nivel de creencia en que ocurra un evento, la formalidad de Kolmogorov debería aplicar a la definición (3) si queremos que esas creencias sean coherentes.

Lo cierto es que esta definición 14 es implacable y precisa, a diferencia de las posibles interpretaciones que le asignemos. Como cualquier formalidad axiomática en matemáticas, los axiomas escogidos por Kolmogorov son la menor cantidad de axiomas que forman un conjunto completo (cualquier proposición que involucre probabilidades puede calificarse como cierta o como falsa), consistente (ninguna proposición podrá ser demostrada como falsa y como verdadera) e independiente (ningún axioma se puede deducir de los demás). Esta formalidad no sólo evita cualquier ambigüedad conceptual sino que enfatiza el pensamiento deductivo con que se debe abordar la teoría de probabilidades: Todos los teoremas se deben derivar de estos tres axiomas (ver definición 17) y es aplicando los axiomas y los teoremas como encontramos las probabilidades de unos eventos en términos de las probabilidades de otros eventos. Por eso aceptamos los axiomas como verdades indiscutibles, algo así como si Kolmogorov los hubiera bajado del monte Sinaí impresos en piedra.

Claro, tenemos la opción de no aceptarlos, en cuyo caso no podemos usar la teoría de probabilidades en nuestros modelos matemáticos de la realidad ingenieril de las redes de comunicaciones y podemos terminar aquí el estudio de este texto. Pero, si optamos por esa opción, nos estaremos privando de una herramienta formidable para nuestro trabajo investigativo y estaremos negándonos a investigar en muchas áreas donde la teoría de las probabilidades es indispensable. Yo, personalmente, creo que la teoría de las probabilidades describe la lógica del pensamiento científico, por lo que no podríamos avanzar muy lejos sin sus axiomas. Sin embargo, para nuestra tranquilidad, enseguida veremos la plausibilidad de los axiomas en las diferentes interpretaciones de la medida de probabilidad.

15. Interpretaciones de la Medida de Probabilidad

Sea un experimento aleatorio $(\Omega,\mathcal{F})$ *y un evento* $A\in\mathcal{F}$. *Una forma de interpretar la probabilidad del evento A es mediante la relación* $\mathbf{P}(A) = \lim_{N\to\infty} f_N(A)$, *donde* $f_N(A)$ *es la frecuencia relativa del evento A en N repeticiones del experimento. Esta interpretación aplica cuando el experimento se puede replicar indefinidamente bajo condiciones idénticas. Otra interpretación muy útil se basa en la relación* $P(A) = |A|/|\Omega|$, *donde* $|A|$ *es la cardinalidad del evento A. Esta interpretación aplica cuando* Ω *es un conjunto de cardinalidad finita y* $P(\{\omega\})=1/|\Omega|$ *para todo* $\omega\in\Omega$. *Sin embargo, sea que ninguna, una o ambas interpretaciones apliquen al experimento aleatorio de interés, para nosotros P(A) es el nivel de certeza que tenemos en que ocurra el evento A cuando ejecutemos el experimento.*

Como mencionamos en la definición 7 sobre la regularidad estadística de un experimento aleatorio, si repetimos N veces un experimento con espacio muestral Ω y contamos en cuántas repeticiones ocurrió el evento $A \subset \Omega$, N_A, definimos la frecuencia relativa del evento A en esas N repeticiones como $f_N(A) = N_A/N$. Obsérvese que el proceso de observación de N_A es, en sí mismo, otro experimento aleatorio, de manera que en diferentes conjuntos de N repeticiones podemos obtener diferentes valores de $f_N(A)$. Sin embargo, la regularidad estadística sugiere que, entre más repeticiones hagamos, el valor de $f_N(A)$ tiende a un valor fijo, independientemente del conjunto particular de N repeticiones que seleccionemos.

Por ejemplo, supongamos que deseamos saber cuál es la probabilidad del evento A = {a un enrutador llegan más de 1000 bytes en un período de 100 ms}. Para esto medimos la frecuencia relativa de dicho evento en 200 períodos consecutivos y la graficamos en función del número de períodos observados. Si hacemos mediciones durante un minuto, obtendremos tres conjuntos distintos, como muestra la Figura 23, en cada uno de los cuales la frecuencia parece tender a un número cercano a 0.4. Si las condiciones del tráfico permanecen estables durante el minuto de observación y son iguales a las condiciones en el período de 100 ms por cuya probabilidad nos interesamos (que podría ser, por ejemplo, el siguiente período que aún no hemos observado), diríamos que la probabilidad de que

lleguen más de 1000 bytes es "cercana" a 0.4[9]. Pues bien, es fácil ver que los axiomas que definen la probabilidad como una medida de los subconjuntos de Ω contenidos en $\mathcal{F}$ están inspirados en propiedades elementales de la frecuencia relativa. En efecto,

(1) $f_N(\Omega) = N / N = 1$
(2) como $N_A \geq 0, f_N(A) \geq 0$
(3) Si $A \cap B = \Phi$, $N_{A \cup B} = N_A + N_B$, de manera que $f_N(A \cup B) = f_N(A) + f_N(B)$.

Permítaseme insistir, porque debemos ser cuidadosos con esto, que el límite de la frecuencia relativa es apenas una interpretación de la probabilidad que puede ser útil para los ingenieros de redes de comunicaciones ya que a nosotros nos es posible tomar muchas mediciones con facilidad y en tiempos razonables (medir el retardo de 10000 paquetes, medir el número de errores en 10000 bits transmitidos, medir la condición de "spam" en 10000 mensajes de correo electrónico, etc.). Sin embargo, en muchos casos, el experimento mismo que queremos modelar ni siquiera es repetible, de manera que no tiene sentido considerar esta interpretación. J. M. Keynes, por ejemplo, era economista y cada uno de sus experimentos podía durar décadas; por esa razón, la interpretación frecuentista, que parece objetiva en cuanto a que muestra resultados verificables "a la larga", lo conduce a expresar su famosa frase: "A la larga, todos estamos muertos". En estas condiciones, lo mejor es considerar la probabilidad como expresión de simetría o como nivel de confianza. Esta es la visión Bayesiana: La probabilidad es un grado subjetivo de creencia. ¿Qué significa, si no, la frase "mañana en la tarde lloverá con probabilidad 0.7"? Puede que el meteorólogo tenga razones frecuentistas para decir algo así, pero para la persona que lo escucha en televisión es sólo una medida de creencia. O ¿a qué se refiere el adolescente cuando afirma "mi novia me ama con probabilidad 0.7"? Con seguridad no se trata de que el 70% de sus novias lo han amado. Finalmente, como mencionamos en la definición 4, verificar la existencia del bosón de Higgs fue un experimento aleatorio que se realizó en el LHC (en realidad un conjunto muy grande de experimentos aleatorios realizados entre el 10 de septiembre de 2008 y el 4 de julio de 2012), que nunca se repetirá porque ya no será aleatorio.

De todas maneras, desde un punto de vista puramente matemático, la interpretación misma pierde relevancia pues la definición es precisa e implacable: la probabilidad es una función que asigna a cada subconjunto de Ω en $\mathcal{F}$ una medida en $\mathbb{R}$ que satisface tres axiomas básicos. Lo cierto es que, cuando uno está inmerso en un problema de modelado probabilístico, a veces resulta muy útil preguntarse "si yo pudiera repetir este experimento muchas veces, ¿qué esperaría que sucediera a la larga?", pues la respuesta puede sugerirnos el siguiente paso en el proceso o puede explicarnos un resultado poco intuitivo. De hecho, en este libro echaremos mano de la interpretación frecuentista liberalmente para justificar muchas definiciones o para interpretar muchos resultados.

[9] De hecho, de acuerdo con la cota de Chebyshev (definición 47(b)), la frecuencia relativa se acerca a la probabilidad a medida que hacemos más y más observaciones, en el sentido de que el evento $B = \{|f_N(A) - \mathbf{P}(A)| > \varepsilon\}$ ocurre con probabilidad menor a $1/(4N\varepsilon^2)$. Más aún, mediante la cota de Chernov (definición 47(c)) se puede demostrar que la aproximación de $f_N(A)$ a $\mathbf{P}(A)$ es exponencial: $P[\{|f_N(A) - \mathbf{P}(A)| > \varepsilon\}] < \exp(-2N\varepsilon^2)$ –desigualdad de Hoeffding–.

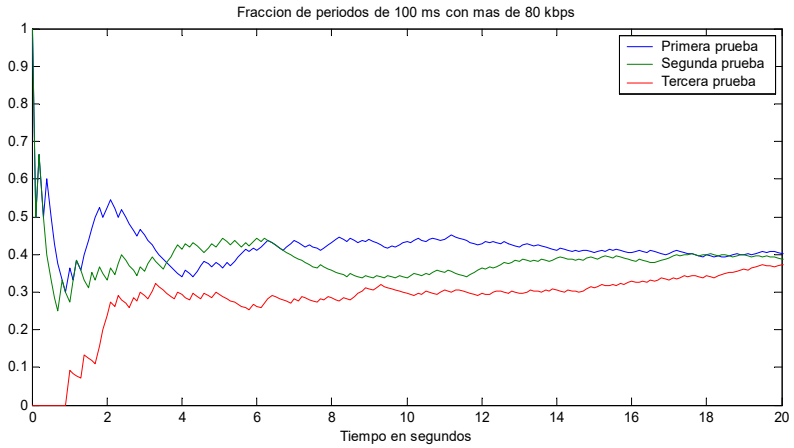


Figura 23. $f_N(A)$ vs N para tres conjuntos distintos de pruebas

Otra interpretación que suele resultar muy útil para estimar las probabilidades de algunos eventos es la interpretación clásica. La probabilidad del evento A es

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\text{Número de resultados favorables al evento A}}{\text{Número total de posibles resultados}}$$

siempre que ninguno de los subconjuntos unitarios de Ω ocurra preferencialmente sobre otros, esto es, siempre que todos los eventos unitarios de Ω sean ¡igualmente probables! Claro, al incluir el concepto a definir dentro de la definición, no estamos definiendo nada. Pero la idea es intuitivamente clara: sacar una carta de la pinta de corazones ocurre con probabilidad 13/52 = 1/4, pues ninguna carta sale preferencialmente sobre las demás. Obtener un par al lanzar un dado bien balanceado ocurre con probabilidad 3/6=1/2, pues ningún lado queda hacia arriba preferencialmente sobre los demás. Los axiomas de Kolmogorov aplican igualmente bien para esta interpretación:

(1) $|\Omega|/|\Omega| = 1$
(2) como $|A| \geq 0$, $|A|/|\Omega| \geq 0$
(3) Si $A \cap B = \Phi$, $|A \cup B| = |A| + |B|$, de manera que $|A \cup B|/|\Omega| = (|A| + |B|)/|\Omega| = |A|/|\Omega| + |B|/|\Omega|$.

Al hacer uso de esta interpretación, debemos ser muy juiciosos con la premisa de los eventos unitarios "igualmente probables". Si el dado está cargado de manera que P({1}) = 0.2 y P({n})=0.16 para n=2,3,4,5,6, el uso de esta interpretación clásica conducirá a resultados desastrosos. Por eso esta definición sólo aplica a una clase muy pequeña de experimentos aleatorios.

A veces el espacio muestral es infinito no contable, pero aún podemos aplicar la interpretación clásica si usamos alguna medida finita del tamaño de Ω, tal como la longitud o el área:

$$P(A) = \frac{\mu(A)}{\mu(\Omega)} = \frac{\text{Medida de la región favorable al evento A}}{\text{Medida total de la región posible}}$$

siempre que todos los subconjuntos de cualquier tamaño dado tengan la misma probabilidad de ocurrir como resultado del experimento:

(1) $\mu(\Omega)/\mu(\Omega) = 1$
(2) como $\mu(A) \geq 0$, $\mu(A)/\mu(\Omega) \geq 0$
(3) Si $A \cap B = \Phi$, $\mu(A \cup B) = \mu(A) + \mu(B)$, de manera que $\mu(A \cup B)/\mu(\Omega) = (\mu(A) + \mu(B))/\mu(\Omega)$
$= \mu(A)/\mu(\Omega) + \mu(B)/\mu(\Omega)$.

Para nosotros, como ingenieros preocupados por problemas técnicos muy precisos, resulta muy cómodo escoger eclécticamente entre cada una de las interpretaciones la que más nos favorezca o la que mejor nos guíe en el proceso de desarrollar un modelo probabilístico para nuestro problema. Por ejemplo, no hemos sabido de ningún ingeniero de comunicaciones al que le quite el sueño el problema filosófico que implica utilizar un medidor de BER (Bit-Error-Rate) para medir la fracción de bits que se dañan durante su transmisión por un canal de comunicaciones (como en el experimento 9) y después utilizar esa medida como su nivel de confianza en que el próximo bit que transmita se dañe en el canal, aunque así esté mezclando las interpretaciones (1) y (3). Dada la facilidad que tenemos para tomar muchas mediciones, esta interpretación frecuentista de la probabilidad la podemos aprovechar ventajosamente desde la formalidad de Kolmogorov, como se describe en la definición 15.

16. Espacio de Probabilidad

> *Un espacio de probabilidad es la tripleta* $(\Omega, \mathcal{F}, \mathbf{P})$ *asociada con un experimento aleatorio, donde* Ω *es el espacio muestral o el conjunto de todos los posibles resultados del experimento,* $\mathcal{F}$ *es un campo-*σ *de subconjuntos de* Ω *construido a partir de una clase de eventos de interés y* **P** *es una función de* $\mathcal{F}$ *en* $\mathbb{R}$ *que satisface los axiomas de la definición* 14. *Como solamente se les puede asignar una medida de probabilidad a los subconjuntos de* Ω *que pertenecen a* $\mathcal{F}$*, a dichos subconjuntos se les denomina "subconjuntos medibles" o "Eventos" (con lo cual cumplimos la promesa de especificar mejor el concepto de evento).*

En cualquier caso en que queramos trabajar sobre modelos probabilísticos de una realidad particular, deberemos partir de la descripción explícita del espacio de probabilidad $(\Omega, \mathcal{F}, \mathbf{P})$ pues, de otra manera, estaremos perdidos: ni siquiera sabremos dónde estamos parados! En efecto, una vez descrito el problema en términos de un espacio de probabilidad, podremos movernos con confianza sobre todos los resultados de la teoría de las probabilidades sabiendo que, mientras seamos consecuentes y rigurosos con ellas, llegaremos a resultados significativos, pues es sobre los axiomas que definen a $\mathcal{F}$ y **P** que se construye TODA la teoría de probabilidades.

Hasta este punto, la teoría de probabilidades sería simplemente una rama de la "teoría de las mediciones", que es el área de la matemáticas que estudia las funciones $\mu: \mathcal{H} \rightarrow \mathbb{R}$ que asignan una medida real $\mu(E)$ a cada conjunto E de una colección de conjuntos $\mathcal{H}$. En teoría de mediciones se estudia formalmente la conveniencia de que $\mathcal{H}$ forme un campo-σ, en cuyo caso $\mu(\cdot)$ es una medida aditivamente contable, como es el caso de las medidas de probabilidad asignadas a los subconjuntos medibles del espacio muestral de un experimento aleatorio. Sin embargo, la definición 21 tratará

sobre la independencia, la cual le dará a la teoría de probabilidades una identidad propia que le permitirá distinguirse de la teoría general de las mediciones.

17. Algunos Resultados Básicos Derivados de los Axiomas de la Probabilidad

Sea $(\Omega,\mathcal{F},\mathbf{P})$ *un espacio de probabilidad en el que hay dos eventos medibles* A *y* $B \in \mathcal{F}$. *Entonces* (1) $\mathbf{P}(A^C) = 1 - \mathbf{P}(A)$, (2) $\mathbf{P}(\Phi) = 0$, (3) $\mathbf{P}(A) \leq 1$, (4) $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$, (5) Si $A \subset B$, $\mathbf{P}(A) \leq \mathbf{P}(B)$.

Los anteriores cinco resultados son apenas una muestra mínima de todas las conclusiones que se pueden sacar de los axiomas de la definición 14 pues, como ya se dijo, de los tres axiomas se deriva TODA la teoría de las probabilidades. Sin embargo, como estos cinco resultados se usan cotidianamente cuando se estudian modelos probabilísticos de cualquier sistema, vale la pena tenerlos tan presentes como los mismos axiomas de los que se derivan:

(1) $\mathbf{P}(A^C) = 1 - \mathbf{P}(A)$
En efecto, como $A \cap A^C = \Phi$ y $A \cup A^C = \Omega$, los axiomas 1 y 3 conducen a $\mathbf{P}(\Omega) = \mathbf{P}(A^C) + \mathbf{P}(A) = 1$. Restando $\mathbf{P}(A)$ a ambos lados obtenemos el resultado deseado.

(2) $\mathbf{P}(\Phi) = 0$
En efecto, $A = A \cup \Phi$, que son eventos mutuamente excluyentes, por lo que podemos aplicar el tercer axioma: $\mathbf{P}(A) = \mathbf{P}(A) + \mathbf{P}(\Phi)$. Restando $\mathbf{P}(A)$ a ambos lados obtenemos el resultado deseado. Es importante notar que $\mathbf{P}(\Phi) = 0$ se refiere al "evento imposible" Φ. Cualquier otro evento distinto de vacío con probabilidad 0 se conoce como "evento nulo" y, aunque improbable, no es imposible.

(3) $\mathbf{P}(A) \leq 1$
En efecto, como ya demostramos que $\mathbf{P}(A) = 1 - \mathbf{P}(A^C)$, basta con aplicar el segundo axioma en A^C, $\mathbf{P}(A^C) \geq 0$, para obtener el resultado deseado. Como en el caso anterior, es importante notar que $\mathbf{P}(\Omega) = 1$ se refiere al "evento seguro" Ω. Cualquier otro evento distinto de Ω con probabilidad 1 se conoce como "evento casi seguro", pues es posible que no ocurra.

(4) $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$
En efecto, podemos expresar $A \cup B$ como la unión de dos eventos mutuamente excluyentes, $A \cup B = A \cup (B \cap A^C)$, de manera que $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B \cap A^C)$. Por otro lado, B también se puede expresar como la unión de dos eventos mutuamente excluyentes, $B = (A \cap B) \cup (B \cap A^C)$, de manera que $\mathbf{P}(B \cup A^C) = \mathbf{P}(B) - \mathbf{P}(A \cap B)$. Remplazando esta expresión de $\mathbf{P}(B \cup A^C)$ en la primera expresión de $\mathbf{P}(A \cup B)$ obtenemos el resultado deseado. Los diagramas de Venn de la Figura 24 representan esta derivación.

(5) Si $A \subset B$, $\mathbf{P}(A) \leq \mathbf{P}(B)$
En efecto, podemos expresar B como la unión de dos eventos mutuamente excluyentes, $B = A \cup (B \cap A^C)$, de manera que $\mathbf{P}(B) = \mathbf{P}(A) + \mathbf{P}(B \cap A^C)$. Y, como $\mathbf{P}(B \cap A^C) \geq 0$ por el segundo axioma, entonces $\mathbf{P}(B) \geq \mathbf{P}(A)$.

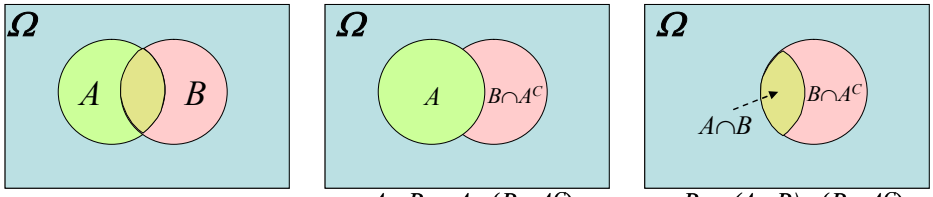


Figura 24. Construcciones para derivar la expresión P($A\cup B$) = P(A) + P(B) - P($A\cap B$)

18. Probabilidad Condicional

Sea (Ω, $\mathcal{F}$, **P**) *un espacio de probabilidad en el que hay dos eventos* A *y* $B \in \mathcal{F}$. *La probabilidad condicional del evento* A *dado que se sabe de la ocurrencia del evento* B *es*

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}, \quad \mathbf{P}(B) > 0$$

Esta definición, como tan "sacada de la manga", en realidad obedece a un concepto muy simple si se le mira desde la interpretación frecuentista de la probabilidad. Supongamos que repetimos N veces el experimento en cuestión y contamos cuántas veces sucedió cada uno de los siguientes eventos: N_A = Número de veces que ocurrió el evento A, N_B = Número de veces que ocurrió el evento B, y $N_{A\cap B}$ = Número de veces que ocurrieron ambos simultáneamente. Ahora consideramos solamente aquellas N_B repeticiones en las que ocurrió el evento B e ignoramos todas las demás. La frecuencia relativa del evento A entre aquellas repeticiones del experimento en que ocurrió B es $f_N(A|B) = N_{A\cap B} / N_B$. Dividiendo el numerador y el denominador por N, obtenemos $f_N(A|B) = f_N(A\cap B)/ f_N(B)$, que es una expresión MUY parecida a la definición de probabilidad condicional.

De acuerdo con el anterior resultado, si pudiésemos definir la probabilidad de un evento como el límite de su frecuencia relativa cuando el número de repeticiones tiende a infinito, tendríamos que el condicionamiento sería simplemente una propiedad más de la probabilidad. Pero como la probabilidad es un concepto más abstracto (una función de $\mathcal{F}$ en $\mathbb{R}$ que satisface 3 axiomas), este resultado frecuentista es apenas un motivo de inspiración para la definición propuesta. De todas formas, la definición no nos debe sorprender porque la teoría de probabilidades quiere modelar, precisamente, el comportamiento de ese límite sin obligarnos a gastar un tiempo infinito en hacer un número infinito de repeticiones del experimento.

Volviendo a nuestra definición axiomática, a veces es posible que la condición sea un evento de probabilidad cero, pues una cosa es un evento improbable –con probabilidad cero– y otra cosa un evento imposible –que no puede ocurrir–. En este caso la probabilidad condicional puede estar indefinida (0/0) o puede que, para cualquier secuencia $\{B_n \in \mathcal{F}\}$ tal que $B=\lim_{n\to\infty} B_n$, exista el límite $P[A|B]=\lim_{n\to\infty} P[A\cap B_n]/P[B_n]$. Estos límites se mencionarán en las definiciones 23 y 25 pero el caso particular de condicionar en eventos de probabilidad cero se mencionará en el próximo capítulo cuando hablemos de variables aleatorias continuas.

Es legítimo preguntarnos por un nuevo espacio de probabilidad en el que la probabilidad condicional sea una medida de probabilidad válida. Lo primero que notamos, por ejemplo, es que, en el nuevo espacio de probabilidades, el espacio muestral debe ser B, pues nos estamos limitando a estudiar los casos en que tenemos certeza absoluta de que el evento B ocurrió. Pero, ¿cuál sería un nuevo campo de eventos apropiado? Como todos los eventos de interés contenidos en $\mathcal{F}$ se ven reducidos a su intersección con B, es razonable pensar en un campo de eventos como $\mathcal{H}= \{A\cap B : A\in\mathcal{F}\}$. ¿Es éste un campo-$\sigma$ de subconjuntos de B? Veamos

(1) $\mathcal{H}$ es no vacío porque por lo menos $\Phi\in\mathcal{F}\Rightarrow\Phi\cap B=\Phi\in\mathcal{H}$:
    $\mathcal{H}$ es no vacío
(2) Si $A\in\mathcal{F}$, entonces $A^C(=\Omega\backslash A)\in\mathcal{F}$, de manera que $A\cap B\in\mathcal{F}$ y $A^C\cap B(=B\backslash(A\cap B))\in\mathcal{F}$, donde $A^C\cap B=B\backslash(A\cap B)$ es el complemento de $A\cap B$ en B:
    Si $X\in\mathcal{H}$, entonces $X^C=B\backslash X\in\mathcal{H}$
(3) Si $A_1\in\mathcal{F}$, $A_2\in\mathcal{F}$, entonces $A_1\cup B\in\mathcal{F}$, $A_2\cup B\in\mathcal{F}$ y, por consiguiente, $A_1\cap B\in\mathcal{H}$, $A_2\cap B\in\mathcal{H}$, entonces $A_1\cap B\in\mathcal{H}$, $A_2\cap B\in\mathcal{H}$, $(A_1\cup A_2)\cap B=(A_1\cap B)\cup(A_2\cap B)\in\mathcal{H}$:
    Si $X\in\mathcal{H}$, $Y\in\mathcal{H}$, entonces $X\cup Y\in\mathcal{H}$
(4) Lo mismo se puede verificar para la uniones contables

¿Y será la probabilidad condicional $\mathbf{Q}(\cdot)=\mathbf{P}(\cdot|B)$ una medida válida en $(B, \mathcal{H})$? Veamos:

(1) $\mathbf{Q}(B)=\mathbf{P}(B|B)=\mathbf{P}(B)/\mathbf{P}(B) = 1$
(2) $\mathbf{Q}(A)=\mathbf{P}(A|B)\geq 0$
(3) Si $(A_1\cap B)\cap(A_2\cap B)=\Phi$, (en realidad no necesitamos que A_1 y A_2 sean excluyentes, pues basta con que no puedan ocurrir simultáneamente con B), entonces $\mathbf{Q}(A_1\cup A_2)=\mathbf{P}((A_1\cup A_2)|B) = \mathbf{P}((A_1\cup A_2)\cap B)/\mathbf{P}(B) = \mathbf{P}((A_1\cap B)\cup(A_2\cap B))/\mathbf{P}(B) = (\mathbf{P}(A_1\cap B) + \mathbf{P}(A_2\cap B))/\mathbf{P}(B) = \mathbf{P}(A_1|B) + \mathbf{P}(A_2|B) = \mathbf{Q}(A_1) + \mathbf{Q}(A_2)$.
(4) Lo mismo se puede verificar para uniones contables.

En conclusión, dado el espacio de probabilidad $(\Omega,\mathcal{F}, \mathbf{P}(\cdot))$, podemos construir un nuevo espacio de probabilidad $(B, \mathcal{H}, \mathbf{Q}(\cdot)=\mathbf{P}(\cdot|B))$ condicionando todos los eventos de $\mathcal{F}$ a la ocurrencia del evento B, donde $B\in\mathcal{F}$ y $\mathbf{P}(B)>0$. Esto es, hemos reducido el espacio original a uno más pequeño.
Esto quiere decir que todo lo que hemos dicho (y diremos) sobre cualquier espacio de probabilidad $(\Omega, \mathcal{F}, \mathbf{P}(\cdot))$ aplica también al espacio de probabilidad condicionado, $(B, \mathcal{H}, \mathbf{Q}(\cdot)=\mathbf{P}(\cdot|B))$. En particular, como **<u>ya lo demostramos</u>** en la definición 17, (1) $\mathbf{Q}(A^C)=\mathbf{P}(A^C|B) = 1 - \mathbf{P}(A|B) = 1- \mathbf{Q}(A)$, (2) $\mathbf{Q}(\Phi) = \mathbf{P}(\Phi|B) = 0$, (3) $\mathbf{Q}(A) = \mathbf{P}(A|B) \leq 1$, (4) $\mathbf{Q}(A\cup C) = \mathbf{P}(A\cup C|B) = \mathbf{P}(A|B) + \mathbf{P}(C|B) - \mathbf{P}(A\cap C|B) = \mathbf{Q}(A) + \mathbf{Q}(C) - \mathbf{Q}(A\cap C)$, (5) Si $A\cap B\subset C\cap B$ (en realidad no necesitamos que $A\subset C$, pues basta con que esta condición ocurra dentro de B), $\mathbf{Q}(A)=\mathbf{P}(A|B)\leq\mathbf{P}(C|B) = \mathbf{Q}(C)$.

Considérese, por ejemplo, el experimento 9, en el que transmitimos un bit y vemos si llegó correctamente a su destino en el otro extremo del canal binario. Si consideramos como parte del experimento observar el bit transmitido, nuestro nuevo espacio muestral será $\Omega= \{(0,0), (0,1), (1,0),$

(1,1)}, donde el resultado (i, j) corresponde a la transmisión del bit i y la recepción del bit j. La probabilidad de que se produzca un error en ese canal es $\mathbf{P}(\{(0,1),(1,0)\})$, que es la probabilidad del evento *ERROR* = {*se recibe un bit distinto al bit transmitido*}. Condicionando en el bit transmitido, tenemos dos tipos de error con las siguientes probabilidades

$$\mathbf{P}(\{Recibir\ 0\} \mid \{se\ transmitió\ 1\}) = \mathbf{P}(\{(1,0\}) \,/\, \mathbf{P}(\{(1,0),(1,1)\})$$
$$\mathbf{P}(\{Recibir\ 1\} \mid \{se\ transmitió\ 0\}) = \mathbf{P}(\{(0,1\}) \,/\, \mathbf{P}(\{(0,0),(0,1)\})$$

Dada la simetría que existe en las técnicas de modulación digital, es de esperar que los dos tipos de error tengan la misma probabilidad, en cuyo caso nos encontramos ante un espacio de probabilidad que modela un Canal Binario Simétrico (BSC, binary symmetric channel). Claramente, al utilizar un medidor de BER –*Bit Error Rate*– sobre un canal BSC, estamos tratando de estimar las probabilidades condicionales descritas anteriormente, por lo que el modelo se puede representar como en la Figura 25,

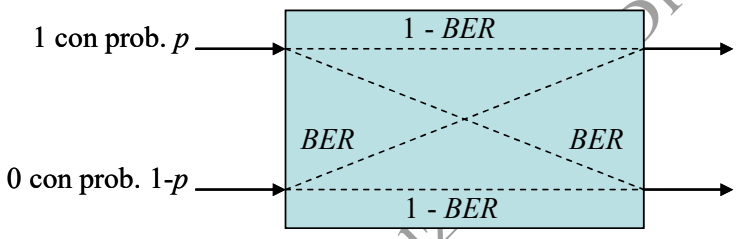


Figura 25. Modelo probabilístico de un canal binario simétrico

donde el espacio de probabilidad que modela el canal es (Ω = {(0,0), (0,1), (1,0), (1,1)}, $\mathcal{F} = \{0,1\}^{\Omega}$, **P**). El conocimiento inicial que tenemos sobre la medida de probabilidad **P** en este modelo probabilístico de un canal de comunicaciones es el siguiente:

$$\mathbf{P}[\{(1,0),(1,1)\}] = 1 - \mathbf{P}[\{(0,0),(0,1)\}] = p$$
$$\mathbf{P}[\{(0,0),(1,0)\} \mid \{(1,0),(1,1)\}] = 1 - \mathbf{P}[\{(0,1),(1,1)\} \mid \{(1,0),(1,1)\}] = \ldots$$
$$\mathbf{P}[\{(0,1),(1,1)\} \mid \{(0,0),(0,1)\}] = 1 - \mathbf{P}[\{(0,0),(1,0)\} \mid \{(0,0),(0,1)\}] = BER$$

Obsérvese en este ejemplo cómo resulta de fácil "medir" la probabilidad condicional BER. En general, ésta es la gran utilidad de la probabilidad condicional: encontrar la probabilidad de un evento A puede ser muy difícil, pero una vez condicionamos el evento de interés a otro evento B (juiciosamente seleccionado), puede resultar muy fácil encontrar la probabilidad condicional de A dado B. Este truco se repite una y otra vez en el modelado probabilístico de redes de comunicaciones, como tendremos oportunidad de ver en breve. Pero, ¿de qué nos sirve la probabilidad condicional de A dado B si lo que queríamos encontrar era la probabilidad de A? El siguiente teorema explica dónde reside la utilidad del "truco".

19. Teorema de la Probabilidad Total

*Sea (Ω, $\mathcal{F}$, **P**) un espacio de probabilidad en el que hay un evento $A \in \mathcal{F}$ y una secuencia de eventos en $\mathcal{F}$, $\{B_k\}$, k=1,2,… que forman una partición de Ω. Entonces*

$$\mathbf{P}(A) = \sum_k \mathbf{P}(B_k)\mathbf{P}(A \mid B_k)$$

Esta relación es más fácil de ver si consideramos la partición más pequeña, constituida por B y B^C, como muestra la Figura 26. En efecto, con esta partición podemos expresar el evento A como la unión de dos eventos mutuamente excluyentes, $A = (A \cap B) \cup (A \cap B^C)$, de manera que $\mathbf{P}(A) = \mathbf{P}(A \cap B) + \mathbf{P}(A \cap B^C)$. Pero, por la definición misma de la probabilidad condicional, $\mathbf{P}(A \cap B) = \mathbf{P}(B)\mathbf{P}(A|B)$ y $\mathbf{P}(A \cap B^C) = \mathbf{P}(B^C)\mathbf{P}(A|B^C)$, de manera que $\mathbf{P}(A) = \mathbf{P}(B)\mathbf{P}(A|B) + \mathbf{P}(B^C)\mathbf{P}(A|B^C)$. La generalización a particiones más numerosas (incluyendo aquellas contablemente infinitas) es inmediata.

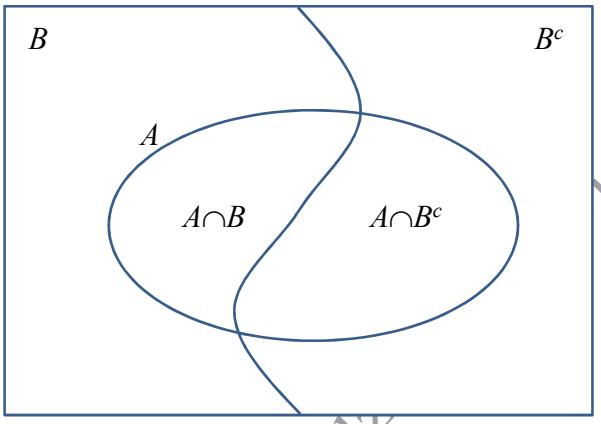


Figura 26. Diagrama de Venn para verificar el teorema de la probabilidad total

En el modelo del canal simétrico binario de la Figura 25, por ejemplo, ¿cuál será la probabilidad de recibir un cero? Podemos condicionar en el bit transmitido, ya que las probabilidades condicionadas en el bit transmitido son conocidas:

$$\begin{aligned}\mathbf{P}(\{Rx\ 0\}) &= \mathbf{P}(\{Tx\ 0\})\mathbf{P}(\{Rx\ 0\} \mid \{Tx\ 0\}) + \mathbf{P}(\{Tx\ 1\})\mathbf{P}(\{Rx\ 0\} \mid \{Tx\ 1\}) \\ &= (1-p)\,(1-BER) + p\,BER = 1 - BER - p + 2pBER\end{aligned}$$

Claro, la probabilidad de recibir un uno debe ser uno menos la probabilidad de recibir un cero, lo cual puede ser verificado mediante la probabilidad total:

$$\begin{aligned}\mathbf{P}(\{Rx\ 1\}) &= \mathbf{P}(\{Tx\ 0\})\mathbf{P}(\{Rx\ 1\} \mid \{Tx\ 0\}) + \mathbf{P}(\{Tx\ 1\})\mathbf{P}(\{Rx\ 1\} \mid \{Tx\ 1\}) \\ &= (1-p)\,BER + p\,(1-BER) = BER + p - 2pBER\end{aligned}$$

Obsérvese que si BER=0 ó si BER=1, no existiría ninguna duda en el receptor sobre el bit transmitido, pues el bit recibido tendrá toda la información necesaria para identificar al primero sin equivocaciones. Cualquier otro valor de BER genera incertidumbre en el receptor, especialmente en el caso extremo en que $BER = 0.5$, pues en este caso obtenemos que $\mathbf{P}(\{Rx\ 1\}) = \mathbf{P}(\{Rx\ 0\}) = 0.5$, independientemente de p, de manera que podemos ahorrarnos el canal y hacer que en el receptor se lance una moneda equilibrada por cada bit transmitido.

Una pregunta de mucho interés para el módem receptor es la siguiente: Dado que recibí cierto símbolo a la salida del canal, ¿cuáles son las probabilidades del respectivo símbolo a la entrada del canal? La siguiente regla es muy útil para este tipo de preguntas.

20. Regla de Bayes

Sea $(\Omega, \mathcal{F}, \mathbf{P})$ *un espacio de probabilidad en el que hay un evento* $A \in \mathcal{F}$ *y una secuencia de eventos* $\{B_k\}$, k=1,2,… *que forman una partición de* Ω *y que también pertenece a* $\mathcal{F}$. *Entonces*

$$\mathbf{P}(B_k \mid A) = \frac{\mathbf{P}(B_k)\mathbf{P}(A \mid B_k)}{\sum_j \mathbf{P}(B_j)\mathbf{P}(A \mid B_j)}$$

Esta regla surge directamente de la definición de la probabilidad condicional, pues $\mathbf{P}(A \cap B_k) = \mathbf{P}(B_k)\,\mathbf{P}(A \mid B_k)$ y $\mathbf{P}(A) = \sum_j \mathbf{P}(B_j)\mathbf{P}(A \mid B_j)$, de manera que la regla de Bayes sólo es una forma de expresar la definición $\mathbf{P}(B_k \mid A) = \mathbf{P}(A \cap B_k) / \mathbf{P}(A)$.

Considérese nuevamente el canal binario simétrico (*BSC*) donde el transmisor se caracteriza por transmitir 1 con probabilidad p y transmitir 0 con probabilidad $1 - p$ y el canal se caracteriza por una tasa de error por bit *BER*, como muestra la Figura 25. Si el destino recibe 0, ¿cuál es la probabilidad de que la fuente haya enviado 1?

$$P\left(\{Tx1\}\middle|\{Rx0\}\right) = \frac{P\left(\{Tx1\}\right)P\left(\{Rx0\}\middle|\{Tx1\}\right)}{P\left(\{Tx1\}\right)P\left(\{Rx0\}\middle|\{Tx1\}\right) + P\left(\{Tx0\}\right)P\left(\{Rx0\}\middle|\{Tx0\}\right)}$$
$$= \frac{p \cdot BER}{p \cdot BER + (1-p)(1-BER)}$$

La regla de Bayes se considera el fundamento del método científico y, en general, de la racionalidad: Yo parto de unas creencias iniciales (las probabilidades a priori $P(B_k)$ que hablan de los posibles estados de la naturaleza, los cuales no puedo medir directamente) y los relaciono con observaciones concretas de la realidad, el evento A (que es un hecho, una realidad objetiva), para lograr unas nuevas creencias (las probabilidades a posteriori $P(B_k|A)$), mejoradas por la consideración de hechos objetivos. La racionalidad consiste, precisamente, en estar dispuesto a cambiar mis creencias para ajustarlas a los hechos observados. Cómo a los posibles estados de la naturaleza (las hipótesis) se les asigna unos niveles de creencia (las probabilidades a priori), para ajustar esas mismas creencias (las probabilidades a posteriori), algunos filósosfos y teóricos de la estadística hablan de la subjetividad que implica la regla de Bayes como algo negativo. Sin embargo, la regla de Bayes da claridad a la toma de decisiones bajo condiciones de escasa información y resultados inciertos. Por ejemplo, la existencia del bosón de Higgs era incierta antes de 2012 pero, dados los resultados de los experimentos del LHC, la probabilidad de la existencia del bosón de Higgs es superior a 0.9999999999990. Este es uno de los resultados científicos que se dan en términos de la regla de

Bayes. Pero igual puede decir el adolescente que no sabe si su compañera de salón está enamorada de él o no, y ajusta su nivel de creencia de acuerdo con las miradas que recibe de ella. O un médico que no sabe si un paciente tiene cierta enfermedad o no y ajusta su nivel de creencia de acuerdo con las pruebas de laboratorio que le ordena. O un juez que no sabe si un acusado es culpable o no y ajusta su nivel de creencia de acuerdo con las evidencias presentadas por la fiscalía y la defensa. Así, la regla de Bayes es el fundamento de los métodos de investigación en casi todas las ciencias y, por supuesto, en ingeniería. Por ejemplo, el usuario de una red puede tener una creencia a priori del ancho de banda disponible para él. Pero si ajusta su creencia de acuerdo con datos experimentales tales como el retardo, las pérdidas y las variaciones de retardo (jitter), obtendrá un estimado más realista del ancho de banda disponible. Todos los anteriores son ejemplos de lo que en ciencias y estadística se conoce indistintamente como prueba de hipótesis, teoría de la detección, toma de decisiones, técnicas de clasificación, etc., y, en todas ellas, la regla de Bayes es la herramienta fundamental.

21. Eventos Independientes

Sea $(\Omega, \mathcal{F}, \mathbf{P})$ *un espacio de probabilidad en el que hay dos eventos* A *y* $B \in \mathcal{F}$. *A y B son independientes si y sólo si* $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$ *o, equivalentemente, si* $\mathbf{P}(A|B) = \mathbf{P}(A)$ *y* $\mathbf{P}(B|A) = \mathbf{P}(B)$.

Tres eventos medibles A, B y C son independientes si se cumplen las siguientes cuatro condiciones: (1) $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$, (2) $\mathbf{P}(A \cap C) = \mathbf{P}(A)\mathbf{P}(C)$, (3) $\mathbf{P}(B \cap C) = \mathbf{P}(B)\mathbf{P}(C)$, y (4) $\mathbf{P}(A \cap B \cap C) = \mathbf{P}(A)\mathbf{P}(B)\mathbf{P}(C)$.

En general, los eventos medibles $\{A_n, n=1,2,\ldots\}$ *forman una secuencia de eventos independientes si* $\mathbf{P}\left(\bigcap_{i\in I} A_i\right) = \prod_{i\in I} \mathbf{P}(A_i) \quad \forall I \subset \{1,2,\ldots\}$

Nuevamente, ésta es simplemente una definición. Pero es una definición muy poderosa, pues nos dice que si A y B son independientes, nuestra incertidumbre respecto a la ocurrencia de A no cambia cuando nos informan que B ocurrió. En términos de nuestra interpretación frecuentista, supongamos que hacemos N repeticiones del experimento en las que observamos que A ocurrió N_A veces, B ocurrió N_B veces, y ambos ocurrieron simultáneamente $N_{A\cap B}$ veces. Supongamos que la frecuencia relativa del evento A en N repeticiones del experimento, N_A/N, tiende al mismo valor al que tiende la frecuencia relativa del evento A en las N_B veces que ocurrió el evento B, $N_{A\cap B}/N_B$, a medida que N tiende a infinito. Siendo así, para estimar $\mathbf{P}(A)$ como el límite de la frecuencia relativa, nos daría igual si nos fijamos en todas las N repeticiones o sólo en aquellas N_B repeticiones en que ocurrió B, pues la restricción del espacio muestral de Ω a B no altera la frecuencia relativa de A.

A veces puede ser fácil identificar si dos eventos son o no son independientes. Por ejemplo sean A = {*me va a ir bien en mi matrimonio*} y B = {*Mi prometida y yo tenemos el mismo nivel de educación*}. Nadie duda que $\mathbf{P}(A|B) > \mathbf{P}(A)$ y que $\mathbf{P}(A|B^C) < \mathbf{P}(A)$, de manera que A y B no son independientes. Sin embargo, si definimos C como el evento {*Yo soy Tauro y mi novia es Libra*}, resulta sorprendente la

cantidad de personas que creen que $\mathbf{P}(A|C) \in \{0,1\}$ independientemente de B. Yo, personalmente, creo que A y C son eventos independientes, de manera que $\mathbf{P}(A|C) = \mathbf{P}(A)$.

En nuestro mundo de las redes de telecomunicaciones, en muchas ocasiones debemos admitir que ciertos eventos no son independientes, aunque preferimos suponer independencia para mantener el análisis matemático tratable. Por ejemplo, muchos resultados útiles suponen que la presencia de errores de transmisión en una trama es independiente de la presencia de errores en la trama inmediatamente anterior. Tal vez en enlaces satelitales o de fibra óptica se pueda argumentar la validez de esa suposición, pero no en enlaces terrestres de radio o de cobre donde los errores se pueden deber, por ejemplo, a la ignición eléctrica de un motor de combustión o a la operación cercana de un horno de microondas. Igualmente, al modelar el tráfico sobre una red, muchas veces preferimos suponer que el tiempo entre la llegada del paquete n-1 y la del paquete n es independiente del tiempo entre la llegada del paquete n y la del paquete n+1. Seguramente, si se trata del punto de acceso a la red de un gran número de usuarios, esta suposición de independencia se pueda justificar. Pero si se trata de paquetes de un mismo flujo o si los paquetes ya han sido sometidos a interacciones debidas a los protocolos de la red, es muy difícil aceptar que sus tiempos entre llegadas puedan ser independientes. Sin embargo, tan poderoso es el concepto de independencia que, aún en estos casos, a veces preferimos suponer independencia con la esperanza de que los resultados obtenidos al final del análisis no estén muy alejados de la realidad.

Por alguna razón muy común (que no he logrado detectar!), muchos estudiantes neófitos de teoría de probabilidad suelen equiparar la independencia de dos eventos con la exclusión mutua entre ellos. Si A y B son mutuamente excluyentes y por lo menos uno de ellos tiene probabilidad mayor que cero, resulta imposible que sean independientes porque $\mathbf{P}(A|B) = 0$ y $\mathbf{P}(B|A) = 0$, de manera que sólo podrían ser independientes si ambos eventos son nulos. De la misma manera, si dos eventos son independientes, resulta imposible que sean mutuamente excluyentes, a menos que ambos sean eventos nulos. Considérese, por ejemplo, el experimento de seleccionar un punto de un rectángulo unitario como el de la figura 9, descrito mediante el siguiente espacio de probabilidad

$$(\Omega = \{(x,y) \in \mathbb{R}^2 : 0 \le x, y \le 1\},\ \mathcal{F} = \mathcal{B}(\mathbb{R}^2) \cap \Omega,\ \mathbf{P}(A) = \text{Área}(A)\ \forall A \in \mathcal{F}).^{10}$$

Sean $A = \{(x,y) \in \Omega : x < 0.5\}$ y $B = \{(x,y) \in \Omega : y < 0.5\}$. Claramente $\mathbf{P}(A) = \mathbf{P}(B) = \mathbf{P}(A^C) = \mathbf{P}(B^C) =$ ½. A y B no son mutuamente excluyentes porque $A \cap B = \{(x,y) \in \Omega : x < 0.5, y < 0.5\} \neq \Phi$. Pero A y B sí son independientes porque el área de $A \cap B$ es ¼, que es la mitad del área de B, de manera que $\mathbf{P}(A|B) =$ ½ $= \mathbf{P}(A)$ o, mejor aún, $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B) =$ ¼. De manera semejante, A y A^C son mutuamente excluyentes, por lo que $\mathbf{P}(A|A^C) = 0 < \mathbf{P}(A) = 0.5$: A y A^C no pueden ser independientes.

[10] Por definición, la "intersección" entre un conjunto y una clase de conjuntos, $A \cap \mathcal{H}$, se refiere a la restricción de la clase $\mathcal{H}$ al evento A, esto es, $A \cap \mathcal{H} = \{ A \cap H : H \in \mathcal{H}\}$ es una nueva clase (reducida) de eventos.

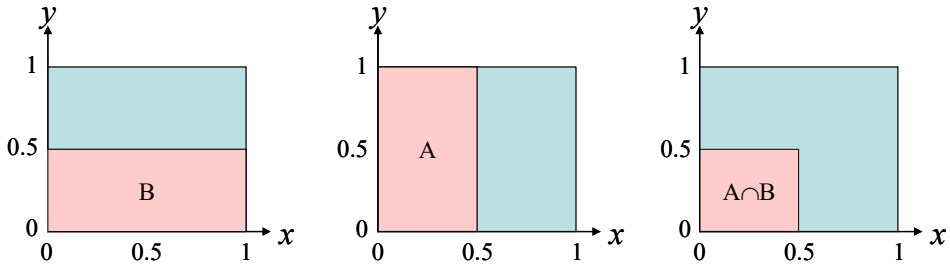


Figura 27. Distinción entre independencia y exclusión mutua

Es importante notar que, si A y B son independientes, también los son A^C y B. En efecto, como B es la unión de dos eventos mutuamente excluyentes, $(A\cap B) \cup (A^C\cap B)$, encontramos que $\mathbf{P}(A^C\cap B) = \mathbf{P}(B) - \mathbf{P}(A\cap B) = \mathbf{P}(B) - \mathbf{P}(A)\mathbf{P}(B) = (1 - \mathbf{P}(A))\mathbf{P}(B) = \mathbf{P}(A^C)\mathbf{P}(B)$. Por la misma razón, si A y B son independientes, también los son A y B^C, así como A^C y B^C.

Tres eventos pueden ser independientes por pares, sin necesidad de que sean tres eventos independientes. Considere, por ejemplo, una fuente de información que es capaz de generar tres símbolos $\{a, b, c\}$ con los cuales puede construir nueve mensajes, $\Omega = \{abc, acb, bac, bca, cab, cba, aaa, bbb, ccc\}$, cada uno con probabilidad 1/9. Sea $A_k = \{$*el k-ésimo símbolo del mensaje es "a"*$\}$, k=1,2,3, de manera que $\mathbf{P}(A_1) = \mathbf{P}(\{abc, acb, aaa\}) = 1/3$, $\mathbf{P}(A_2) = \mathbf{P}(\{bac, cab, aaa\}) = 1/3$ y $\mathbf{P}(A_3) = \mathbf{P}(\{bca, cba, aaa\}) = 1/3$. Claramente, A_1, A_2 y A_3 son independientes por pares porque $\mathbf{P}(A_i\cap A_j) = \mathbf{P}(\{aaa\}) = 1/9 = \mathbf{P}(A_i)\mathbf{P}(A_j)$ si $i\neq j$. Sin embargo no son tres eventos independientes porque $\mathbf{P}(A_1\cap A_2\cap A_3) = \mathbf{P}(\{aaa\}) = 1/9 \neq \mathbf{P}(A_1)\mathbf{P}(A_2)\mathbf{P}(A_3) = 1/27$.

El caso contrario también puede ocurrir: $\mathbf{P}(A\cap B\cap C) = \mathbf{P}(A)\mathbf{P}(B)\mathbf{P}(C)$ pero $\mathbf{P}(A\cap B) \neq \mathbf{P}(A)\mathbf{P}(B)$. Un ejemplo trivial pero definitivo puede ser el siguiente: Se mide el tiempo de transferencia ftp de un archivo de 100 kbytes y se definen los siguientes eventos $A = \{$*menos de* 1 *segundo*$\}$, $B = \{$*menos de* 100 *ms*$\}$ y $C = \{0$ *segundos*$\}$. Claramente, $C \subset B \subset A$ de manera que $0 = \mathbf{P}(C) < \mathbf{P}(B) < \mathbf{P}(A) < 1$, por lo que $\mathbf{P}(A\cap B) = \mathbf{P}(B) > \mathbf{P}(A)\,\mathbf{P}(B)$, por lo cual A y B no son independientes, pero $\mathbf{P}(A\cap B\cap C) = \mathbf{P}(C) = \mathbf{P}(A)\mathbf{P}(B)\mathbf{P}(C) = 0$, de manera que la probabilidad de la intersección de los tres eventos es igual al producto de las tres probabilidades individuales.

Como ejemplo del poder del concepto de independencia considere la red de la Figura 28 en la que cada enlace falla con probabilidad p, independientemente de los otros enlaces. ¿Cuál es la probabilidad de que exista una ruta desde A hasta C? Sean $E_i = \{$El enlace e_i está bueno$\}$, i=1,2,3,4,5, y $R = \{$Existe una ruta entre A y $C\}$. Considerando la partición del espacio muestral dada por E_5 y E_5^C, podemos aplicar el teorema de la probabilidad total así:

$$\mathbf{P}(R) = \mathbf{P}(E_5)\mathbf{P}(R \mid E_5) + \mathbf{P}(E_5^C)\,\mathbf{P}(R \mid E_5^C)$$

Donde $\mathbf{P}(E_5) = 1 - p$, $\mathbf{P}(R \mid E_5) = 1$, y $\mathbf{P}(E_5^C) = p$, de manera que

$$\mathbf{P}(R) = 1 - p + p\,\mathbf{P}(R \mid E_5^C)$$

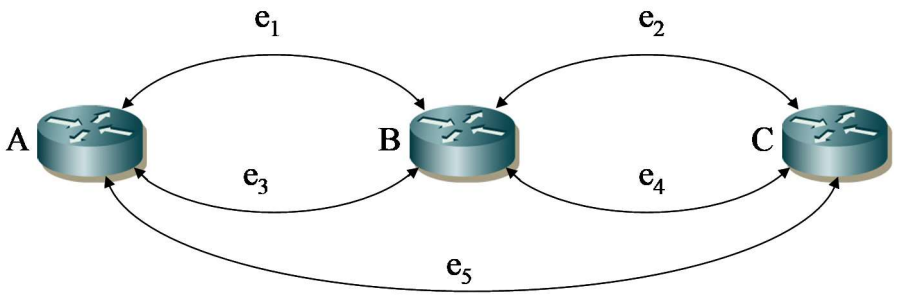


Figura 28. Red para análisis de confiabilidad

Si e_5 está dañado, A y C sólo se podrán comunicar a través de B, para lo cual se requiere que de A se pueda llegar a B ($E_1 \cup E_3$) y que de B se pueda llegar a C ($E_2 \cup E_4$):

$$\mathbf{P}(R \mid E_5^C) = \mathbf{P}((E_1 \cup E_3) \cap (E_2 \cup E_4))$$

Aquí es cuando la suposición de independencia facilita enormemente el problema pues, si E_1 y E_3 son independientes de E_2 y E_4, $\mathbf{P}(R \mid E_5^C)$ será el producto de las dos probabilidades que, por simetría, son idénticas:

$$\mathbf{P}(R \mid E_5^C) = \mathbf{P}(E_1 \cup E_3)^2$$

Pero $E_1 \cup E_3 = (E_1^C \cap E_3^C)^C$, por lo que podemos aplicar nuevamente la independencia de E_1 y E_3: $\mathbf{P}(E_1 \cup E_3) = 1 - \mathbf{P}(E_1^C \cap E_3^C) = 1 - \mathbf{P}(E_1^C)\mathbf{P}(E_3^C) = 1 - p^2$. Reemplazando,

$$\mathbf{P}(R) = 1 - p + p(1 - p^2)^2$$

Si los enlaces no fallaran independientemente unos de otros, la solución del problema sería enormemente compleja.

Dos eventos pueden no ser independientes, a menos que se condicionen a un tercer evento: $\mathbf{P}(A \cap B) \neq \mathbf{P}(A)\mathbf{P}(B)$ pero $\mathbf{P}(A \cap B|C) = \mathbf{P}(A|C)\mathbf{P}(B|C)$, en cuyo caso se dice que A y B son condicionalmente independientes. Considérese, por ejemplo, un enlace en tiempo discreto donde la unidad de tiempo es el tiempo de transmisión de un paquete. Sea $X[n]$ = Número de paquetes transmitidos hasta el instante n, con $X[0] = 0$. Definamos los siguientes eventos:

$A_2 = \{X[2] = 1\} = \{01, 10\}$, donde 01 indica 0 paquetes transmitidos en el primer slot y 1 paquete transmitido en el segundo slot.

$A_3 = \{X[3] = 2\} = \{011, 101, 110\}$

$A_4 = \{X[4] = 2\} = \{0011, 0101, 0110, 1001, 1010, 1100\}$.

Si en cada slot se transmite un paquete con probabilidad p independientemente de los slots vecinos, tenemos

$$\mathbf{P}(A_2) = 2p(1\text{-}p), \quad \mathbf{P}(A_3) = 3p^2(1\text{-}p), \quad \mathbf{P}(A_4) = 6p^2(1\text{-}p)^2$$

Obsérvese que $\mathbf{P}(A_2 \cap A_4) = \mathrm{P}(\{0101, 0110, 1001, 1010\}) = 4p^2(1\text{-}p)^2 \neq \mathbf{P}(A_2)\mathbf{P}(A_4) = 12p^3(1\text{-}p)^3$, de manera que A_2 y A_4 no son independientes. Sin embargo consideremos las siguientes probabilidades:

$\mathbf{P}(A_2 \mid A_3) = \mathbf{P}(\{011, 101\}) / \mathbf{P}(\{011, 101, 110\}) = 2p^2(1\text{-}p) / 3p^2(1\text{-}p) = 2/3$
$\mathbf{P}(A_4 \mid A_3) = \mathbf{P}(\{0110, 1010, 110\}) / \mathbf{P}(\{011, 101, 110\}) = 3p^2(1\text{-}p)^2 / 3p^2(1\text{-}p) = 1 - p$
$\mathbf{P}(A_2 \cap A_4 \mid A_3) = \mathbf{P}(\{0110, 1010\}) / \mathbf{P}(\{011, 101, 110\}) = 2p^2(1\text{-}p)^2 / 3p^2(1\text{-}p) = 2(1 - p)/3$

Claramente, $\mathbf{P}(A_2 \cap A_4 \mid A_3) = \mathbf{P}(A_2 \mid A_3)\ \mathbf{P}(A_4 \mid A_3)$, de manera que A_2 y A_4 son condicionalmente independientes dado A_3.

El anterior ejemplo es una Cadena de Markov (ver definición 122), cuya principal características es que, aunque el futuro depende del pasado, el futuro resulta condicionalmente independiente del pasado cuando se conoce el presente. Esta propiedad me parece un principio importante para aplicar a una vida positiva: Todo mi futuro depende solamente de quién soy yo en este momento, independientemente de cómo llegué a ser lo que soy. Mi futuro sólo dependerá de mi pasado si yo no sé quién soy en este momento.

22. Experimentos Compuestos

Sean $(\Omega_1, \mathcal{F}_1, \mathbf{P}_1)$ y $(\Omega_2, \mathcal{F}_2, \mathbf{P}_2)$ dos espacios de probabilidad que corresponden a dos experimentos aleatorios. Si realizamos los dos experimentos y vemos el par de resultados como un único resultado de un experimento aleatorio compuesto, tendremos un nuevo espacio de probabilidad conjunto en el que el espacio muestral es el producto cartesiano $\Omega_1 \times \Omega_2$, correspondiente a los pares ordenados $\{(\omega_1,\omega_2) : \omega_1 \in \Omega_1,\ \omega_2 \in \Omega_2\}$. Los eventos de interés, en general, pertenecerán al producto cartesiano $\mathcal{F}_1 \times \mathcal{F}_2$, correspondiente a los conjuntos de pares ordenados $\{(\omega_1,\omega_2) \in A_1 \times A_2\},\ \forall A_1 \in \mathcal{F}_1,\ \forall A_2 \in \mathcal{F}_2$[11]. Sin embargo, $\mathcal{F}_1 \times \mathcal{F}_2$ no necesariamente forma un campo sigma de subconjuntos de $\Omega_1 \times \Omega_2$, por lo que será necesario cerrarlo para los complementos y las uniones contables. Así, un campo sigma apropiado para el experimento compuesto es $\sigma(\mathcal{F}_1 \times \mathcal{F}_2)$, el mínimo campo sigma de subconjuntos de $\Omega_1 \times \Omega_2$ que incluye los eventos que pertenencen a $\mathcal{F}_1 \times \mathcal{F}_2$, que son los pares (ω_1,ω_2) en los que ω_1 pertenece a algún evento A_1 en $\mathcal{F}_1$ y ω_2 pertenece a algún evento A_2 en $\mathcal{F}_2$.

Por ejemplo, sean $\Omega_1 = \{a,b\}$ con $\mathcal{F}_1 = \{\phi, \{a\}, \{b\}, \{a,b\}\}$ y $\Omega_2 = \{1,2,3\}$ con $\mathcal{F}_2 = \{\phi, \{1,2\}, \{3\}, \{1,2,3\}\}$. El espacio muestral del experimento compuesto es $\Omega_1 \times \Omega_2 = \{a1, a2, a3, b1, b2, b3\}$ y el campo sigma del experimento compuesto debe incluir a los conjuntos $\mathcal{F}_1 \times \mathcal{F}_2 = \{\phi, \{a3\}, \{b3\}, \{a1,a2\}, \{a3,b3\}, \{b1,b2\}, \{a1,a2,a3\}, \{b1,b2,b3\}, \{a1,a2,b1,b2\}, \{a1,a2,a3,b1,b2\}, \{a1,a2,b1,b2,b3\}, \{a1,a2,a3,b1,b2,b3\}\}$. Evidentemente, éste no es un campo sigma de subconjuntos de $\Omega_1 \times \Omega_2$ aunque $\mathcal{F}_1$ y $\mathcal{F}_2$ sean campos sigma de subconjuntos de Ω_1 y Ω_2, respectivamente. Para completar un campo sigma apropiado para el experimento compuesto, debemos cerrar $\mathcal{F}_1 \times \mathcal{F}_2$ con los complementos y las uniones, lo que implica añadir seis eventos compuestos adicionales:

$\sigma(\mathcal{F}_1 \times \mathcal{F}_2) = \{\phi, \{a3\}, \{b3\}, \{a1,a2\}, \{a3,b3\}, \{b1,b2\}, \{a1,a2,a3\}, \{a1,a2,b3\}, \{a3,b1,b2\}, \{b1,b2,b3\},$
$\{a1,a2,a3,b3\}, \{a1,a2,b1,b2\}, \{a3,b1,b2,b3\}, \{a1,a2,a3,b1,b2\}, \{a1,a2,b1,b2,b3\},$
$\{a1,a2,a3,b1,b2,b3\}\}$

Nótese que $|\mathcal{F}_1 \times \mathcal{F}_2| = 10$, $|\sigma(\mathcal{F}_1 \times \mathcal{F}_2)| = 16$ y $|\{0,1\}^{\Omega_1 \times \Omega_2}| = 64$.

[11] Donde $\phi \times A = \phi\ \forall A \subset \Omega$

Falta considerar la medida de probabilidad en este nuevo espacio medible $(\Omega_1\times\Omega_2,\sigma(\mathcal{F}_1\times\mathcal{F}_2))$, que podría ser tan simple como $P(A_1,A_2) = P_1(A_1)P_2(A_2)$ si los dos experimentos que componen el experimento compuesto son independientes[12]. En general, puede haber alguna dependencia entre ellos, de manera que $P(A_1,A_2) = P_1(A_1)P_{21}(A_2|A_1) = P_2(A_2)P_{12}(A_1|A_2)$, donde las probabilidades condicionales $P_{21}(A_2|A_1)$ y $P_{12}(A_1|A_2)$ involucrarían las dependencias que puedan ocurrir entre los dos experimentos (esto es, considerando la naturaleza de los experimentos, en cuyo caso las probabilidades condicionales hablan sobre los procesos de observación correspondientes).

Sigamos considerando el ejemplo anterior añadiendo medidas de probabilidad. El primer espacio de probabilidad es $(\Omega_1=\{a,b\}$, $\mathcal{F}_1=\{\phi,\{a\},\{b\},\{a,b\}\}$, $\mathbf{P}_1=\{P_1(\phi)=0,P_1(\{a\})=p$, $P_1(\{b\})=1\text{-}p$, $P_1(\{a,b\})=1\})$, que corresponde a escoger una caja entre dos cajas, una marcada con la letra *a* y otra marcada con la letra *b*, y lees la letra marcada en la caja escogida. El segundo espacio de probabilidad es $(\Omega_2=\{1,2,3\}$, $\mathcal{F}_2=\{\phi,\{1,2\},\{3\},\{1,2,3\}\}$, $\mathbf{P}_2=\{P_2(\phi)=0,P_2(\{1,2\})=q,P_2(\{3\})=1\text{-}q$, $P_2(\{1,2,3\})=1\})$, que corresponde a escoger una bola de billar de una bolsa oscura y leer el número marcado en la bola, donde algunas bolas están marcadas con el número 1, otras con el número 2 y las demás con el número 3. Sabemos que el espacio de probabilidad asociado con el experimento compuesto es $(\Omega_3=\Omega_1\times\Omega_2=\{a1,a2,a3,b1,b2,b3\},\mathcal{F}_3=\sigma(\mathcal{F}_1\times\mathcal{F}_2)=\{\phi$, $\{a3\}$, $\{b3\}$, $\{a1,a2\}$, $\{a3,b3\}$, $\{b1,b2\}$, $\{a1,a2,a3\}$, $\{a1,a2,b3\}$, $\{a3,b1,b2\}$, $\{b1,b2,b3\}$, $\{a1,a2,a3,b3\}$, $\{a1,a2,b1,b2\}$, $\{a3,b1,b2,b3\}$, $\{a1,a2,a3,b1,b2\}$, $\{a1,a2,b1,b2,b3\}$, $\{a1,a2,a3,b1,b2,b3\}\}$, $\mathbf{P}_3)$, pero ¿cómo es la medida de probabilidad $\mathbf{P}_3$? Depende de la naturaleza de la composición entre los experimentos. Si la bola se escoge de una bolsa oscura que contiene todas las bolas y la caja se escoge de una bolsa oscura que contiene las dos cajas, el resultado del experimento 1 no tiene nada que ver con el resultado del experimento 2, de manera que los experimentos son independientes:
$P_3(\phi)=0,P_3(\{a3\})=p(1-q)$, $P_3(\{b3\})=(1-p)(1-q)$, $P_3(\{a1,a2\})=pq$, $P_3(\{a3,b3\})=1-q$, $P_3(\{b1,b2\})=(1-p)q$, $P_3(\{a1,a2,a3\})=p$, $P_3(\{a1,a2,b3\})=1+2pq-p-q$, $P_3(\{a3,b1,b2\})=p+q-2pq$, $P_3(\{b1,b2,b3\})=1-p$, $P_3(\{a1,a2,a3,b3\})=1-q(1-p)$, $P_3(\{a1,a2,b1,b2\})=q$, $P_3(\{a3,b1,b2,b3\})=1-pq$, $P_3(\{a1,a2,a3,b1,b2\})=p+q-pq$, $P_3(\{a1,a2,b1,b2,b3\})=1-p(1-q)$, $P_3(\{a1,a2,a3,b1,b2,b3\})=1$.

Pero si el experimento compuesto es tal que la bola se obtiene de la caja seleccionada, la probabilidad del experimento compuesto dependerá de la distribución de las bolas en cada una de las dos cajas. Si la fracción de bolas marcadas con 3 en la caja *a* es *r* y la fracción de bolas marcadas con 3 es la caja *b* es *s*, el experimento sólo tendrá sentido si $q = 1 - pr - (1-p)s$, lo cual demuestra la dependencia de los dos experimentos.

Es fácil generalizar esta idea a una secuencia de experimentos, $\{(\Omega_n, \mathcal{F}_n, \mathbf{P}_n), n = 1,2,\ldots,N\}$, donde N puede ser cualquier número natural o, inclusive, infinito. Por ejemplo, los intentos repetidos de un experimento generarán nuevos espacios de probabilidad en los que

[12] Aquí debemos ser cuidadosos. El concepto de independencia que invocamos aquí no es exactamente el mismo de la definición 21 porque los eventos no están definidos en el mismo espacio de probabilidad. Por eso necesitamos esta definición 22. Los eventos

a. el espacio muestral para n repeticiones es el n-ésimo producto cartesiano del espacio muestral del experimento individual, $\Omega^{(n)} = \Omega \times \Omega \times \ldots \times \Omega$;
b. el campo de eventos medibles será el mínimo campo-σ que incluya los n-ésimos productos cartesianos del campo de eventos del experimento individual, $\mathcal{F}^{(n)} = \sigma(\mathcal{F} \times \mathcal{F} \times \ldots \times \mathcal{F})$;
c. la medida de probabilidad de un evento medible en el nuevo espacio dependerá de las posibles dependencias entre repeticiones del experimento. Si son independientes, por ejemplo, la probabilidad de un evento medible en el nuevo espacio será el producto de las medidas de probabilidad de los eventos respectivos en cada repetición individual.

La tercera característica del nuevo espacio de probabilidad es la razón por la que en teoría de probabilidades se le da tanto énfasis y tanta importancia al concepto de independencia. Varios autores coinciden que es la independencia la que hace de la teoría de la probabilidad un área de la matemática aparte de la teoría de la medida.

El ejemplo más típico es el de la repetición de experimentos de Bernoulli. En cada repetición el campo de eventos es $\mathcal{F}=\{\phi, A, A^C, \Omega\}$ para algún $A \subset \Omega$, como en el caso de lanzar una moneda ($A=\{\text{cara}\}$), ver el estado de ocupación de un enlace ($A=\{\text{ocupado}\}$) o medir el retardo de un paquete y ver si supera 100 ms ($A = \{\omega \in \Omega | \omega > 0.1\}$). Cuando hacemos dos repeticiones, el espacio muestral es Ω^2 y los eventos de interés pertenencen a $\mathcal{F}^2$, y, en general, cuando hacemos n repeticiones, el espacio muestral es Ω^n y los eventos de interés pertenencen a $\mathcal{F}^n$. Aquí hay dos interpretaciones importantes para el experimento compuesto: En una interpretación estamos viendo estadísticas de un único experimento desde un punto de vista frecuentista, de manera que si en las n repeticiones obtenemos n_A veces el evento A, diríamos que la probabilidad de A es cercana a n_A/n. Pero ahora hacemos una segunda interpretación: No se trata de n veces un experimento de Bernoulli sino de un único nuevo experimento cuyo único proceso de observación consiste en repetir n veces un experimento de Bernoulli. Un posible resultado es que en las primeras n_A repeticiones obtengamos A y en las restantes $n-n_A$ obtengamos A^C, cuya probabilidad es $p^{n_A}(1-p)^{n-n_A}$ si las repeticiones son independientes. Ahora las preguntas sobre el experimento (único) se refieren a cuáles repeticiones obtuvieron A y cuáles obtuvieron A^C. Por ejemplo, una pregunta válida en ese único experimento es cuántas de las n repeticiones fueron A, en cuyo caso obtendríamos la frecuencia relativa asociada con la interpretación anterior del experimento compuesto en donde hacemos n experimentos diferentes.

Como otro ejemplo consideremos una "regla de parada" tal como lanzar una moneda repetidamente hasta que aparezca una cara y, entonces, detenerse. El espacio muestral más sencillo sería $\cup_{n=0,1,2,\ldots} \{\{s\}^n, c\}$ que es el conjunto de secuencias de cero o más sellos seguidos por una cara. Por supuesto, surge la inquietud si debemos incluir el caso $n=\infty$. Si consideramos que la probabilidad de que eventualmente aparezca una cara es 1, tal vez sería útil eliminar el caso $n=\infty$ de nuestro espacio muestral pues, después de todo, ¿cómo podríamos llegar a observar ese resultado? Si la probabilidad de cara es mayor a cero, no importa cuan pequeña sea, eventualmente ocurrirá una cara con probabilidad 1. Las siguientes definiciones aclararán esta idea fundamental del la ocurrencia "eventual" de un evento.

23. Continuidad de la Medida de Probabilidad

Sea $(\Omega, \mathcal{F}, \mathbf{P})$ *un espacio de probabilidad en el que hay una secuencia creciente de eventos* $A_1 \subseteq A_2 \subseteq A_3 \subseteq \ldots$ y *una secuencia decreciente de eventos* $B_1 \supseteq B_2 \supseteq B_3 \supseteq \ldots$. *contenidos en* $\mathcal{F}$. *Entonces*

$$P\left[\lim_{i\to\infty} A_i\right] = \lim_{i\to\infty} P[A_i] \quad y \quad P\left[\lim_{i\to\infty} B_i\right] = \lim_{i\to\infty} P[B_i]$$

Para una secuencia creciente de eventos sabemos por la definición 9 que el límite de la secuencia es la unión de todos los eventos

$$\lim_{i\to\infty} A_i = \bigcup_{n=1}^{\infty} A_n$$

la cual se puede expresar como unión de eventos mutuamente excluyentes,

$$\lim_{i\to\infty} A_i = A_1 \cup (A_2 \setminus A_1) \cup (A_3 \setminus A_2) \cup (A_4 \setminus A_3) \cup \cdots$$

para aplicar el tercer axioma de la definición 14:

$$P\left[\lim_{i\to\infty} A_i\right] = P[A_1] + P[A_2 \setminus A_1] + P[A_3 \setminus A_2] + P[A_4 \setminus A_3] + \cdots$$

$$P\left[\lim_{i\to\infty} A_i\right] = P[A_1] + \sum_{n=2}^{\infty} P[A_n \setminus A_{n-1}] = P[A_1] + \lim_{i\to\infty} \sum_{n=2}^{i} P[A_n \setminus A_{n-1}]$$

$$P\left[\lim_{i\to\infty} A_i\right] = P[A_1] + \lim_{i\to\infty} \sum_{n=2}^{i} \left(P[A_n] - P[A_{n-1}]\right) = P[A_1] + \lim_{i\to\infty} \left(P[A_i] - P[A_1]\right)$$

$$P\left[\lim_{i\to\infty} A_i\right] = \lim_{i\to\infty} P[A_i]$$

Igualmente, el límite de una secuencia decreciente de eventos es la intersección de todos los eventos (definición 9)

$$\lim_{i\to\infty} B_i = \bigcap_{n=1}^{\infty} B_n$$

de manera que, según la ley de DeMorgan, la secuencia de complementos es una secuencia creciente:

$$\left(\lim_{i\to\infty} B_i\right)^C = \left(\bigcap_{n=1}^{\infty} B_n\right)^C = \bigcup_{n=1}^{\infty} B_n{}^C = \lim_{i\to\infty} B_i{}^C$$

a la que aplica el resultado anterior:

$$P\left[\lim_{i\to\infty} B_i{}^C\right] = \lim_{i\to\infty} P\left[B_i{}^C\right]$$

esto es,

$$1 - P\left[\lim_{i\to\infty} B_i\right] = 1 - \lim_{i\to\infty} P[B_i]$$

o, lo que es lo mismo,

$$P\left[\lim_{i\to\infty} B_i\right] = \lim_{i\to\infty} P[B_i]$$

Consideremos, por ejemplo, una secuencia de lanzamientos de una moneda equilibrada. ¿Con qué probabilidad obtendremos eventualmente una cara? Si definimos el evento A_n como el evento en el que ha habido al menos una cara en las primeras n lanzadas habremos creado una secuencia creciente

de eventos, ya que $A_1 \subset A_2 \subset A_3 \subset \ldots$ Claramente, el evento "obtendremos eventualmente una cara" corresponde al evento $\lim_{n\to\infty} A_n$. Por la continuidad de la medida de probabilidad sabemos que $P[\lim_{n\to\infty} A_n] = \lim_{n\to\infty} P[A_n]$. La probabilidad de A_1 es 1/2, la probabilidad de A_2 es 3/4, la probabilidad de A_3 es 7/8 y, en general, la probabilidad de A_n es $(2^n-1)/2^n$. Esto es, $\lim_{n\to\infty} P[A_n] = 1$, por lo que $P[\lim_{n\to\infty} A_n] = 1$: Eventualmente obtendremos una cara con probabilidad 1. Claro, como la probabilidad de cara es tan alta, es de esperar que veamos una cara con bastante frecuencia. Pero si apostamos al baloto en todas sus rifas, también es cierto que eventualmente lo ganaremos pues cada vez que lo juguemos ganaremos con probabilidad $p=(1/16)(5/43)(4/42)(3/41)(2/40)(1/39) = 1/15'401.568 > 0$), de manera que el evento {Ganaremos al menos una vez en n intentos} tiene probabilidad $1-(1-p)^n \to 1$. El problema es que esta vez no veremos que dicho evento ocurra con mucha frecuencia pues, si participamos en todas las apuestas, ganaremos en promedio una vez cada ciento cincuenta mil años[13]. Sin embargo, es interesante notar que si, podemos seguir jugando infinitas veces al baloto, lo ganaremos un número infinito de veces, como lo demuestra el segundo lema de Borel-Cantelli que veremos en la definición 25, después de la siguiente breve introducción.

24. Límite del supremo y límite del ínfimo

Sea $(\Omega, \mathcal{F}, \mathbf{P})$ *un espacio de probabilidad en el que hay una secuencia de eventos* $\{A_1, A_2, A_3, \ldots\}$. *Definimos el límite del supremo y el límite del ínfimo de la secuencia mediante las siguientes relaciones:*

$$\limsup_{n\to\infty} A_n = \bigcap_{m=1}^{\infty} \bigcup_{k=m}^{\infty} A_k = \{\omega \in \Omega : \forall n \exists k > n : \omega \in A_K\}$$

$$\liminf_{n\to\infty} A_n = \bigcup_{m=1}^{k} \bigcap_{k=m}^{\infty} A_k = \{\omega \in \Omega : \exists n : \forall k > n : \omega \in A_K\}$$

Cuando ocurre el evento $\limsup_{n\to\infty} A_n$ *decimos que ocurrió un número infinito de eventos de la secuencia* $\{A_n, n=1,2,\ldots\}$, *por lo que el límite del supremo también se conoce como* $\{A_n$, i.o.$\}$ (*infinitely often*). *Cuando ocurre el evento* $\liminf_{n\to\infty} A_n$ *decimos que ocurrieron todos los eventos de la secuencia* $\{A_n, n=1,2,\ldots\}$, *con la posible excepción de un número finito de ellos, por lo que el límite del ínfimo también se conoce como* $\{A_n$, e.a.$\}$ (*eventually always*).

Claramente, $\liminf_{n\to\infty} A_n \subseteq \limsup_{n\to\infty} A_n$. *Si también ocurre que* $\lim_{n\to\infty}\sup A_n \subseteq \lim_{n\to\infty}\inf A_n$, *entonces la secuencia de eventos* $A_1, A_2, A_3, \ldots$ *tiene como límite el evento* L *dado por*

$$L = \lim_{n\to\infty}\sup A_n = \lim_{n\to\infty}\inf A_n$$

Para empezar, consideremos primero una secuencia de números reales $x_1, x_2, x_3, \ldots$ El supremo de la secuencia $\{x_m, m \geq n\}$ es la cota superior más pequeña de $\{x_m, m \geq n\}$ (por eso al supremo también se le conoce como LUB, *least upper bound*). Dicho de otra manera, el supremo de $\{x_m, m \geq n\}$ es el mínimo $x \in \mathbb{R}$ tal que $x \geq x_m \ \forall \ m \geq n$. Nótese que el supremo no necesariamente pertenece a la

[13] Ver definición 43(b)

secuencia $\{x_m, m \geq n\}$. Si así fuera, al supremo le diríamos también el máximo, pues en este caso coinciden. En otras palabras, el máximo debe hacer parte de la secuencia mientras que el supremo no está necesariamente contenido en la secuencia. Por ejemplo, $\sup_{n\geq 0}\left(1-e^{-n}\right)=1$ aunque $\max_{n\geq 0}\left(1-e^{-n}\right)$ no existe porque la secuencia tiende a 1 sin llegar jamás a ese valor. Sin embargo $\sup_{n\geq 0}\left((-1)^n\right)=\max_{n\geq 0}\left((-1)^n\right)=1$. De manera análoga, el ínfimo de una secuencia de números es la máxima cota inferior o GLB, *greatest lower bound*: el ínfimo de $\{x_m, m \geq n\}$ es el máximo $x \in \mathbb{R}$ tal que $x \leq x_m$ $\forall$ $m \geq n$. Para secuencias finitas, el ínfimo es el mismo mínimo como el supremo es el mismo máximo. Pero en secuencias infinitas pueden darse los casos en que el ínfimo y el mínimo coincidan, o que sólo exista el ínfimo o que no exista ninguno de ellos. Por ejemplo, el ínfimo de la secuencia $X = \{1/n, n \in \mathbb{N}\}$ es 0, pero no es el mínimo de X porque $0 \notin X$. El supremo de X es 1, que es a su vez el máximo de X porque $1 \in X$.

Por supuesto, $s_n = \sup\{x_m, m \geq n\}$ es una secuencia no creciente en n mientras que $i_n = \inf\{x_m, m \geq n\}$ es una secuencia no decreciente en n, como muestra la Figura 29.

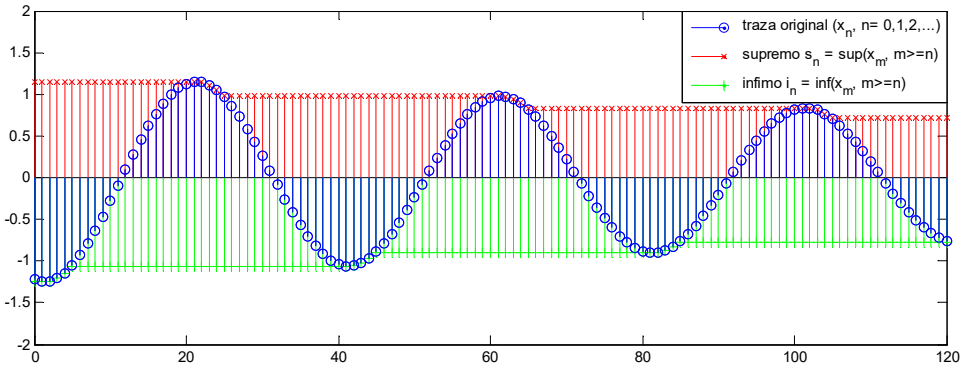


Figura 29. Supremo e ínfimo de una secuencia de números reales

Consideremos ahora el límite de la secuencia decreciente $s_n = \sup\{x_m, m \geq n\}$ cuando n tiende a infinito, conocida como *lim sup*,

$$\limsup_{n\to\infty} x_n = \lim_{n\to\infty}\left(\sup_{m\geq n} x_m\right)$$

Si l_s es el límite del supremo de $\{x_n\}$, para cualquier $\varepsilon>0$ existe un natural N tal que $x_n < l_s + \varepsilon$ para todo $n>N$, y no hay un número menor que l_s que satisfaga esta misma propiedad. En otras palabras, cualquier número mayor que l_s eventualmente se convertirá en una cota superior de la secuencia $\{x_n\}$, de manera que sólo un número finito de elementos de la secuencia pueden ser mayores a $l_s + \varepsilon$.

Igualmente se puede hablar del límite de la secuencia creciente $i_n = \inf\{x_m, m \geq n\}$ cuando n tiende a infinito, o *lim inf*,

$$\liminf_{n\to\infty} x_n = \lim_{n\to\infty}\left(\inf_{m\geq n} x_m\right)$$

Si l_i es el límite del ínfimo de $\{x_n\}$, para cualquier ε>0 existe un natural N tal que $x_n > l_i - \varepsilon$ para todo $n>N$, y no hay un número mayor que l_i que satisfaga esta misma propiedad. En otras palabras, cualquier número menor que l_i eventualmente se convertirá en una cota inferior de la secuencia $\{x_n\}$, de manera que sólo un número finito de elementos de la secuencia pueden ser menores a $l_i - \varepsilon$. En la Figura 29, el límite de la señal roja cuando n tiende a inifnito es el *lim sup* de la secuencia $\{x_n\}$ y el límite de la señal verde cuando n tiende a inifnito es el *lim inf* de la secuencia $\{x_n\}$. Si los dos límites coinciden, la secuencia converge al valor común de los límites del supremo y del ínfimo, esto es, si existe $L = \lim_{n\to\infty} x_n$, entonces $L = \lim\inf_{n\to\infty} x_n = \lim\sup_{n\to\infty} x_n$, como muestra la Figura 30.

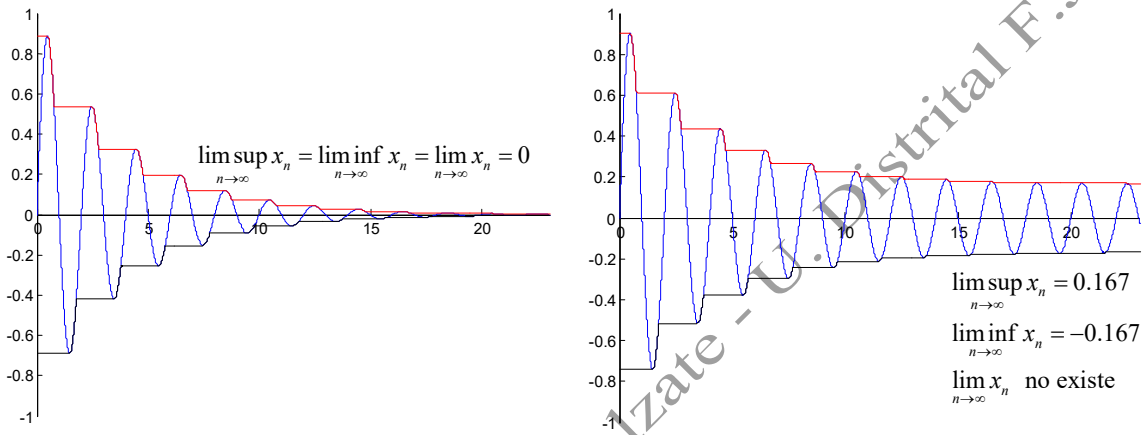


Figura 30. Los límites del supremo y del ínfimo pueden ser diferentes. Si son iguales, corresponden al límite de la secuencia (se cambió `stem` por `plot` para mejor visualización).

Permítame insistir en algunas propiedades evidentes y muy interesantes del limsup y el liminf de una secuencia de números son las siguientes:

- $\liminf_{n\to\infty}\{x_n\} \le \limsup_{n\to\infty}\{x_n\}$
- Si $\limsup_{n\to\infty}\{x_n\}<\infty$, $\forall\ \varepsilon>0\ \exists N$ tal que $\forall m>N\ x_m < \varepsilon+\limsup_{n\to\infty}\{x_n\}$ (decimos que, $\forall\ \varepsilon>0$, $x_m < \varepsilon+\limsup_{n\to\infty}\{x_n\}$ "para todos los m suficientemente grandes" o *eventually always* (e.a.))
- Si $\limsup_{n\to\infty}\{x_n\}<\infty$, $\forall\ \varepsilon>0\ \forall N\ \exists m>N$ tal que $x_m > -\varepsilon+\limsup_{n\to\infty}\{x_n\}$ (decimos que, $\forall\ \varepsilon>0$, $x_m > -\varepsilon+\limsup_{n\to\infty}\{x_n\}$ "para un número infinito de subíndices m" o *infinitely often* (i.o.))
- Si $\liminf_{n\to\infty}\{x_n\}<\infty$, $\forall\ \varepsilon>0\ \exists N$ tal que $\forall m>N\ x_m > -\varepsilon+\liminf_{n\to\infty}\{x_n\}$ (decimos que, $\forall\ \varepsilon>0$, $x_m > -\varepsilon+\liminf_{n\to\infty}\{x_n\}$ "para todos los m suficientemente grandes" o *eventually always* (e.a.))
- Si $\liminf_{n\to\infty}\{x_n\}<\infty$, $\forall\ \varepsilon>0\ \forall N\ \exists m>N$ tal que $x_m < \varepsilon+\liminf_{n\to\infty}\{x_n\}$ (decimos que, $\forall\ \varepsilon>0$, $x_m < \varepsilon+\liminf_{n\to\infty}\{x_n\}$ "para un número infinito de subíndices m" o *infinitely often* (i.o.))

Estos conceptos de limsup y liminf para secuencias de números reales son una introducción perfecta para entender los mismos conceptos asociados con una secuencia de eventos $\{A_1, A_2, A_3,\ldots\}$. En este caso podemos estar interesados en saber cuántos de los eventos A_n ocurren en una realización del experimento aleatorio. El supremo y el ínfimo de la secuencia son dos nuevas secuencias, la primera decreciente y la segunda creciente, definidas como

$$\sup A_{k\ge n} = \bigcup_{k=n}^{\infty} A_k \qquad\qquad \inf A_{k\ge n} = \bigcap_{k=n}^{\infty} A_k$$

(ver definición 9). El evento $\sup\{A_{k\geq n}\}$ ocurre cuando ocurre al menos uno de los $\{A_k, k\geq n\}$. El evento $\inf\{A_{k\geq n}\}$ ocurre cuando ocurren todos los $\{A_k, k\geq n\}$. El límite del supremo de la secuencia de eventos es el evento en el que ocurre un número infinito de eventos de la secuencia. Igualmente, el límite del ínfimo es el evento en que ocurren todos los eventos de la secuencia con excepción de un número finito de ellos:

$$\limsup_{n\to\infty} A_n = \bigcap_{m=1}^{\infty}\bigcup_{k=m}^{\infty} A_k \qquad \liminf_{n\to\infty} A_n = \bigcup_{m=1}^{\infty}\bigcap_{k=m}^{\infty} A_k$$

En efecto, $\omega\in\limsup_{n\to\infty}A_n$ quiere decir que $\omega\in\cup_{k=1..\infty}A_k$, y $\omega\in\cup_{k=2..\infty}A_k$, y $\omega\in\cup_{k=3..\infty}A_k$ y, en general, $\omega\in\cup_{k=m..\infty}A_k$ para todo $m\geq 1$. Esto implica que debe existir algún $i\geq 1$ tal que $\omega\in A_i$. Pero tambien es cierto que $\omega\in\cup_{k=(i+1)..\infty}A_k$, por lo que debe existir algún $j>i$ tal que $\omega\in A_j$. Pero, como $\omega\in\cup_{k=(j+1)..\infty}A_k$, debe existir algún $l>j$ tal que $\omega\in A_l$. Continuando de esta manera es fácil ver que existe una secuencia infinita $i<j<l<\ldots$ de eventos en $\{A_n, n=1,2,3,\ldots\}$ a los que ω pertenece, es decir, cuando ocurre $\limsup_{n\to\infty}A_n$ ocurre un número infinito de eventos en $\{A_n, n=1,2,3,\ldots\}$. Por eso al evento $\limsup_{n\to\infty}A_n$ también se le conoce como $\{A_n$, i.o.$\}$, donde i.o. representa la expresión en inglés *infinitely often*.

De otro lado, $\omega\in\liminf_{n\to\infty}A_n$ quiere decir que $\omega\in\cap_{k=1..\infty}A_k$, o que $\omega\in\cap_{k=2..\infty}A_k$, o que $\omega\in\cap_{k=3..\infty}A_k$ o, en general, que $\omega\in\cap_{k=m..\infty}A_k$ para algún m, de donde $\omega\in A_i$ para todo $i\geq m$. En consecuencia, ω pertenece a todos los A_n con la posible excepción de un número finito de ellos. Por eso al evento $\liminf_{n\to\infty}A_n$ también se le conoce como $\{A_n$, e.a.$\}$, donde e.a. representa la expresión en inglés *eventually always*.

Sea, por ejemplo, el espacio de probabilidad $([0,1], \mathcal{B}([0,1]), P(A)=\text{longitud}(A))$. Definamos en él las siguientes secuencias de eventos: $\{A_n=[1/n,1],\ n=1,2,\ldots\}$, $\{B_n=[(n-1)/4n,(3n+1)/4n],\ n=1,2,\ldots\}$, $\{C_n=[(n-1)/4n,(3n-1)/4n],\ n=1,2,\ldots\}$. Claramente, $P[A_n]=1-1/n\to 1$, $P[B_n]=(n+1)/2n\to 1/2$, $P[C_n]=1/2$ para todo n. En efecto, $\{A_n\}$ es una secuencia creciente con $\lim_{n\to\infty}A_n=\cup_{n=1..\infty}A_n=(0,1]$, de manera que $P[\lim_{n\to\infty}A_n]=\lim_{n\to\infty}P[A_n]$. $\{B_n\}$, en cambio, es una secuencia decreciente con $\lim_{n\to\infty}B_n=\cap_{n=1..\infty}B_n=[1/4,3/4]$, de manera que $P[\lim_{n\to\infty}B_n]=\lim_{n\to\infty}P[B_n]$. Hasta aquí sólo hemos verificado la continuidad de la probabilidad. Sin embargo, $\{C_n\}$ no es una secuencia creciente ni decreciente. Podemos construir la secuencia $S_n=\cup_{m=n..\infty}C_m=[(n-1)/4n,3/4]$ cuya probabilidad es $P[S_n]=(2n+1)/4n\to 1/2$. En efecto, $\text{limSup}_{n\to\infty}C_n=\lim_{n=1..\infty}S_n=[1/4,3/4]$, cuya probabilidad es ½. Tanbién podemos construir la secuencia $I_n=\cap_{m=n..\infty}C_m=[1/4,(3n-1)/4n]$ cuya probabilidad es $P[I_n]=(2n-1)/4n\to 1/2$. En efecto, $\liminf_{n\to\infty}C_n=\lim_{n=1..\infty}I_n=[1/4,3/4]$, cuya probabilidad es ½. En este caso, $\lim_{n\to\infty}C_n=\text{limInf}_{n\to\infty}C_n=\text{limSup}_{n\to\infty}C_n=[1/4,\ 3/4]$, y se cumple nuevamente que $P[\lim_{n\to\infty}C_n]=\lim_{n\to\infty}P[C_n]$.

Las secuencias de eventos $\{A_1, A_2, A_3,\ldots\}$ y los eventos correspondientes limsup $A_n=\{A_n$, i.o.$\}$ y liminf $A_n=\{A_n$, e.a.$\}$ son de gran importancia en redes de telecomunicaciones. Por ejemplo, si A_n es el evento en el que el buffer de un nodo de comunicaciones se encuentra vacío en el instante n, el evento $\{A_n$, i.o.$\}$ indica que el buffer volverá a estar desocupado una y otra vez, indicando que no hay

congestión. El evento $\{A_n^C$, e.a.$\}$, en cambio, indica que desde algún momento el buffer dejará de estar desocupado, lo que implica que la congestión será inevitable.

25. Lemas de Borel-Cantelli

Sea $(\Omega, \mathcal{F}, \mathbf{P})$ *un espacio de probabilidad en el que hay una secuencia de eventos* A_1, A_2, A_3,... *Sea* $\{An$, i.o.$\}$ *el evento en el que un número infinito de los* A_n *ocurre,*

$$\{A_n, \text{i.o.}\} = \limsup_{k\to\infty} A_k = \bigcap_{n=1}^{\infty}\bigcup_{m=n}^{\infty} A_m$$

El primer lema de Borel-Cantelli establece que

$$P(\{A_n, \text{i.o.}\})=0 \text{ si } \Sigma_n P(A_n)<\infty$$

El segundo lema de Borel-Cantelli establece que

$$P(\{A_n, \text{i.o.}\})=1 \text{ si } \Sigma_n P(A_n)=\infty \text{ y los } A_n \text{ son eventos independientes}$$

(*En el primer lema de Borel-Cantelli, los* A_n *no necesariamente deben ser independientes*)

Para demostrar el primer lema de Borel/Cantelli debemos recordar primero que si $s_n = \Sigma_{k=1..n}\, x_k$ converge a $s<\infty$ cuando n tiende a infinito, entonces x_n converge a cero cuando n tiende a infinito. En efecto, $x_n = s_n - s_{n-1}$ de manera que $\lim_{n\to\infty} x_n = \lim_{n\to\infty} s_n - \lim_{n\to\infty} s_{n-1} = s - s = 0$. Siendo así, la suposición $\Sigma_n P(A_n)<\infty$ implica que $\lim_{n\to\infty} P(A_n) = 0$. Ahora, como $A=\limsup_{n\to\infty} A_n \subseteq \cup_{m=n..\infty} A_m$ para todo n, entonces, para cualquier n,

$$P(A) \le P\left(\bigcup_{m=n}^{\infty} A_m\right) \le \sum_{m=n}^{\infty} P(A_m) \underset{n\to\infty}{\to} 0 \quad \text{cuando} \quad \sum_{n=1}^{\infty} P(A_n) < \infty$$

Para demostrar el segundo lema consideremos el complemento de A,

$$A^C = \bigcup_{n=1}^{\infty}\bigcap_{m=n}^{\infty} A_m^C$$

Por la continuidad de la medida de probabilidad (definición 23),

$$P\left[\lim_{N\to\infty}\bigcap_{m=n}^{N} A_m^C\right] = \lim_{N\to\infty} P\left[\bigcap_{m=n}^{N} A_m^C\right] = P\left[\bigcap_{m=n}^{\infty} A_m^C\right]$$

y, como suponemos independencia,

$$P\left[\lim_{N\to\infty}\bigcap_{m=n}^{N} A_m^C\right] = \prod_{m=n}^{\infty}\left(1-P(A_m)\right)$$

pero es bien sabido que para cualquier $x\ge 0$, $1 - x \le e^{-x}$, por lo que

$$P\left[\lim_{N\to\infty}\bigcap_{m=n}^{N} A_m^C\right] \le \prod_{m=n}^{\infty}\exp\left(-P(A_m)\right) = \exp\left(-\sum_{m=n}^{\infty} P(A_m)\right) = 0$$

donde la igualdad a cero se debe a que estamos suponiendo que $\Sigma_{n=1..\infty} P(A_n)=\infty$. Entonces

$$P\left[A^C\right] = \lim_{n\to\infty} P\left[\bigcap_{m=n}^{\infty} A_m^C\right] = 0 \quad \Rightarrow \quad P[A] = 1$$

Estos lemas son muy poderosos, pues la ocurrencia de $\{A_n$, i.o$\}$ se puede establecer con probabilidad 1 o con probabilidad 0 con sólo ver la convergencia de una secuencia de números. El ejemplo básico es una secuencia infinita de lanzadas de una una moneda bien balanceada: Obtendremos sello un

número infinito de veces con probabilidad 1 porque las lanzadas son independientes y $\Sigma_n P(\textit{Sello en } n) = \Sigma_n ½ = \infty$. Aún si la moneda se va deteriorando de manera que la probabilidad de sello en la n-ésima lanzada es $1/(n+1)$, obtendremos sello un número infinito de veces con probabilidad 1 por la divergencia de la serie armónica, aunque cada vez los sellos estarán más espaciados. Sin embargo, si la moneda se deteriora tanto con cada lanzada que la probabilidad de sello en la n-ésima lanzada es $1/(n+1)^2$, finalmente dejaremos de ver sellos porque $\Sigma_n 1/(n+1)^2 = \pi^2/6 - 1 < \infty$.

En próximos capítulos, los lemas de Borel-Cantelli serán muy útiles al determinar la convergencia de secuencias de variables aleatorias, al determinar la recurrencia de los estados de una cadena de Markov, etc. En machine learning los lemmas de Borel-Cantelli son muy útiles para determinar si, eventualmente, los algoritmos de aprendizaje convergirán a un error tolerable. En redes de comunicaciones es muy importante saber que algunos eventos se repetirán indefinidamente (se produce una transmisión sin errores), o que se producirá eventualmente un evento (se transmite correctamente un paquete específico), etc.

26. Modelo Probabilístico

> *Cuando representamos el comportamiento de un sistema físico mediante un experimento aleatorio, al espacio de probabilidad correspondiente se le denomina Modelo Probabilístico.*

Como ingenieros de redes de telecomunicaciones, diariamente nos enfrentamos a problemas tecnológicos particulares caracterizados por nuestra incertidumbre sobre los resultados de las mediciones que no nos es posible observar, ya sea porque son mediciones en el futuro ("Diseñe una red en el que el máximo retardo de un paquete sea menor a 50 ms el 99% de las veces") o porque no tenemos acceso directo a ellas ("Determine el ancho de banda disponible a lo largo de una ruta para el flujo entre dos usuarios dados"). Esta situación se presenta de manera mucho más explícita cuando nuestra actividad profesional está asociada con la investigación y el desarrollo, como se espera que suceda con los estudiantes de postgrado que estudian este libro. En estos casos, se hará necesario especificar el problema en términos de un experimento aleatorio al que se le debe asociar un espacio de probabilidad $(\Omega, \mathcal{F}, \mathbf{P})$ adecuadamente seleccionado. Como este espacio de probabilidad, aunque descrito con toda la formalidad de la teoría de probabilidades, ya no se refiere a un concepto completamente abstracto e inmaterial del mundo de las matemáticas teóricas sino a un modelo matemático de un sistema real, cuyos resultados deben ser aplicables nuevamente a la realidad de donde provino, se le denomina Modelo Probabilístico.

El objetivo del modelado de sistemas, en general, es representar una realidad altamente compleja mediante un modelo matemático que sea lo más sencillo posible pero que capture los aspectos más relevantes que afectan el desempeño del sistema en estudio, de manera que la complejidad no imposibilite el análisis del modelo y que dicho análisis nos permita comprender mejor el comportamiento del sistema real. En el caso del modelado probabilístico, se trata de reformular la pregunta que nos queremos responder sobre el sistema real en términos de algunas características de

un experimento aleatorio. Una vez se ha hecho explícito el experimento, debemos especificar un espacio de probabilidad correspondiente, que se convertirá en un Modelo Probabilístico del Sistema Real.

Ya discutimos en qué consiste el proceso de determinar el espacio muestral y el conjunto de eventos de interés con los que debemos construir el campo-σ de eventos medibles. Sin embargo especificar la medida de probabilidad es un proceso algo más elaborado. El espacio de probabilidad exige que determinemos $\mathbf{P}(A)$ para todo $A \in \mathcal{F}$. Lo que podemos hacer en un primer paso es determinar la probabilidad de algunos eventos en $\mathcal{F}$ mediante algún proceso inductivo que tenga en cuenta la naturaleza precisa del "proceso de observación" de la definición 4. En este primer paso podemos usar mediciones de frecuencias relativas (en cuyo caso debemos ser muy juiciosos con el diseño estadístico de los experimentos) o, en el peor de los casos, elucubrar mediante argumentos razonables sobre independencias, simetrías, uniformidades u otras propiedades que podamos derivar del conocimiento inicial que tengamos de la naturaleza del experimento aleatorio. A partir de las probabilidades de estos eventos iniciales, en un segundo paso podemos calcular las probabilidades de otros eventos mediante el uso juicioso de las herramientas con que dotamos al lector en este libro, tales como los axiomas de la definición 14, los resultados de la definición 17, el teorema de la probabilidad total, la regla de bayes, la composición de experimentos, etc. Mientras el primer paso es un proceso inductivo asociado con la naturaleza del experimento aleatorio, el segundo paso es un proceso deductivo asociado con la rigurosidad matemática de la teoría de probabilidades. En un tercer paso, debemos inferir las implicaciones que los resultados deducidos en el paso dos tengan sobre la realidad compleja que estamos modelando. Este es otro proceso inductivo que le da validez al modelo probabilístico si las implicaciones son correctas y útiles (o lo invalida si las implicaciones son incorrectas o inútiles).

Desafortunadamente, este libro no es sobre modelado probabilístico de redes de telecomunicaciones (cómo seleccionar un modelo para una realidad tecnológica compleja dada) sino sobre cómo analizar un modelo probabilístico dado. De hecho, obsérvese que a partir de la definición 16 (espacio de probabilidad) todas las definiciones subsecuentes (17 a 25) siempre han empezado con la frase "Sea $(\Omega, \mathcal{F}, \mathbf{P})$ un espacio de probabilidad en el que …". Para nosotros, en este libro, el modelo siempre va a estar dado y simplemente nos dedicaremos a obtener herramientas para el proceso deductivo del segundo paso en el modelado probabilístico. Más aún, para ahorrarnos tener que reescribir esa frase en todas las definiciones que siguen, en el siguiente capítulo definiremos la variable aleatoria con el propósito de dejar explícito un modelo probabilístico particular $(\mathbb{R}, \mathcal{B}(\mathbb{R}), F(\cdot))$ que será el que usemos casi siempre de ahí en adelante. En este libro apenas vamos a llenar la caja de herramientas del lector con un buen conjunto de aparatos útiles, entre los cuales hemos puesto ya los tres axiomas, cinco de las innumerables conclusiones que se pueden sacar de ellos, el teorema de la probabilidad total y la regla de Bayes: diríamos que ya pusimos en la caja el serrucho y el martillo. En los siguientes capítulos colocaremos en ella algunas herramientas más sofisticadas. Sin embargo, sólo la experiencia le permitirá al lector aprender a "construir muebles" con las herramientas de su caja a partir de "la madera" proporcionada por las redes de comunicaciones. Lo más que podemos hacer (y esa es la razón de este libro) es escoger ejemplos del mundo de las redes de comunicaciones, pues pienso que el aprendizaje de las técnicas de modelado sólo es posible mediante el estudio de ejemplos y la sorprendente capacidad de generalización que exhibe la inteligencia humana. Claro, como nuestros

ejemplos son puramente pedagógicos, invitamos al lector a que se afilie a IEEE, ACM y tantas otras sociedades técnicas que publican excelentes revistas en las que cada artículo es un ejemplo realista de un proceso de modelado probabilístico llevado a cabo por uno o varios expertos.

27. ¿Por qué es necesario definir un espacio de probabilidad $(\Omega,\mathcal{F},P)$?

No es posible definir una medida de probabilidad sobre todos los subconjuntos del espacio muestral [0,1] *de manera que satisfaga los tres axiomas.*

Supongamos un experimento aleatorio en el que el espacio muestral Ω es el intervalo unitario [0,1], de manera que la probabilidad de que el resultado del experimento caiga en cualquier intervalo de Ω es igual a la longitud del intervalo: P[[0,1/2)] = 1/2, P[(1/4,1/2]] = 1/4, P[(1/8,1/3)∪[2/3, 8/9]] = 31/72, P[{1/2}]=0, P[{1/2}∪{3/4}]=0, P[{1/2}∪(2/3, 8/9)]=2/9, etc.

Si A_1, A_2, A_3, … son subconjuntos disyuntos de [0,1], sabemos que $P[A_1\cup A_2\cup A_3\cup\ldots] = P[A_1] + P[A_2]+P[A_3]+\ldots$ Pero es imposible extender esta propiedad a una suma incontable de eventos disyuntos pues, si la extendiéramos, deberíamos esperar que:

$$P([0,1]) = \sum_{x\in[0,1]} P[\{x\}]$$

lo cual es evidentemente falso: mientras la expresión de la izquierda es igual a uno, la expresión de la derecha es igual cero[14]. Esta es la razón por la cual debemos restringirnos a considerar únicamente operaciones contables y es por eso que necesitamos definir la probabilidad sobre un campo sigma, esto es, cerrado para los complementos y las uniones ***contables***.

Consideremos la misma medida de probabilidad en el mismo espacio muestral (el intervalo unitario [0,1], donde la probabilidad de cualquier intervalo es igual a la longitud del intervalo). El resultado fundamental de esta definición 27 (que no es posible definir una medida de probabilidad para todos los subconjuntos del espacio muestral [0,1] que satisfaga los tres axiomas) se puede demostrar recurriendo a los conjuntos de Vitali que, como mencionamos, no son conjuntos de Borel. Supongamos que se puede definir $P[A]$ para cualquier $A\subset[0,1]$ y derivemos una contradicción de dicha suposición. Para esto definamos una relación de equivalencia[15] en [0,1] de la siguiente manera: $x\sim y$ si y sólo si la diferencia $y–x$ es racional. Esta relación de equivalencia genera una partición de Ω. Sea C un conjunto formado con uno y sólo un elemento de cada clase de equivalencia[16] que no

[14] La suma sobre un conjunto incontable $|A| = \aleph_1$, $x = \Sigma_{\alpha\in A}x_\alpha$, se define como el supremo de $\Sigma_{i\in I}x_i$, entre todos los subconjuntos contables I contenidos en A. Así es como sabemos que la suma no contable de ceros es igual a cero.

[15] Una relación en un conjunto Ω es una función booleana en $\Omega\times\Omega$ (dados $x,y\in\Omega$ sólo puede ocurrir que x esté relacionada con y ($x\sim y$) o no). Una relación es de equivalencia si (1) es reflexiva ($x\sim x\ \forall x\in\Omega$), (2) es simétrica ($x\sim y \Leftrightarrow y\sim x$) y (3) es transitiva ($x\sim z$ si $x\sim y$ y $y\sim z$). Dada una relación de equivalencia $\sim$ y un elemento $x\in\Omega$, la clase equivalente de x es el conjunto de todos los y tales que $x\sim y$. Claramente, dos clases equivalentes sólo pueden ser idénticas o disyuntas. Dada una relación de equivalencia $\sim$, el conjunto de clases de equivalencia forman una partición de Ω.

[16] La posibilidad de escoger uno y sólo un elemento de cada subconjunto se conoce como el "axioma de la selección". Parece obvio si tenemos un número finito de conjuntos finitos (tome exactamente un músico de cada una de las bandas de rock en Bogotá). Pero cuando tenemos infinitos conjuntos, cada uno de ellos de cardinalidad infinita, el axioma de selección no es tan obvio y debe proponerse, precisamente, como un axioma.

contenga 0 (si escogimos cero, lo podemos cambiar, por ejemplo, por 1/2, que es equivalente a cero). Sea $C\oplus r$ el desplazamiento circular de C por r :

$$C \oplus r = \{c + r : c \in C, c + r < 1\} \cup \{c + r - 1 : c \in C, c + r > 1\}$$

Claramente, P[$A\oplus r$] = P(A) $0 \le r \le 1$. Además, cada punto en (0,1] pertenece a la unión $\cup_{r\in[0,1)\cap\mathbb{Q}}$ ($C\oplus r$) de desplazamientos circulares de C. Más aún, como C contiene exactamente un elemento de cada clase equivalente, todos los conjuntos $C\oplus r$ son disyuntos para $r\in[0,1)\cap\mathbb{Q}$. De esta manera, el tercer axioma exige que

$$P[(0,1]] = \sum_{r\in[0,1)\cap\mathbb{Q}} P[C \oplus r]$$

Pero P[$C\oplus r$] = P[C], de donde

$$1 = P[(0,1]] = \sum_{r\in[0,1)\cap\mathbb{Q}} P[C]$$

Esta es la contradicción que estábamos buscando: Al sumar un número contable infinito de veces el mismo número no negativo sólo podemos obtener 0 si P[C]=0 u obtener ∞ si P[C]>0, pero nunca podremos obtener 1.En conclusión, si las probabilidades han de satisfacer propiedades razonables, no podemos definirlas para todos los posibles subconjuntos de [0,1] sino que debemos restringirnos a algunos conjuntos medibles.

28. Sobre los Conceptos de Aleatoriedad y Probabilidad

> *El concepto de aleatoriedad presenta muchas dificultades intuitivas, que aún son materia de controversia entre filósofos y que pueden confundir a los estudiantes neófitos. En el análisis de modelos probabilísticos debemos usar con precaución la intuición, sólo como una guía que siempre debe ser corroborada por el formalismo axiomático de Kolmogorov, pues en muchas ocasiones la intuición puede fallar drásticamente. De todas maneras, dada la naturaleza de la mayoría de experimentos que se refieren a redes de comunicaciones, en los que casi siempre están involucrados o un gran número de usuarios, o un gran número de paquetes, o un gran número de bits, etc., la intuición basada en la interpretación de la probabilidad como frecuencia relativa suele sugerir caminos acertados en el proceso hacia el objetivo del modelado probabilístico en redes de comunicaciones. Después de todo, como dijo Laplace, "En el fondo, la teoría de las probabilidades no es más que sentido común reducido al cálculo".*

Todos los seres humanos se encuentran con el azar en cada momento de sus vidas: en el noticiero recibimos estadísticas económicas, encuestas políticas y predicciones del clima; desde niños jugamos con dados, cartas y monedas; hacemos filas en los bancos y supermercados; padecemos trancones en las avenidas; al decidir dónde invertir nuestros ahorros debemos evaluar el riesgo y ponderarlo respecto a los posibles rendimientos de cada una de las opciones; como usuarios de las redes modernas de comunicaciones sufrimos tiempos excesivos de respuesta, falta de disponibilidad en los recursos de la red, ruidos y recortes en las señales que recibimos, etc. Sin embargo, a pesar de nuestra vasta experiencia con el azar, el concepto de aleatoriedad sigue estando muy alejado del común de la gente

e, inclusive, de muchas personas muy bien preparadas en distintas profesiones (incluyendo las ciencias y la ingeniería: Me consta!). No es sólo por la necesidad de entender conceptos tan abstractos como los asociados con la teoría de medidas (definiciones 10, 11, 12, 13, 23, 25 y 27) sino por las múltiples alternativas de interpretación de la medida de probabilidad.

Y es que el concepto de aleatoriedad se va alcanzando gradual y muy lentamente. Por ejemplo, cuando mi hija de cinco años y mi hijo de cuatro años deseaban resolver alguna disputa de manera "justa", recurrían a un juego infantil muy tradicional en Colombia: Uno de ellos canta "Pi NU no, pin DOS, pin TRES, pin CUA tro, pin CIN co, pin SEIS, pin SIE te, PIno, CHIto, SErás, TÚ" mientras que con el dedo índice se señala a sí mismo o al hermanito, alternando la dirección con cada sílaba en mayúsculas. María Alejandra iniciaba señalándose ella misma, mientras que Juan Diego parecía escoger *al azar* si iniciaba señalándose él o señalando a la hermanita. Aunque María Alejandra fue una fervorosa defensora del método de conciliación, Juan Diego empezó a perder la confianza en él porque, extrañamente, sólo ganaba la mitad de las veces en que él mismo contaba.

La humanidad misma parece haber seguido ese mismo proceso gradual y lento que sigue cada ser humano individualmente, pues sólo hasta el siglo XVI se empezó a formalizar un concepto que, hasta entonces, era sólo el mecanismo de expresión de las voluntades divinas. ¡Y qué útil resultaba ser el vocero de los dioses cuando se usaban dados no balanceados! Lo cierto es que los seres humanos (y la humanidad entera en su conjunto) primero aprendemos mediante la intuición y, después, sobre esa base, empezamos a formalizar conceptos. Cuando la intuición es correcta, ese proceso es formidable porque ayuda profundamente en la comprensión de temas difíciles. Yo mismo, como profesor de Procesamiento Digital de Señales, Sistemas de Comunicación, Redes de Comunicaciones y Control de Sistemas Dinámicos, me preocupo por presentar los conceptos de manera que los estudiantes primero los capturen de manera intuitiva antes de aprenderlos desde las formalidades matemáticas o tecnológicas. Pero jamás intento hacer eso como profesor de Probabilidades, Variables Aleatorias y Procesos estocásticos porque, en los problemas asociados a estos temas, la intuición suele fallar miserablemente! Por eso he dejado esta discusión para el final de este capítulo, cuando ya hemos visto la formulación axiomática de Kolmogorov, pues en este caso es mucho mejor presentar la formalidad antes que la intuición. Ya en el siglo XVIII DeMoivre lo mencionó: Los problemas que tienen que ver con el azar suelen parecer fácilmente solucionables mediante el sentido común, cuando en realidad casi nunca es así.

Veamos algunos ejemplos:

(1) Un presentador de un concurso de televisión le ofrece que escoja una de tres puertas sabiendo que sólo una de ellas conduce a un gran premio mientras que las otras dos sólo llevan a pequeños premios de consolación. Llamemos *a* a la puerta que usted escoge. Una vez usted escogió su puerta, el presentador le revela una de las otras dos puertas que conducía a un premio de consolación. Llamemos *b* a la puerta revelada por el presentador y *c* a la otra puerta. Ahora el presentador le pregunta: "¿Desea quedarse con *a* o prefiere cambiarse a *c*?". ¿Cuál sería la mejor estrategia en este juego? He aquí el razonamiento que hace la gran mayoría de personas:

> Al principio cada puerta tenía una probabilidad 1/3 de conducir al premio mayor, independientemente de la que yo escogiera. Sin embargo, una vez el presentador me revela que *b* no tiene el premio, me quedan sólo dos puertas, *a* y *c*. Como sólo una de las dos conduce al premio, la probabilidad de que cada una de ellas conduzca al premio es ½. El hecho de que yo haya escogido antes la puerta *a* no cambia el nuevo hecho de que ahora tengo dos puertas, una de las cuales conduce al premio grande y otra al premio de consolación. Luego da igual si me quedo con *a* o si me cambio a *c*, pues en ambos casos ganaré o perderé con probabilidad ½.

Deténgase usted, señor lector, a pensar un poco en el análisis anterior antes de seguir leyendo. Ahora sí, he aquí el análisis (correcto) que hacen muy pocas personas y que, seguramente, fue el que usted hizo:

> Sean los eventos A = {*a* tiene el premio}, B = {*b* tiene el premio} y C = {*c* tiene el premio}. En un principio, $\mathbf{P}(A) = \mathbf{P}(B) = \mathbf{P}(C) = 1/3$. Si el presentador me hubiera informado de la ocurrencia del evento B^C antes de que yo escogiera la puerta *a*, tendría el caso del análisis anterior, $\mathbf{P}(A|B^C_{antes}) = \mathbf{P}(C|B^C_{antes}) = ½$. Sin embargo el protocolo del juego me permitió escoger primero, cuando todavía $\mathbf{P}(A)$ era un tercio. Un hecho cierto en ese momento era que al menos una de las otras dos puertas debía conducir a un premio de consolación, de manera que cuando el presentador me revela que era *b* la que conducía al premio de consolación no me dio ninguna información nueva: La probabilidad de A cuando yo la escogí era de 1/3, independientemente de que se me revelara B^C después de mi escogencia, $\mathbf{P}(A|B^C_{después}) = \mathbf{P}(A) = 1/3$. Una vez informado de la ocurrencia del evento B^C, la única alternativa al evento A es el evento C, de manera que $\mathbf{P}(C|B^C_{después}) = 1 - \mathbf{P}(A|B^C_{después}) = 1 - \mathbf{P}(A) = 2/3$. Es mejor estrategia cambiarse a la puerta *c*.

Nótese que la solución correcta no concuerda con la solución intuitiva porque hay un asunto de protocolo que pasa inadvertido. Si el presentador le hubiera revelado una puerta mala antes de que usted hubiera escogido la puerta *a*, él tendría dos posibles puertas para escoger. Cuando él lo deja escoger primero, usted le reduce sus posibilidades de dos a una, con probabilidad 1/3. Un gran amigo mío lo pone de la siguiente manera para ver la bondad intuitiva de la solución correcta: Suponga que no son tres sino mil puertas, de las cuales una sola conduce al premio. Usted escoge una, el presentador le revela 998 que no tenían premio… ¿Se quedaría con la que escogió primero? Es "casi seguro" que el premio está en la otra puerta!

(2) Supongamos que cada nuevo bebé que llega al mundo es niño o niña con probabilidad ½, independientemente del resto de la humanidad y, en particular, independientemente de sus hermanitos y hermanitas. Bajo esta suposición consideremos los siguientes problemas:
Se encuentra con una amiga a quien no veía desde hacía diez años y sostienen la siguiente conversación:
Usted: "¡Hola! ¿tienes hijos?"
Amiga: "Sí. Tengo dos"
Usted: "¿alguna niña?"
Amiga: "Sí"
Usted: "¡Adiós!"

¿Cuál es la probabilidad de que su amiga tenga dos niñas?

Al otro día se encuentra otra vez con su amiga y ve que lleva a una niña de la mano:
Usted: "¡Hola! ¿Esta niña tan preciosa es hija tuya?"
Amiga: "Sí"
Usted: "¡Adiós!"
¿Cuál es la probabilidad de que su amiga tenga dos niñas?

A diferencia del ejemplo anterior, en este caso distintas personas hacen diferentes razonamientos. Sin embargo casi todo el mundo da por hecho que la pregunta es la misma y, por lo tanto, exige la misma respuesta. Después de todo, dicen, desde el primer día yo ya sabía que mi amiga tenía por lo menos una niña, así que en el segundo día no obtengo ninguna información nueva al ver exactamente eso: una niña hija de mi amiga, que yo ya sabía que existía! Como verla o no verla no hace ninguna diferencia, la pregunta (y la respuesta) no cambia de un día para otro.

Deténgase usted, seños lector, a pensar un poco en el análisis anterior antes de seguir leyendo. Ahora sí, he aquí el análisis (correcto) que hacen muy pocas personas y que, seguramente, fue el que usted hizo:

Cada hijo de mi amiga puede ser una niña (f) o un niño (m). El primer día supe que mi amiga tenía dos hijos, de manera que el espacio muestral del experimento consistente en observar el género de cada uno de los hijos de mi amiga es $\Omega = \{(f,f),(f,m),(m,f),(m,m)\}$, donde cada evento unitario tiene probabilidad ¼. Pero también supe que el evento $\{(m,m)\}$ no ocurrió. Luego la pregunta que me hago el primer día es $\mathbf{P}(\{(f,f)\} \mid \{(f,f),(f,m),(m,f)\}) = \mathbf{P}(\{(f,f)\}) / (1 - \mathbf{P}(\{(m,m)\}) = 1/3$. Sin embargo el segundo día me estoy preguntando por la probabilidad de que el otro hijo (el que no estoy viendo) sea una niña, lo cual ocurre con probabilidad 1/2. ¿La probabilidad de que mi amiga tenga dos niñas aumentó de 1/3 a 1/2 solamente porque pude observar a una niña de mi amiga, siendo que yo ya sabía que tenía por lo menos una niña? No. Es solamente que la pregunta que me estoy haciendo es distinta, ¡a pesar de que la formulación parece idéntica!

Los dos ejemplos anteriores son paradojas que confunden inclusive a algunos expertos en probabilidades (tal vez usted mismo haya releído cada caso varias veces y haya tomado papel y lápiz para ver con detalle qué es lo que está pasando). Hay otros muchos ejemplos que, aunque no confundan a los expertos, si resultan paradójicos para el común de la gente, como el siguiente:

(3) Ante un juez presentan a una persona muy honorable, x, acusada de haber cometido un crimen. La reputación de x la precede, de manera que el evento $X = \{x$ cometió el crimen$\}$ tiene una probabilidad muy bajita: $\mathbf{P}(X) = 0.001$. La fiscalía presenta dos testigos muy confiables, y y z, cada uno de los cuales dice la verdad con probabilidad 0.9. Más aún, estos testigos son independientes pues no se conocen entre sí y observaron los hechos desde sitios distintos. Durante el juicio se presentan los eventos $Y = \{y$ dice que x cometió el crimen$\}$ y $Z = \{z$ dice que x cometió el crimen$\}$. Según la más antigua tradición legislativa y judicial (Moisés, Hamurabi, Nemqueteba) y la opinión

de la mayoría de personas, *x* es condenado porque el testimonio de dos testigos independientes muy confiables es evidencia suficiente. ¿Se hizo justicia?

Veamos cómo se afecta la probabilidad de *X* cuando la condicionamos a los eventos *Y* y *Z*:

$$P(X \mid Y \cap Z) = \frac{P(X)P(Y \cap Z \mid X)}{P(X)P(Y \cap Z \mid X) + P(X^C)P(Y \cap Z \mid X^C)} = \frac{P(X)P(Y \mid X)P(Z \mid X)}{P(X)P(Y \mid X)P(Z \mid X) + P(X^C)P(Y \mid X^C)P(Z \mid X^C)}$$

$$P(X \mid Y \cap Z) = \frac{(0.001)(0.9)(0.9)}{(0.001)(0.9)(0.9) + (0.999)(0.1)(0.1)} = \frac{3}{40}$$

Ciertamente los testigos aumentan la probabilidad de *X* 75 veces (desde 0.001 hasta 0.075), pero condenar a alguien cuando la probabilidad de que haya cometido el crimen es menos de 0.1 me parece muy injusto!

Lo que pasa, como lo anunció DeMoivre, es que la intuición suele fallar estrepitosamente en asuntos de aleatoriedad. ¿Cuántas veces lo ha abordado un voceador de lotería en la calle anunciándole un dato valiosísimo : "Hace cinco semanas que el número 3 no sale en la lotería y aquí le tengo un billetico terminado en 3". Él espera convencerlo de que compre el billete porque el hecho de que el 3 no haya salido cinco veces seguidas aumenta la probabilidad de que salga la próxima vez. De hecho haga una encuesta entre sus amigos no ingenieros ni científicos y observará que muchos de ellos creen que, si se ha lanzado cinco veces una moneda bien equilibrada y en las cinco ocasiones se ha obtenido cara, aumentan las probabilidades de que en la próxima lanzada se obtenga sello pues, después de todo, el sello debería aparecer en la mitad de las lanzadas[17].

Las paradojas y las concepciones populares y erróneas sólo indican que el concepto de aleatoriedad parece obvio, cuando en realidad no lo es. De hecho, su misma interpretación ha enfrentado a importantes científicos. Ya 500 años antes de Cristo Leucipo había manifestado que nada ocurre por azar sino que todo obedece a la razón y a la necesidad. Así pues, lo que denominamos azar se refiere a los efectos de causas escondidas que están fuera de nuestro conocimiento o fuera de nuestro control, como sostuvo Demócrito, discípulo de Leucipo. Los sofistas reafirmaron este concepto en contra de Epicuro, quien sostenía que si todos los eventos tenían una causa conocible, el hombre carecería de libre albedrío… El cristianismo ayudó a afianzar el concepto sofista, pues el resultado final de todos los experimentos debía obedecer, necesariamente, a la voluntad de Dios: En un universo en el que todo está sometido a la voluntad de Dios (y a su plan de salvación) sólo nuestra ignorancia puede abrirle espacios al azar. Con la aparición de la mecánica Newtoniana pocos años después, se terminó de consolidar la visión determinista: La voluntad de Dios se manifestaba en leyes del movimiento

17 Existe un muy reconocido locutor y comentarista deportivo en Colombia que debe ser muy bien formado porque le dicen "doctor". A él le escuché decir en una ocasión lo siguiente: "La historia muestra que el equipo *A* le ha ganado al equipo *B* en el 70% de los encuentros que han disputado. Sin embargo, en los últimos tres meses se han enfrentado cinco veces y en todas ellas ha ganado *B*. Luego, si la teoría de las probabilidades no falla, *A* debería ganar en el próximo partido". Claro, *A* perdió el partido de esa tarde contra *B* y no por una falla de la teoría de las probabilidades sino porque, al menos desde hacía tres meses, *A* parecía ser un equipo de "troncos". Si el locutor doctor quiso ser optimista, debió suponer que cada partido es independiente de los demás, en cuyo caso la probabilidad de que *A* ganara esa tarde sería 0.7. Sin embargo, parece mucho más correcto pensar que la probabilidad de que *A* le ganara esa tarde a *B* dado que *A* llevaba cinco partidos seguidos perdiendo contra *B* era un número muy cercano a cero.

que eran asequibles para el hombre a través de las matemáticas. Según Laplace, sólo necesitamos conocer la masa, la posición y la velocidad de cada partícula del Universo en un instante dado para predecir con precisión su destino último y su pasado más remoto. Siendo el comportamiento del universo tan determinístico y predecible, ¿cómo pudo, entonces, desarrollarse tan profunda y aceleradamente la teoría de la probabilidades durante este período? Porque si bien el mundo estaba sometido al determinismo de la voluntad inmutable de Dios, nuestra capacidad de observación era limitada: El hombre comete errores y la probabilidad nos permite cuantificar el error. Ya Galileo y Tycho Brae formularon proposiciones fascinantes sobre el error en las mediciones astronómicas: es inevitable, es simétrico y entre más pequeño sea más probable es. Estas ideas sugieren tomar muchas mediciones y promediarlas, de donde surgen las leyes de los grandes números. Thomas Simpson fue el primero en introducir la teoría de los juegos de azar, cuando en 1756 dijo que si cada fuente de error se comportaba como un dado, el error total se debe comportar como la suma de muchos dados. En 1808 Gauss usó la famosa campana $\exp(-x^2/2)/\sqrt{2}$, justificado por el teorema del límite central, propuesto por Laplace 2 años después. Tendremos oportunidad en este libro de divertirnos un buen rato con los tres conceptos: la campana gaussiana, la ley de los grandes números y el teorema del límite central.

Si bien el determinismo se remonta a los atomistas como Leucipo y Demócrito, es precisamente el estudio de las partículas subatómicas el que reivindica a Epicuro: La naturaleza puede ser inherentemente aleatoria, ¡puede ser que haya una indeterminación básica en el universo! En 1900 Max Planck explicó (exitosamente) por qué los cuerpos calientes no irradiaban en todas las posibles frecuencias, diciendo que la radiación se daba en "cuantos" de energía. Lo que sorprendió a Planck es que esta idea resultó más que un truco matemático cuando muchos físicos empezaron a encontrar más comportamientos cuánticos en las partículas subatómicas. En 1926 Heisenberg fue el primero en advertir que, siendo así, jamás nos sería dado conocer el estado del universo como lo proponía Laplace, pues para observar una pequeña partícula debemos iluminarla con, por lo menos, un cuanto de luz, alterando irremediablemente el estado que queríamos observar. Esto conduce al principio de incertidumbre como una de las leyes básicas de la naturaleza: si conocemos la posición exacta de una partícula no podemos saber nada sobre su velocidad y viceversa. Schrödinger, de hecho, describe las partículas mediante una ecuación de onda con la que evalúa *la probabilidad* de que una partícula se encuentre, en un instante de tiempo particular, en un punto dado del espacio! Albert Einstein, quien contribuyó notablemente al desarrollo de la mecánica cuántica con su estudio de la radiación de cuerpo negro, nunca se sintió cómodo con esta idea pues él era abiertamente determinista. Una conversación entre Einstein y Max Born pudo haber ocurrido hace 2500 años entre Demócrito y Epicuro: Einstein (o Demócrito): "Dios no juega a los dados". Born (o Epicuro): "Y ¿quiénes somos nosotros para decidir a qué puede o no puede jugar Dios?"[18]

¿Dónde está, pues, la aleatoriedad de una secuencia de resultados obtenidos al repetir muchas veces un experimento? Si creemos que la naturaleza sufre de una indeterminación básica, la aleatoriedad de la secuencia está en el experimento mismo que la generó; si creemos que las leyes que rigen el

[18] En una carta a Max Born Einstein escribió "La mecánica cuántica es realmente imponente. Pero una voz interior me dice que aún no es la buena. La teoría dice mucho, pero no nos aproxima realmente al secreto del 'viejo'. Yo, en cualquier caso, estoy convencido de que Él no tira dados".

experimento son conocidas pero muy complejas y difíciles de evaluar, la aleatoriedad de la secuencia está en nuestra incapacidad computacional para calcular el siguiente resultado; y si creemos que las leyes existen pero no las conocemos, la aleatoriedad de la secuencia está en nuestra ignorancia. En cualquiera de los tres casos, la aleatoriedad se caracteriza por nuestra incapacidad para predecir el resultado del siguiente experimento, como propusimos en la primera definición. En los dos últimos casos, sin embargo, la aleatoriedad se vuelve un asunto subjetivo: alguien que esté mejor capacitado que nosotros para predecir el resultado del siguiente experimento encontrará que la secuencia es menos aleatoria de lo que nosotros creemos. Consideremos, por ejemplo, las siguientes secuencias de números:

1	2	3	4	5	6	7	8	9	10	…
2	4	6	8	10	12	14	16	18	20	…
4	7	2	6	4	7	2	6	4	7	…
3	1	4	1	5	9	2	6	5	3	…

En cada una de ellas es fácil predecir el siguiente número, ¿cierto? 11 para la primera secuencia, porque los números parecen ir de uno en uno; 22 para la segunda secuencia porque los números parecen ir de dos en dos; 2 para la tercera secuencia porque los números parecen tener un período de longitud 4; ¿identifica usted fácilmente el siguiente número de la cuarta secuencia? Parece una secuencia aleatoria, hasta que reconocemos en ella la expansión decimal de π: el siguiente número es 5. Se diría que las cuatro secuencias anteriores son completamente determinísticas, aunque un estudiante de primaria que no haya visto trigonometría *podría considerar* que la cuarta secuencia es aleatoria ¿Qué tal la siguiente secuencia?

7	5	7	9	6	3	9	2	7	0	…

Difícil adivinar el siguiente número, ¿cierto? Pero se trata de una secuencia completamente determinística! Iniciando con $Z_0 = 7182$, hago $Z_{i+1} = [floor(Z_i^2/100)]_4$, donde $floor(x)$ es la parte entera de x, y $[x]_4$ es el número compuesto por los cuatro dígitos menos significativos de x (las unidades, decenas, centenas y unidades de mil en x). El i-ésimo número de la secuencia corresponde a las unidades de mil en Z_i. La siguiente tabla muestra cómo se construye la secuencia completa. Nótese que a partir del último cero en la secuencia mostrada, continúa una cadena infinita de ceros. Sin embargo, hasta donde se mostró, parecía una secuencia completamente aleatoria, aunque no lo sería *para quien conozca el algoritmo y el valor de* Z_0. ¿Dónde está, entonces, la aleatoriedad?

i	Z_i	Z_i^2
0	7182	51581124
1	5811	33767721
2	7677	58936329
3	9363	87665769
4	6657	44315649
5	3156	09960336
6	9603	92217609

7	2176	04734976
8	7349	54007801
9	0078	00006084
10	0060	00003600
11	0036	00001296
12	0012	00000144
13	0001	00000001
14	0000	00000000

Nadie duda que las lanzadas consecutivas de un dado generan una secuencia aleatoria porque la única manera de conocer el siguiente número de la secuencia es lanzando el dado una vez más. Pero, si no conociéramos un algoritmo para calcular π con cualquier precisión deseada y nos muestran su expansión decimal a partir del dígito 100, ¿no consideraríamos la secuencia como aleatoria? En 1888, Venn verificó que los primeros 707 dígitos de la expansión decimal de π satisfacen criterios importantes de aleatoriedad: cada dígito aparece en la secuencia aproximadamente el mismo número de veces que los demás dígitos, sin ninguna estructura aparente. 101 años después, Gregory y David Chudnovsky verificaron el mismo comportamiento para más de mil millones de dígitos en la expansión de π. ¿Qué más podría uno esperar de una secuencia aleatoria? Pero sabemos, por supuesto, que la expansión decimal de π es completamente determinística! (ver definición 100.d).

Como la teoría de la probabilidad no se puede construir sobre la subjetividad que da la ignorancia, Kolmogorov mismo, en su esfuerzo por formalizar la teoría, llegó a un concepto muy interesante de aleatoriedad: Si la complejidad de una secuencia está dada por la longitud del programa de computador más pequeño capaz de generarla (en un modelo computacional particular, tal como una máquina de Turing), una secuencia es aleatoria cuando su complejidad es máxima, esto es, cuando el único algoritmo que la puede generar es el algoritmo que la menciona, elemento por elemento. Por supuesto, π resulta poco complejo, pues el algoritmo es sencillo: basta contar cuántos diámetros caben en una circunferencia[19], para lo cual el computador puede evaluar iterativamente una serie de potencias. Pero una secuencia de lanzadas de un dado es aleatoria, porque necesitamos lanzar los dados para poder especificar la secuencia. Esta definición pone el concepto de aleatoriedad en términos formales, muy al estilo de Kolmogorov, pues la aleatoriedad de una secuencia ya no depende de la apreciación del observador sino que es una medida objetivamente cuantificable (más adelante tendremos oportunidad de introducirnos brevemente en la teoría de la complejidad y la teoría de la información que sustentan este concepto).

Pero, igualmente, queda casi sin resolver la pregunta de cómo asignar las probabilidades a los eventos del campo-σ escogido para analizar un modelo probabilístico de una realidad compleja. Fundamentalmente, se trata de evaluar nuestra confianza en que el evento suceda cuando realicemos el experimento, para lo cual podemos valernos de la frecuencia relativa observada en experimentos

[19] Es fácil calcular rápidamente millones de dígitos de π mediante, por ejemplo, el algoritmo de Chudnovsky:

$$\frac{1}{12\pi} = \sum_{n=0}^{\infty} (-1)^n \frac{(6n)!\ (3591409\ +\ 545140134 \cdot n)}{(3n)!(n!)^3 \cdot 640320^{(3 \cdot n + 3/2)}}$$

anteriores o en razonamientos plausibles sobre la naturaleza del experimento, tales como las simetrías proporcionadas por eventos equiprobables (ver la definición 26). En los capítulos siguientes formularemos diferentes modelos para muchos casos típicos de experimentos que surgen una y otra vez en el mundo de las redes de comunicaciones, los cuales podremos usar para generar hipótesis sobre las probabilidades de algunos eventos y, con nuestras herramientas, encontrar las probabilidades de otros eventos o algunas estadísticas de interés.

Por último, vale la pena mencionar la teoría de las posibilidades como una formalidad reciente para tratar con cantidades inciertas, alternativa (y a veces complementaria) a la teoría de las probabilidades. Sobre los eventos pertenecientes a un campo aditivo $\mathcal{F}$ de subconjuntos de un espacio muestral Ω, se define la posibilidad del evento $A \in \mathcal{F}$ como una función $\mathbf{Po} : \mathcal{F} \rightarrow \mathbb{R}$ que satisface los siguientes postulados: $\mathbf{Po}(\Phi)=0$, $\mathbf{Po}(\Omega)=1$, $\mathbf{Po}(A \cup B) = \max(\mathbf{Po}(A), \mathbf{Po}(B))$ si A y B son conjuntos disyuntos en $\mathcal{F}$. Este último postulado se puede extender igualmente a campos-σ infinitamente aditivos, como en la definición14. Esta medida de posibilidad está más asociada con la función de membresía de un elemento en un conjunto difuso, esto es, un conjunto al que se puede pertenecer con cierto grado de pertenencia en el rango [0, 1], en cuyo caso la posibilidad de un evento es el máximo entre las funciones de membresía de sus miembros. Como un ejemplo revelador de la diferencia entre los dos conceptos, considere que lleva muchos días perdido en el desierto y encuentra dos botellas llenas de un líquido de apariencia deliciosa. La etiqueta en una botella dice que la probabilidad de que su contenido sea potable es 0.9, mientras la etiqueta en la otra botella dice que la posibilidad de que sea potable es 0.9. ¿Cuál líquido consumiría usted? Es muy probable que el contenido de la primera botella sea agua pura, aunque, en el peor de los casos, la primera botella podría contener ácido sulfúrico o cianuro. En cambio puede estar seguro que la segunda botella no contiene ni agua pura (porque su grado de membresía en el conjunto de los líquidos potables sería 1) ni ácido sulfúrico (porque su grado de membresía en el conjunto de los líquidos potables sería cero) sino, talvez, gaseosa, en cuyo caso su consumo podría tener algún efecto negativo en su salud, aunque insignificante ante la alternativa de morir de sed.

De cualquier manera, la incertidumbre es un aspecto ineludible en la gran mayoría de situaciones que enfrenta el ser humano. Como mencionamos en la definición 14, hay muchas razones para considerar la incertidumbre en los modelos matemáticos de la realidad que nos interesa. En muchas ocasiones, por ejemplo, no conocemos el verdadero estado del sistema de interés porque tenemos sólo una observación parcial del mismo. Un médico, por ejemplo, puede observar los síntomas de su paciente, pero casi nunca podrá observar su enfermedad y, ciertamente, nunca podrá observar su prognosis, aunque debe tomar decisiones importantes respecto a la enfermedad y su prognosis para recomendar un tratamiento. Por otro lado, las observaciones de aquellos aspectos observables de nuestro sistema suelen ser ruidosas y distorsionadas, por lo que debemos considerar, además, el error de nuestras observaciones. Adicionalmente, el estado del sistema ni siquiera tiene una relación determinística con nuestras observaciones, de manera que, aunque ellas fueran completas y exactas, puede existir aún alguna incertidumbre con respecto al estado real del sistema. Continuando con el ejemplo anterior, un conjunto de síntomas puede asociarse con diferentes enfermedades, aunque la experiencia puede indicar que algunas enfermedades son más probables que otras dado el conjunto de síntomas. La incertidumbre debida a estos aspectos generadores de incertidumbre (dificultad para observar el

estado real del mundo, dificultad para medir con precisión los aspectos medibles del mundo y dificultad para modelar con precisión la dependencia entre lo observado y lo inobservable) podría reducirse con los avances en el conocimiento científico y en el desarrollo tecnológico. Sin embargo, aún nisiquiera es claro que el universo sea realmente determinístico cuando se estudia con un nivel suficiente de detalle. En el dominio actual de la mecánica cuántica, ciertamente no lo es; pero tampoco lo es con respecto a nuestra comprensión actual de muchos otros aspectos del mundo (sin mencionar, por ejemplo, el caos).

Esta presencia inevitable de la incertidumbre sobre el estado real del mundo nos exige considerar todos los posibles estados, aunque nuestras observaciones no nos permitan eliminar muchas de esas posibilidades. Y, en estas condiciones, resulta muy conveniente argumentar no sólo sobre los estados posibles, sino sobre las probabilidades de cada uno de esos posibles estados. De esta manera, las decisiones que tomemos (por ejemplo, decisiones de ingeniería, de tratamiento médico, de planeación económica, etc.) será mucho más significativas y mucho más acertadas. En este contexto, la inutilidad de la intuición en la teoría de las probabilidades a la que se refería DeMoivre se puede matizar con otras tres ideas íntimamente relacionadas. La primera es del gran biólogo Stephen Jay Gould: “el desconocimiento de la teoría de las probabilidades puede ser el mayor impedimento para el conocimiento científico”. La segunda es de Pierre Simon Laplace: “En el fondo, la teoría de las probabilidades no es más que sentido común reducido al cálculo”. La tercera es de Louis Pasteur : “La suerte sólo favorece a las mentes preparadas”.

III. Conceptos Básicos de Variables Aleatorias

29. Variable Aleatoria

Dado un espacio de probabilidad $(\Omega, \mathcal{F}, \mathbf{P})$*, una variable aleatoria (va) es una función* $X: \Omega \to \mathbb{R}$ *tal que,* $\forall x \in \mathbb{R}$*, el evento* $A(x)$ *definido como* $\{\omega \in \Omega : X(\omega) \leq x\}$ *es un evento medible* $(A(x) \in \mathcal{F})$

La variable aleatoria le asigna a cada elemento del espacio muestral un número real, de manera tal que las imágenes de los eventos en $\mathcal{F}$ resultan ser conjuntos de Borel en $\mathbb{R}$, como sugiere la Figura 31.

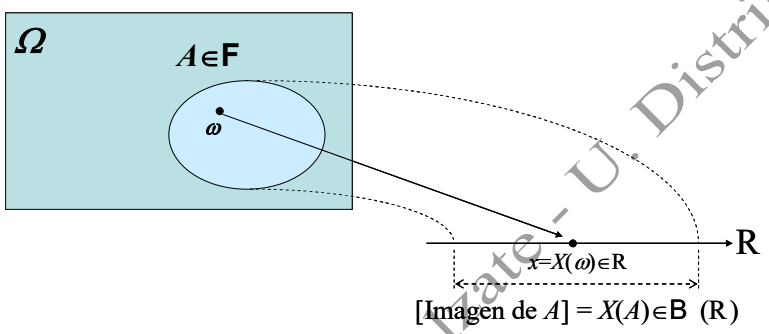


Figura 31. Concepto de Variable Aleatoria

Lo primero que podemos ver es que una variable aleatoria no es una variable sino una función; y no es aleatoria sino determinística: a cada $\omega \in \Omega$ le corresponde uno y sólo un valor real, $X(\omega)$. El nombre (aparentemente inapropiado) de variable aleatoria se debe a razones históricas, pero se convierte en un buen truco mnemotécnico: una variable aleatoria no es una variable sino una función y no es aleatoria sino determinística.

Considere el ejemplo 4 en el que observamos la ocupación de un canal de comunicaciones. El espacio de probabilidad del experimento está dado por el espacio muestral $\Omega = \{libre, ocupado\}$, el campo-$\sigma$ de eventos $\mathcal{F} = \{0,1\}^{\Omega} = \{\Phi, \{libre\}, \{ocupado\}, \Omega\}$, y la medida de probabilidad $P: \mathcal{F} \to \mathbb{R}$ dada por $P(\Phi)=0$, $P(\{libre\}) = 1-p$, $P(\{ocupado\}) = p$ y $P(\Omega)=1$. Sobre este espacio de probabilidad $(\Omega, \mathcal{F}, P)$ podemos definir una función determinística $X: \Omega \to \mathbb{R}$ dada por los valores $X(libre)=0$ y $X(ocupado)=1$. ¿Es esta función una variable aleatoria? Claramente,

$A(x) = \{\omega \in \Omega : X(\omega) \leq x\} = \Phi \in \mathcal{F} \quad \forall\, x<0$

$A(x) = \{\omega \in \Omega : X(\omega) \leq x\} = \{libre\} \in \mathcal{F} \quad \forall\, 0 \leq x<1$

$A(x) = \{\omega \in \Omega : X(\omega) \leq x\} = \Omega \in \mathcal{F} \quad \forall\, x \geq 1$

Como $A(x) = \{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F} \quad \forall\, x \in \mathbb{R}$, X es una variable aleatoria. Es más, podemos graficar la probabilidad de $A(x)$ para cada $x \in \mathbb{R}$, pues $\mathbf{P}(A(x)) = \mathbf{P}(\Phi) = 0$ si $x>0$, $\mathbf{P}(A(x)) = \mathbf{P}(\{libre\}) = 1-p$ si $0 \leq x<1$, y $\mathbf{P}(A(x)) = \mathbf{P}(\Omega) = 1$ si $x \geq 1$. La Figura 32 muestra esta probabilidad $\mathbf{P}(A(x))$ como función de x.

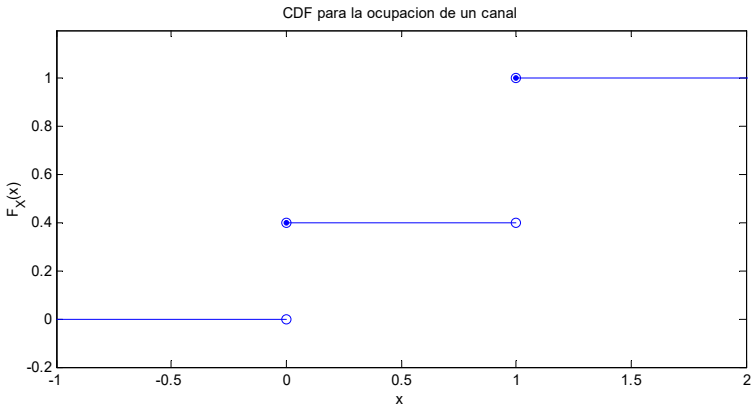


Figura 32. P($\{\omega\in\Omega : X(\omega)\leq x\}$) en el ejemplo de la ocupación de un canal con X = Número da canales ocupados, cuando P($\{ocupado\}$)=0.6

Claro, no cualquier función $X: \Omega\rightarrow\mathbb{R}$ es una *va*. Por ejemplo, consideremos el espacio de probabilidad ($\Omega = \{a_1, a_2, a_3, a_4\}$, $\mathcal{F}= \{\Phi, \{a_1\}, \{a_2\}, \{a_1, a_2\}, \{a_3, a_4\}, \{a_1, a_3, a_4\}, \{a_2, a_3, a_4\}, \Omega\}$, $\mathbf{P} = \{0, p, q, p+q, 1-p-q, 1-q, 1-p, 1\}$). Este es un espacio de probabilidad bien definido pues $\mathcal{F}$ es una clase de subconjuntos de Ω cerrado para el complemento y para la unión, y $\mathbf{P}$ es una medida de probabilidad de $\mathcal{F}$ en $\mathbb{R}$ que satisface los tres axiomas. Definamos la función $X(a_i) = i$, i=1,2,3,4. ¿Es X una *va*? ¡Esta vez no! En efecto, para $x<1$, $A(x) = \Phi\in\mathcal{F}$ con $\mathbf{P}(A(x)) = 0$; para $1\leq x<2$, $A(x) = \{a_1\}\in\mathcal{F}$ con $\mathbf{P}(A(x)) = p$; para $2\leq x<3$, $A(x) = \{a_1, a_2\}\in\mathcal{F}$ con $\mathbf{P}(A(x)) = p+q$; para $x\geq 4$, $A(x) = \Omega\in\mathcal{F}$ con $\mathbf{P}(A(x)) = 1$; pero para $3\leq x<4$, $A(x) = \{a_1, a_2, a_3\}\notin\mathcal{F}$… Y la definición de *va* exige que $A(x)$ debe ser medible <u>para todo</u> $x\in\mathbb{R}$. Pero es fácil verificar que, por ejemplo, $Y(a_i) = \lceil i/2\rceil$ sí es una variable aleatoria en este espacio de probabilidad ($\lceil z\rceil$ es el techo de z, esto es, el mínimo entero mayor o igual a z). Efectivamente, $Y(a_1) = Y(a_2) = 1$, $Y(a_3) = Y(a_4) = 2$, de manera que para $y<1$, $\mathbf{P}(A(y)) = \mathbf{P}(\Phi) = 0$; para $1\leq y<2$, $\mathbf{P}(A(y)) = \mathbf{P}(\{a_1, a_2\}) = p+q$; y para $2\leq y$, $\mathbf{P}(A(y)) = \mathrm{P}(\Omega) = 1$.

En general, si $|\Omega|$ es finita o contable y $\mathcal{F} = \{0,1\}^{\Omega}$, cualquier función de Ω en $\mathbb{R}$ es una *va*. Pero si $|\Omega|$ es infinita incontable, podemos convertir cualquier función $X: \Omega\rightarrow\mathbb{R}$ en una *va* si construimos un espacio de probabilidad para ella con el mínimo campo-σ que contiene los eventos $A(x) = \{\omega\in\Omega : X(\omega)\leq x\}$, $\forall x\in\mathbb{R}$, y asignamos alguna medida de probabilidad a dichos eventos. De esta manera nos aseguramos que la imagen de los eventos en el espacio de probabilidad sean conjuntos de Borel en $\mathbb{R}$.

Observe que en muchos casos el espacio muestral mismo está contenido en (o es igual a) el conjunto de los números reales, de manera que $X(\omega)=\omega$ es una *va* perfectamente válida, como lo podría ser cualquier otra función de $\mathbb{R}$ en $\mathbb{R}$. Este es el caso de los ejemplos 2, 5, 6, 7, 10, 11, 12, 13 y 14 de la definición 5. En el caso del ejemplo 4, como acabamos de ver, la función bivaluada $X(libre)$=0, $X(ocupado)$=1 es una variable aleatoria. En el experimento 9, ver si un bit se recibe con error o no, la *va* dada por $X(si)$=1, $X(no)$=0, tiene un comportamiento muy parecido al caso del ejemplo 4.

En el caso del ejemplo 8 (verificar el estado de ocupación de cada canal de una trama E1) podríamos definir una *va* en ese experimento asociando cada uno de los elementos del espacio muestral con el número binario de 32 bits conformado de la siguiente manera: asignamos un cero a cada canal libre y un uno a cada canal ocupado y decimos que cada dígito representa la potencia de dos asociada con la posición del canal en la trama, $X(\omega)=\sum_{i=0}^{31}2^i 1\left(el\ i\text{-}ésimo\ canal\ en\ \omega\ está\ ocupado\right)$[20]. Siendo así, el rango de la *va* será el conjunto de los números enteros desde 0 hasta 2^{32}-1 = 4.294'967.295. Entre muchas otras variables aleatorias que podríamos imaginar en este mismo experimento se puede mencionar el ancho de banda libre en la trama, que es un múltiplo de 64 kbps: $Y(\omega)=64000\sum_{i=0}^{31}1\left(el\ i\text{-}ésimo\ canal\ en\ \omega\ está\ libre\right)$.

30. Función de Distribución de Probabilidad Acumulativa, *CDF*

> *Sea* $(\Omega, \mathcal{F}, \mathbf{P})$ *un espacio de probabilidad sobre el cual se define una variable aleatoria* $X: \Omega\to\mathbb{R}$. *La Función acumulativa de distribución de probabilidad de X es la función* $F_X: \mathbb{R}\to\mathbb{R}$ *definida como* $F_X(x) = \mathbf{P}(\{\omega\in\Omega: X(\omega)\le x\})$, $\forall x\in\mathbb{R}$. *Le diremos la CDF por la sigla en inglés de Cumulative Distribution Function.*

Obsérvese de dónde surge la importancia de que los eventos $A(x)=\{\omega\in\Omega: X(\omega)\le x\}$ sean medibles en el espacio de probabilidad en que se define la función (*va*) *X*: Si no fuera así no se podría definir la *CDF* de *X* (al menos no con dominio en todos los reales).

No hay manera de destacar suficientemente la importancia de esta función en lo que resta de nuestro estudio en este libro. Para empezar, obsérvese que, mientras $\mathbf{P}(\cdot)$ es una medida de conjuntos (una función de $\mathcal{F}$ en $\mathbb{R}$) y $X(\cdot)$ es una función de Ω en $\mathbb{R}$, $F_X(\cdot)$ es, por primera vez en este libro, un función de los reales en los reales. Con teoría de conjuntos fueron pocas las herramientas que pudimos guardar en nuestra caja de herramientas: tres axiomas, algunas propiedades derivadas de ellos, probabilidad total y regla de Bayes. Pero ahora, con funciones de $\mathbb{R}$ en $\mathbb{R}$, podemos echar mano del análisis real para atiborrar nuestra caja de herramientas.

En el ejemplo 14, el tiempo entre llegadas puede tomar cualquier valor no negativo y menor que infinito, de manera que una posible *CDF* (de hecho, una *CDF* típica en estos casos) es la que se muestra en la Figura 33, donde *x* está medida en ms.

[20] Recordemos que $1(s)$ es la *función indicadora* de la sentencia *s*, igual a 1 si la sentencia *s* es cierta e igual a 0 si la sentencia *s* es falsa, como se dijo en la definición 8.

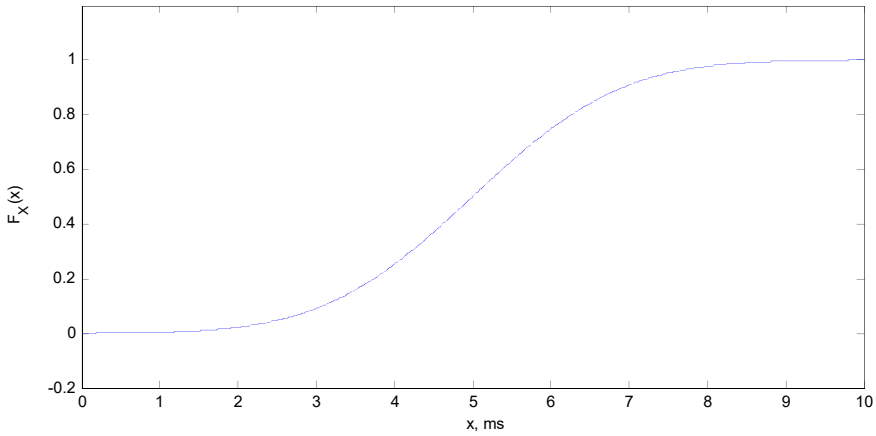


Figura 33. *CDF* para el tiempo entre llegadas de paquetes consecutivos

31. Propiedades de la *CDF*

Sea $(\Omega, \mathcal{F}, \mathbf{P})$ *un espacio de probabilidad sobre el cual se define una variable aleatoria* $X: \Omega \to \mathbb{R}$ *con CDF* $F_X(\cdot)$. *Entonces,*

(a) La CDF es no-negativa: $F_X(x) \geq 0 \ \forall \ x \in \mathbb{R}$.

(b) La CDF es no-decreciente: si $x_1 < x_2$ *entonces* $F_X(x_1) \leq F_X(x_2)$

(c) La CDF es acotada: $F_X(-\infty) = 0$, $F_X(\infty) = 1$.

(d) La CDF es continua por la derecha: $F_X(x^+) = F_X(x)$.

En efecto, como la *CDF* $F_X(x)$ es una medida de probabilidad de un evento indicado por el número real x, las propiedades anteriores son las formas que toman algunas propiedades de la medida de probabilidad. Para cada número real x definamos $A(x)$ como el evento $\{\omega \in \Omega : X(\omega) \leq x\}$, medible en el espacio de probabilidad $(\Omega, \mathcal{F}, \mathbf{P})$.

(a) El segundo axioma de la definición 14 exige que $\mathbf{P}(A(x)) \geq 0$, de donde surge la no-negatividad de $F_X(x)$.

(b) Si $x_1 < x_2$, $A(x_2) = A(x_1) \cup \{\omega \in \Omega : x_1 < X(\omega) \leq x_2\}$, de manera que $A(x_1) \subseteq A(x_2)$ y, de acuerdo con el quinto resultado de la definición 17, $\mathbf{P}(A(x_1)) \leq \mathbf{P}(A(x_2))$, por lo que $F_X(x)$ debe ser no-decreciente.

(c) Como $A(-\infty) = \Phi$, el segundo resultado de la definición 17 exige que $F_X(-\infty) = 0$. Y como $A(\infty) = \Omega$, el primer axioma de las probabilidades exige que $F_X(\infty) = 1$. Dado que $F_X(x)$ es no decreciente, estos resultados implican que $F_X(x)$ está acotada en el rango [0, 1].

(d) Para cualquier $n \in \mathbb{N}$, $A(x+1/n) = A(x) \cup \{\omega \in \Omega : x < X(\omega) \leq x+1/n\} = A(x) \cup B_n(x)$, donde definimos $B_n(x)$ como $\{\omega \in \Omega : x < X(\omega) \leq x+1/n\}$. Como $A(x)$ y $B_n(x)$ son mutuamente excluyentes, aplica el tercer axioma de la definición 14, $\mathbf{P}(A(x+1/n)) = \mathbf{P}(A(x)) + \mathbf{P}(B_n(x))$, de donde $\mathbf{P}(B_n(x)) = F_X(x+1/n) - F_X(x)$. A medida que n tiende a infinito, la cota superior en el intervalo que define a $B_n(x)$ tiende a x, pero x está por fuera del intervalo por la cota inferior, que es abierta en x, por lo

que $B_n(x)$ tiende a Φ. Formalmente, $\lim_{n\to\infty} B_n(x) = \bigcap_{k=1}^{\infty} \{\omega \in \Omega : x < X(\omega) \le x + \frac{1}{k}\} = \Phi$, de manera que $\lim_{n\to\infty} \mathbf{P}\left(B_n(x)\right) = \mathbf{P}(\Phi) = 0$. En consecuencia, $\lim_{n\to\infty} F_X(x+\frac{1}{n}) = F_X(x)$, que es la definición de continuidad por la derecha, $F_X(x^+) = F_X(x)$.

Obsérvese que las funciones de la Figura 32 y la Figura 33 satisfacen las cuatro propiedades anteriores. Veamos otros dos ejemplos sencillos en los que definimos un espacio de probabilidad (Ω, $\mathcal{F}$, **P**) y una función determinística X: $\Omega \to \mathbb{R}$, de manera que podemos deducir su CDF, $F_X(x)$, a partir de la medida de probabilidad **P**, y verificar las propiedades anteriores.

Primero consideremos el estado de ocupación de dos enlaces, donde cada uno está ocupado con probabilidad p independientemente del otro. El espacio muestral de este experimento aleatorio es

$$\Omega = \{(libre,libre),(libre,ocupado),(ocupado,libre),(ocupado,ocupado)\}$$

Y la probabilidad de cada subconjunto unitario se puede calcular fácilmente por la suposición de independencia:

$$\mathrm{P}(\{(l,l)\}) = (1-p)^2, \qquad \mathrm{P}(\{(l,o)\}) = \mathrm{P}(\{(o,l)\}) = p(1-p), \qquad \mathrm{P}(\{(o,o)\}) = p^2$$

Con esta asignación de probabilidad podemos considerar todos los subconjuntos de Ω en nuestro campo-σ de eventos y calcular fácilmente sus probabilidades sumando las probabilidades de sus respectivos subconjuntos unitarios, pues son mutuamente excluyentes:

(*l*,*l*)	**(*l*,*o*)**	**(*o*,*l*)**	**(*o*,*o*)**	***A*⊆Ω**	**P(*A*)**
0	0	0	0	Φ	0
0	0	0	1	$\{(o,o)\}$	p^2
0	0	1	0	$\{(o,l)\}$	$p(1-p)$
0	0	1	1	$\{(o,l),(o,o)\}$	$p(1-p)+p^2 = p$
0	1	0	0	$\{(l,o)\}$	$p(1-p)$
0	1	0	1	$\{(l,o),(o,o)\}$,	$p(1-p)+p^2 = p$
0	1	1	0	$\{(l,o),(o,l)\}$	$2p(1-p)$
0	1	1	1	$\{(l,o),(o,l),(o,o)\}$	$2p(1-p)+p^2 = 1-(1-p)^2$
1	0	0	0	$\{(l,l)\}$	$(1-p)^2$
1	0	0	1	$\{(l,l),(o,o)\}$	$p^2+(1-p)^2 = 1 - 2p(1-p)$
1	0	1	0	$\{(l,l),(o,l)\}$	$p(1-p)+(1-p)^2 = 1-p$
1	0	1	1	$\{(l,l),(o,l),(o,o)\}$	$p(1-p)+p^2+(1-p)^2 = 1-p(1-p)$

1	1	0	0	$\{(l,l),(l,o)\}$	$p(1-p)+(1-p)^2 = 1-p$
1	1	0	1	$\{(l,l),(l,o),(o,o)\}$	$p(1-p)+p^2+(1-p)^2 = 1-p(1-p)$
1	1	1	0	$\{(l,l),(l,o),(o,l)\}$	$2p(1-p)+(1-p)^2 = 1-p^2$
1	1	1	1	Ω	1

Teniendo completamente especificado nuestro espacio de probabilidad (Ω, $\mathcal{F}$, **P**), podemos definir la función determinística $X:\Omega\to\mathbb{R}$ que asigna a cada elemento del espacio muestral el número de canales ocupados:

$$X((l,l)) = 0, \quad X((l,o)) = X((o,l)) = 1, \quad X((o,o)) = 2$$

Entonces podemos definir los eventos $A(x) = \{\omega\in\Omega: X(\omega) \le x\}$ $\forall x\in\mathbb{R}$ y sus respectivas probabilidades, las cuales corresponden a la CDF, $F_X(\cdot)$

$$A(x) = \begin{cases} \Phi & x<0 \\ \{(l,l)\} & 0\le x<1 \\ \{(l,l),(l,o),(o,l)\} & 1\le x<2 \\ \Omega & 2\le x \end{cases} \Rightarrow F_X(x) = P\big(A(x)\big) = \begin{cases} 0 & x<0 \\ (1-p)^2 & 0\le x<1 \\ 1-p^2 & 1\le x<2 \\ 1 & 2\le x \end{cases}$$

Como muestra la Figura 34, esta función satisface las cuatro propiedades de una CDF: no-negativa, no-decreciente, acotada y continua por la derecha.

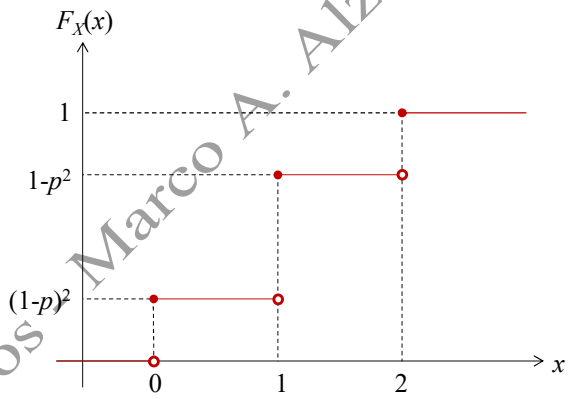


Figura 34. CDF del número de enlaces ocupados entre dos enlaces cuando cada uno se encuentra ocupado con probabilidad p, independientemente del otro

En el segundo ejemplo escogeremos un punto al azar del cuadrado unitario,

$$\Omega=\{(a,b)\in\mathbb{R}^2 : a\in[0,1],\ b\in[0,1]\}$$

Por supuesto, en este experimento no podemos escoger el conjunto potencia de Ω como dominio de la probabilidad, pues sería un dominio demasiado grande. Por eso escogemos como campo-σ de eventos el campo de Borel de $\mathbb{R}^2$, restringido al cuadrado unitario (ver la definición 55 y la nota de pie de página de la definición 21). En ese espacio muestral y con ese campo de eventos, una medida de probabilidad adecuada para cada subconjunto medible es su área, como muestra la Figura 35.

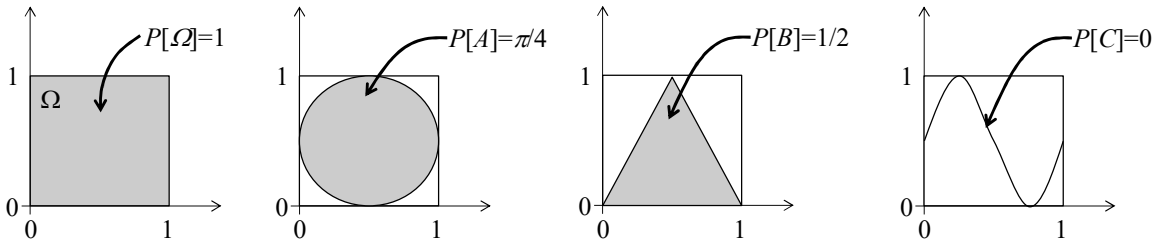


Figura 35. Probabilidad de algunos conjuntos de Borel en el cuadrado unitario de $\mathbb{R}^2$

Ahora que tenemos la especificación completa del espacio de probabilidad (Ω, $\mathcal{F}$, **P**), podemos definir una función determinística $X:\Omega\to\mathbb{R}$. En este caso, escogemos la siguiente función: $X(a,b)=\min(a,b)$. ¿Cuáles son los conjuntos $A(x) = \{\omega\in\Omega : X(\omega) \le x\}\ \forall x\in\mathbb{R}$, y cuáles son sus probabilidades? La Figura 36(a) muestra estos conjuntos: todos los puntos (a,b) del cuadrado unitario tales que $\min(a,b)\le x$. Su probabilidad se puede expresar como la suma de las áreas de dos subconjuntos mutuamente excluyentes: Uno de área x y otro de área $x(1-x)$ para un área total de $x(2-x)$. Evidentemente, si $x<0$, $A(x)=\Phi$ y, si $x>1$, $A(x)=\Omega$. Por consiguiente, la función de distribución acumulativa de esta variable aleatoria es

$$F_X(x) = \begin{cases} 0 & x<0 \\ x(2-x) & 0\le x<1 \\ 1 & x\ge 1 \end{cases}$$

Como se muestra en la Figura 36(b).

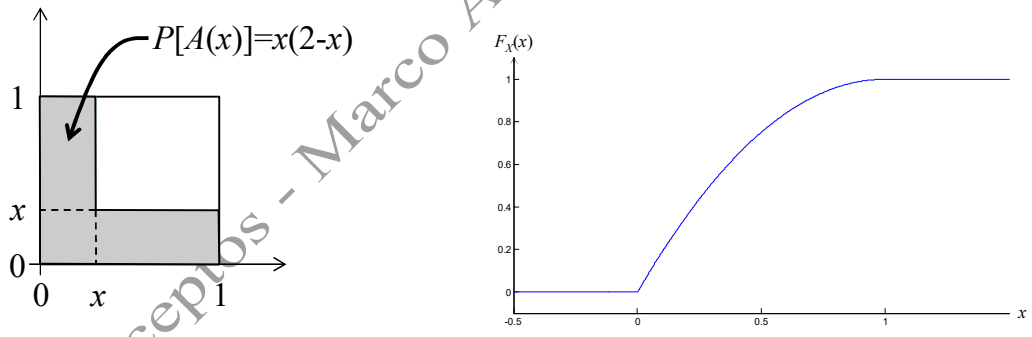


Figura 36. $A(x)=\{(a,b)\in\Omega : \min(a,b) \le x\}$, $F_X(x)=P[A(x)] = x(2-x)$

Nuevamente, esta función satisface las cuatro propiedades de una CDF: es no-negativa, no-decreciente, acotada y continua por la derecha.

Una de las razones por las que la *CDF* es un concepto tan fundamentalmente importante es que cualquier función de $\mathbb{R}$ en $\mathbb{R}$ que satisfaga las anteriores cuatro propiedades es una *CDF* válida en el espacio de probabilidad $(\mathbb{R}, \mathcal{B}(\mathbb{R}), F_X(\cdot))$. Esto es, no es estrictamente necesario considerar un espacio de probabilidad sobre el cual podamos definir una *va* para la cual construiríamos la respectiva *CDF* de acuerdo con las probabilidades de los eventos medibles en el espacio original, como hicimos en los dos ejemplos anteriores. Podemos tomar el camino inverso: Considerar una *CDF* y definir con

ella una *va* apropiada en el espacio $(\mathbb{R}, \mathcal{B}(\mathbb{R}), F_X(\cdot))$. Siendo así, para especificar completamente una variable aleatoria basta con describir su *CDF*: Decir qué valores toma y cómo se distribuye la probabilidad sobre esos valores. ¡No hace falta definir ningún otro espacio de probabilidad subyacente!

Por ejemplo, considere la función $g(x) = (1 - e^{-\lambda x})u(x)$, donde $u(x)$ es el escalón unitario que vale 1 si $x \geq 0$ y vale 0 en otro caso, y λ es un número real positivo. Esta es una función no-negativa, no-decreciente, acotada y continua. Por lo tanto, podemos suponer la existencia de una *va* X que toma valores en los reales no negativos y asignarle la *CDF* $F_X(x) = g(x)$, con lo que construiríamos un espacio de probabilidad formalmente definido. Si establecemos la hipótesis de que dicho espacio modela el tiempo de vida útil de los componentes de una red, por ejemplo, podríamos construir y evaluar así un modelo probabilístico de confiabilidad. A las variables aleatorias con $F_X(x) = 1 - e^{-\lambda x}$, $x \geq 0$, $\lambda > 0$, se les conoce como variables aleatorias exponenciales, como se describe en la definición 44.

Como de ahora en adelante vamos a trabajar casi exclusivamente con variables aleatorias, vamos a despreocuparnos desde ahora por la definición explícita de un espacio de probabilidad, pues tácitamente dejaremos que dicho espacio sea $(\mathbb{R}, \mathcal{B}(\mathbb{R}), F_X(\cdot))$. Tanto es así, que de ahora en adelante nos tomaremos muchas libertades en la notación. Por ejemplo, en vez de hablar de $\mathbf{P}(\{\omega \in \Omega : X(\omega) \in B\})$, donde B es un conjunto de Borel, diremos solamente $\mathbf{P}(X \in B)$. Por supuesto, formalmente este es un error gramatical que podría confundirse con un garrafal error conceptual porque las probabilidades no se asignan a sentencias lógicas sino a subconjuntos medibles de Ω. Pero como ya no necesitamos hacer referencia a un espacio muestral subyacente, es simplemente nuestra convención para referirnos a la probabilidad del evento medible B en el espacio muestral $\mathbb{R}$[21]. Es importante insistir en lo que decíamos en el primer capítulo: Si no tenemos perfectamente definido un espacio de probabilidad para nuestro modelo, no sabremos dónde estamos parados. Sólo estamos diciendo que, mientras nuestro modelo probabilístico se base en una variable aleatoria, el correspondiente espacio de probabilidad puede dejarse implícitamente definido. Por esta razón, otra libertad en la notación será la de cambiar la frase "*Sea* $(\Omega, \mathcal{F}, \mathbf{P})$ *un espacio de probabilidad sobre el cual se define una va* X*:* $\Omega \to \mathbb{R}$ *con CDF* $F_X(x)$" por la frase "*Sea* $F_X(x)$ *la CDF de alguna va* X" (a menos, claro, que necesitamos referirnos explícitamente al espacio de probabilidad subyacente).

32. Probabilidad de Algunos Subconjuntos de $\mathbb{R}$

Sea $F_X(\cdot)$ *la CDF de alguna va* X*. Por simplicidad, denotemos* $\mathbf{P}(X \in B)$ *como* $\mathbf{P}(B)$ *para cualquier* $B \in \mathcal{B}(\mathbb{R})$*. Entonces*

(a) $\mathbf{P}((-\infty, a]) = F_X(a) \ \forall a \in \mathbb{R}$

(b) $\mathbf{P}((a, \infty)) = 1 - F_X(a) \ \forall a \in \mathbb{R}$

(c) $\mathbf{P}((a, b]) = F_X(b) - F_X(a) \ \forall a, b \in \mathbb{R}, a < b$

(d) $\mathbf{P}([a]) = F_X(a^+) - F_X(a^-) \ \forall a \in \mathbb{R}$

[21] Ahora un evento medible es, simplemente, un conjunto de Borel en los reales, $B \in \mathscr{B}(\mathbb{R})$.

(e) $\mathbf{P}((-\infty,a)) = F_X(a) - \mathbf{P}([a]) = F_X(a^-)\ \forall a\in\mathbb{R}$
(f) $\mathbf{P}([a,\infty)) = 1 - F_X(a) + \mathbf{P}([a]) = 1- F_X(a^-)\ \forall a\in\mathbb{R}$
(g) $\mathbf{P}((a,b)) = F_X(b) - F_X(a) - \mathbf{P}([b]) = F_X(b^-) - F_X(a)\ \forall a,b\in\mathbb{R},\ a<b$
(h) $\mathbf{P}([a,b]) = F_X(b) - F_X(a) + \mathbf{P}([a]) = F_X(b) - F_X(a^-)\ \forall a,b\in\mathbb{R},\ a<b$
(i) $\mathbf{P}([a,b)) = (F_X(b) - \mathbf{P}([b]) - (F_X(a) - \mathbf{P}([a])) = F_X(b^-) - F_X(a^-)\ \forall a,b\in\mathbb{R},\ a<b$

Como de costumbre, estas propiedades surgen de los tres axiomas de la probabilidad, como mostraremos a continuación:

(a) Esta es la definición 30
(b) Este es el resultado 1 de la definición 17, aplicado a (a)
(c) $(-\infty,b] = (a,b]\cup(-\infty,a]$ son dos eventos disyuntos, por lo que aplica el tercer axioma: $F_X(b) = \mathbf{P}((a,b]) + F_X(a)$. Restando $F_X(a)$ a ambos lados se obtiene el resultado.
(d) De acuerdo con el resultado anterior, $\mathbf{P}((a-1/n,a]) = F_X(a) - F_X(a-1/n)$ para todo entero n mayor o igual a 1. En el límite cuando n tiende a infinito, el evento $(a-1/n,a]$ tiende a $\lim_{n\to\infty}(a-1/n,a] = \bigcap_{j=1}^{\infty}(a-1/j,a] = [a]$, mientras que $F_X(a-1/n)$ tiende a $F_X(a^-)$. Por la propiedad (d) de la definición 31, $F_X(a)=F_X(a^+)$. Poniendo los tres resultados juntos obtenemos $\mathbf{P}([a]) = F_X(a^+) - F_X(a^-)\ \forall a\in\mathbb{R}$.
(e) $(-\infty,a] = (-\infty,a)\cup[a]$ son dos subconjuntos mutuamente excluyentes, por lo que aplica el segundo axioma de la definición 14: $F_X(a)= \mathbf{P}([a])+\mathbf{P}((-\infty,a))$. Restando $\mathbf{P}([a])$ a ambos lados obtenemos el resultado.
(f) $[a,\infty) = (-\infty,a)^C$, donde el superíndice C se refiere al complemento respecto al conjunto de los reales. Aplicando el primer resultado de la definición 17 al resultado (e) anterior, $\mathbf{P}([a,\infty)) = 1 - F_X(a) + \mathbf{P}([a])\ \forall a\in\mathbb{R}$.
(g) Como $(a,b] = (a,b)\cup[b]$, $F_X(b) - F_X(a) = \mathbf{P}((a,b)) + \mathbf{P}([b])$. Restando $\mathbf{P}([b])$ a ambos lados se obtiene el resultado.
(h) Como $[a,b] = (a,b]\cup[a]$, $\mathbf{P}([a,b]) = F_X(b) - F_X(a) + \mathbf{P}([a])$.
(i) Como $[a,b] = [a,b)\cup[b]$, podemos aplicar el tercer axioma y el resultado (h) anterior, $\mathbf{P}([a,b]) = F_X(b) - F_X(a) + \mathbf{P}([a]) = \mathbf{P}([a,b)] + \mathbf{P}([b])$. Restando $\mathbf{P}([b])$ obtenemos $\mathbf{P}([a,b)) = (F_X(b) - \mathbf{P}([b]) - (F_X(a) - \mathbf{P}([a]))$

Cada uno de estos resultados tiene interpretaciones importantes. En particular, consideremos el punto (*d*): Si un punto individual x_0 de $\mathbb{R}$ tiene una probabilidad diferente de cero, la *CDF* de la correspondiente *va* debe tener una discontinuidad en ese punto, de manera que $F_X(x_0) = F_X(x_0^-) + \mathbf{P}([x_0])$. Por otro lado, si la *CDF* de una variable aleatoria X es una función continua, la probabilidad de cada punto individual es cero, pues la continuidad significa que $F_X(x^+) = F_X(x^-)$ para todo x. Este efecto lo podemos ver con claridad en la Figura **37**, donde graficamos una *CDF* con dos puntos de discontinuidad y seleccionamos cuatro subconjuntos en el eje horizontal para los cuales graficamos sus respectivas probabilidades en el eje vertical.

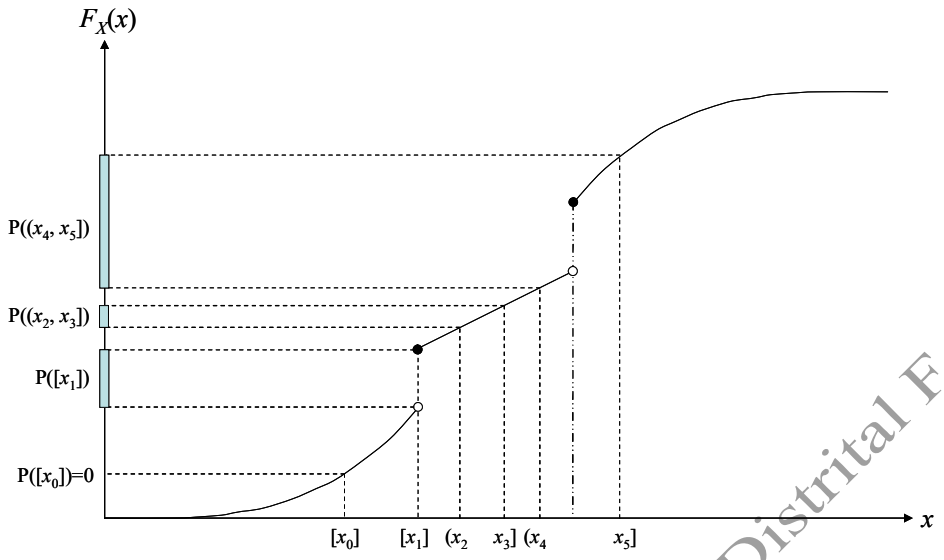


Figura 37. Probabilidad de algunos intervalos

Claramente, x_0 es un punto en el que $F_X(\cdot)$ es continua y, por lo tanto, como subconjunto unitario de $\mathbb{R}$, tiene una probabilidad igual a cero. A diferencia de x_0, x_1 es un punto de discontinuidad, donde la discontinuidad corresponde a un salto de longitud $\mathbf{P}([x_1])$: el evento unitario $[x_1]$ puede suceder con probabilidad mayor que cero. Obsérvese que x_0 también puede suceder, a pesar de que su probabilidad es cero! De hecho, todos los valores en el rango mostrado en la figura pueden suceder, aunque sólo dos de ellos con probabilidad diferente de cero. De muchas maneras, nuestra vida está construida a partir de eventos que, aunque tenían probabilidad cero, ocurrieron para hacer de nosotros lo que somos hoy: casi todo lo que ocurre a nuestro alrededor ocurre a pesar de tener probabilidad cero. Se diría que cada uno de nosotros es un milagro! Este es un aspecto importante por considerar con las regiones en que la *CDF* de una *va* es continua. Consideremos, por ejemplo, el intervalo $(x_2, x_3]$: Cada punto individual de ese intervalo tiene probabilidad cero, aunque la probabilidad de que la *va* tome algún valor dentro de ese intervalo es $\mathbf{P}((x_2, x_3]) = F_X(x_3)-F_X(x_2) > 0$. De acuerdo con la figura, esta probabilidad es pequeña comparada con la probabilidad de que la *va* tome un valor en el intervalo $(x_4, x_5]$, el cual contiene un punto de discontinuidad (llamémosle x_a), de manera que $\mathbf{P}((x_4, x_5]) = \mathbf{P}((x_4, x_a)) + \mathbf{P}((x_a, x_5])+\mathbf{P}([x_a])$. Todos los demás puntos del intervalo tienen probabilidad cero aunque, en conjunto, tienen una probabilidad mayor que la de $[x_a]$ (ver definición 27).

Como acabamos de ver, en general, una *CDF* puede tener puntos de discontinuidad, regiones monótonamente crecientes y regiones donde toma un valor constante. Sin embargo, a veces resulta conveniente describir estas *CDF* generales como la combinación convexa de dos CDFs, una continua en todo el rango $\mathbb{R}$ y otra que es constante en intervalos delimitados por un número contable de discontinuidades. Por ejemplo, si $F_1(x)$ toma una forma semejante a la de la Figura **33** y $F_2(x)$ toma una forma semejante a la de la Figura **32**, la combinación convexa $F_X(x) = \alpha F_1(x) + (1-\alpha)F_2(x)$, $0 \leq \alpha \leq 1$, tomaría una forma semejante a la de la Figura **37**. El hecho de que pueda haber puntos de discontinuidad con probabilidad mayor que cero y puntos de continuidad con probabilidad igual a cero motiva la definición de variables aleatorias continuas y discretas aunque, estrictamente, no es

una definición necesaria. Se dice que X es una variable aleatoria continua si $F_X(x)$ es una función continua para todo $x\in\mathbb{R}$. Se dice que X es una variable aleatoria discreta si la imagen de Ω es un subconjunto contable de $\mathbb{R}$, en cuyo caso la CDF toma la forma de una suma acumulada de escalones, $F_X(x)=\sum_k p_k u(x-x_k)$, donde $u(x)$ es el escalón unitario que vale 0 si $x<0$ y vale 1 si $x\geq 0$, de manera que $p_k = F_X(x^+) - F_X(x^-) = \mathrm{P}[X=x_k]$. En otro caso, se dice que X es una variable aleatoria híbrida o mixta (Figura 38).

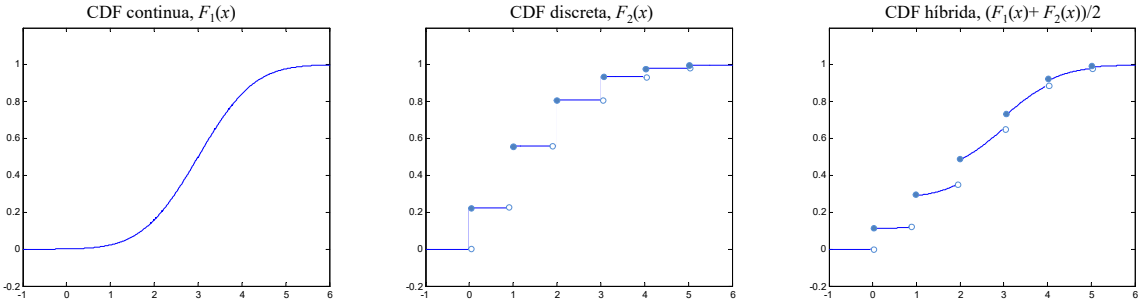


Figura 38. Funciones de distribución acumulativa continua, discreta e híbtrida

Volvamos a los primeros seis ejemplos de la definición 6:

1. Lanzar una moneda y ver qué lado queda hacia arriba. El espacio de probabilidad de este experimento es ($\Omega=\{cara, sello\}$, $\mathcal{F}=\{0,1\}^\Omega$, $\mathbf{P}(\{cara\})=\mathbf{P}(\{sello\})=0.5\}$), de donde podemos definir la variable aleatoria discreta X dada por $X(cara)=0$ y $X(sello)=1$, cuya *CDF* se grafica en la Figura **39**(a).
2. Lanzar un dado y contar los puntos en la cara que queda hacia arriba: $\Omega=\{1,2,3,4,5,6\}$. En este caso $X(\omega)=\omega$ es una variable aleatoria discreta en el que cada posible valor ocurre con probabilidad 1/6, como se muestra en la Figura **39**(b).
3. Escoger una carta de la baraja de naipes: El espacio muestral de este experimento es $\Omega=\{(f,n)$: $f\in\{picas, tréboles, corazones, diamantes\}$, $n\in\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K\}\}$. Sea Y una *va* que asigna a cada palo un número entero así: $Y(picas) = 0$, $Y(tréboles)=1$, $Y(corazones)=2$, $Y(diamantes)=3$. Esta es una variable aleatoria discreta en la que cada posible valor ocurre con probabilidad ¼. Sea Z otra *va* que asigna a cada figura un número entero así: $Z(n) = n$-1 si $n\in\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$, $Z(J)=10$, $Z(Q)=11$ y $Z(K)=12$. Esta es otra variable aleatoria discreta en la que cada posible valor ocurre con probabilidad 1/13. La variable aleatoria $X(f,n) = 13Y(f) + Z(n)$ toma valores en el rango de números enteros [0, 51], donde 0 le corresponde al as de picas y 51 le corresponde al rey de diamantes. Cada posible valor en el rango de X ocurre con probabilidad 1/52, de manera que X es una variable aleatoria discreta cuya *CDF* es como se muestra en la Figura **39**(c).
4. Medir la fracción de paquetes perdidos durante una hora en una red IP: $\Omega=\{\omega\in\mathbb{R} : 0\leq\omega\leq 1\}$. Nuevamente, $X(\omega)=\omega$ es una *va* adecuada para la cual quisiéramos que el valor $X=0$ ocurriera con una probabilidad significativa. Por consiguiente se trata de una *va* mixta cuya *CDF* tiene un punto de discontinuidad en el origen, como muestra la Figura **39**(d). La forma particular de esta *CDF*

puede ser diferente, dependiendo de las condiciones particulares de la red. En el caso que se muestra, se trata de la fracción de pérdidas en un simple enrutador que conecta una pequeña red local con Internet. La probabilidad de que no hayan pérdidas durante una hora es 0.8 y la probabilidad de que se pierdan menos de un cuarto de los paquetes es, para efectos prácticos, uno.

5. Medir el retardo experimentado por un paquete de datos mientras transita por una red IP. Como el espacio es $\Omega = \mathbb{R}^+ = \{x \in \mathbb{R} : x > 0\}$, una *va* perfectamente válida es $X(\omega)=\omega$. El quinto capítulo desarrollaremos varios modelos probabilísticos para este experimento, uno de los cuales conduce a la *CDF* mostrada en la Figura **39**(e). Se trata de una variable continua cuya distribución muestra que el 50% de los paquetes tardan menos de 100 ms y el otro 50% tarda entre 100 y 200 ms.
6. Verificar el estado de ocupación de un canal de comunicaciones: $\Omega = \{libre, ocupado\}$. Aquí, la variable definida en el ejemplo 1 resulta válida. La Figura **39**(f) muestra la *CDF* cuando la probabilidad de ocupación es 0.8.

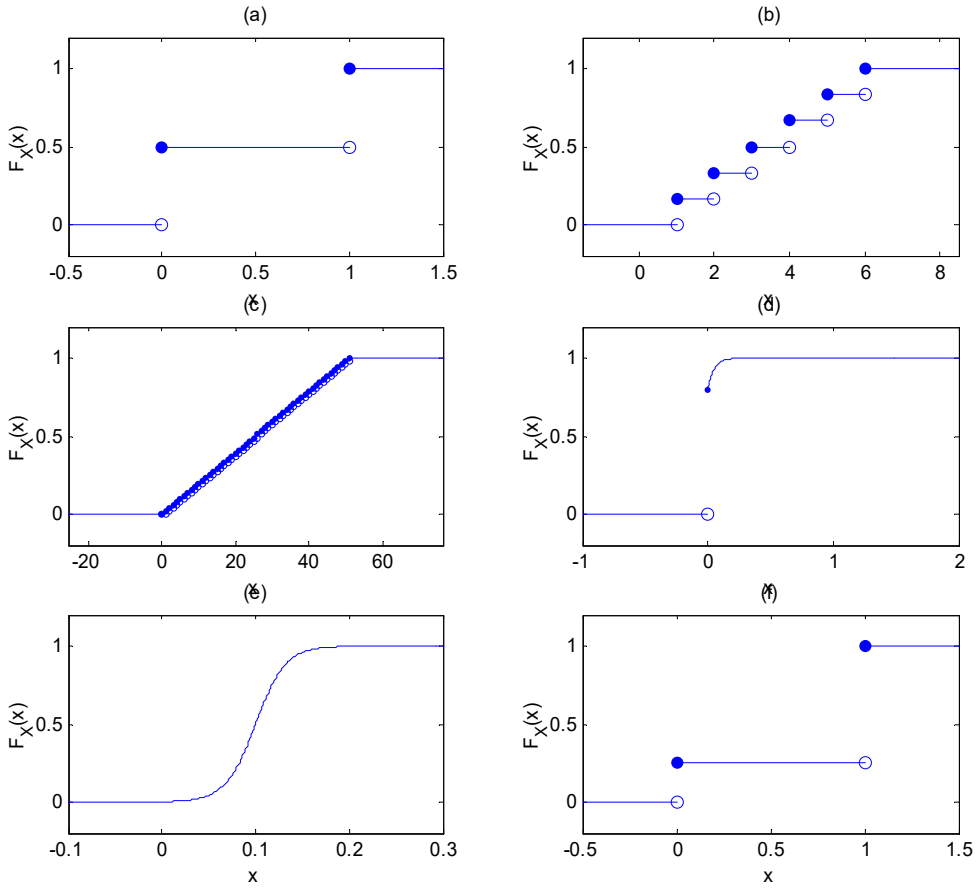


Figura 39. Función de Distribución Acumulativa (CDF) de las variables aleatorias mencionadas en los primeros seis ejemplos de la definición 6

33. Función de Densidad de Probabilidad, *pdf*, y Función de Distribución de Probabilidad, *pmf*

Sea $F_X(\cdot)$ la CDF de alguna va X. La ***función de densidad de probabilidad*** *de X (**pdf** por "probability density function"), $f_X(x)$, se define como la derivada de $F_X(x)$, esto es,*

$$f_X(x) = \frac{d}{dx} F_X(x), \forall x \in \mathbb{R}$$

Si X toma sólo un conjunto contable de posibles valores $\{x_1, x_2, x_3, \ldots\}$, la CDF toma la forma de una suma acumulada de escalones, $F_X(x) = \sum_k p_k u(x - x_k)$, donde $u(x)$ es el escalón unitario que vale 0 si $x<0$ y vale 1 si $x\geq 0$ y $p_k = P[X=x_k]$ es la ***función de distribución de probabilidad*** *de X, (**pmf** por "probability mass function"). En este caso la pdf toma la forma $f_X(x) = \sum_{k=1}^{\infty} p_k \delta(x - x_k)$, donde $p_k = \mathbf{P}(X = x_k)$, y $\delta(x)$ es el impulso de Dirac*

Después de haber pasado por la situación extraña para un ingeniero de tener que manipular una función que va de una clase de subconjuntos a los reales, $P:\mathcal{F}\rightarrow\mathbb{R}$, logramos una forma menos rara de función que va del conjunto muestral a los reales, $X:\Omega\rightarrow\mathbb{R}$. Estas funciones eran raras porque a X habría que evaluarla sobre el color de una bola de billar o sobre la figura de una carta de naipes, mientras que a P habría que evaluarla sobre subconjuntos de objetos como esos. Con la CDF tenemos por primera vez en este libro una función que va de $\mathbb{R}$ en $\mathbb{R}$ y, por supuesto, no podía pasar mucho tiempo antes de que un ingeniero sintiera la necesidad de derivarla (para tranquilidad de nuestros lectores, pronto hallaremos oportunidad hasta de aplicar transformadas de Fourier ☺).

Por ejemplo, la *pdf* de una variable aleatoria exponencial, como la definida en el ejemplo de la definición 31, es $f_X(x) = \frac{d}{dx} F_X(x) = \frac{d}{dx}\left(1 - e^{-\lambda x}\right) = \lambda e^{-\lambda x}, x \geq 0$.

Recordemos que para los puntos $x\in\mathbb{R}$ en los que la *CDF* $F_X(x)$ es continua, la probabilidad $\mathbf{P}(X=x)$ era cero. Sin embargo, sabemos por la definición 32(c) que la probabilidad de que X caiga en un pequeño intervalo $(x, x+\Delta x]$ es $\mathbf{P}(x < X \leq x+\Delta x) = F_X(x+\Delta x) - F_X(x)$. Así pues, la *pdf* se puede considerar como el límite de $\mathbf{P}(x < X \leq x+\Delta x) / \Delta x$ cuando Δx tiende a cero, lo cual justifica su nombre como *densidad* de probabilidad:

$$f_X(x)\Delta x \approx \mathbf{P}(x < X \leq x+\Delta x)$$

Esto es, si bien la *va X* toma el valor x con probabilidad cero, $f_X(x)\Delta x$ nos dice cuál es la probabilidad de un intervalo muy pequeño cercano a x, que resulta un valor proporcional a la longitud del intervalo (si Δx es suficientemente pequeño), con $f_X(x)$ como factor de proporcionalidad, según muestra la Figura 40. Por esta razón, si $F_X(x)$ es continua en $x=x_0$, la probabilidad de que X tome el valor x_0 es cero, aunque x_0 es un valor posible si $f_X(x)$ es mayor a cero en $x=x_0$; sin embargo, si $F_X(x)$ es constante en un intervalo alrededor de x_0, de manera que $f_X(x)$ es cero en $x=x_0$, entonces x_0 no sólo es improbable sino que también es imposible.

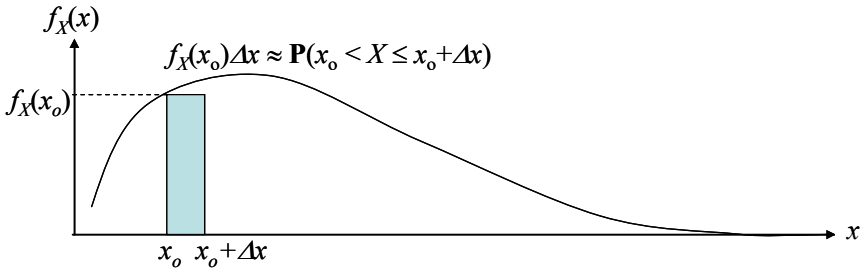


Figura 40. Interpretación de la *pdf*

Es de anotar que la interpretación anterior también tiene una aplicación práctica, pues sugiere una técnica de estimación de la *pdf* de una variable aleatoria mediante el cálculo de la frecuencia relativa de intervalos pequeños en una larga secuencia de muestras de la variable aleatoria. Al dividir cada término de la frecuencia relativa por la longitud del intervalo correspondiente, tendremos un estimado de la *pdf*.

La definición de la *pdf* como la derivada de la *CDF* puede ser muy general en cuanto puede aplicarse a cualquier tipo de variable aleatoria si aceptamos que la *pdf* puede tener discontinuidades (cuando la *CDF* es continua pero no derivable) y singularidades (cuando la *CDF* $F_X(x)$ tiene discontinuidades). En particular, si X es una *va* discreta (que toma sus valores en el conjunto contable $\{x_1, x_2, x_3, \ldots\}$), su derivada será cero en todo punto excepto en los de discontinuidad, en los cuales la derivada se hace singular. En consecuencia, la *pdf* de una variable aleatoria discreta es un tren de impulsos de Dirac[22],

$$f_X(x) = \sum_{k=1}^{\infty} \mathbf{P}(X = x_k)\delta(x - x_k)$$

en el que el área debajo de cada impulso corresponde a la respectiva función de distribución de probabilidad, o *pmf,* $p_k = \mathbf{P}(X = x_k) = F_X(x_k^+) - F_X(x_k^-)$. En este libro hablaremos en general de la *pdf* y, sólo cuando sea estrictamente necesario o conveniente, particularizaremos para la *pmf*. Por ejemplo, cuando verificábamos el estado de ocupación de un canal de comunicaciones y definíamos X(*libre*)=0 y X(*ocupado*)=1, obteníamos la *CDF* mostrada en la Figura **39**(f) si la probabilidad del evento {*ocupado*} fuese 0.8. Evidentemente, se trata de una *va* discreta con la *pmf* mostrada en la Figura 41.

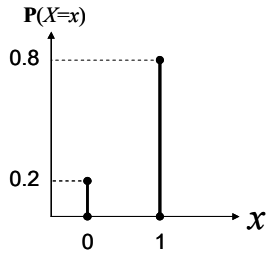


Figura 41. *pmf* de la *va* generada por la ocupación de un canal, cuya CDF aparece en la Figura 39(f)

[22] Recordemos que el impulso de Dirac $\delta(x)$ vale cero en cualquier valor $x \in \mathbb{R}$, excepto en $x=0$, y que $\int_{-\infty}^{\infty} \delta(x)dx = 1$.

34. Propiedades de la *pdf* y la *pmf*

Sea $f_X(\cdot)$ la pdf de alguna va X. Entonces

$(a)\ f_X(x) \geq 0 \ \forall x \in \mathbb{R}$

$(b)\ F_X(x) = \int_{-\infty}^{x} f_X(a)da$

$(c)\ \int_{-\infty}^{\infty} f_X(a)da = 1$

de manera que, para variables discretas, las anteriores propiedades se pueden reescribir en términos de la pmf así:

$(a')\ p_k \geq 0$

$(b')\ F_X(x) = \sum_{k:x_k \leq x} p_k$

$(c')\ \sum_k p_k = 1$

La primera propiedad se debe a que la *CDF* es no decreciente. La segunda propiedad es, simplemente, el teorema fundamental del cálculo. Y la tercera propiedad, que surge de evaluar la segunda en el punto $x=\infty$, es simplemente el primer axioma de las probabilidades definido en 14(a). La propiedad (a') es el segundo axioma de las probabilidades, y las propiedades (b') y (c') surgen de (b) y (c) evaluando la integral

$$\int_a^b f_X(x)dx = \int_a^b \sum_k p_k \delta(x - x_k)\,dx = \sum_k p_k \int_a^b \delta(x - x_k)\,dx = \sum_{k:a<x_k\leq b} p_k$$

Obsérvese que, a la luz de las anteriores propiedades y la interpretación sugerida por la Figura 40, podemos evaluar varias probabilidades en términos de la *pdf* (o la *pmf*) así:

(1) $P(a < X \leq b) = \int_{a^+}^{b^+} f_X(x)dx$ $\quad\left(P(a < X \leq b) = \sum_{k:a<x_k\leq b} p_k\right)$

(2) $P(a \leq X \leq b) = \int_{a^-}^{b^+} f_X(x)dx$ $\quad\left(P(a \leq X \leq b) = \sum_{k:a\leq x_k\leq b} p_k\right)$

(3) $P(a \leq X < b) = \int_{a^-}^{b^-} f_X(x)dx$ $\quad\left(P(a \leq X < b) = \sum_{k:a\leq x_k< b} p_k\right)$

(4) $P(a < X < b) = \int_{a^+}^{b^-} f_X(x)dx$ $\quad\left(P(a < X < b) = \sum_{k:a<x_k< b} p_k\right)$

donde las diferencias sutiles en los límites de la integral se refieren a la necesidad de incluir o excluir posibles impulsos de Dirac. Por supuesto, si X es una *va* continua (esto es, si la CDF es continua), los

cuatro intervalos contemplados en la columna izquierda tienen la misma probabilidad, pues en ese caso la probabilidad de cada punto individual es cero.

Nuevamente, como de la *pdf* (o la *pmf*) de una variable aleatoria se puede obtener la correspondiente *CDF*, cualquier función $g:\mathbb{R}\to\mathbb{R}$ que satisfaga las propiedades anteriores es la *pdf* de alguna variable aleatoria.

Como se nota en las expresiones anteriores, por brevedad basta con referirse solamente la *pdf*, pues todo lo que digamos de ella se extiende inmediatamente a la *pmf* en el caso de variables discretas, a través de la expresión $\int_a^b f_X(x)dx = \sum_{k:a<x_k\le b} p_k$. Más aún, como la necesidad de distinguir entre las propiedades (a) y (a'), (b) y (b'), y (c) y (c') obedecen al uso de la integral de Riemann en las expresiones de probabilidad basadas en la *pdf*, simplificaremos la terminología y la notación si usamos la integral de Lebesgue (o, para este caso, su forma menos general de Riemann-Stieltjes) para evitar referirnos separadamente a la *pdf* o a la *pmf*. En efecto, usando la integral de Lebesgue, la probabilidad del evento $X\in A$ para algún conjunto de Borel A se expresa como $\int_A dF_X(x)$, tanto para variables continuas como para variables discretas o mixtas, donde

$$\mathbf{P}\left(X\in A\right)=\int_A dF_X(x)=\begin{cases}\int_A f_X(x)dx & \text{si } X \text{ es continua}\\ \sum_{k:x_k\in A} p_k & \text{si } X \text{ es discreta}\end{cases}$$

Para los lectores poco familiarizados con la teoría de mediciones o con el análisis real, baste pensar que el término $\mathbf{P}\left(X\in A\right)=\int_A dF_X(x)$ es, simplemente, una notación sencilla para referirse indistintamente a cualquiera de las dos expresiones $\int_A f_X(x)dx$ o $\sum_{k:x_k\in A} p_k$, según corresponda. Por supuesto, en muchas ocasiones será necesario hacer la distinción correspondiente, en cuyo caso volveremos a la sumatoria basada en la *pmf* o a la integral de Riemann basada en la *pdf*, que corresponden a la respectiva integral de Lebesgue en cada caso. Sin embargo, haremos una brevísima mención de esta notación, pues es la más común en la buena literatura técnica sobre redes de comunicaciones.

35. Integrales de Riemann e integrales de Riemann-Stieltjes

Sea $g:\mathbb{R}\to\mathbb{R}$ una función acotada definida en el intervalo $(a,b]$. Subdividamos dicho intervalo mediante algunos valores crecientes de x así $\{x_i\} = \{a = x_0 < x_1 < x_2 < \ldots < x_n = b\}$. Para cada una de estas subdivisiones del intervalo podemos definir las siguientes dos sumas:

$$S=\sum_{i=1}^{n} M_i\cdot\left(x_i-x_{i-1}\right) \qquad\qquad s=\sum_{i=1}^{n} m_i\cdot\left(x_i-x_{i-1}\right)$$

donde

$$M_i=\sup_{x_{i-1}<x\le x_i} g(x) \qquad\qquad m_i=\inf_{x_{i-1}<x\le x_i} g(x)$$

La integral superior de Riemann de g en $[a,b]$ es el ínfimo de los S sobre todas las posibles subdivisiones $\{x_i\}$, mientras que la integral inferior de Riemann de g en $[a,b]$ es el supremo de los s sobre todas las posibles subdivisiones $\{x_i\}$. La integral superior siempre será mayor o igual a la integral inferior pero, si las dos son iguales, el valor común es la integral de Riemann de g en $[a,b]$, que se denota así:

$$\int_a^b g(x)dx = \inf_{\{x_i\}} \sum_{i=1}^{n} \left[(x_i - x_{i-1}) \cdot \sup_{x_{i-1}<x\le x_i} g(x) \right] = \sup_{\{x_i\}} \sum_{i=1}^{n} \left[(x_i - x_{i-1}) \cdot \inf_{x_{i-1}<x\le x_i} g(x) \right]$$

En la integral de Riemann-Stieltjes hacemos la misma partición del mismo intervalo, pero en vez de considerar la longitud del subintervalo $(x_i - x_{i-1})$, consideramos el incremento de una función integradora (en nuestro caso, una función de distribución acumulativa):

$$\int_a^b g(x)dF_X(x) = \inf_{\{x_i\}} \sum_{i=1}^{n} \left[\left(F(x_i) - F(x_{i-1})\right) \cdot \sup_{x_{i-1}<x\le x_i} g(x) \right] = \sup_{\{x_i\}} \sum_{i=1}^{n} \left[\left(F(x_i) - F(x_{i-1})\right) \cdot \inf_{x_{i-1}<x\le x_i} g(x) \right]$$

La notación que se usa para referirnos a la integral de Riemann o a la integral de Riemann-Stieltjes es bastante clara:

$\int_a^b g(x)dx$	Integral de Riemann : El diferencial dx se refiere al límite de la longitud de cada intevalo, $\Delta x_i = x_i - x_{i-1} \to dx$
$\int_a^b g(x)dF_X(x)$	Integral de Riemann-Stieltjes : El diferencial $dF_X(x)$ se refiere al límite del incremento de la función integradora que, en el caso de una CDF, corresponde a la probabilidad de cada intevalo, $\Delta F_X(x_i) = F_X(x_i) - F_X(x_{i-1}) \to dF_X(x)$

Nótese que con la integral de Riemann-Stieltjes, el concepto de integral supera la noción geométrica de "área bajo la curva", apropiado para la integral de Riemann. Ahora estamos considerando otra medida de los intervalos diferente a su longitud. Por otro lado, si describimos el diferencial $dF_X(x)$ mediante la *pdf*, $dF_X(x) = f_X(x)dx$, la integral Riemann-Stieltjes de $g(x)$ con respecto a la función integradora $F_X(x)$ equivale a la integral de Riemann de la función $g(x)f_X(x)$.

Si consideramos que la longitud de un intervalo es una medida del intervalo y que la probabilidad de que una variable aleatoria caiga en un intervalo es otra medida del intervalo, tanto la integral de Riemann como la de Riemann-Stieltjes son integrales con respecto a una medida particular de intervalos.

De todas maneras, el lector desinteresado puede simplemente considerar un asunto de notación:

$$\int_A g(x)dF_X(x) = \begin{cases} \int_A g(x) f_X(x)dx & \text{si } X \text{ es continua} \\ \sum_{k:x_k \in A} g(x_k) p_k & \text{si } X \text{ es discreta} \\ \alpha \int_A g(x) f_1(x)dx + (1-\alpha) \sum_{k:x_k \in A} g(x_k) p_k & \text{si } X \text{ es mixta} \end{cases}$$

En el tercer caso, nos referimos a una variable aleatoria mixta

$$F_X(x) = \alpha F_1(x) + (1-\alpha) F_2(x)$$

donde $F_1(x)$ es la CDF de una variable aleatoria continua con $pdf f_1(x)$, $F_2(x)$ es la CDF de una variable aleatoria discreta con *pmf* p_k y $0 \leq \alpha \leq 1$ (Figura 38).

36. Valor Esperado de una Variable Aleatoria

Sea $F_X(\cdot)$ la CDF de alguna va X. El Valor esperado de X se define como $E[X] = \int_{\mathbb{R}} x dF_X(x)$. Al valor esperado también se le conoce como media, esperanza o primer momento de X.

Supongamos que obtenemos N calificaciones parciales en un curso de procesos estocásticos, $\{X_1, X_2, \ldots, X_N\}$, cada una de ellas en el rango {0.0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0}. Como cada calificación parcial puede tener, de alguna manera, algún componente aleatorio, el profesor querría tomar muchas muestras. Pero como, de todas maneras, al final del semestre el departamento le exige un solo número, el profesor entrega el promedio, $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$. Esto puede ser injusto, porque debe haber circunstancias distintas entre alguien que obtiene tres en todas las notas parciales y alguien que obtiene cinco en el 60% de ellas y cero en el 40% restante, aunque ambos obtienen un mismo promedio de 3.0. Sin embargo, como toca representar toda la secuencia $\{X_1, X_2, \ldots, X_N\}$ mediante un solo número, casi nadie duda que el promedio $\overline{X}$ es la mejor selección posible, especialmente si N es un número grande. Lo ideal sería presentar "la calificación final" como un histograma con la frecuencia relativa de cada posible valor de las calificaciones parciales, pero la administración de semejante proceso de calificación sería muy costosa para el departamento.

¿Cómo se relaciona ese número mágico $\overline{X}$, el promedio, con la distribución de la *va* X? Consideremos la suma que se usa para el promedio y recalculémosla usando la asociatividad de la suma, así

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i = \frac{1}{N}\sum_{k=1}^{11} N_k x_k = \sum_{k=1}^{11} \frac{N_k}{N} x_k$$

donde x_k es el k-ésimo posible valor de X (en este caso $x_k = (k\text{-}1)/2$ para k=1,2,...11), y N_k es el número de veces que se obtuvo la calificación x_k entre las N calificaciones parciales. La máxima justicia de esa calificación final se obtendría cuando el número de calificaciones parciales tendiera a infinito, en cuyo caso, de acuerdo con la definición 15,

$$\lim_{N\to\infty} \overline{X} = \sum_{k=1}^{11} x_k \left(\lim_{N\to\infty} \frac{N_k}{N} \right) = \sum_{k=1}^{11} x_k \mathbf{P}(X = x_k) = \sum_{k=1}^{11} x_k p_k \equiv \int_{\mathbb{R}} x dF_X(x)$$

Si X fuese una variable aleatoria continua de la que tomamos N muestras, bastaría con discretizar el rango de posibles valores en M subintervalos de longitud Δx, de manera que

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i = \frac{1}{N}\sum_{k=1}^{M} N_k x_k = \sum_{k=1}^{M} \frac{N_k}{N} x_k$$

donde N_k es el número de muestras X_i que caen en el k-ésimo intervalo y x_k es algún punto dentro del k-ésimo intervalo que satisface la igualdad de la suma (el cual existe por el teorema del valor medio). Si hacemos que el número de muestras N tienda a infinito, la relación N_k/N tiende a la probabilidad de que X caiga en el k-ésimo intervalo (según nuestra pragmática interpretación frecuentista de la definición15) que, de acuerdo con la definición 33, equivale aproximadamente a $f_X(x_k)\Delta x$, si Δx es suficientemente pequeño:

$$\lim_{N\to\infty}\overline{X}=\sum_{k=1}^{M}x_k\lim_{N\to\infty}\frac{N_k}{N}\approx\sum_{k=1}^{M}x_k f_X(x_k)\Delta x$$

Ahora sólo basta con considerar el límite en el que Δx tiende a cero (en cuyo caso M debe tender a infinito) para que la aproximación sea exacta:

$$\lim_{\substack{M\to\infty\\ \Delta x\to 0}}\lim_{N\to\infty}\overline{X}=\int_{\mathbb{R}}xf_X(x)dx\equiv\int_{\mathbb{R}}xdF_X(x)$$

Así pues, el valor esperado no es más que una generalización del promedio numérico cuando consideramos un número infinito de muestras de la variable aleatoria. Más aún, de acuerdo con el ejemplo de las calificaciones, si el promedio es la estadística más sencilla que mejor resume la secuencia total de calificaciones, el valor esperado es la estadística más sencilla que mejor resume la distribución de una variable aleatoria.

Obsérvense también, en la interpretación anterior, las formas particulares que toma la integral de Lebesgue cuando se aplica a variables continuas y discretas separadamente, en cuyo caso utilizamos explícitamente la *pmf* o la *pdf* en vez de un diferencial general de la *CDF*:

$$E[X]=\int_{\mathbb{R}}xdF_X(x)\equiv\begin{cases}\sum_k x_k p_k & \text{si } X \text{ es discreta}\\ \int_{-\infty}^{\infty}xf_X(x)dx & \text{si } X \text{ es continua}\end{cases}$$

Podemos derivar otra expresión muy útil para $E[X]$ cuando $X\geq 0$:

$$E[X]=\int_0^\infty xdF_X(x)=\int_0^\infty dF_X(x)\int_0^u du=\int_{x=0}^{\infty}\int_{u=0}^{x}dF_X(x)du$$

La Figura 42 muestra cómo podemos realizar el cambio del orden de integración:

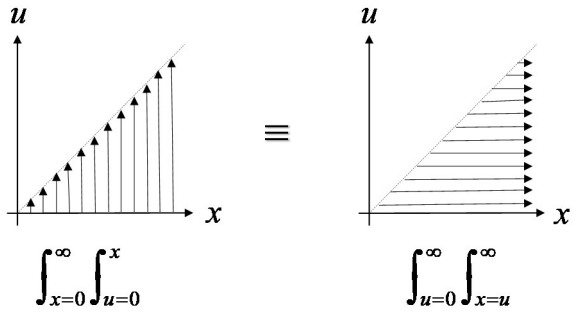


Figura 42. Cambio del orden de integración para $E[X]$, $X\geq 0$

Entonces

$$E[X]=\int_{x=0}^{\infty}\int_{u=0}^{x}dF_X(x)du=\int_{u=0}^{\infty}\left(\int_{x=u}^{\infty}dF_X(x)\right)du=\int_0^\infty P[X>u]du$$

De donde podemos concluir que

$$\text{Si } X \geq 0,\ E[X] = \int_0^\infty \left(1 - F_X(x)\right) dx$$

Que es la expresión útil a la que nos referíamos.

Una segunda estadística importante para resumir la distribución de una *va* es la varianza pero, para definirla, debemos conocer el valor esperado de una función de una variable aleatoria, lo cual requiere tres definiciones previas.

37. Función de una Variable Aleatoria

> *Sea* $(\Omega, \mathcal{F}, \mathbf{P})$ *un espacio de probabilidad sobre el cual se define una variable aleatoria* $X: \Omega \to \mathbb{R}$. *Sea* $g: \mathbb{R} \to \mathbb{R}$ *una función de los reales en los reales. Sea* $Y: \Omega \to \mathbb{R}$ *una función del espacio muestral en los reales tal que a cada* $\omega \in \Omega$ *le asigna la cantidad real* $Y(\omega) = g(X(\omega))$. *Si* $\forall y \in \mathbb{R}$, *el evento* $B(y)$ *definido como* $\{\omega \in \Omega : Y(\omega) \leq y\}$ *es un evento medible* (*esto es, si* $B(y) \in \mathcal{F}$), *entonces* Y *es una nueva variable aleatoria.*

La relación entre las funciones $X(\omega)$ y $Y(\omega) = g(X(\omega))$ se muestran en la Figura 43. Bajo ambas transformaciones, las imágenes inversas de cualquier conjunto de Borel deben corresponder a eventos medibles en $\mathcal{F}$. Esto es, si $B \in \mathcal{B}(\mathbb{R})$, entonces $X^{-1}(B) \in \mathcal{F}$ y $Y^{-1}(B) \in \mathcal{F}$. Como X está definida en el espacio de probabilidad $(\mathbb{R}, \mathcal{B}(\mathbb{R}), F_X(\cdot))$, cualquier función $g(\cdot)$ que transforme conjuntos de Borel en conjuntos de Borel generará una variable aleatoria válida $Y = g(X)$. Basta con verificar que la imagen inversa de cualquier intervalo $(-\infty, y]$, $y \in \mathbb{R}$, sea un conjunto de Borel, $B_y = g^{-1}((-\infty, y]) \in \mathcal{B}(\mathbb{R})$ porque, siendo así, $F_Y(y) = \mathrm{P}[X \in B_y]$ y la variable aleatoria Y queda completamente definida[23].

[23] g es una función de los reales en los reales, $g: \mathbb{R} \to \mathbb{R}$, de manera que el argumento de g debe ser un número real. Sin embargo, podemos hablar de la imagen de un subconjunto $A \subset \mathbb{R}$ bajo la transformación g, $B = g(A) = \{y = g(x) : x \in A\} \subset \mathbb{R}$, así como de la imagen inversa de un conjunto $B \subset \mathbb{R}$, $A = g^{-1}(B) = \{x = g^{-1}(y) : y \in B\} \subset \mathbb{R}$. Para que $Y = g(X)$ sea una variable aleatoria se necesita que $B \in \mathcal{B}(\mathbb{R}) \Leftrightarrow A \in \mathcal{B}(\mathbb{R})$.

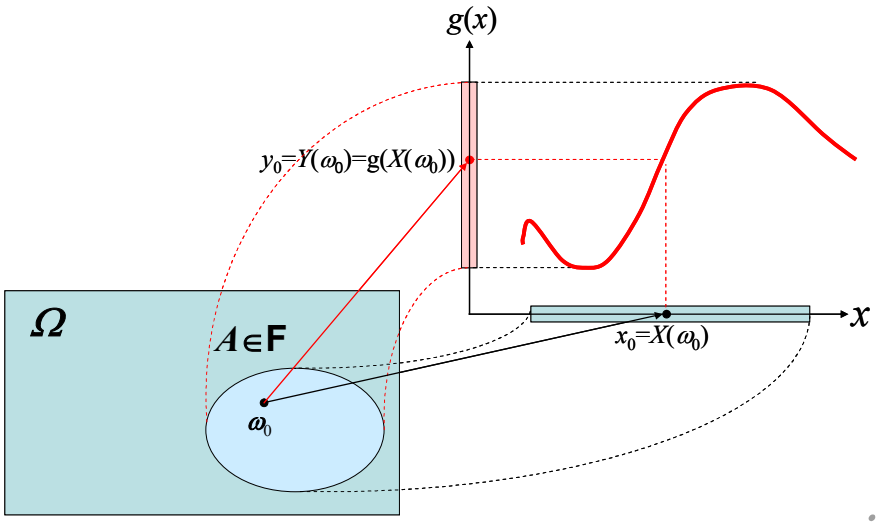


Figura 43. Concepto de Función de una Variable Aleatoria

Por ejemplo, si L es una variable aleatoria que representa la longitud en bits de un paquete de datos que se transmite por un enlace de capacidad C bps, el tiempo de transmisión del paquete será una nueva variable aleatoria dada por $T = (L+h)/C$, donde h es la longitud total de los encabezados que se le añaden al paquete en capas inferiores de la pila de protocolos. Definiendo $g(l) = (l+h)/C$, tenemos que $g^{-1}(t) = Ct-h$, de manera que $g^{-1}((-\infty,t]) = (-\infty,Ct-h]$ y, en consecuencia, g transforma conjuntos de Borel en conjuntos de Borel: $T=g(L)$ es una variable aleatoria válida.

Así como es difícil imaginar subconjuntos de los reales que no sean conjuntos de Borel, es difícil imaginar funciones $g:\mathbb{R}\to\mathbb{R}$ que no definan una nueva variable aleatoria. Por ejemplo, podría parecer que $g(x) = sin(1/x)$ no definiría una variable aleatoria por su extraño comportamiento cerca de $x=0$, como muestra la Figura 44. Pero la verdad es que, para cualquier $y_0\in\mathbb{R}$, el evento $\{\omega\in\Omega: Y(\omega)\le y_0\}$ corresponde a una unión contable de conjuntos medibles definidos por X mediante intervalos, $\cup_{n\in\mathbb{Z}}\{\omega\in\Omega: 1/(2n\pi+\theta_2) \le X(\omega) \le 1/(2n\pi+\theta_1)\}$, donde $[\theta_1,\theta_2]$ es el primer intervalo con $\theta_1 \ge 0$ en el que $sin(\alpha) \le y_0$ si $\alpha\in[\theta_1,\theta_2]$.

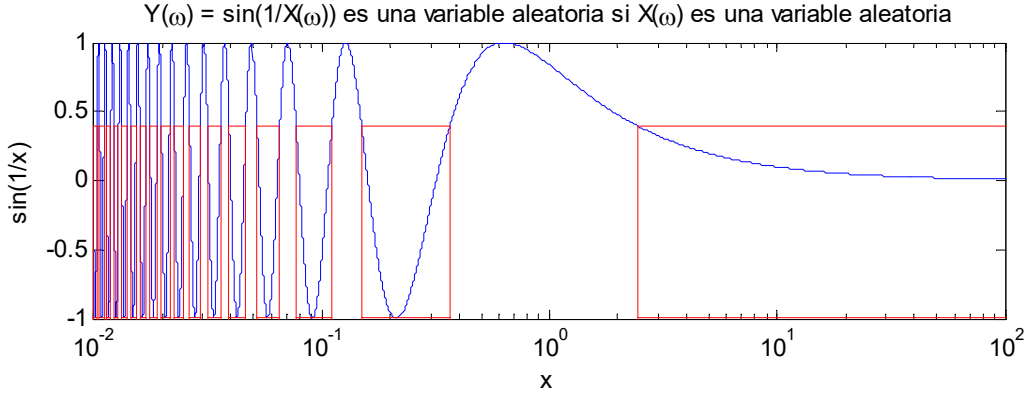


Figura 44. $Y(\omega)=g(X(\omega))$ es una variable aleatoria si $g:\mathbb{R}\to\mathbb{R}$ transforma conjuntos de Borel en conjuntos de Borel

Para otro ejemplo interesante, consideremos la secuencia "tienda de campaña",

$$x_{n+1} = \begin{cases} 3x_n & x_n < 1/2 \\ 3(1-x_n) & x_n \geq 1/2 \end{cases}$$

y definamos la función g:[0,1]→ℝ así: $g(x_0) = 1$ si la secuencia $\{x_0, x_1, x_2, \ldots\}$ permanece siempre dentro del intervalo unitario ($x_n \in [0,1]\ \forall n \in \mathbb{N}$) y $g(x_0) = 0$ en otro caso. Es fácil ver que, si x_0 inicia en el intervalo (1/3, 2/3), ya en el primer paso x_1 habrá excedido el intervalo unitario, como muestra la gráfica de la izquierda de la Figura 45, de manera que $g(x_0)=0$ para $x_0 \in (1/3, 2/3)$. De la misma manera, si $x_0 \in (1/9, 2/9) \cup (7/9, 8/9)$, ya en el segundo paso x_2 habrá excedido el intervalo unitario, como muestra la gráfica central de la Figura 45, de manera que $g(x_0)=0$ para $x_0 \in (1/9, 2/9) \cup (7/9, 8/9)$. La gráfica de la derecha de la Figura 45 muestra lo que ocurre en el tercer paso. En los intervalos (1/27,2/27), (7/27,8/27), (19/27,20/27) y (25/27,26/27), $g(x_0)=0$. Siguiendo de esta manera, queda claro que $g(x_0)$ vale uno en el conjunto de Cantor y vale cero fuera del conjunto de Cantor, esto es, $g(x)$ es la función indicadora del conjunto de Cantor.… ¿Es $Y=g(X)$ una v.a. si X está uniformemente distribuida en [0,1]?... En adelante supondremos que la expresión $Y=g(X)$ siempre definirá una nueva variable aleatoria Y cuando X sea una variable aletoria.

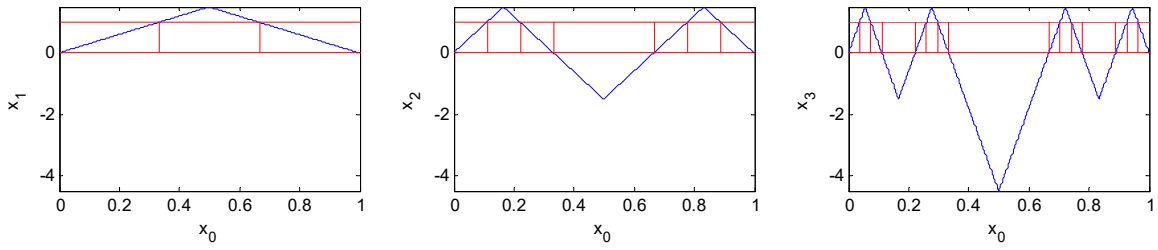


Figura 45. Funcion tienda de campaña g:[0,1]→ℝ

38. *pdf* de una Función de una Variable Aleatoria

Sea $F_X(\cdot)$ la CDF de alguna va X y sea Y otra va definida mediante $Y=g(X)$, donde g es una función de los reales en los reales. Entonces la CDF de Y, $F_Y(y)$, satisface

$$dF_Y(y) = \sum_{i=1}^{n} dF_X(x_i)$$

donde $\{x_1, x_2, \ldots, x_n\}$ son las raíces de la ecuación $y=g(x)$.

Si X ex continua con pdf $f_X(\cdot)$ y g es una función diferenciable en todo punto, la pdf de Y está dada por

$$f_Y(y) = \sum_{i=1}^{n} f_X(x_i) \frac{1}{|g'(x_i)|}$$

donde $g'(x)$ es la derivada de $g(x)$. Si X es una va discreta, la pmf de Y está dada por

$$\mathbf{P}(Y = y_j) = \sum_{i:y_j = g(x_i)} \mathbf{P}(X = x_i)$$

En efecto, considérese el ejemplo mostrado en la Figura 46 en la que, para el punto y seleccionado, existen tres raíces de la ecuación $y=g(x)$, pues $g(x_1)=g(x_2)=g(x_3)=y$. Por el tercer axioma de la definición 14,

$$\mathbf{P}(y<Y\leq y+\Delta y) = \mathbf{P}(x_1<X\leq x_1+\Delta x_1) + \mathbf{P}(x_2+\Delta x_2<X\leq x_2) + \mathbf{P}(x_3<X\leq x_3+\Delta x_3)$$

donde todos los incrementos son positivos con la excepción de Δx_2, que es menor que cero. A medida que Δy se hace más y más pequeña, obtenemos la expresión original de la definición, $dF_Y(y) = \sum_{i=1..3} dF_X(x_i)$. Si X es discreta, esta suma se interpreta como $\mathbf{P}(Y=y) = \mathbf{P}(X=x_1) + \mathbf{P}(X=x_2) + \mathbf{P}(X=x_3)$. Ahora veamos la interpretación de la suma si X es continua.

Si Δy es suficientemente pequeño, $dF_Y(y) = \sum_{i=1..3} dF_X(x_i)$ se puede reescribir de la siguiente manera:

$$f_Y(y)\Delta y \approx f_X(x_1)\Delta x_1 + f_X(x_2)|\Delta x_2| + f_X(x_3)\Delta x_3$$

donde la aproximación se hace exacta a medida que Δy tiende a cero. En términos generales, si existen n raíces, tenemos

$$f_Y(y) \approx \frac{1}{\Delta y}\sum_{i=1}^{n} f_X(x_i)|\Delta x_i| = \sum_{i=1}^{n} \frac{f_X(x_i)}{\Delta y/|\Delta x_i|}$$

donde, en el límite cuando Δy tiende a cero, obtenemos exactitud en la igualdad:

$$f_Y(y) = \lim_{\Delta y\to 0} \sum_{i=1}^{n} \frac{f_X(x_i)}{|\Delta y/\Delta x_i|} = \sum_{i=1}^{n} f_X(x_i)\frac{1}{|g'(x_i)|}$$

Nótese que si la ecuación $y=g(x)$ no tiene raíces, $dF_Y(y)=0$, como muestra la Figura 47(a). Igualmente, si las raíces forman un continuo, la *va* Y puede tener un componente discreto aunque X sea continua, como muestra la Figura 47(b).

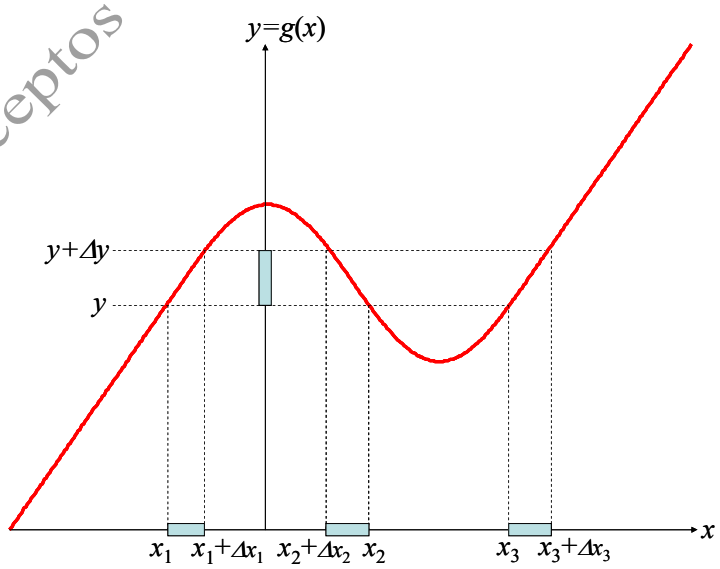


Figura 46. Construcción para encontrar $f_Y(y)$ cuando $Y=g(X)$

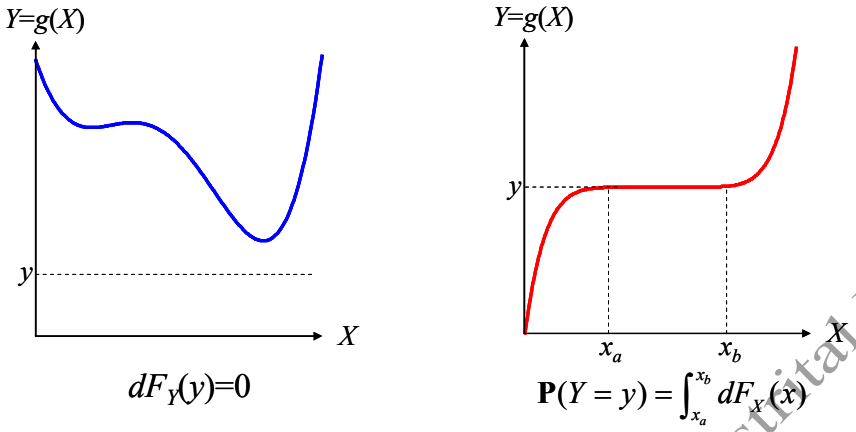


Figura 47. Casos en que $y=g(x)$ no tiene raíces (azul) y en que las raíces de $y=g(x)$ forman un continuo (rojo)

A manera de ejemplo, considérese la transmisión de un archivo desde un servidor ftp a través de un enlace de C bps. Si la longitud del archivo, L, tiene una *pdf* $f_L(l)$, ¿cuál será la *pdf* $f_T(t)$ del tiempo de transferencia, T? Digamos que $T = L/C + t_0$, donde t_0 es el tiempo de establecimiento de la conexión ftp. Si definimos $g(l) \equiv l/C + t_0$ obtenemos que la única raíz de $t=g(l)$ es $l = C\,(t - t_0)$. En este caso la derivada de $g(l)$ es constante, $g'(l) = 1/C$. Consecuentemente, $f_T(t) = C\,f_L(C\,(t - t_0))$.

Como un segundo ejemplo, considérese la eficiencia en la transmisión de un paquete cuya longitud es una variable aleatoria L con *pdf* $f_L(l)$, cuando se le añade un encabezado de h bits: $E = g(L) = L/(L+h)$. La única raíz de $e=g(l)$ es $l = h\,e/(1\text{-}e)$ y la derivada de $g(l)$ es $g'(l) = h/(h+l)^2$. En consecuencia, $f_E(e) = \dfrac{h}{(1-e)^2} f_L\left(\dfrac{h\,e}{1-e}\right)$. La Figura 48 muestra las respectivas distribuciones de L y E cuando $h = 192$ bits y L tiene una distribución exponencial $f_L(l) = \exp(-l/1024)/1024$.

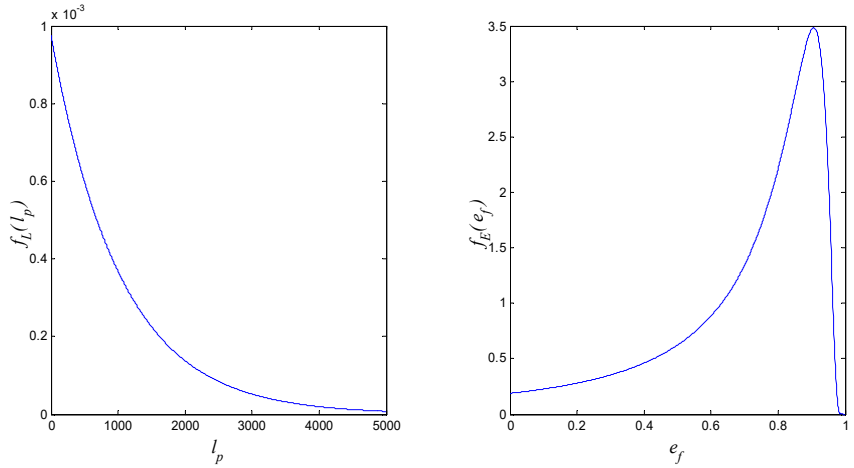


Figura 48. Funciones de densidad de probabilidad de la longitud de un paquete (a) y de la eficiencia en la transmisión cuando se añaden 192 bits de encabezado (b)

39. Valor Esperado de una Función de una Variable Aleatoria

Sea $F_X(\cdot)$ la CDF de alguna va X y sea Y otra va definida mediante $Y=g(X)$, donde g es una función de los reales en los reales. Entonces el valor esperado de Y está dado por

$$E[Y] = \int_{\mathbb{R}} g(x) dF_X(x)$$

En efecto, si por simplicidad suponemos que $g(\cdot)$ es una función monótonamente creciente, de la definición 38 sabemos que $dF_X(x) = dF_Y(g(x))$, por lo que

$$E[Y] = \int_{\mathbb{R}} y dF_Y(y) = \int_{\mathbb{R}} g(x) dF_X(x) = \begin{cases} \int_{-\infty}^{\infty} g(x) f_X(x) dx & \text{si } X \text{ es continua} \\ \sum_k g(x_k) p_k & \text{si } X \text{ es discreta} \end{cases}$$

Para los valores de y en los que $y=g(x)$ tenga varias raíces, la expresión es la misma por asociatividad.

En el ejemplo de la transmisión de un archivo desde un servidor ftp a través de un enlace de C bps cuando la longitud del archivo, L, tiene una *pdf* $f_L(l)$, ¿cuál será el valor esperado del tiempo de transferencia $T = L/C + t_0$? Acabamos de ver que $f_T(t) = C f_L(C(t - t_0))$, de donde podemos verificar que $E[T] = \int_0^{\infty} (t_0 + l/C) f_L(l) dl = t_0 + E[L]/C$.

40. Varianza de una Variable Aleatoria

Sea X una va con valor esperado $E[X]$. La varianza de X, $V[X]$, se define como $V[X] = E[(X - E[X])^2]$. La desviación estándar de X, σ_X, se define mediante la relación $V[X] = \sigma_X^2$.

Supongamos que mandamos medir a uno de nuestros técnicos más brillantes una variable aleatoria X. El técnico es brillante pero perezoso y tramposo, por lo que decide inventarse algún número a y decir que ése fue el valor que midió. Como la equivocación será X-a, él quisiera escoger a de manera que la diferencia X-a sea lo más cercana a cero posible. Para conseguir esto, el brillante técnico querría minimizar $(X\text{-}a)^2$ pero, como ésta es una función de una variable aleatoria, decide escoger el valor de a que minimiza $E[(X\text{-}a)^2]$. Por supuesto, la manera simple de encontrar el valor apropiado de a es observando la derivada de $E[(X\text{-}a)^2]$ respecto a a:

$$\frac{d}{da} E\left[(X-a)^2\right] = E\left[\frac{d}{da}(X-a)^2\right] = E\left[2(a-X)\right] = 2\int_{\mathbb{R}} (a-x) dF_X(x)$$
$$= 2a\int_{\mathbb{R}} dFX(x) - 2\int_{\mathbb{R}} x dF_X(x) = 2(a - E[X])$$

Debido a la convexidad de la función $g(a)=E[(X\text{-}a)^2]$, el único valor extremo corresponde a un mínimo, así que basta con igualar la anterior derivada a cero para obtener el valor de a que minimiza el error cuadrado promedio (MSE *–Mean Square Error–*), $a=E[X]$. Así pues, cuando remplazamos una variable aleatoria por su valor esperado minimizamos el MSE, el cual es, precisamente, la varianza de X, $V[X] = E[(X - E[X])^2]$. Correspondientemente, la desviación estándar σ_X es una medida de qué tan dispersos están los valores observados de X respecto a su valor medio, $E[X]$.

41. Propiedades del Valor Esperado y la Varianza de una Variable Aleatoria

Sea X una variable aleatoria con valor esperado $E[X]$ y varianza $V[X]$ y c una constante entre los reales. Entonces,

(a) $E[X+c] = E[X]+c$
(b) $E[\mathrm{c}\,X] = c\,E[X]$
(c) $V[X+c] = V[X]$
(d) $V[\mathrm{c}\,X] = c^2\,V[X]$
(e) $V[X] = E[X^2] - E[X]^2$

Estas propiedades son muy fáciles de verificar:

(a) $E[X+c] = \int_{\mathbb{R}} (x+c)dF_X(x) = \int_{\mathbb{R}} xdF_X(x) + c\int_{\mathbb{R}} dF_X(x) = E[X]+c$

(b) $E[cX] = \int_{\mathbb{R}} cxdF_X(x) = c\int_{\mathbb{R}} xdF_X(x) = cE[X]$

(c) $V[X+c] = E[((X+c) - (E[X]+c))^2] = E[(X - E[X])^2] = V[X]$
(d) $V[\mathrm{c}X] = E[(\mathrm{c}X - \mathrm{c}E[X])^2] = E[c^2(X - E[X])^2] = c^2E[(X - E[X])^2] = c^2V[X]$
(e) $V[X] = E[(X - E[X])^2] = E[X^2 - 2XE[X] + E[X]^2] = E[X^2] - E[X]^2$

donde la demostración de la propiedad (c) hace uso de la propiedad (a), la demostración de la propiedad (d) hace uso de la propiedad (b) y la demostración de la propiedad (e) hace uso de las propiedades (a) y (b). Estas propiedades se usarán con tanta cotidianidad que, finalmente, deberán ser recordadas como conceptos fundamentales de las variables aleatorias.

42. Momentos de una Variable Aleatoria

El n-ésimo momento de una variable aleatoria X es $E[X^n]$. El n-ésimo momento central es $E[(X-E[X])^n]$.

De acuerdo con lo anterior, el valor esperado es el primer momento y la varianza es el segundo momento central. El *skewness* es una cantidad muy útil relacionada con el tercer momento central, $S[X] = E[(X-E[X])^3]/V[X]^{3/2}$, que mide la simetría de la *pdf* de X alrededor de su valor medio (si $S[X]=0$, la *pdf* de X es simétrica alrededor de $E[X]$; si $S[X]<0$, la *pdf* "se recuesta" hacia la izquierda; y si $S[X]>0$, la *pdf* "se recuesta" hacia la derecha). El *kurtosis* es otra cantidad relacionada con el cuarto momento central, $K[X] = E[(X-E[X])^4]/V[X]^2 - 3$, que mide qué tan plana o puntuda es la *pdf* de X (entre más negativo es $K[X]$, la *pdf* de X tiende a ser más plana; entre más positivo, la *pdf* tiende a ser más puntuda. La referencia $K[X]=0$ corresponde a la distribución gaussiana, descrita en las definiciones 44 y 45. La Figura 49 muestra las características de una *pdf*, representadas por los cuatro primeros momentos de X.

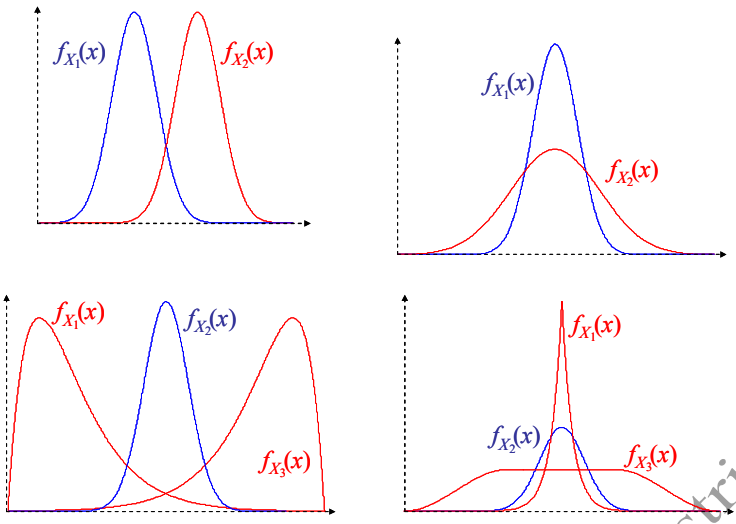


Figura 49. Efecto de los cuatro primeros momentos en la *pdf* de una *va*. En la parte superior izquierda, $E[X_1] < E[X_2]$. En la parte superior derecha, $V[X_1] < V[X_2]$. En la parte inferior izquierda, $S[X_1] < S[X_2]=0 < S[X_3]$. Y en la parte inferior derecha, $K[X_1] > K[X_2] = 0 > K[X_3]$

43. Algunas Variables Aleatorias Discretas

(a) *Una variable aleatoria de Bernoulli con parámetro* $p \in [0,1]$ *toma dos posibles valores*, $X \in \{0,1\}$, *con* $\mathbf{P}[X=1]=1-\mathbf{P}[X=0]=p$. *Su valor esperado es p y su varianza es* $p(1-p)$.

(b) *Una variable aleatoria geométrica con parámetro* $p \in [0,1]$ *toma valores enteros positivos,* $X \in \{1,2,3,\ldots\}$, *de manera que* $\mathbf{P}[X=k]=p^{k-1}(1-p)$. *Su valor esperado es* $1/(1-p)$ *y su varianza es* $p/(1-p)^2$.

(c) *Una variable aleatoria binomial con parámetros* (n,p), *donde n es un entero positivo y p un real en el intervalo* $[0,1]$, *toma valores enteros no negativos en el rango* $\{0,1,2,\ldots,n\}$, *de manera que* $\mathbf{P}[X=k]=\binom{n}{k}p^k(1-p)^{n-k}$. *Su valor esperado es np y su varianza es* $np(1-p)$.

(d) *Una variable aleatoria de Poisson con parámetro* $\rho>0$ *toma valores enteros no negativos,* $X \in \{0,1,2,\ldots\}$, *de manera que* $\mathbf{P}[X=k]=\frac{\rho^k}{k!}e^{-\rho}$. *Tanto su valor esperado como su varianza son iguales a* ρ.

(e) *Una variable aleatoria uniforme discreta con parámetros* (m,n), *donde m y n son enteros tales que* $m \le n$, *toma valores en el rango de números enteros* $\{m, m+1, m+2,\ldots, n-1, n\}$, *de manera que* $\mathbf{P}[X=k] = 1/(n-m+1)$ *si k está en el rango mencionado. Su valor esperado es* $(m+n)/2$ *y su varianza es* $(n-m)(n-m+2)/12$.

A continuación damos algunos ejemplos de modelos probabilísticos en redes de comunicaciones basados en las anteriores variables aleatorias, y demostramos los resultados obtenidos respecto a la

media y la varianza de cada una de ellas. Todos ellos son de gran importancia práctica en telecomunicaciones, en especial el modelo de tráfico Poisson.

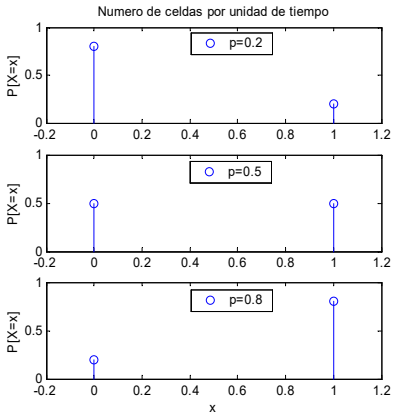
(a) Muchos fenómenos aleatorios en el estudio de redes de comunicaciones pueden modelarse mediante variables aleatorias de Bernoulli, como ya se ha mencionado previamente.

Sea X=1 si un enlace se encuentra ocupado y X=0 si el mismo enlace se encuentra desocupado. Entonces X es una variable aleatoria de Bernoulli, donde el parámetro p es la utilización del enlace.

Sea X=1 si un bit transmitido sobre un enlace de radio punto-a-punto llega con error al otro extremo del enlace, y X=0 si el bit llega correctamente. Entonces X es una variable aleatoria de Bernoulli, donde el parámetro p es la tasa de error del canal, BER (Bit Error Rate).

El siguiente es el primero de una serie de modelos de tráfico que estudiaremos en este libro. Hay un enlace por el que se transmiten celdas ATM (Asynchronous Transfer Mode), donde el tiempo se discretiza en unidades correspondientes al tiempo de transmisión de una celda. En cada unidad de tiempo puede llegar una celda con probabilidad p o no llegar ninguna celda con probabilidad 1-p. Dada una unidad particular de tiempo, sea X=1 si llega una celda en esa unidad y X=0 si no llega ninguna celda. Entonces X es una variable aleatoria de Bernoulli, donde p es la tasa de llegadas, en celdas/unidad de tiempo.

En cualquiera de los tres casos tenemos que $E[X] = 1\cdot p + 0\cdot (1-p) = p$, $E[X^2] = 1\cdot p + 0\cdot (1-p) = p$ y $V[X]= E[X^2] - E[X]^2 = p - p^2 = p(1-p)$, lo cual cobra mucho sentido a la luz de los ejemplos propuestos. En el modelo de tráfico, por ejemplo, si p=0, el promedio es 0 con varianza 0 pues en ninguna unidad de tiempo llegan paquetes; si p=1, el promedio es 1 con varianza 0 pues cada unidad de tiempo trae un paquete. La máxima varianza se da con p = ½, que corresponde a la máxima incertidumbre sobre la llegada de paquetes: si p es menor que ½, tenemos mayor certeza de que no llegará un paquete y, si p es mayor que ½, tenemos mayor certeza de que llegará un paquete. La figura 2.13 muestra la distribución, el promedio y la varianza del número de celdas que llegan por unidad de tiempo en función de la tasa de llegadas, p.



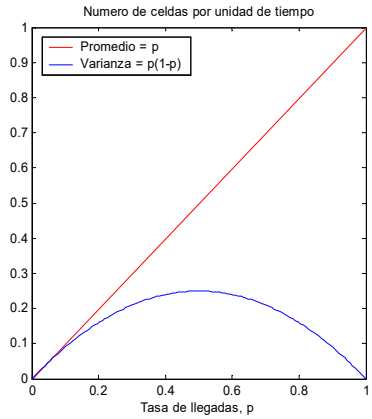


Figura 2.13 Distribución, promedio y varianza del número de celdas que llegan por unidad de tiempo bajo un modelo de Tráfico de Bernoulli

(b) La variable aleatoria geométrica surge de repetir un experimento de Bernoulli hasta que se obtenga uno de los resultados deseados, asegurando que cada experimento sea independiente de los demás. De hecho, existen cuatro formas posibles de definir una distribución geométrica, dependiendo si el experimento se repite hasta obtener un cero o hasta obtener un uno y, en cada caso, si el experimento exitoso se cuenta o no:

- Repetir hasta obtener 1 y no contar el 1: $\mathbf{P}[X=k]=p(1-p)^k$, k=0,1,2,…
- Repetir hasta obtener 1 y contar el 1: $\mathbf{P}[X=k]=p(1-p)^{k-1}$, k=1,2,3,…
- Repetir hasta obtener 0 y no contar el 0: $\mathbf{P}[X=k]=p^k(1-p)$, k=0,1,2,…
- Repetir hasta obtener 0 y contar el 0: $\mathbf{P}[X=k]=p^{k-1}(1-p)$, k=1,2,3,…

En la definición 43(b) se escogió el cuarto caso, que puede corresponder al siguiente ejemplo: Al transmitir un paquete por un enlace no-confiable se produce un error detectable con probabilidad p. El paquete se retransmite cuantas veces sea necesario hasta que llegue sin errores detectables al otro extremo del enlace, y se cuenta el número de transmisiones que se requieren, X. La probabilidad de tener que hacer una sola transmisión es la misma probabilidad de que el paquete llegue sin errores, $\mathbf{P}[X=1] = \mathbf{P}$[Primera transmisión exitosa] = $1-p$. Será necesario hacer dos transmisiones si hay un error en la primera transmisión y la segunda resulta exitosa, lo cual ocurre con probabilidad $\mathbf{P}[X=2] = \mathbf{P}$[Primera transmisión con error]$\mathbf{P}$[Segunda transmisión exitosa | Primera con error]. Si la presencia de errores es independiente de una transmisión a otra, la anterior probabilidad condicional es igual a la correspondiente probabilidad incondicional, $\mathbf{P}[X=2] = \mathbf{P}$[Primera transmisión con error]$\mathbf{P}$[Segunda transmisión exitosa] = $p(1-p)$. En general, será necesario hacer k transmisiones si las primeras k-1 transmisiones sufren algún error y la k-ésima llega sin errores detectables. Dada la suposición de independencia, este evento sucede con probabilidad $\mathbf{P}[X=k] = p^{k-1}(1-p)$. El número promedio de transmisiones será

$$E[X]=\sum_{k=1}^{\infty} kp_k = (1-p)\sum_{k=0}^{\infty} kp^{k-1} = (1-p)\sum_{k=0}^{\infty}\frac{d}{dp}p^k$$
$$=(1-p)\frac{d}{dp}\sum_{k=0}^{\infty}p^k = (1-p)\frac{d}{dp}\frac{1}{1-p} = \frac{1}{1-p}$$

El segundo momento se puede calcular igualmente fácil

$$E[X^2]=\sum_{k=1}^{\infty} k^2 p_k = (1-p)p\sum_{k=0}^{\infty} k^2 p^{k-2} = (1-p)p\sum_{k=0}^{\infty}\left(\frac{d^2}{dp^2}p^k + kp^{k-2}\right) = (1-p)p\frac{d^2}{dp^2}\sum_{k=0}^{\infty}p^k + E[X]$$
$$=(1-p)p\frac{d^2}{dp^2}\frac{1}{1-p}+\frac{1}{1-p} = \frac{2p}{(1-p)^2}+\frac{1}{1-p} = \frac{1+p}{(1-p)^2}$$

de donde $V[X] = E[X^2] - E[X]^2 = p/(1-p)^2$.

En efecto, si la probabilidad de error es cero, el número promedio de transmisiones es uno y la varianza es cero, pues con probabilidad uno sólo se necesita una transmisión. A medida que aumenta p, tanto el promedio como la varianza aumentan, aunque la varianza aumenta más

rápidamente. Si p es uno, se requerirá un número infinito de transmisiones o, lo que es lo mismo, el paquete nunca llegará bien si el enlace es un lazo de cabuya.

Un segundo modelo de tráfico basado en la *va* geométrica puede construirse a partir del modelo anterior (basado en la *va* de Bernoulli) si contamos el número de unidades de tiempo (o *slots*) que debemos esperar hasta ver la llegada de la siguiente celda. Si en el primer slot que observamos llegó una celda, lo cual ocurre con probabilidad p, debimos esperar cero unidades. Para esperar una unidad de tiempo será necesario que en el primer slot no venga ninguna celda y en el segundo venga una celda, lo cual ocurre con probabilidad $p(1-p)$ si cada unidad de tiempo es independiente de las demás. En general, si en los primeros k slots no llegaron celdas y la primera celda llegó en el slot $k+1$, debimos esperar k unidades, lo cual, bajo la suposición de independencia, ocurre con probabilidad $\mathbf{P}[X=k] = p(1-p)^k$, $k=0,1,2\ldots$ Nótese que ésta es otra de las cuatro maneras de definir una distribución geométrica. Haciendo $Y = X+1$ y $q=1-p$, notamos que $\mathbf{P}[Y=k] = q^{k-1}(1-q)$, $k=1,2,3\ldots$, como en el ejemplo anterior, de manera que $E[Y] = 1/(1-q)$ y, por consiguiente, usando la definición 41(a), $E[X] = 1/p - 1 = (1-p)/p$. Similarmente, por la definición 41(c), como $V[Y] = q/(1-q)^2$, entonces $V[X] = (1-p)/p^2$. La Figura 50 muestra estas cantidades.

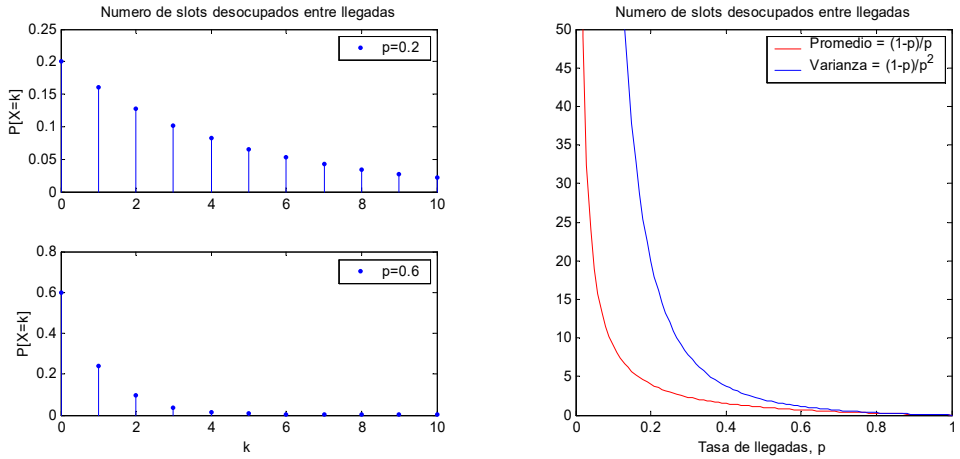


Figura 50. Distribución, promedio y varianza del número de slots desocupados entre llegadas bajo un modelo de tráfico geométrico

Nótese que en estos modelos geométricos es absolutamente necesario que los distintos experimentos de Bernoulli se realicen de manera independiente entre ellos. En el modelo de errores de transmisión, ¿será posible que la presencia de errores en la transmisión de una trama sea independiente de las tramas anteriores o siguientes? Si la transmisión se hace por un par trenzado no blindado de baja categoría y los errores se deben a la ignición de un motor eléctrico, los errores NO son independientes. Pero si la transmisión se hace a través de un satélite en horas nocturnas y los errores se deben al ruido galáctico, los errores SI pueden ser independientes: El modelo exige independencia y el analista deberá determinar si el modelo es aplicable o no. Igualmente, en el modelo de tráfico, si las celdas vienen de un gran número de fuentes independientes en las que cada una participa con una fracción muy pequeña del tráfico de manera que ninguna de ellas pueda generar celdas en unidades de tiempo cercanas entre sí, la suposición de independencia puede tener sentido. Pero si es un número pequeño de fuentes, cada una de las

cuales puede generar ráfagas de celdas en breves instantes de tiempo, será necesario revisar cuidadosamente la validez de la suposición de independencia.

(c) La variable aleatoria binomial surge de hacer n repeticiones independientes de un experimento de Bernoulli y contar cuántas veces sucedió el resultado 1. Supongamos, por ejemplo, que se transmite una trama de n bits sobre un enlace no confiable y se cuenta el número de bits que llegan con error cuando los errores se dan en cada bit independientemente con probabilidad p. La probabilidad de que no haya ningún error es, claramente, $\mathbf{P}[X=0] = (1-p)^n$. La probabilidad de que solamente se dañe el i-ésimo bit es $p(1-p)^{n-1}$, de manera que la probabilidad de que se dañe un solo bit es

$$\mathbf{P}[X=1] = P\left[\bigcup_{i=1}^{n}\{\text{error sólo en el bit } i\}\right] = \sum_{i=1}^{n} p(1-p)^{n-1} = np(1-p)^{n-1}$$

donde la segunda igualdad obedece al tercer axioma de las probabilidades. De la misma manera, la probabilidad de que sólo se dañen los bits i y j es $p^2(1-p)^{n-2}$, por lo que la probabilidad de que se dañen exactamente dos bits es

$$\mathbf{P}[X=2] = P\left[\bigcup_{i=1}^{n-1}\bigcup_{j=i+1}^{n}\{\text{error sólo en los bits } (i,j)\}\right] = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} p^2(1-p)^{n-2} = \frac{n(n-1)}{2}p^2(1-p)^{n-2}$$

En general, una combinación particular de sólo k errores se da con probabilidad $p^k(1-p)^{n-k}$, que es la misma probabilidad de que se dañen los k primeros bits o los k últimos, o los k de la mitad, o k de ellos tomados de dos en dos, etc. Como hay $\binom{n}{k} = n!/(k!(n-k)!)$ formas posibles de combinar k bits con errores entre n bits trasmitidos, $\mathbf{P}[X=k] = \binom{n}{k}p^k(1-p)^{n-k}$.

El número promedio de bits recibidos con error es

$$E[X] = \sum_{k=0}^{n} kp_k = \sum_{k=1}^{n} k\binom{n}{k}p^k(1-p)^{n-k} = \sum_{k=1}^{n} n\binom{n-1}{k-1}p^k(1-p)^{n-k} = np\sum_{k=1}^{n}\binom{n-1}{k-1}p^{k-1}(1-p)^{(n-1)-(k-1)} = np$$

De manera semejante podemos calcular el segundo momento,

$$E[X^2] = \sum_{k=1}^{n} k^2\binom{n}{k}p^k(1-p)^{n-k} = np\sum_{k=1}^{n} k\binom{n-1}{k-1}p^{k-1}(1-p)^{(n-1)-(k-1)} = np\sum_{j=0}^{m}(j+1)\binom{m}{j}p^j(1-p)^{m-j}\Big|_{\substack{j=k-1\\ m=n-1}} = np((n-1)p+1)$$

de donde la varianza del número de bits equivocados es
$V[X] = [(np)^2 + np(1-p)] - (np)^2 = np(1-p)$

Nótese que la variable aleatoria binomial es la suma de n variables aleatorias de Bernoulli independientes. Como veremos en el próximo capítulo, eso justifica el hecho de que la media y la varianza de la distribución binomial sean n veces la media y la varianza de la distribución de Bernoulli, respectivamente.

Siguiendo con la serie de modelos de tráfico, podemos considerar una trama TDM (Time Division Multiplexing) de n slots, donde cada slot se comporta según los modelos de tráfico descritos en los modelos Bernoulli y geométrico. Entonces la variable aleatoria X=Número de celdas en una trama, está binomialmente distribuida con parámetros (n,p). Igualmente, si consideramos un multiplexor que concentra n enlaces ATM como los descritos antes, donde el tráfico en cada

enlace es independiente de los demás enlaces, el número de celdas que llegan por unidad de tiempo tiene una distribución binomial con los mismos parámetros. La Figura 51 muestra la distribución, el promedio y la varianza del número de celdas que llegan en una trama de 32 slots.

Como en el caso de la variable aleatoria geométrica, cada vez que se quiera aplicar el modelo de la variable aleatoria binomial debemos justificar la suposición de independencia de los experimentos de Bernoulli subyacentes.

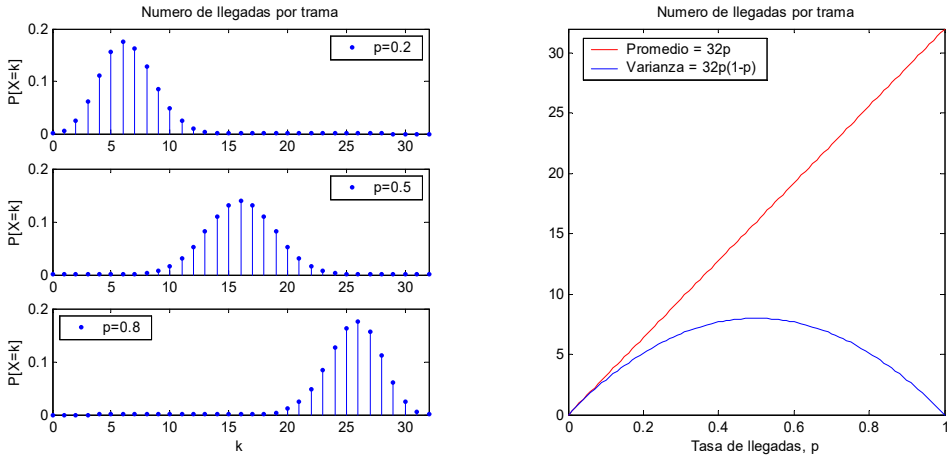


Figura 51. Distribución, promedio y varianza del número de celdas que llegan en una trama de 32 slots bajo un modelo de tráfico binomial

(d) Considérese un multiplexor que concentra un gran número de usuarios, de manera que los paquetes de datos pueden llegar en cualquier instante (modelo de tiempo continuo). Definamos la *va X* como el número de llegadas que hay en un período de t segundos. Para caracterizar la *va X*, dividimos el intervalo de t segundos en n subintervalos contiguos y no sobrelapados de longitud Δt, donde $t = n\Delta t$, y hacemos dos suposiciones básicas:

- a medida que la longitud del subintervalo Δt se hace más y más pequeña, la probabilidad de más de una llegada en un subintervalo tiende a cero y la probabilidad de una sola llegada en un subintervalo se hace proporcional a la longitud del intervalo, con factor de proporcionalidad λ:

$$\mathbf{P}\left[k \text{ llegadas en } \Delta t\right] = \begin{cases} \lambda\Delta t + o(\Delta t) & k = 1 \\ 1 - \lambda\Delta t + o(\Delta t) & k = 0 \\ o(\Delta t) & k > 1 \end{cases}$$

donde $o(\Delta t)$ -ómicron de Δt - es cualquier función que tienda a cero más rápidamente que Δt:

$$\lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0$$

de manera que $o(\Delta t) \pm o(\Delta t) = o(\Delta t)$, $o(\Delta t) \cdot o(\Delta t) = o(\Delta t)$, $\Delta t \cdot o(\Delta t) = o(\Delta t)$, etc. La distribución anterior indica que las llegadas simultáneas son improbables y que en cada subintervalo infinitesimal tenemos un experimento de Bernoulli en el que puede haber una llegada con probabilidad $\lambda\Delta t$ o ninguna llegada con probabilidad $1 - \lambda\Delta t$.

- El número de llegadas en un intervalo de tiempo dado es independiente del número de llegadas en cualquier otro intervalo de tiempo no sobrelapado con el primero. En particular, el número de llegadas en cualquiera de los subintervalos de longitud Δt en que dividimos el tiempo es independiente del número de llegadas en cualquier otro intervalo anterior o posterior.

Para que X tome el valor k puede ocurrir que en k de los n subintervalos haya habido una sola llegada y en los restantes n-k subintervalos no hayan habido llegadas, o que las k llegadas hayan sucedido en menos de k subintervalos. En este último caso, hubo más de una llegada en por lo menos un subintervalo, lo cual sucede con alguna probabilidad que tiende a cero más rápidamente que Δt, $o(\Delta t)$:

$$\mathbf{P}[X=k] = o(\Delta t) + \mathbf{P}\left[\bigcup_{I\subset\{1,2,\ldots,n\},|I|=k} \left\{\begin{array}{c}\text{una llegada en c/u de los subintervalos de } I,\\ \text{cero llegadas en los restantes } n-k \text{ subintervalos}\end{array}\right\}\right]$$

Dada la suposición de independencia en intervalos no sobrelapados, así se trate de intervalos infinitesimales, la anterior expresión toma la siguiente forma:

$$\mathbf{P}[X=k] = o(\Delta t) + \binom{n}{k}\left(\lambda\Delta t + o(\Delta t)\right)^k \left(1-\lambda\Delta t + o(\Delta t)\right)^{n-k}$$

Expandiendo las potencias en el segundo término de la derecha y agrupando todas las funciones $o(\Delta t)$ en una sola,

$$\mathbf{P}[X=k] = o(\Delta t) + \binom{n}{k}(\lambda\Delta t)^k (1-\lambda\Delta t)^{n-k} = o\left(\frac{t}{n}\right) + \frac{n!}{k!(n-k)!}\left(\frac{\lambda t}{n}\right)^k \left(1-\frac{\lambda t}{n}\right)^{n-k}$$

$$= o\left(\frac{t}{n}\right) + \frac{n!}{n^k(n-k)!}\frac{(\lambda t)^k}{k!}\frac{1}{\left(1-\lambda t/n\right)^k}\left(1-\frac{\lambda t}{n}\right)^n = o\left(\frac{t}{n}\right) + \left(\frac{n}{n}\cdot\frac{n-1}{n}\cdot\frac{n-2}{n}\cdots\frac{n-k+1}{n}\right)\frac{(\lambda t)^k}{k!}\frac{1}{\left(1-\lambda t/n\right)^k}\left(1-\frac{\lambda t}{n}\right)^n$$

Tomando el límite cuando n tiende a infinito (y Δt tiende a cero de manera que $t=n\Delta t$ siga constante), obtenemos $o(t/n)\to 0$, $(n\text{-}i)/n\to 1$, $(1-\lambda t/n)^k\to 1$ y $(1-\lambda t/n)^n\to e^{-\lambda t}$, de manera que

$$\mathbf{P}[X=k] = \frac{(\lambda t)^k}{k!}e^{-\lambda t}$$

Esto es, bajo las suposiciones anteriores, el número de llegadas en t segundos tiene una distribución Poisson con parámetro $\rho = \lambda t$. Durante cerca de un siglo éste ha sido el modelo de tráfico por excelencia en el diseño y análisis de redes de comunicaciones, aunque en las últimas dos décadas se ha acumulado una gran cantidad de evidencia que muestra que, en redes modernas de comunicaciones conmutadas por paquetes, la suposición de independencia en intervalos no sobrelapados no es válida cuando se habla de la llegada de paquetes (aunque aún puede serlo cuando se habla del establecimiento de flujos o sesiones). Más aún, en muchos casos hay evidencia empírica que muestra cierta dependencia aún entre intervalos muy separados en el tiempo, fenómeno conocido como "dependencia de largo rango", LRD –long range dependence–. Sin embargo, como veremos más adelante, la simplicidad del modelo Poisson (que supone independencia aún a nivel infinitesimal) permite obtener expresiones cerradas para muchas medidas de desempeño, gracias a lo cual sigue siendo utilizado como una primera aproximación en el dimensionamiento de la capacidad de las redes de comunicaciones y en el diseño de algoritmos de control para las mismas.

El valor esperado del número de paquetes que llegan en t segundos es

$$E[X]=\sum_{k=0}^{\infty}kp_k=\sum_{k=1}^{\infty}k\frac{\rho^k}{k!}e^{-\rho}=\rho\sum_{k=1}^{\infty}\frac{\rho^{k-1}}{(k-1)!}e^{-\rho}=\rho$$

que, con ρ definido como λt, indica que λ es el número promedio de llegadas por segundo o la tasa promedio de llegadas, que es uno de los parámetros más importantes en la caracterización de tráfico (el único parámetro en el caso de modelos de Poisson). El segundo momento es

$$E[X^2]=\sum_{k=0}^{\infty}k^2\frac{\rho^k}{k!}e^{-\rho}=\sum_{k=1}^{\infty}k\big((k-1)+1\big)\frac{\rho^k}{k!}e^{-\rho}=\rho\sum_{k=1}^{\infty}\big((k-1)+1\big)\frac{\rho^{k-1}}{(k-1)!}e^{-\rho}$$
$$=\rho\sum_{k=1}^{\infty}(k-1)\frac{\rho^{k-1}}{(k-1)!}e^{-\rho}+\rho\sum_{k=1}^{\infty}\frac{\rho^{k-1}}{(k-1)!}e^{-\rho}=\rho E[X]+\rho=\rho^2+\rho$$

de donde $V[X] = E[X^2]-E[X]^2 = \rho$. La varianza de una variable aleatorio de Poisson es igual a su valor medio.

La Figura 52 muestra la distribución del número de llegadas en un segundo cuando el tráfico obedece a un modelo de Poisson. La gráfica de la media y la varianza respecto a ρ son sólo dos líneas a 45º.

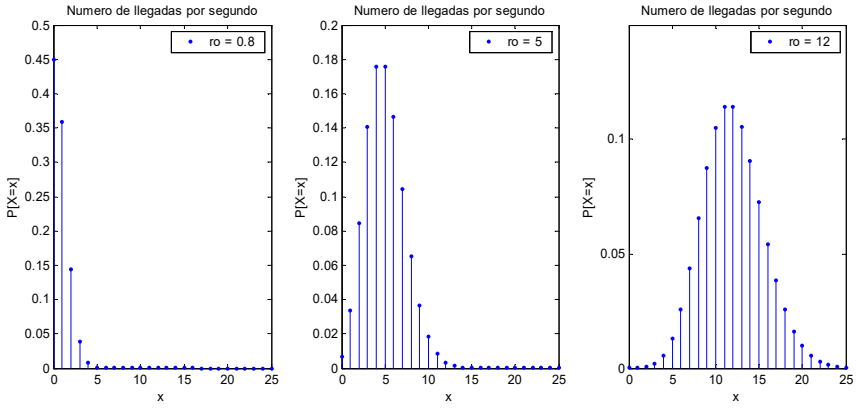


Figura 52. Distribución del número de llegadas en un segundo bajo un modelo de tráfico Poisson

(e) Una trama TDM tiene n slots, numerados de 0 a n-1. Los paquetes llegan al multiplexor TDM en instantes completamente aleatorios, independientemente de la sincronización de la trama. Sea la *va* X el slot que se está transmitiendo de la trama que se está transmitiendo en el instante en que llega un paquete. Como no hay ninguna razón que permita imaginar que un paquete tenga alguna preferencia por un slot o un grupo de slots particular[24], parece razonable suponer que $\mathbf{P}[X=k] = 1/n$, k=0,1,…,n-1.

[24] Consideramos cada paquete independientemente de los demás. Dado un proceso de llegadas particular, podría haber alguna preferencia si condicionamos en el slot que le correspondió al paquete anterior, por ejemplo.

Su valor promedio es $E[X]=\sum_{k=0}^{n-1} k\frac{1}{n}=\frac{1}{n}\frac{n(n-1)}{2}=\frac{n-1}{2}$ y su segundo momento es $E[X^2]=\frac{1}{n}\sum_{k=0}^{n-1} k^2=\frac{(2n-1)(n-1)}{6}$, de manera que su varianza es $V[X] = E[X^2]-E[X]^2 = (n^2-1)/12$.

Nótese la naturaleza de los ejemplos anteriores en los que cada *va* se convierte en un modelo probabilístico adecuado. Los experimentos de Bernoulli consisten en observar uno de dos posibles resultados, a cada uno de los cuales se les asocia el valor 0 ó 1 (error o no-error en un bit, ocupación o desocupación de un enlace, falla u operatividad de un dispositivo, presencia o ausencia de un paquete, etc.). Las variables binomial, geométrica y de Poisson modelan repeticiones independientes de un experimento de Bernoulli. En el modelo geométrico, se repite el experimento independientemente hasta obtener alguno de los dos resultados. En el modelo binomial el experimento se repite independientemente n veces y se cuenta el número de ocasiones en que ocurrió el resultado favorable. El modelo de Poisson es el límite consistente en un número infinito de repeticiones independientes durante un período finito de tiempo. El modelo uniforme obedece al principio de la máxima incertidumbre: Si tenemos un conjunto de proposiciones excluyentes a las cuales queremos asignar una distribución de probabilidad, debemos tener en cuenta qué sabemos de ellas. Si conocemos cuál es la cierta, le debemos asignar un valor de probabilidad igual a uno y las demás proposiciones tendrán probabilidad cero, pues no tenemos ninguna incertidumbre. Si algún conocimiento previo nos permite favorecer algunas proposiciones más que otras, podremos asignarles mayor probabilidad. Pero si no tenemos ninguna información que nos permita favorecer a ninguna de las proposiciones sobre las demás, nuestra incertidumbre será máxima y lo más conveniente será asignar las probabilidades uniformemente. Si lo hiciésemos de otra manera, estaríamos suponiendo una información que no poseemos.

44. Algunas Variables Aleatorias Continuas

(a) *Una variable aleatoria X uniformemente distribuida tiene parámetros reales* (a,b), *toma valores en el intervalo* $[a,b]$, *y su pdf es* $f_X(x) = 1/(b-a)$, $x\in[a,b]$. *Su valor esperado es* $(a+b)/2$ *y su varianza es* $(b-a)^2/12$.

(b) *Una variable aleatoria X exponencialmente distribuida tiene un parámetro real positivo,* $\lambda>0$, *toma valores entre los reales no negativos, y su pdf es* $f_X(x)= \lambda e^{-\lambda x}$, $x\geq 0$. *Su valor esperado es* $1/\lambda$ *y su varianza es* $1/\lambda^2$.

(c) *Una variable aleatoria X Normalmente* (*o Gaussianamente*) *distribuida tiene parámetros* (μ,σ^2), *donde* μ *es un número real y* σ^2 *es un número real no negativo, toma valores en los reales, y su pdf es*

$$f_X(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right),\ x\in\mathbb{R}$$

Su valor esperado es μ *y su varianza es* σ^2.

(d) *Una variable aleatoria X con distribución de Pareto tiene parámetros positivos (a,b), toma valores en los reales mayores o iguales a b, y su pdf es*

$$f_X(x) = \frac{a}{b}\left(\frac{b}{x}\right)^{a+1}, \; x \geq b$$

Si a>1, su valor esperado es ab/(a-1); si no, su valor esperado es infinito. Si a>2, su varianza es $ab^2/((a-2)(a-1)^2)$; si no, su varianza es infinita.

(e) *Una variable aleatoria X con distribución de Cauchy tiene parámetros reales (a,b), b>0, toma valores reales, y su pdf es*

$$f_X(x) = \frac{1}{\pi}\frac{b}{(x-a)^2 + b^2}, \; x \in \mathbb{R}$$

Ni la media ni la varianza de la distribución de Cauchy están definidas.

(f) *Una variable aleatoria X con distribución de Laplace tiene un parámetro real positivo a, toma valores reales, y su pdf es*

$$f_X(x) = \frac{a}{2}e^{-a|x|}, \; x \in \mathbb{R}$$

Su valor esperado es cero y su varianza es $2a^{-2}$.

(g) *Una variable aleatoria X con distribución de Erlang tiene parámetros (n,λ), donde n es un entero positivo y λ es un real positivo. Toma valores reales no negativos y su pdf es*

$$f_X(x) = \frac{\lambda(\lambda x)^{n-1}e^{-\lambda x}}{(n-1)!}, \; x \geq 0$$

Su valor esperado es n/λ y su varianza es n/λ^2.

(h) *Una variable aleatoria X con distribución Gamma tiene parámetros reales positivos (a,λ), toma valores reales no negativos y su pdf es*

$$f_X(x) = \frac{\lambda(\lambda x)^{a-1}e^{-\lambda x}}{\Gamma(a)}, x \geq 0 \quad donde \quad \Gamma(a) = \int_0^\infty s^{a-1}e^{-s}ds \;\; (función\ Gamma)$$

Su valor esperado es a/λ y su varianza es a/λ^2.

(i) *Una variable aleatoria X con distribución de Weibull tiene dos parámetros (a,λ), ambos reales positivos, toma valores reales no negativos y su pdf es*

$$f_X(x) = a\lambda^a x^{a-1}\exp(-(\lambda x)^a), \; x \geq 0$$

Su valor esperado es $\Gamma((a+1)/a)/\lambda$ y su varianza es $((\Gamma((a+2)/a) - \Gamma((a+1)/a))^2/\lambda^2$.

(j) *Una variable aleatoria X con distribución Chi-cuadrado (χ^2) tiene un parámetro real positivo, a, toma valores reales no negativos y su pdf es*

$$f_X(x) = \frac{x^{a/2-1}\exp(-x/2)}{2^{a/2}\Gamma(a/2)}, \; x \geq 0$$

Su valor esperado es a y su varianza es 2a.

(k) *Una variable aleatoria X con distribución de t de Student tiene un parámetro real positivo a, toma valores reales y su pdf es*

$$f_X(x)=\frac{\Gamma\left(\frac{a+1}{2}\right)}{\sqrt{a\pi}\,\Gamma\left(\frac{a}{2}\right)}\left(1+\frac{x^2}{a}\right)^{-\frac{a+1}{2}},\ x\in\mathbb{R}$$

Su valor esperado es 0 y su varianza es $a/(a-2)$ *para* $a>2$.

(l) *Una variable aleatoria X con distribución de Rayleigh tiene un parámetro real positivo, a, toma valores no negativos y su pdf es*

$$f_X(x)=\frac{x}{a^2}\exp\left(-\frac{1}{2}\left(\frac{x}{a}\right)^2\right),\ x\geq 0$$

Su valor esperado es $a(\pi/2)^{1/2}$ *y su varianza es* $(2-\pi/2)a^2$.

(m) *Una variable aleatoria X con distribución Rice tiene dos parámetros reales positivos,* (ν,σ)*, toma valores no negativos y su pdf es*

$$f_X(x)=\frac{x}{\sigma^2}\exp\left(-\frac{x^2+\nu^2}{2\sigma^2}\right)I_0\left(\frac{x\nu}{\sigma^2}\right),\quad x\geq 0$$

donde $I_0(z)$ *es la función modificada de Bessel del primer tipo y de orden cero. Su valor esperado es* $\sigma\sqrt{\pi/2}L_{1/2}\left(-\nu/2\sigma^2\right)$ *y su varianza es* $2\sigma^2+\nu^2-(\pi\sigma^2/2)L_{1/2}^2\left(-\nu/2\sigma^2\right)$*, donde* $L_{1/2}(x)$ *es el polinomio de Laguerre,* $L_{1/2}(x)=\exp\left(x/2\right)\left[(1-x)I_0(-x/2)-xI_1(-x/2)\right]$.

Las anteriores distribuciones son la base de algunos de los modelos más ampliamente usados y, por tal motivo, es importante que el lector aprenda a usar estos modelos en los contextos adecuados en los que se pueden utilizar. A continuación damos algunos ejemplos de los ocho primeros modelos probabilísticos en redes de comunicaciones y demostramos los resultados obtenidos respecto a la media y la varianza.

(a) A un multiplexor estadístico llegan paquetes de longitud fija en instantes aleatorios e independientes de tiempo. En el instante de su llegada, el paquete *b* encuentra el enlace de salida ocupado transmitiendo el paquete *a*, y una larga cola de paquetes delante de él esperando ser transmitidos. Se mide el tiempo que transcurre desde la llegada de *b* hasta que *a* termina de ser transmitido, *X*, o "tiempo residual de servicio de *a*". Como *a* y *b* no son paquetes consecutivos (hubo un gran número de llegadas entre ellos) y como lo único que conocemos respecto al proceso de llegadas es que los tiempos entre llegadas son aleatorios e independientes, parece razonable suponer que *b* no tiene ninguna preferencia por llegar hacia el comienzo, el final o la mitad del tiempo de servicio de *a*. Y, como el rango de posibles valores de *X* es el intervalo $[0,T]$, donde *T* es el tiempo de transmisión de un paquete, el principio de máxima incertidumbre sugiere escoger la distribución uniforme para *X*, $f_X(x) = 1/T$ para $0 \leq x \leq X$. El valor medio de esta distribución es $\frac{1}{T}\int_0^T t\,dt=\frac{t^2}{2T}\Big|_0^T=\frac{T}{2}$ y el segundo momento es $\frac{1}{T}\int_0^T t^2\,dt=\frac{t^3}{3T}\Big|_0^T=\frac{T^2}{3}$, de

manera que su varianza es $(T^2/3) - (T^2/4) = T^2/12$. La Figura 53 muestra la *pdf*, la media y la varianza del tiempo residual de servicio de un paquete.

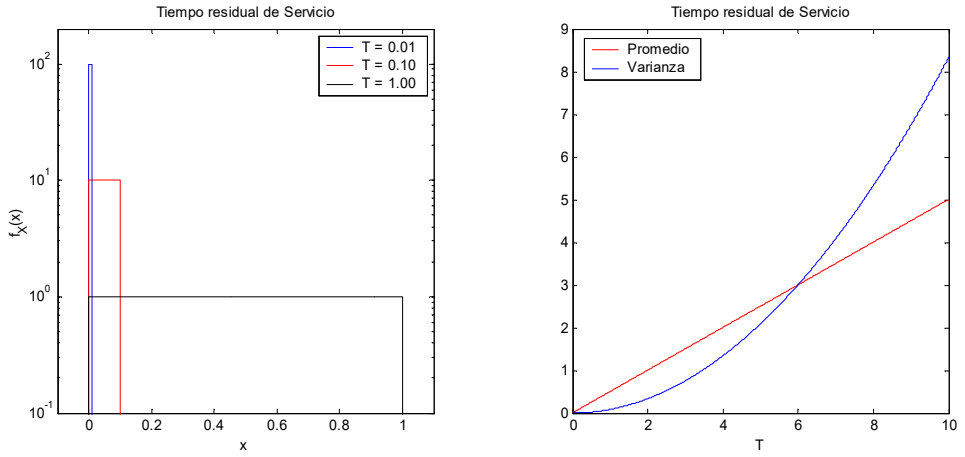


Figura 53. Función de densidad de probabilidad, promedio y varianza del tiempo residual de servicio según un modelo uniforme

(b) A un multiplexor llegan paquetes según un proceso de Poisson como el descrito en la definición 43(d), es decir, la probabilidad de que hayan k llegadas en t segundos es $\mathbf{P}[X = k] = \frac{(\lambda t)^k}{k!} e^{-\lambda t}$. Sea T la variable aleatoria "tiempo que toca esperar hasta ver la próxima llegada". Consideremos el evento $T>t$, que corresponde al caso en el que, desde que empezamos a ver, han transcurrido t segundos sin que haya llegado aún ningún paquete. La probabilidad de dicho evento es la misma probabilidad de que en t segundos haya habido cero llegadas que, de acuerdo con la suposición de llegadas tipo Poisson, corresponde a $\mathbf{P}[T>t] = e^{-\lambda t}$. La probabilidad del evento complementario es $F_T(t) = \mathbf{P}[T \le t] = 1 - e^{-\lambda t}$. La derivada de esta *CDF* es $f_T(t) = \lambda e^{-\lambda t}$, que es la *pdf* de una variable aleatoria exponencial. Su valor esperado es $E[T] = \lambda \int_0^\infty t e^{-\lambda t} dt = \lambda \left[-\frac{1+\lambda t}{\lambda^2} e^{-\lambda t} \right]_0^\infty = \frac{1}{\lambda}$ y su 2do momento es $E[T^2] = \lambda \int_0^\infty t^2 e^{-\lambda t} dt = \left[-\left(t^2 + (2t/\lambda) + (2/\lambda^2) \right) e^{-\lambda t} \right]_0^\infty = \frac{2}{\lambda^2}$, de manera que su varianza es $1/\lambda^2$. Estos resultados refuerzan la idea de que el parámetro λ es la tasa promedio de llegada de paquetes.

Obsérvese que esta variable aleatoria es el modelo probabilístico de los tiempos entre llegadas cuando el tráfico obedece a un proceso de Poisson que, como dijimos, es el modelo de tráfico preferencialmente utilizado en redes de comunicaciones. Por consiguiente, la variable aleatoria exponencial es uno de los modelos probabilísticos más usados en redes de comunicaciones. La razón de su amplio uso es fácil de ver en la misma derivación que acabamos de hacer: nótese que medimos el tiempo que tardamos en ver la llegada del próximo paquete ¡sin tener en cuanto hace cuánto tiempo llegó el paquete anterior! La variable aleatoria exponencial es la única variable

continua que no tiene memoria: la distribución del tiempo que falta para ver la llegada del siguiente paquete es la misma independientemente del tiempo que ha transcurrido desde la llegada del paquete anterior. Esto es, la distribución de T sigue siendo $f_T(t)=\lambda e^{-\lambda t}$ así hayamos empezado a medir desde que llegó el paquete anterior o desde media hora después de que llegó el paquete anterior (dado que en esa media hora no ha llegado ningún paquete, por supuesto). Esta falta de memoria, que demostraremos formalmente en la definición 51, facilita enormemente el análisis de redes de comunicación, como veremos en el capítulo de teoría de colas.

La Figura 54 muestra la *pdf*, la media y la varianza del tiempo entre llegada de paquetes cuando el número de llegadas en cierto período de tiempo se modela mediante una distribución de Poisson.

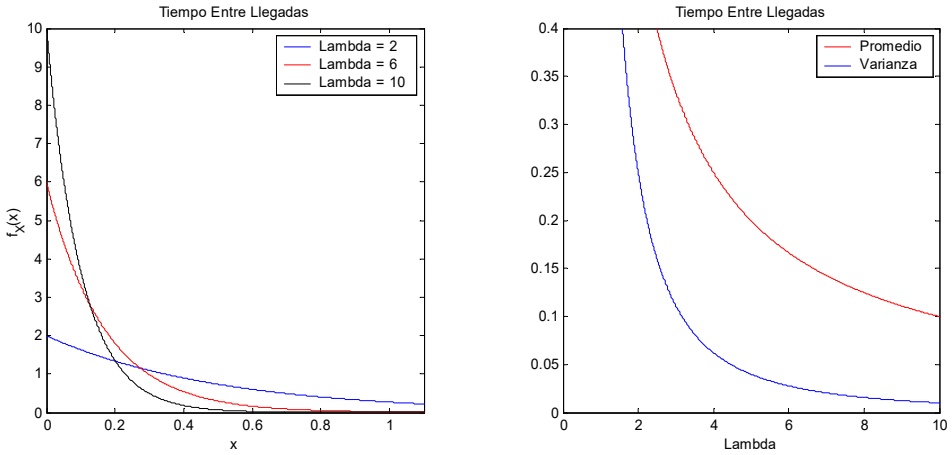


Figura 54. Función de densidad de probabilidad, promedio y varianza del tiempo entre llegada de paquetes según un modelo exponencial

(c) Cuando una resistencia metálica de R ohmios se encuentra a una temperatura de T kelvins, sus electrones se mueven de manera aleatoria generando un voltaje de *ruido térmico* con nivel DC (media) cero y potencia (varianza) $2R(\pi kT)^2/3h$ W, donde k es la constante de Boltzmann y h es la constante de Planck (sugiriendo la presencia de fenómenos termodinámicos cuánticos). Supongamos, de una manera muy simplificada, que el movimiento de cada electrón en una resistencia R de 6.37 megohmios a 290 kelvins produce una caída de $+\Delta$ voltios con probabilidad 0.5 y $-\Delta$ voltios con probabilidad 0.5 y que cada electrón se mueve independientemente de los demás. Si existen n electrones libres en la resistencia, el voltaje producido será $V = (2X - n)\Delta$, donde X es una variable aleatoria binomial con parámetros (n, ½), correspondiente al número de electrones que producen $+\Delta$ voltios. Aplicando los resultados de la definición 41 y de la definición 43(c), el valor medio del ruido térmico es cero y la varianza es $n\Delta^2$. Si hacemos que n crezca y Δ disminuya de manera que $n\Delta^2 = 2R(\pi kT)^2/3h = 1$ V^2, como predice la física, la probabilidad de obtener un voltaje de $(2k\text{-}n)\Delta$ voltios, con $0 \le k \le n$, es $\binom{n}{k}2^{-n}$. Dividiendo esta probabilidad por $2\Delta = 2/\sqrt{n}$, que es el mínimo cambio en el voltaje, obtenemos la siguiente densidad de probabilidad:

$$\frac{P\left[(2k-n)/\sqrt{n} \le V < (2(k+1)-n)/\sqrt{n}\right]}{2/\sqrt{n}} = \sqrt{n}\binom{n}{k}2^{-(n+1)}$$

la cual se muestra en la Figura 55 (barras) y que se compara con la expresión $f_V(v) = \frac{1}{\sqrt{2\pi}}\exp\left(-v^2/2\right)$ (línea continua), conocida como "Campana de Gauss", que es la *pdf* de una variable Gaussiana con media 0 y varianza 1.

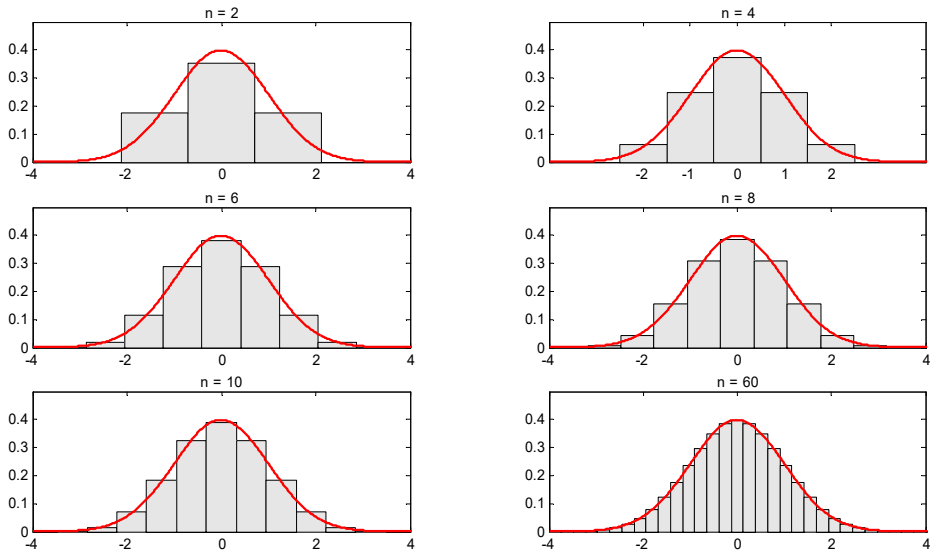


Figura 55. Función de densidad de probabilidad del ruido térmico producido por *n* partículas, donde cada partícula genera +1/√*n* voltios con probabilidad ½ o -1/√*n* voltios con probabilidad ½. Se compara con la función de densidad de probabilidad Gaussiana

Claramente, a medida que consideramos más y más electrones, la *pdf* del ruido térmico se hace más cercana a la distribución Gaussiana. Por supuesto, lo más razonable es considerar un número infinito de electrones, cada uno participando con un infinitésimo del voltaje de ruido, de manera que el modelo Gaussiano resulta apenas natural para modelar el ruido térmico en una resistencia metálica, tal como la impedancia de entrada del amplificador de radiofrecuencia en un sistema de comunicaciones.

Como en el ejemplo anterior, si X representa la suma de N componentes aleatorios independientes en la que cada componente contribuye con una pequeña fracción de la suma, la *pdf* de X se aproxima a la distribución Gaussiana a medida que N aumenta, ¡independientemente de la distribución de los componentes individuales! Este es el teorema del límite central propuesto por Laplace en 1810, que estudiaremos con cuidado en la definición 82. De hecho, dado el determinismo que imperaba en esa época, la aleatoriedad sólo se usaba para modelar los errores experimentales de medición que, en términos de observaciones astronómicas, Gauss asoció con su famosa "campana" pues, evidentemente, se trataba de la suma de muchos errores debidos a la dispersión y la difracción de la luz con cada partícula de

la atmósfera. Debido a la naturaleza de estos modelos de error, la distribución Gaussiana resultó la más normal de las distribuciones y, por esa razón, también se le conoce como distribución normal, $\mathcal{N}(\mu,\sigma^2)$. Hoy se sabe que las distribuciones más "normales" en la naturaleza son las que tienen colas pesadas (definición 48), como la distribución de Pareto que se describirá a continuación de ésta distribución Gaussiana.

El valor medio de una variable $X\sim\mathcal{N}(\mu,\sigma^2)$ (que se lee "normalmente distribuida con parámetros μ y σ^2") es

$$E[X]=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}x\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)dx=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}(\sigma y+\mu)\exp(-y^2/2)dy\Big|_{y=\frac{x-\mu}{\sigma}}$$

$$=\frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{\infty}y\,\mathrm{e}^{-y^2/2}\,dy+\mu\left(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\mathrm{e}^{-y^2/2}\,dy\right)=\mu$$

Pues, en la última expresión, la primera integral es cero por tratarse de un función con simetría impar y la expresión entre paréntesis del segundo término es la probabilidad total de una variable $\mathcal{N}(0,1)$. Para hallar la varianza de X partamos de la probabilidad total:

$$\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)dx=1$$

Multipliquemos ambos lados por $\sigma\sqrt{(2\pi)}$ y derivemos respecto a σ:

$$\int_{-\infty}^{\infty}\frac{(x-\mu)^2}{\sigma^3}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)dx=\sqrt{2\pi}$$

Y, finalmente, multipliquemos a ambos lados por $\sigma^2/\sqrt{(2\pi)}$ para obtener

$$V[X]=E\left[(X-\mu)^2\right]=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}(x-\mu)^2\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)dx=\sigma^2$$

De donde los parámetros de una variable normal son su media y su varianza. La figura 2.20 muestra algunas *pdf*s Gaussianas, donde se nota el efecto del valor esperado μ como un parámetro de posición y el efecto de la varianza σ^2 como un parámetro de forma que describe la concentración de la distribución alrededor de su valor esperado.

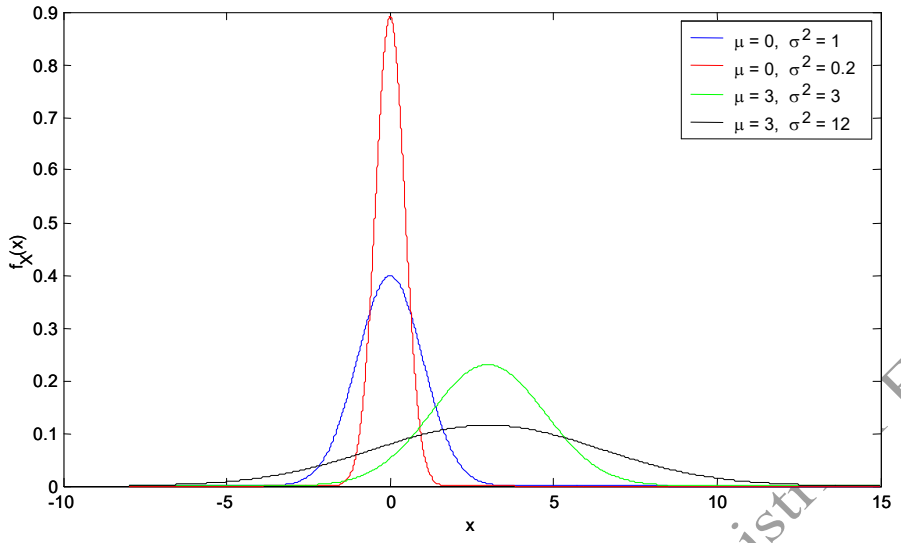


Figura 56. Funciones de densidad de probabilidad Gaussianas

(d) Las distribuciones que hemos visto hasta ahora se caracterizan porque la probabilidad de que las variables aleatorias tomen valores muy grandes es muy pequeña, de manera que el efecto total de dichos valores es despreciable. Sin embargo, en redes modernas de comunicaciones (como en todos los sistemas que recientemente se han caracterizado como "complejos"), se hacen cada vez más comunes algunas variables aleatorias que pueden tomar valores muy grandes con probabilidad no despreciable, de manera que, cuando finalmente se presentan estos valores, su efecto puede ser determinante. Este es el caso del tamaño de los archivos que se intercambian por la red (la gran mayoría son muy pequeños pero la pequeña fracción de archivos grandes son los que consumen la mayoría de recursos en la red), la duración de una conexión http (la gran mayoría de conexiones son breves, pero las pocas conexiones duraderas son las que más ocupan a los servidores web), etc. De estas cantidades se dice que tienen "cola pesada" (ver definiciones 0, 48 y 51), y una de las distribuciones más utilizadas para modelarlas probabilísticamente es la distribución de Pareto, la cual se usó originalmente para describir la concentración de riquezas (la gran mayoría de personas son pobres, pero las pocas personas ricas que existen poseen la gran mayoría de la riqueza del mundo)[25]. En efecto, la Figura 57 compara una distribución exponencial con parámetro $\lambda = 1/3$ y una distribución de Pareto con parámetros a=1.5 y b=1, de manera que ambas tienen el mismo valor promedio μ=3, aunque la segunda tiene varianza infinita. Un cálculo muy simple muestra que la probabilidad de que la variable de Pareto sea superior a n veces su valor esperado es $[(a\text{-}1)/(na)]^a = (3n)^{-3/2}$, mientras que la probabilidad de que la variable exponencial sea superior a n veces su valor esperado es $\exp(-n)$. Esto es, aunque la probabilidad de que la variable exponencial supere la media es casi el doble de que la variable Pareto también lo haga, la probabilidad de que la variable exponencial supera 9 veces la media es menos de una millonésima de la probabilidad de que la variable Pareto también lo haga!

[25] Este fenómeno de cola pesada (o "ley-de-potencia" en la cola de la distribución) ha resultado tan ubicuo, que muchos científicos empiezan a considerar una explicación general basada en la auto-organización en puntos críticos al borde del caos o basada en la tolerancia altamente optimizada,

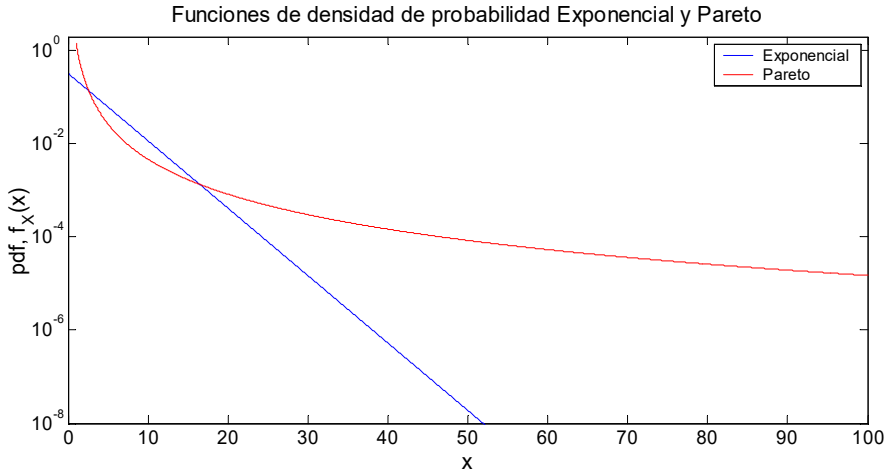


Figura 57. Funciones de densidad de probabilidad Exponencial y de Pareto

El valor medio de una variable $X\sim Pareto(a,b)$ es

$$E[X]=\int_b^\infty x\frac{a}{b}\left(\frac{b}{x}\right)^{a+1}dx=ab^a\int_b^\infty x^{-a}dx=-\frac{ab^a}{a-1}(x^{-a+1})\Big|_b^\infty=\begin{cases}\dfrac{ab}{a-1} & a>1\\ \infty & a\le 1\end{cases}$$

Y su segundo momento es

$$E[X^2]=\int_b^\infty x^2\frac{a}{b}\left(\frac{b}{x}\right)^{a+1}dx=-\frac{ab^a}{a-2}(x^{-a+2})\Big|_b^\infty=\begin{cases}\dfrac{ab^2}{a-2} & a>2\\ \infty & a\le 2\end{cases}$$

De manera que la varianza es

$$V[X]=E[X^2]-E^2[X]=\begin{cases}\dfrac{ab^2}{(a-2)(a-1)^2} & a>2\\ \infty & a\le 2\end{cases}$$

Obsérvese que, en el rango $1<a\le 2$, una v.a. de Pareto tiene media finita y varianza infinita, de donde surgen las características de ley de potencia que hacen tan interesante esta distribución para representar los fenómenos de complejidad observados recientemente en redes de comunicaciones.

(e) La distribución de Cauchy (o de Lorentz, como se le conoce en física) resuelve la ecuación diferencial que describe algunos sistemas de resonancia forzada, tales como el ensanchamiento de las líneas espectroscópicas debido a fenómenos de resonancia. En redes de telecomunicaciones el interés en la distribución de Cauchy es de tipo estadístico porque, al ser semejante a la distribución normal cerca al máximo de la distribución (el *modo*), como muestra la Figura 58, la robustez de las pruebas de hipótesis que asumen normalidad se puede probar con datos tomados de una distribución Cauchy. Además, la razón X/Y de dos v.a. gaussianas independientes X y Y tiene una distribución Cauchy.

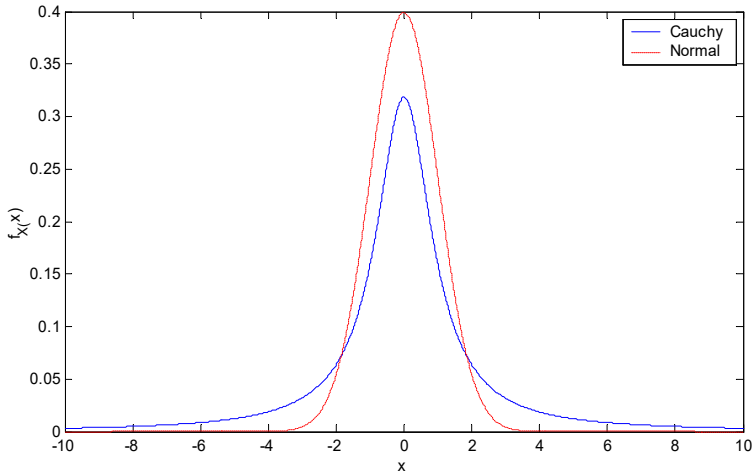


Figura 58. La distribución de Cauchy es la versión con cola pesada de la distribución Gaussiana

(f) Considere dos paquetes de longitud exponencialmente distribuida con promedio L bits, independientes entre ellos, que empiezan a transmitirse simultáneamente en dos canales de la misma capacidad, C. Sea T_1 el tiempo de transmisión del primer paquete y T_2 el tiempo de transmisión del segundo paquete. La diferencia entre los tiempos de transmisión, $T = T_1 - T_2$, tiene una distribución Laplaciana:

$$f_T(t) = \int_{\max(0,-t)}^{\infty} f_{T_1}(t+s) f_{T_2}(s) ds = \lambda^2 e^{-\lambda t} \int_{\max(0,-t)}^{\infty} e^{-2\lambda s} ds = \frac{\lambda}{2} e^{-\lambda t} e^{-2\lambda \max(0,-t)} = \frac{\lambda}{2} e^{-\lambda |t|}$$

donde $1/\lambda = L/C$ es el tiempo promedio de transmisión de un paquete. En efecto, más adelante veremos cómo la independencia de los eventos asociados con cada variable hace que la *pdf* conjunta $f_{T1,T2}(t_1,t_2)$ sea el producto de las *pdf* marginales $f_{T1}(t_1)f_{T2}(t_2)$, de manera que la expresión anterior es, sencillamente, la evaluación de la probabilidad total (definición 19). La distribución de Laplace se muestra en la Figura 59.

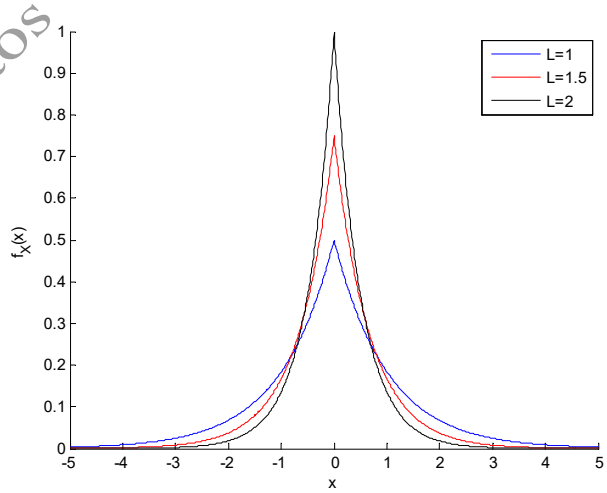


Figura 59. La distribución de Laplace es una versión simétrica de la distribución exponencial

El valor medio de una variable Laplaciana es

$$E[T]=\frac{\lambda}{2}\int_{-\infty}^{\infty}te^{-\lambda|t|}dt=\frac{\lambda}{2}\left(\int_{0}^{\infty}te^{-\lambda t}dt+\int_{-\infty}^{0}te^{-\lambda|t|}dt\right)=0$$

y su segundo momento es

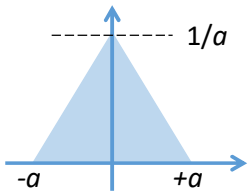
$$V[T]=E[T^2]=\frac{\lambda}{2}\int_{-\infty}^{\infty}t^2e^{-\lambda|t|}dt=\lambda\int_{0}^{\infty}t^2e^{-\lambda t}dt=\frac{2}{\lambda^2}$$

(g & h) Cómo vimos, el tiempo entre llegadas consecutivas de paquetes que obedecen a un proceso de Poisson es una variable exponencial. Cabe preguntarse por el tiempo que tomará la llegada de n paquetes, que corresponderá a la suma de n variables aleatorias exponenciales independientes e idénticamente distribuidas. La suma obedecerá a la distribución de una variable aleatoria de Erlang(n,λ), donde λ es el parámetro del proceso Poisson subyacente. Si el número de términos que se suman se puede interpolar a valores no enteros, se obtiene la distribución Gamma.

45. La distribución Gaussiana

En 1904, Henri Poincaré intentó explicar con humor el uso generalizado de la distribución Gaussiana: "Los físicos creen que la distribución Gaussiana fue demostrada por las matemáticas y los matemáticos creen que la distribución Gaussiana fue descubierta experimentalmente por los físicos". Lo cierto es que diferentes formas de esta distribución han sido propuestas desde el siglo XVIII, por lo que ha habido cerca de 300 años para estudiarla, aprovecharla y hasta para abusar de ella. De hecho, en 1733, DeMoivre la descubrió como una fórmula rápida para calcular la probabilidad del número de caras y sellos en una gran cantidad de monedas (Figura 55), que era un empleo típico de los matemáticos en las cortes europeas. Sin embargo, pasaron más de 70 años antes de que la distribución Gaussiana cobrara toda su importancia. Entonces se suponía que el uso de la aleatoriedad era exclusivamente para considerar errores de medición, como la dispersión de los rayos de luz provenientes de una estrella cuando atraviesan las microturbulencias de la atmósfera terrestre antes de llegar al lente de un telescopio. En 1755 Thomas Simpson propuso una función de densidad de probabilidad triangular para el error de medición, que tiene su base en el intervalo [-a, a] y su altura máxima, 1/a, en cero. Con esta distribución, el error se podía reducir tomando muchas muestras y promediándolas. En 1770 Laplace generalizó esta propuesta postulando que la distribución del error debe ser simétrica alrededor de cero y debe tender a cero a medida que el valor absoluto del error tiende a infinito, algo que ya Galileo había intuido 150 años antes (que los errores de medición eran simétricos y que los errores más pequeños eran más frecuentes). En 1780 Laplace mismo introdujo la distribución de Laplace como función de densidad del error para capturar sus postulados de manera concreta. La Figura 60 muestra las propuestas de Simpson y Laplace para las distribuciones de los errores de medición.

$$f_E(x)=\frac{1}{a}máx\left[0,mín\left(1+x/a,1-x/a\right)\right]$$



$$f_E(x)=\frac{\lambda}{2}e^{-\lambda|x|}$$

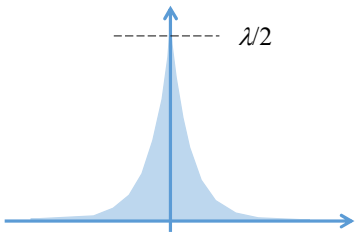


Figura 60. Funciones de densidad de probabilidad para el error de Simpson y Laplace

En 1805 Legendre publicó el método de los mínimos cuadrados para regresión lineal. Dado un conjunto de medidas $\{(x_i,y_i),i=1,2,\cdots,n\}$ que se quieren ajustar a una línea recta $\hat{y}_i=ax_i+b$, se considera el error cuadrado promedio de las medidas, $MSE=\frac{1}{n}\sum_{i=1}^{n}(y_i-(ax_i+b))^2$, y se encuentran los valores de *a* y *b* que minimizan el *MSE*, $(a^*,b^*)=\arg\min_{(a,b)}\frac{1}{n}\sum_{i=1}^{n}(y_i-ax_i-b)^2$. Al derivar el MSE con respecto a *a* y a *b* e igualar a cero se obtiene el mínimo (pues el MSE es una parábola como función de *a* y de *b*), $\frac{\partial}{\partial a}MSE=\frac{\partial}{\partial b}MSE=0$, lo que conduce a la solución

$$\begin{bmatrix}a^*\\b^*\end{bmatrix}=\begin{bmatrix}\frac{1}{n}\sum_{i=1}^{n}x_i & 1\\ \frac{1}{n}\sum_{i=1}^{n}x_i^2 & \frac{1}{n}\sum_{i=1}^{n}x_i\end{bmatrix}^{-1}\begin{bmatrix}\frac{1}{n}\sum_{i=1}^{n}y_i\\ \frac{1}{n}\sum_{i=1}^{n}x_iy_i\end{bmatrix}$$

Con los resultados de Legendre y Laplace, y la hermosa función de DeMoivre, Gauss realizó el siguiente análisis:

Supongamos que $\{X_i,\ i=1\ldots n\}$ son *n* medidas erróneas de una cantidad *Y*, de manera que a cada medida corresponde un error $E_i=Y-X_i$. Supongamos también que todos los errores son independientes e idénticamente distribuidos con *pdf* $\phi(x)$, que es una función que satisface los postulados de Laplace. Por independencia, la pdf de la secuencia de errores, $E=\{E_i,\ i=1\ldots n\}$ es $f(E)=\prod_{i=1}^{n}\phi(E_i)$ (ver definiciones 65 y 75). Se quiere que el valor promedio propuesto por Simpson, $\overline{x}=\frac{1}{n}\sum_{i=1}^{n}X_i$, sea el punto de las mediciones donde se produzca el error más probable, esto es, $\frac{\partial}{\partial\overline{x}}f(E)=0$ o, lo que es lo mismo, $\frac{\partial}{\partial\overline{x}}\log\left[f(E)\right]=0$. Expandiendo *f*(*E*), la condición necesaria para maximizar la probabilidad del error con el promedio de las mediciones es

$$\frac{1}{\phi(E_1)}\frac{\partial}{\partial x}\phi(E_1)+\frac{1}{\phi(E_2)}\frac{\partial}{\partial x}\phi(E_2)+\cdots+\frac{1}{\phi(E_n)}\frac{\partial}{\partial x}\phi(E_n)=0$$

Mediante la regla de la cadena,

$$\left(\frac{\partial}{\partial x}\phi(E_i)=\frac{\partial}{\partial E_i}\phi(E_i)\frac{\partial}{\partial x}E_i=\phi'(E_i)\frac{\partial}{\partial x}E_i=-\frac{1}{n}\phi'(E_i)\right)$$

Obtenemos

$$\frac{\phi'(E_1)}{\phi(E_1)}+\frac{\phi'(E_2)}{\phi(E_2)}+\cdots+\frac{\phi'(E_n)}{\phi(E_n)}=0$$

Como las medidas X_2, X_3, X_4,… aparecen únicamente en la suma $\sum_{i=2}^{n}X_i$, Gauss consideró la suma como una constante, de manera que los errores son $E_1=(n-1)d$, $Ei=-d$, $i=2,...,n$ y, por consiguiente,

$$\frac{\phi'\big((n\text{-}1)d\big)}{\phi\big((n\text{-}1)d\big)}=(1\text{-}n)\frac{\phi'\big(\text{-}d\big)}{\phi\big(\text{-}d\big)}$$

de donde Gauss dedujo que para todo $x\in\mathbb{R}$ y algunas constantes a y c,

$$\frac{\phi'(x)}{\phi(x)}=cx \quad\Rightarrow\quad \phi(x)=ae^{cx^2/2}$$

Para satisfacer las condiciones de Laplace, es necesario que $c<0$, que en notación de hoy escribimos $c=-1/\sigma^2$. La constante a se obtiene por normalización:

$$\int_{-\infty}^{\infty}\phi(x)dx=1 \Rightarrow \quad \phi(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2\right)$$

que es la distribución Gaussiana o normal. Resumiendo, la distribución Gaussiana fue diseñada para que el promedio aritmético de las medidas sea el que minimice el error cuadrado promedio, suponiendo errores simétricos. Este modelo resulta muy apropiado para conisderar errores de medición con pequeñas varianzas (la ganancia de un jugador de dados, la altura física de los seres humanos, el ruido térmico en una resisitencia eléctrica, etc.). Sin embargo, son muchos los ejemplos que se pueden encontrar en los que este modelo diseñado no aplica: el número de personas afectadas por un apagón, el tamaño de los incendios forestales, las cascadas de los eventos de congestión en tráfico aéreo, los impactos de meteoritos, las muertes y pérdidas económicas por desastres naturales o artificiales, las variaciones en el mercado de valores, el uso de las palabras del español, la población de las ciudades, los ingresos y riqueza de las compañías y los individuos, la citación de artículos, las publicaciones por autor, los tamaños de los archivos en un disco duro, la utilización de la CPU en un computador, etc. Todos estos son ejemplos de variables aleatorias cuyas distribuciones exhiben colas pesadas y, por lo tanto, la alta variabilidad no corresponde con las condiciones para las que se diseñó la distribución gaussiana. Nunca podremos subestimar la importancia de esta distribución, como veremos al estudiar, por ejemplo, el teorema del limite central (definición *xx*). Sin embargo, es importante tener en cuenta que es bastante fácil (y bastante común) abusar de ella.

46. Algunos Ejemplos muy Simplificados de Modelos Probabilísticos de Eficiencia en Redes de Comunicaciones Basados en Variables Aleatorias

(a) *Con tráfico tipo Poisson, longitud fija de paquetes y un gran número de usuarios, la eficiencia del protocolo Aloha en la utilización efectiva del enlace es* $\rho e^{-2\rho}$, *donde* ρ *es la intensidad de tráfico. Esta eficiencia tiene un valor máximo de* 0.184 *cuando* ρ=0.5.

(b) *Bajo las mismas condiciones, la eficiencia del protocolo Aloha ranurado es* $\rho e^{-\rho}$, *que tiene un valor máximo de* 0.368 *cuando* ρ=1.

(c) *Si en el protocolo Aloha ranurado se considera un número finito de usuarios, n, la máxima eficiencia que se puede conseguir es* $[(n-1)/n]^{n-1}$, *cuando la intensidad de tráfico es* 1. *Esta eficiencia tiende a* 0.368 *a medida que n tiende a infinito.*

(d) *La máxima eficiencia del protocolo de retransmisión Stop&Wait es* $(L(1-BER)^{L+2h})/(L + h + (h+2Ct_p))$, *donde L es la longitud (constante) de los paquetes en bits, h es el número de bits en el encabezado que se les añade, y el canal se caracteriza por la tasa de errores, BER, el retardo de propagación,* t_p, *y la velocidad de transmisión, C.*

(e) *Bajo las mismas condiciones, la máxima eficiencia del protocolo de retransmisión GoBack-N es* $(L(1-BER)^{L+2h})/(L + h + p(h+2Ct_p))$, *donde* p=$1-(1-BER)^{L+2h}$ *es la probabilidad de que se dañe al menos un bit de una trama o de su reconocimiento. Y la máxima eficiencia del protocolo de retransmisión Selective-Repeat es* $(L(1-BER)^{L+2h})/(L + h)$.

Como hemos mencionado, el modelado probabilístico no es un formalismo matemático que se pueda aprender como aprendimos, por ejemplo, los axiomas que definen la probabilidad. Al contrario, es un tipo de arte que sólo se puede llegar a conocer después de haber estudiado muchos ejemplos y haber acumulado mucha práctica, a veces con éxito y a veces no. En los ejemplos sencillos que vienen a continuación notaremos algunos de los aspectos fundamentales de este modelado. Por ejemplo, es necesario conocer muy bien el sistema que se quiere modelar y establecer con claridad y sin ambigüedades el problema que se quiere resolver respecto a dicho sistema. Luego se deben identificar los elementos que participan en el sistema, incluyendo las variables que describen tanto a los elementos como a las interacciones entre ellos. En esta etapa se incluyen todas las restricciones y las suposiciones que se deban hacer para que el modelo sea analíticamente tratable sin alejarlo demasiado de la realidad. Entonces es cuando se aplican los aspectos particulares de las teorías matemáticas apropiadas. Los siguientes pretenden ser ejemplos muy sencillos de este proceso, en los que se desea estimar la eficiencia en el uso de un canal de comunicaciones, ya sea compartido (ejemplos (a), (b), y (c)) o dedicado (ejemplos (d) y (e)).

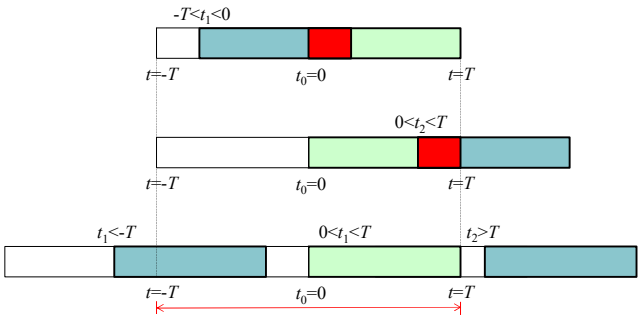
(a) Muchos medios de comunicación no conectan dos puntos de manera exclusiva, sino que deben compartirse entre diferentes usuarios, por lo que se conocen como medios de acceso múltiple (redes satelitales, algunas redes de área local, redes de radio, etc.). Para controlar el acceso a este tipo de medios existe la capa MAC (Medium Access Control), para la cual se han desarrollado diferentes tipos de protocolos que van desde algunos perfectamente organizados como *round-robin TDM* (en el que cada nodo tiene una ranura de tiempo exclusiva para sus transmisiones) hasta otros completamente aleatorios como *Aloha* (en el que cada nodo envía sus paquetes en el momento en que se generan, con la esperanza de que no "colisionen" con las transmisiones de otros paquetes). Este último, el más simple de todos los protocolos de acceso múltiple, se adoptó en la Universidad de Hawaii en la década de los 70´s y es la base de la mayoría de protocolos de acceso aleatorio más utilizados en diferentes tipos de redes. El protocolo se basa en que los usuarios, que transmiten un paquete cada vez que desean, puedan detectar si la interferencia generada por otros nodos afectó la recepción de su paquete, en cuyo caso esperan un tiempo aleatorio antes de reintentar la retransmisión del mismo paquete.

Supongamos las siguientes condiciones: El enlace tiene una capacidad de C bps y lo comparte un gran número de usuarios, cada uno participando con una fracción muy pequeña del tráfico total, de manera que la suma de las transmisiones nuevas y las retransmisiones forma un proceso de Poisson con un promedio de λ paquetes por segundo. Todos los paquetes tienen la misma longitud, L, y el mismo tiempo de transmisión, $T = L/C$. Nos preguntamos por la eficiencia con que se puede explotar el enlace, donde definimos la eficiencia como la tasa efectiva de paquetes por segundo que puede transmitir el canal, normalizada por la máxima tasa que se podría lograr si no hubieran colisiones y el enlace permaneciera ocupado transmitiendo paquetes,

$$Eficiencia = \frac{Tasa\ efectiva}{Tasa\ ideal} = \frac{\lambda P[éxito]}{C/L}$$

En efecto, si el medio puede permanecer el 100% del tiempo transmitiendo paquetes, lograría transmitir C/L paquetes por segundo. Pero, en realidad, de los λ paquetes que en total producen los usuarios, sólo transmite efectivamente aquellos que no sufren colisiones, esto es, aquellos que tienen éxito, $\lambda P[éxito]$. Como el tiempo de transmisión es $T=L/C$, la forma que toma la expresión anterior para la eficiencia es $E = \rho P[éxito]$, donde $\rho=\lambda T$ es la intensidad de tráfico. Así pues, sólo falta determinar la probabilidad de que un paquete no sufra colisiones, $P[éxito]$.

Para que un paquete tenga éxito es necesario que nadie haya empezado a transmitir durante los T segundos anteriores al inicio de su transmisión (con eso él no colisiona con nadie que ya haya estado transmitiendo) ni durante los T segundos que dura su transmisión (con eso nadie colisiona con él durante su transmisión). Esto es, para que un paquete tenga éxito, hay un intervalo de $2T$ segundos durante los cuales no debe iniciarse la transmisión de ningún otro paquete: $P[éxito] = P[0$ paquetes inicien transmisión en $2T$ segundos] (verFigura 61).



a) Para que un paquete tenga éxito es necesario que nadie haya empezado a transmitir durante los T segundos anteriores al inicio de su transmisión (con eso él no colisiona con nadie que ya haya estado transmitiendo)

b) Para que un paquete tenga éxito es necesario que nadie haya empezado a transmitir durante los T segundos que dura su transmisión (con eso nadie colisiona con él durante su transmisión).

c) Esto es, para que un paquete tenga éxito, hay un intervalo de $2T$ segundos durante los cuales no debe iniciarse la transmisión de ningún otro paquete

Figura 61. Para que un paquete que inicia transmisión en t_0=0 tenga éxito es necesario que nadie más intente empezar a transmitir en el intervalo [-T, T]

De acuerdo con la suposición de tráfico Poisson,

$$P[\acute{e}xito] = [(\lambda\, 2T)^0/0!]e^{-\lambda\, 2T} = e^{-2\rho}$$

En consecuencia, la eficiencia del protocolo Aloha bajo las condiciones mencionadas depende de la intensidad de tráfico, ρ, así:

$$\text{Eficiencia} = \rho\, e^{-2\rho}.$$

Esta eficiencia, que se grafica en la Figura 62, tiene un máximo en el punto en el que $\frac{d}{d\rho} Eficiencia = (1-2\rho)e^{-2\rho} = 0$, que corresponde a $\rho = 1/2$, donde la eficiencia toma el valor $e^{-1}/2 = 0.184$.

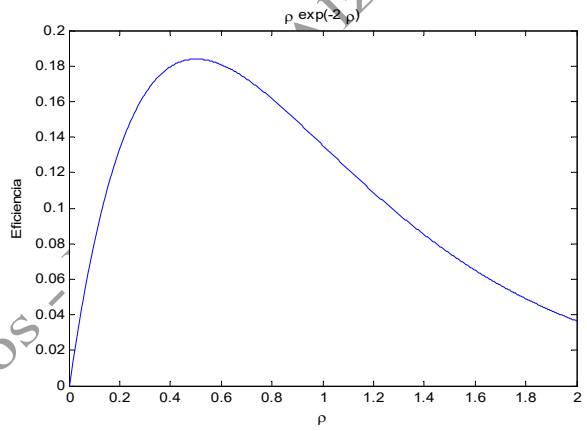


Figura 62. Eficiencia del protocolo Aloha como función de la intensidad de tráfico

(b) Supongamos que, en el ejemplo anterior, hacemos que una estación emita una señal cada T=L/C segundos indicando que quienes quieran transmitir pueden empezar a hacerlo. Cuando un usuario quiera transmitir un paquete, en vez de hacerlo inmediatamente, espera a la señal de temporización. Esto es, el canal se ha dividido en franjas de tiempo (slots) que supondremos exactamente iguales al tiempo de transmisión de un paquete, T. De esta manera, para que un paquete tenga éxito, es necesario que nadie haya decidido empezar a transmitir durante el período de T segundos anteriores al inicio de su transmisión: $P[\acute{e}xito] = [(\lambda T)^0/0!]e^{-\lambda T} = e^{-\rho}$. La eficiencia es, entonces, $E = \rho P[\acute{e}xito] = \rho e^{-\rho}$, que se maximiza cuando $\frac{d}{d\rho} Eficiencia = (1-\rho)e^{-\rho} = 0$, es decir, cuando ρ=1, en cuyo caso la eficiencia es e^{-1}=0.368, como muestra la Figura 63. Este es el protocolo *Aloha ranurado*.

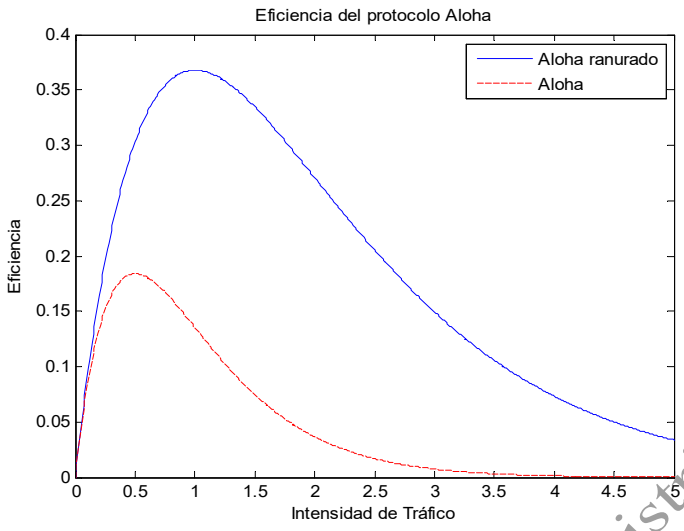


Figura 63. Eficiencia del protocolo Aloha ranurado

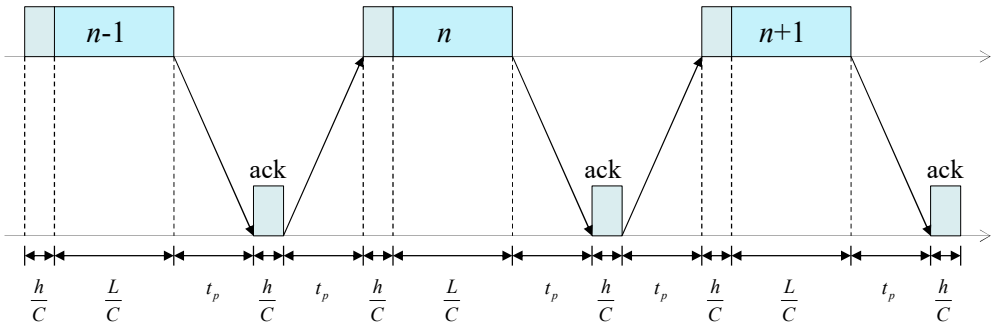
(c) En los dos ejemplos anteriores supusimos un número infinito de usuarios, cada uno participando con un infinitésimo del tráfico. ¿Qué pasa si consideramos un número finito, n, de usuarios? Por supuesto, debemos considerar otro modelo de tráfico. Por ejemplo, si el canal se divide en ranuras de tiempo, podemos suponer que cada usuario genera un paquete en cada slot con probabilidad p, independientemente entre usuarios e independientemente entre slots. En este caso, en cada slot pueden ocurrir tres casos diferentes: Que el slot se pierda porque nadie lo use, que el slot se aproveche efectivamente porque sólo un usuario decide usarlo, o que el slot se pierda porque más de un usuario lo quiere usar y se produzca una colisión. Sea N el número de usuarios que deciden transmitir en un slot. Como vimos, N tiene una distribución binomial con parámetros (n, p). La eficiencia es la fracción de tiempo que aprovechamos el canal para la transmisión efectiva de paquetes, esto es, la fracción de slots que se aprovechan exitosamente. En nuestra interpretación frecuentista, esta fracción de slots tiende exactamente a la probabilidad de que un slot se aproveche correctamente: $E = \mathrm{P}[N=1] = \binom{n}{1}p(1-p)^{n-1} = np(1-p)^{n-1}$. Esta es una función cóncava en p, así que para encontrar el valor de p que maximiza la eficiencia basta con igualar la derivada de E respecto a p a cero y despejar p:

$$\frac{d}{dp}E = n(1-p)^{n-1} - n(n-1)p(1-p)^{n-2} = n(1-p)^{n-2}\left(1-np\right) = 0$$

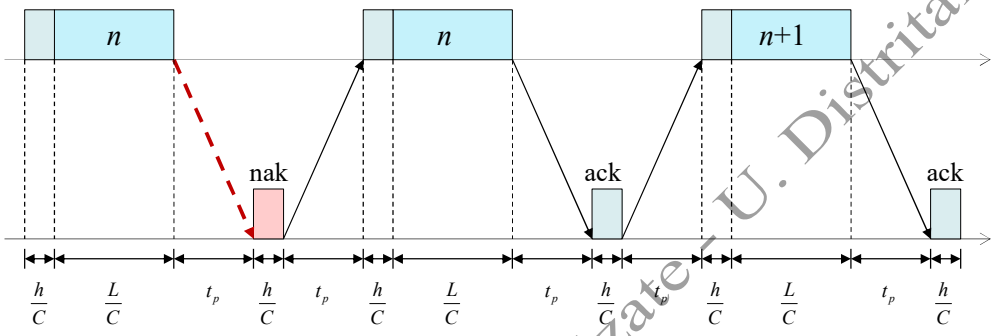
Que tiene dos soluciones, $p=1/n$ y $p=1$, de las cuales la de interés para nosotros es $p=1/n$. Remplazando este valor de p en la expresión de la eficiencia, obtenemos la eficiencia máxima, $E = [(n\text{-}1)/n]^{n-1}$.

Nótese que esta expresión se basa en suponer un modelo binomial de tráfico, mientras que la expresión anterior para Aloha ranurado surgió de suponer un modelo Poisson. Sin embargo, si en el modelo binomial hacemos tender n a infinito y p a cero de manera que np permanezca igual a la intensidad de tráfico deseada, obtenemos un modelo de Poisson. En efecto, si hacemos $n\to\infty$ con $np=1$, obtenemos $[(n\text{-}1)/n]^{n-1}\to e^{-1}=0.368$, como corresponde al modelo anterior de aloha ranurado.

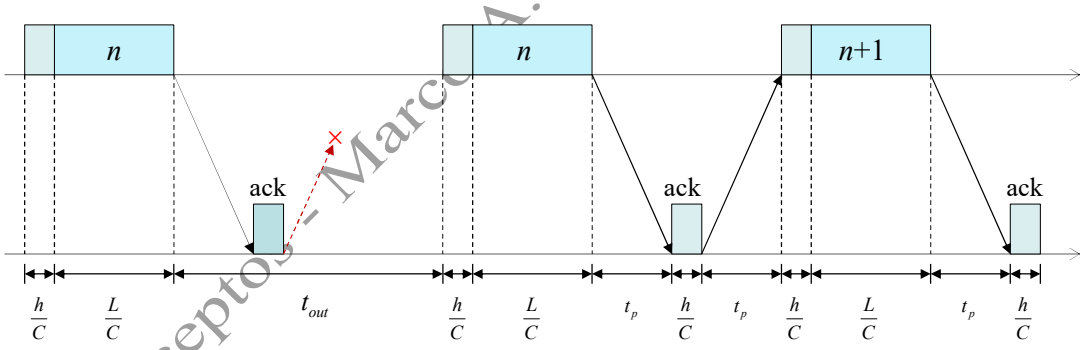
(d) Consideremos un enlace dedicado punto a punto caracterizado por una capacidad de transmisión, C bits/segundo, una tasa de errores de bit, BER, y un tiempo de propagación, t_p segundos. Sobre ese enlace enviamos paquetes de L bits desde el extremo transmisor hasta el extremo receptor. Supongamos que usamos un código detector de errores (CRC, por ejemplo) capaz de detectar todos los posibles errores (¡!). Dicho código, junto con el campo de numeración de secuencia de paquetes y otros campos de control, requieren un encabezado de h bits para cada paquete, con lo que se construye una trama de $L+h$ bits.
Una forma simple y efectiva de corregir los errores de transmisión es a través de los protocolos de solicitud automática de retransmisión (ARQ): el módulo receptor detecta las tramas con errores y solicita automáticamente la retransmisión del paquete correspondiente. El protocolo ARQ más simple de todos se denomina *Stop&Wait* y se basa en la idea de que no se debe transmitir ningún nuevo paquete hasta no estar completamente seguro que el paquete anterior llegó correctamente. Para esto, por cada trama recibida correctamente, el nodo receptor devuelve un reconocimiento positivo indicando que recibió bien el último paquete y espera el siguiente paquete. Este reconocimiento se convierte en un permiso para que el nodo transmisor transmita el siguiente paquete (Figura 64(a)). Si algún paquete llega con errores, el nodo receptor devuelve un reconocimiento negativo solicitando la retransmisión del último paquete (Figura 64(b)). Y, si después de un tiempo prudente, t_{out}, el transmisor no recibe ningún reconocimiento positivo ni negativo, supone que hubo un error y retransmite el último paquete enviado (Figura 64(c)). Los reconocimientos son tramas de control que no contienen ningún paquete y, por lo tanto, sólo tienen h bits de longitud.



(a) Sin errores



(b) Con errores



(c) Con pérdida de paquetes (o reconocimientos)

Figura 64. Secuencia de eventos en el protocolo *Stop&Wait*

Claramente, es necesario que t_{out} sea significativamente mayor que $2t_p + h/C$ para evitar retransmisiones innecesarias. Sin embargo, para encontrar la máxima eficiencia de este protocolo, supondremos que t_{out} es exactamente igual a $2t_p + h/C$. De esta manera, una transmisión, exitosa o no, toma un tiempo $t_1 = (L+2h)/C + 2t_p$. La probabilidad de que no haya errores es $(1\text{-}BER)^{L+2h}$, suponiendo que la presencia de errores en un bit es independiente de la presencia de errores en los bits vecinos (lo cual no puede ser cierto en la mayoría de casos, pero es una libertad que nos tomamos para mantener el modelo tratable. Entonces, la probabilidad de que una trama se dañe y sea necesario retransmitirla es p=1- $(1\text{-}BER)^{L+2h}$ y, como se vio en la

definición 43(b), se necesitan en promedio $1/(1-p)$ transmisiones para que una trama llegue correctamente. En consecuencia, el tiempo promedio que toma la transmisión correcta de una trama es $t_T = t_1/(1-p)$. Si pudiéramos transmitir bits de usuario a la máxima velocidad del canal, resulta que hubiéramos enviado Ct_T bits en el tiempo en el tiempo en el que, en realidad, solamente enviamos L bits. Por lo tanto, la eficiencia de este protocolo es $E = L/Ct_T$. Multiplicando arriba y abajo por $L+h$ y expandiendo los términos de t_T y p, encontramos la siguiente expresión para la eficiencia del protocolo *Stop&Wait*:

$$E_{S\&W} = \frac{L}{L+h}\cdot\left(1-BER\right)^{L+2h}\cdot\frac{L+h}{L+h+\left(h+2t_pC\right)}$$

Esta expresión está compuesta por tres términos: El primero, $L/(L+h)$, se refiere a la reducción en eficiencia debida al encabezado, para la cual no se puede hacer nada más que reducir al máximo la longitud del encabezado que se añade en el nivel de enlace a cada paquete del nivel de red. El segundo término, $(1-BER)^{L+2h}$, corresponde a la reducción en eficiencia debida a los errores de transmisión, para la cual no se puede hacer nada más que reducir el BER mediante la selección apropiada de las técnicas de modulación y codificación a nivel físico. El término que queda es la reducción en eficiencia debida propiamente al protocolo *Stop&Wait* y se caracteriza por un parámetro muy importante en el análisis de desempeño de los sistemas de comunicación: El "*bandwidth-delay product*", $2t_pC$, que indica cuántos bits podríamos transmitir durante el tiempo de propagación. Con el propósito de ver el efecto de este parámetro, consideremos 5 tipos de enlaces, así:

Tipo de enlace	Tiempo de propagación (Distancia, velocidad)	Capacidad	Tasa de errores
Línea telefónica	5 μs (1 km a $2x10^8$ m/s)	56 Kbps	10^{-5}
ADSL	5 μs (1 km a $2x10^8$ m/s)	512 Kbps	10^{-5}
LAN inalámbrica	1 μs (300 m a $3x10^8$ m/s)	10 Mbps	10^{-4}
Fibra óptica	5 ms (100 km a $2x10^8$ m/s)	1 Gbps	10^{-6}
Satélite	240 ms (72000 km a $3x10^8$ m/s)	100 Mbps	10^{-5}

Suponiendo un encabezado de 64 bits, podemos graficar la eficiencia de *Stop&Wait* para cada uno de esos cinco enlaces, en función de la longitud del paquete, como muestra la Figura 65

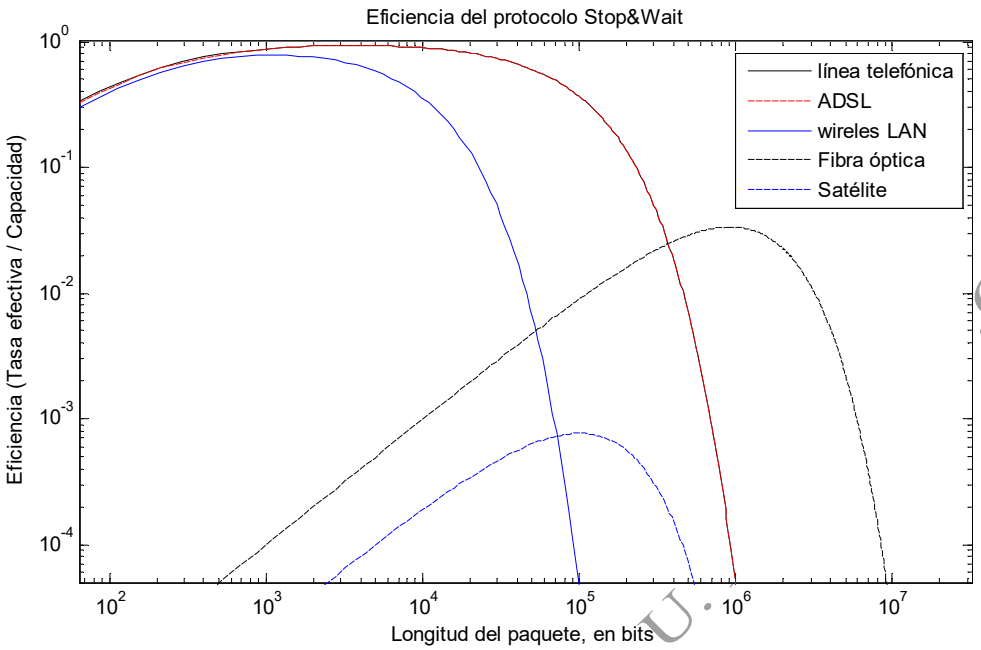
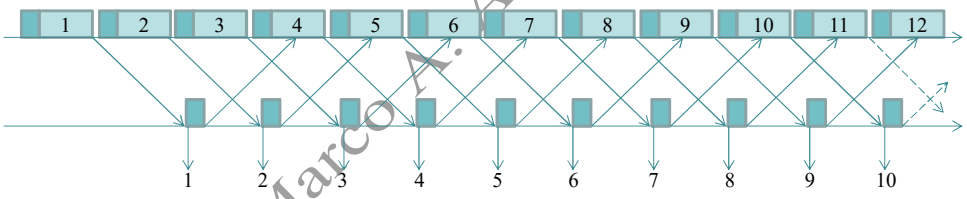


Figura 65. Eficiencia del protocolo *Stop&Wait* para los cinco enlaces de ejemplo

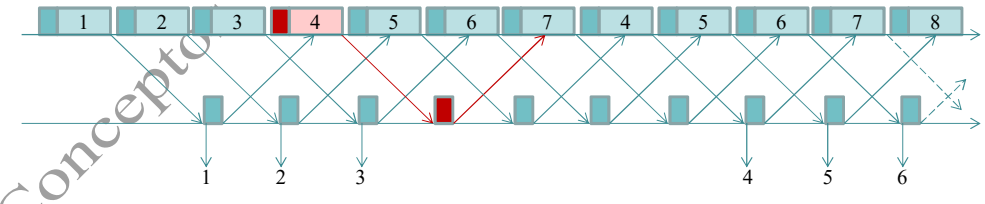
En cada caso hay una longitud óptima de paquete, por debajo de la cual dominan la ineficiencia debida al encabezado (si $2t_pC$ es pequeña comparada con el encabezado) o la ineficiencia debida al protocolo (si $2t_pC$ es grande comparada con el encabezado), y por encima de la cual domina la ineficiencia debida a los errores. Para paquetes entre 2500 y 5000 bits de longitud, la eficiencia del protocolo *Stop&Wait* puede ser hasta del 93% en la línea telefónica y la línea ADSL, en las que $2t_pC$ es de sólo algunos pocos bits (línea ADSL) o una fracción de bit (línea telefónica). En el enlace *wireless* LAN, la máxima eficiencia que se alcanza es cercana al 80% con paquetes cercanos a 1000 bits, donde $2t_pC$ es de algunas decenas de bits. Pero en el enlace de fibra óptica se alcanzaría a transmitir 10 millones de bits durante el tiempo de propagación, mientras que a través del satélite se alcanzaría a transmitir 48 millones de bits. Con paquetes de datos de semejante longitud, la probabilidad de error es casi uno. Por eso en el canal de fibra óptica apenas se alcanza una máxima eficiencia del 3% con paquetes de un millón de bits, y en el canal satelital la máxima eficiencia alcanzable es de menos del 0.1% con paquetes de cerca de cien mil bits. Claramente, para enlaces con un gran *bandwidth-delay product* se hace necesario diseñar protocolos más apropiados.

(e) Puesto que la ineficiencia del protocolo *stop&wait* (cuando el *bandwidth-delay product*, $2t_pC$, no es un número despreciable de bits) se debe a que durante los tiempos de propagación el canal se mantiene ocioso, una alternativa sería permitir que el transmisor siguiera transmitiendo nuevos paquetes mientras espera un reconocimiento. Sin embargo, ante la presencia de un error, el transmisor debe saber cuáles tramas debe reenviar, por lo cual debe mantener un buffer con los paquetes que ya ha transmitido pero que no han sido reconocidos por el receptor. Cada reconocimiento positivo libera un espacio en el buffer (porque ya existe seguridad de que el paquete correspondiente llegó correctamente) y cada paquete transmitido ocupa un espacio en el buffer (porque debemos tener disponible ese paquete por si llegara a hacer falta retransmitirlo).

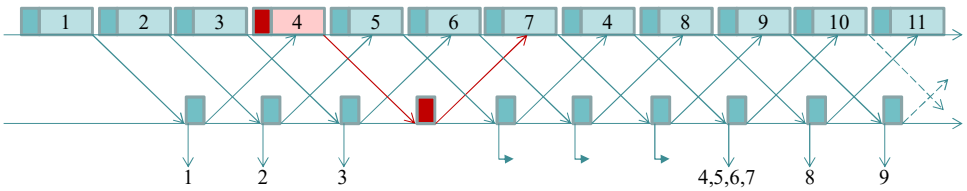
Así pues, el buffer se maneja como una "ventana deslizante" (*sliding window*). El manejo de este buffer aumenta la complejidad del protocolo en el transmisor pero permite aprovechar mejor los tiempos de propagación. En el receptor podemos hacer dos cosas: (1) Aceptar sólo los paquetes que llegan en orden de manera que, si llega un paquete fuera de secuencia, se descarta y se envía un reconocimiento negativo. (2) Aceptar todos los paquetes que lleguen sin errores de manera que, si llega un paquete fuera de secuencia, lo almacena para entregarlo posteriormente al nivel superior, cuando haya completado todos los paquetes intermedios. En el primer caso, cada paquete que el receptor acepta puede entregarlo inmediatamente al nivel superior, por lo que no necesita ninguna ventana deslizante que almacene los paquetes recibidos fuera de orden, aunque, con cada error, el transmisor debe volver a enviar la ventana entera (*GoBack-N*). En el segundo caso, como al nivel superior se deben entregar los paquetes en orden, aquellos que lleguen fuera de secuencia deben almacenarse en un buffer local, con el fin de entregarlos al nivel superior cuando se complete una secuencia ordenada de paquetes. Por eso el receptor también necesita un buffer para almacenar los paquetes recibidos que no ha podido entregar al nivel superior, el cual se maneja como una ventana deslizante que se abre con cada paquete que entrega al nivel superior y se cierra con cada paquete que acepta en desorden. La ventaja es que, en caso de error, en vez de retransmitir toda la ventana, el transmisor sólo debe retransmitir el paquete en problemas y continuar donde iba (*SelectiveRepeat*). El manejo del buffer en el receptor aumenta la complejidad del protocolo, pero permite aprovechar las transmisiones exitosas que se hayan hecho entre el paquete que se dañó y la notificación de dicho error en el transmisor. La Figura 66 muestra el comportamiento de estos dos protocolos.



(a) *GoBack-N* y *SelectiveRepeat* sin errores de transmisión



(b) *GoBack-N* con un error de transmisión



(c) *SelectiveRepeat* con un error de transmisión

Figura 66. Secuencias de eventos en los protocolos *Go-Back-N* y *SelectiveRepeat*

En *GoBack-N*, una transmisión con error toma $t_a = (L+2h)/C + 2t_p$ mientras que una transmisión exitosa sólo toma $t_b = (L+h)/C$. Como la probabilidad de error es p=1- $(1\text{-}BER)^{L+2h}$, de las $1/(1\text{-}p)$ transmisiones que toca hacer en promedio hasta que un paquete llegue bien, una corresponde a la transmisión correcta y las otras $p/(1\text{-}p)$ corresponden a transmisiones con errores, de manera que el tiempo promedio que toma la transmisión de un paquete hasta que llegue correctamente a su destino es $t_T=t_b+pt_a/(1\text{-}p)=[L+h+p(h+2t_pC)]/[C(1\text{-}p)]$. Entonces, en el tiempo en que se podrían transmitir t_TC bits de usuario, sólo se transmiten L bits, por lo que la eficiencia es L/t_TC, ó

$$E_{GB-N} = \frac{L}{L+h}\cdot\left(1-BER\right)^{L+2h}\cdot\frac{L+h}{L+h+p\left(h+2t_pC\right)}$$

Comparando con la eficiencia de *Stop&Wait*, los dos primeros términos no cambian pues los errores y el encabezado son inevitables, pero lo que si logramos hacer es que la ineficiencia debida al *bandwidth-delay product* sólo se haga presente cuando haya errores de transmisión, esto es, con probabilidad p=1- $(1\text{-}BER)^{L+2h}$. Con respecto a *SelectiveRepeat*, tanto las transmisiones correctas como las transmisiones con errores toman un tiempo $t = (L+h)/C$, por lo que la el término en la eficiencia debido al protocolo es uno:

$$E_{SR} = \frac{L}{L+h}\cdot\left(1-BER\right)^{L+2h}$$

La figura 2.29 muestra la eficiencia de *GoBack-N* y de *SelectiveRepeat*, en función de la longitud del paquete, para los mismos enlaces de la figura 2.27.

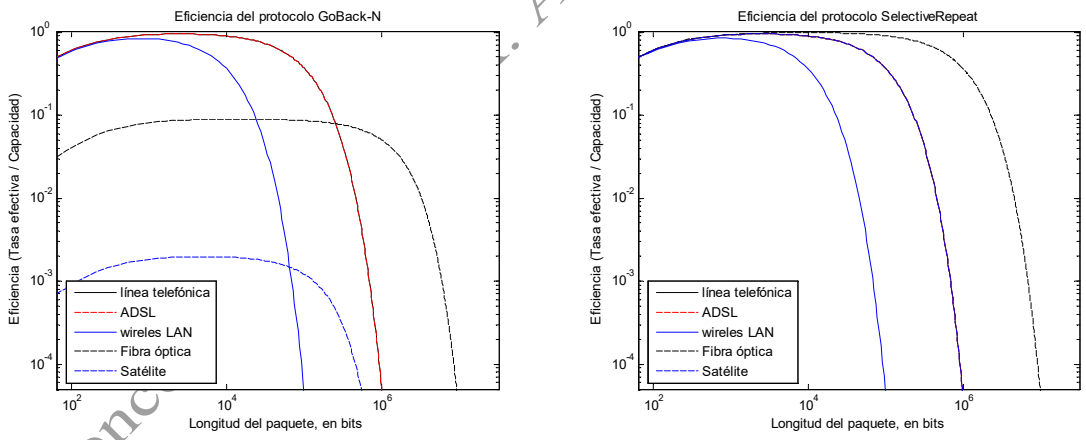


Figura 67. Eficiencia de los protocolos *GoBack-N* y *SelectiveRepeat*

Nótese cómo la diferencia en las expresiones de eficiencia para los tres protocolos se encuentra en la manera como afecta el término asociado con el *bandwidth-delay-product*, $2t_pC+h$. En *Stop&Wait* aparece multiplicado por uno, ya que su efecto se sufre en todas las transmisiones. En *GoBack-N* aparece multiplicado por p, ya que su efecto se sufre sólo en presencia de errores, que ocurren con probabilidad p. Y en *SelectiveRepeat* aparece multiplicado por cero, ya que su efecto nunca se sufre.

47. Desigualdades en la Cola de una Distribución

Sea X una v.a. cualquiera. A las probabilidades de la forma P[X≥a] ó P[|X|≥a] se les conoce como "Cola de la Distribución de X". Estas colas se pueden acotar sin necesidad de calcularlas exactamente:

(a) Desigualdad de Markov: Sea X una variable aleatoria no negativa con valor esperado E[X]<∞. Para cualquier α > 0, se cumple que

$$P\left[X > \alpha\right] \leq \frac{E[X]}{\alpha}$$

(b) Desigualdad de Chebyshev: Sea X una variable aleatoria con valor esperado E[X]<∞ y varianza V[X]<∞.Para cualquier α > 0, se cumple que

$$P\left[\left|X - E[X]\right| > \alpha\right] \leq \frac{V[X]}{\alpha^2}$$

(c) Cota de Chernoff: Sea X una variable aleatoria. Para cualquier α > 0, se cumple que

$$P\left[X > \alpha\right] \leq \min_{s>0} e^{-\alpha s} E\left[e^{sX}\right]$$

Entre sus muchas aplicaciones, las cotas de la cola de una distribución resultan muy útiles para determinar garantías de calidad de servicio en redes con servicios diferenciados. Por ejemplo, una medida de calidad de servicio indicaría que menos del 1% de los paquetes sufrirá un retardo superior a 100 ms; esto es, si *D* es la variable aleatoria que representa el retardo sufrido por un paquete, la medida de calidad de servicio está dada como una cota en la cola de su distribución: $P[D \geq 0.1] \leq 0.01$. Por esta razón, las cotas en la cola de la distribución de una v.a. resultan de gran importancia, al punto que un desarrollo teórico tan fundamental en ingeniería de redes con QoS como la "Capacidad equivalente" de una fuente de tráfico, por ejemplo, se basa en las cotas de Chernoff (ver definición ??).

Resulta fácil verificar la validez de estas cotas. En efecto, para una variable aleatoria no negativa, podemos partir de la definición misma del valor esperado, *E*[*X*], así:

$$E[X] = \int_0^\infty x dF_X(x) = \int_0^\alpha x dF_X(x) + \int_\alpha^\infty x dF_X(x)$$
$$\geq \int_\alpha^\infty x dF_X(x) \geq \alpha \int_\alpha^\infty dF_X(x) = \alpha P[X > \alpha]$$

que es la desigualdad de Markov. Esta desigualdad es muy interesante pues, por ejemplo, permite ver que para cualquier variable aleatoria no negativa con valor esperado finito, la probabilidad de que la variable exceda *n* veces su valor esperado siempre será menor o igual a 1/*n*.

Ahora, para cualquier variable *X* con valor medio y varianza finitos, podemos construir la variable no negativa $(X\text{-}E[X])^2$, cuyo valor esperado es la varianza de *X*, y aplicarle la desigualdad de Markov, con lo que obtenemos la desigualdad Chebyshev:

$$P\left[\left(X - E[X]\right)^2 \geq \alpha^2\right] \leq \frac{V[X]}{\alpha^2} \Rightarrow P\left[\left|X - E[X]\right| \geq \alpha\right] \leq \frac{V[X]}{\alpha^2}$$

La cota de Chernoff tiene una gran aplicabilidad en la teoría de las grandes desviaciones, de donde deriva gran parte de la formalidad de la ingeniería de redes con calidad de servicio. Para verificarla podemos partir de una relación muy general,

$$e^{s(x-a)} \geq u(x-a), \quad s>0$$

donde $u(x)$ es el escalón unitario, igual a cero para valores negativos de x e igual a 1 para otros valores de x, como se aprecia en la Figura 68.

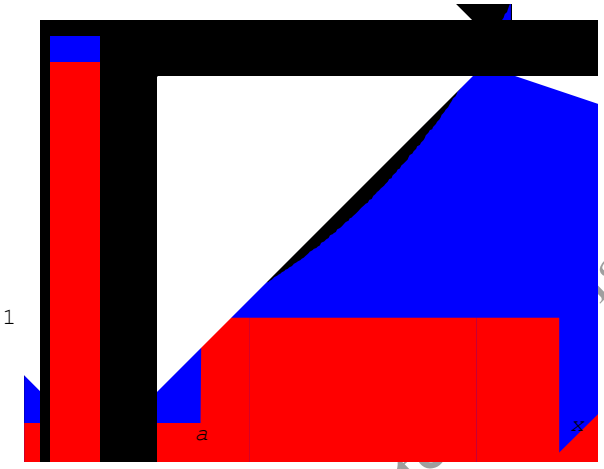


Figura 68. $e^{s(x-a)} \geq u(x)$ para cualquier $s>0$

De donde podemos verificar las siguientes expresiones:

$$P[X>a] = \int_a^{\infty} dF_X(x) = \int_{-\infty}^{\infty} u(x-a)dF_X(x) \leq \int_{-\infty}^{\infty} e^{s(x-a)}dF_X(x) = e^{-as}E\left[e^{sX}\right]$$

Como dicha desigualdad es válida para cualquier valor no-negativo de s, podemos minimizar con respecto a $s>0$ para obtener la cota de Chernoff.

Aunque tendremos oportunidad de usar estas desigualdades recurrentemente, es interesante mostrar un ejemplo sencillo para notar su poder y sus limitaciones: Se quiere asegurar que, en promedio, de cada mil paquetes no más de uno experimente un retardo superior o igual a 100 ms. ¿Cuánto debe ser el promedio del retardo de los paquetes para garantizar esta condición? Como el retardo D es una v.a. no negativa, aplica la desigualdad de Markov según la cual $P[D\geq 0.1]\leq 10E[D]=0.001$, por lo que podemos ofrecer la garantía de retardo máximo si diseñamos para un retardo promedio de 0.1 ms, independientemente de la distribución de D. Pero si el promedio resulta ser, digamos, de 0.2 ms, ¿Cuánto debe valer la desviación estándar para garantizar la misma condición? Según la desigualdad de Chebyshev, $P[D\geq 0.1] = P[D-0.0002\geq 0.0998] = P[|D-0.0002|\geq 0.0998] \leq V[D]/0.0998^2 = 0.001$, por lo que necesitamos que la desviación estándar no supere los 3.2 ms, independientemente de la distribución de D (nótese que la positividad del retardo nos permitió la segunda igualdad, pues $|D-0.0002|\geq 0.0998$ significa que $D\geq 0.1$ o que $D\leq -0.0996$, pero la segunda desigualdad es imposible por la positividad del retardo).

Claramente, estas cotas pueden ser muy poco estrictas, en la medida en que el servicio que realmente estemos ofreciendo puede ser mucho mejor a la garantía que ofrecemos. Si conocemos la distribución del retardo, podríamos hacer algún diseño más eficiente usando cotas de Chernoff. Por ejemplo, si

sabemos que el retardo obedece a una distribución normal[26] con media μ y desviación estándar $\mu/3$, basta con tener un retardo promedio inferior a 44.7 ms para asegurar que la probabilidad de superar los 100 ms no sea mayor que 1/1000. En efecto, la cota de Chernoff resulta ser $\exp(s(\mu(90+5\mu s)-9)/90)$, que se minimiza con $s=9(1-10\mu)/10\mu^2$, lo que conduce a una cota mínima igual a $\exp(-9(10\mu-1)^2/(200\mu^2))$, que es igual a 1/1000 cuando $\mu=0.0447$.

Para este último caso podemos calcular exactamente la probabilidad de superar los 100 ms de retardo, pues para variables $N(\mu,\sigma^2)$ se puede calcular fácilmente la cola de la distribución mediante la función $Q[(x-\mu)/\sigma] = 1 - F_X(x)$, ampliamente tabulada o fácilmente calculable numéricamente: Para un retardo promedio de 44.7 ms y una desviación estándar de 14.9 ms, la probabilidad de superar los 100 ms es 0.000103 < 0.001. En este caso, aún la cota de Chernoff resulta ser poco estricta.

48. Distribuciones con Cola Pesada

Sea X una v.a. con CDF F. Se dice que X tiene una distribución con cola pesada si $\lim_{x\to\infty} e^{\mu x}\left(1-F(x)\right)=\infty \quad \forall \mu>0$, *esto es, si el decrecimiento de la cola de la distribución (la probabilidad de que la variable tome valores mayores a x para valores grandes de x,* $P[X>x]=1-F(x)$*) es más lento que exponencial. Como un decrecimiento hiperbólico es más lento que exponencial, a veces el concepto de cola pesada se particulariza al caso en que* $1-F(x)$ *toma la forma* cx^{-a} *cuando* $x\to\infty$, *para* $0<a<2$ *y* $c>0$.

Nótese cómo, por ejemplo, la distribución exponencial tiene una cola que decae exponencialmente rápido:

$$\Pr[X>x] = 1-F(x) = \int_x^\infty \lambda e^{-\lambda u}du = e^{-\lambda x}, \quad x\geq 0$$

mientras que la distribución Gaussiana tiene una cola que decae aún más rápidamente:

$$Q(x) \triangleq \Pr[X>x] = 1-F(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-u^2/2}du \leq \frac{1}{2}e^{-x^2/2}$$

Sin embargo, considere la distribución de Pareto:

$$\Pr[X>x] = 1-F(x) = \int_x^\infty \frac{a}{b}\left(\frac{b}{u}\right)^{a+1} du = \left(\frac{b}{x}\right)^a, \quad x\geq b$$

que tiene exactamente la forma hiperbólica mencionada en la definición, lo cual implica una forma mucho más lenta de decaer. Con esta distribución podemos ver algunas de las características principales de las distribuciones con cola pesada. Por ejemplo, como se vio en la definición 44(d), nótese que si $a<2$, la varianza de la distribución es infinita y, si $a<1$, el valor esperado de la distribución también es infinito. Esta es una característica fundamental de las distribuciones con cola pesada: Una altísima variabilidad. En otras palabras, la cola pesada conduce a valores extremadamente grandes con una probabilidad no despreciable, de manera que al tomar muestras de

[26] Para una variable $X\sim N(\mu,\sigma^2)$ no es difícil encontrar que $E[e^{sX}] = exp(\mu s+\sigma^2 s^2/2)$. Ver definición 53.

una distribución así, la mayoría de ellas serán "pequeñas" pero algunas pocas de ella tendrán valores muy grandes. A medida que el parámetro a tiende a uno por la derecha, se reduce la velocidad con que la media muestral tiende al valor esperado.

Podemos verificar la ubicuidad de este fenómeno analizando la longitud de los archivos en nuestro disco duro. Invitamos al lector a que lea las longitudes de todos los archivos en su disco duro y grafique las frecuencias relativas de los eventos $A(k)$ = {Un archivo tiene una longitud mayor que o igual a $1024\cdot(2^k - 1)$ bytes y menor que $1024\cdot(2^{k+1} - 1)$ bytes}, $k \in \{0,1,2,\ldots\}$. La Figura 69 muestra el resultado obtenido en el computador portátil del autor. Del espacio ocupado en el disco, la mitad la ocupan los 12118 archivos más grandes (el 5.3% de los archivos) y la otra mitad la ocupan los 218607 archivos más pequeños (el 94.7% de los archivos). Si esta distribución caracteriza los archivos que intercambian por internet, se empezaría a explicar por qué las características del tráfico moderno son tan variables. De aquí la importancia que tiene el estudio de este tipo de variables aleatorias en el modelado de redes de comunicaciones. De hecho, muchas otras medidas en estas redes, tales como la duración de una sesión TCP, la longitud de un período de silencio en una conversación VoIP, o los tiempos de actividad e inactividad de una sesión http, tienen distribuciones con cola pesada. Las conclusiones que se hagan sobre el desempeño de la red con base en muestras de este tipo de variables aleatorias pueden ser equivocadas si no se hace un muy juicioso estudio estadístico para determinar la significancia de los resultados. Otras variables con cola pesada incluyen el número de personas afectadas por un apagón, el tamaño de los incendios forestales, los eventos de congestión en cascada del tráfico aéreo, el impacto de meteoritos en la superficie terrestre, las muertes en desastres naturales, las variaciones del mercado de valores, el uso de las palabras del español, la población de las ciudades, la riqueza de los individuos, la citación de artículos científicos, etc.

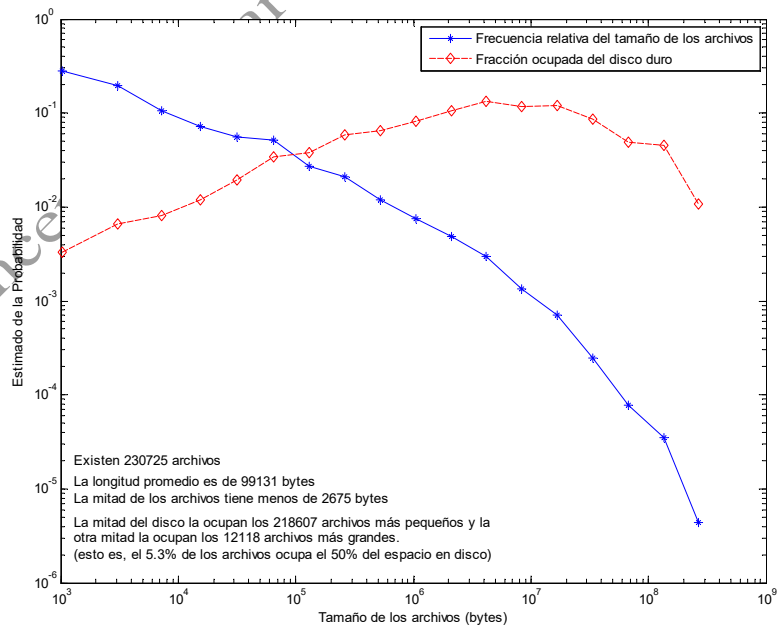


Figura 69. Frecuencia relativa de la longitud de los archivos en un disco duro

49. Distribución Condicional

Sea $(\Omega, \mathcal{F}, \mathbf{P})$ *un espacio de probabilidad donde se encuentra definida la variable aleatoria* $X:\Omega\to\mathbb{R}$ *y el evento* $B\in\mathcal{F}$. *La función de distribución acumulativa condicional de X dado B es* $F_{X|B}(x|B) = \mathbf{P}(\{X\le x\}\cap B)/\mathbf{P}(B)$. *Su derivada es la pdf condicional de X dado B y, si X es discreta, las áreas debajo de los impulsos de la pdf condicional es la pmf condicional de X dado B:*

$$f_{X|B}(x\,|\,B) = \frac{d}{dt}F_{X|B}(x\,|\,B) \qquad\qquad p_{k|B} = \frac{P\left[\{X = x_k\}\cap B\right]}{P[B]}$$

El evento condicional B bien puede estar definido en términos de la misma variable aleatoria X aunque, en general, puede ser cualquier evento del espacio de probabilidad original, $(\Omega, \mathcal{F}, \mathbf{P})$.

Como se demostró en la definición 18, $F_{X|B}(x|B)$ es una función de distribución acumulativa con las mismas propiedades que cualquier otra CDF:

(a) La CDF condicional es no-negativa: $F_{X|B}(x|B)\ge 0\ \forall\, x\in\mathbb{R}$
(b) La CDF condicional es no-decreciente: si $x_1 < x_2$ entonces $F_{X|B}(x_1|B) \le F_{X|B}(x_2|B)$
(c) La CDF condicional es acotada: $F_{X|B}(-\infty|B) = 0$, $F_{X|B}(\infty|B)=1$.
(d) La CDF condicional es continua por la derecha: $F_{X|B}(x^+|B) = F_{X|B}(x|B)$.

Lo mismo podemos decir de las funciones condicionales de densidad y distribución de probabilidad:

$(a)\ f_{X\|B}(x\|B)\ge 0\ \forall x\in\mathbb{R}$	$p_{k\|B}\ge 0$
$(b)\ F_{X\|B}(x\|B) = \int_{-\infty}^{x} f_{X\|B}(a\|B)da$	$F_{X\|B}(x\|B) = \sum_{k:x_k\le x} p_{k\|B}$
$(c)\ \int_{-\infty}^{\infty} f_{X\|B}(a\|B)da = 1$	$\sum_k p_{k\|B} = 1$

En efecto, al condicionar en el evento B sólo se ha restringido el espacio muestral de $(\Omega,\mathcal{F}, \mathbf{P})$ al espacio $(B, \mathcal{H}=\mathcal{F}\cap B, Q(\cdot)=P(\cdot|B))$, de manera que los eventos $A(x)=\{\omega\in\Omega:X(\omega)\le x\}, x\in\mathbb{R}$ se restringen a $A_B(x)=\{\omega\in B:X(\omega)\le x\}$ con probabilidad $Q(A_B(x))=P(A(x)\cap B)/P(B)=P(A_B(x))/P(B) = F_{X|B}(x|B)$.

A manera de ejemplo, supongamos que el evento B es $\{\omega\in\Omega : a < X(\omega) \le b\}$ o, en nuestra notación simplificada, $B = \{a < X \le b\}$. ¿Cuánto es $F_{X|B}(x|B)$?

$$F_{X|B}(x|B)=P\left[X\le x\middle|a<X\le b\right]=\frac{P\left[X\le x,a<X\le b\right]}{P\left[a<X\le b\right]}=\begin{cases}\dfrac{P[\Phi]}{F_X(b)-F_X(a)}=0 & x\le a\\ \dfrac{P\left[a<X\le x\right]}{P\left[a<X\le b\right]}=\dfrac{F_X(x)-F_X(a)}{F_X(b)-F_X(a)} & a<x\le b\\ \dfrac{P\left[a<X\le b\right]}{P\left[a<X\le b\right]}=\dfrac{F_X(\mathrm{b})-F_X(a)}{F_X(b)-F_X(a)}=1 & x>b\end{cases}$$

Derivando respecto a x,

$$f_{X|a<X\le b}(x|a<X\le b)=\begin{cases}0 & x\le a\\ \dfrac{f_X(x)}{\int_a^b f_X(s)ds} & a<x\le b\\ 0 & x>b\end{cases}$$

o, en el caso de una variable discreta,

$$p_{k|a<X\le b}=\begin{cases}0 & x_k\le a\\ \dfrac{p_k}{\sum_{a<x_j\le b}p_j} & a<x_k\le b\\ 0 & x_k>b\end{cases}$$

Por ejemplo, si lanzamos un dado y sabemos que cayó un número par, ¿Cuál es la distribución de probabilidad del número obtenido? P[X=k|X es par] = P[X=k]/(1/2) = 0.3 si k es par o cero si k es impar.

Como un último ejemplo, si X es gaussiana con media μ y varianza σ^2, ¿cuál será su distribución dado que $|X-\mu|\le k\sigma$? Sabemos que P[$|X-\mu|\le k\sigma$] = P[$\mu-k\sigma\le X\le\mu+k\sigma$] = $F_X(\mu+k\sigma)$-$F_X(\mu-k\sigma)$. Esta es la misma probabilidad de que una variable gaussiana con media cero y varianza uno esté en el intervalo [$-k$,k]. Si llamamos $G(x)$ a la CDF de la variable $\mathcal{N}(0,1)$, la pdf de X dado $|X-\mu|\le k\sigma$ es

$$f_X\left(x\middle\|X-\mu\right|\le k\sigma\left.\right)=\begin{cases}\dfrac{\exp\left(-(x-\mu)^2/2\sigma^2\right)}{\sigma(2G(k)-1)\sqrt{2\pi}} & \text{si } x\in[\mu-k\sigma,\mu+k\sigma]\\ 0 & \text{otro caso}\end{cases}$$

que se conoce como la distribución gaussiana truncada.

50. Momentos condicionales

Sea $(\Omega, \mathcal{F}, \mathbf{P})$ *un espacio de probabilidad donde se encuentra definida la variable aleatoria* $X:\Omega\to\mathbb{R}$ *y el evento* $B\in\mathcal{F}$. *El n-ésimo momento condicional de la variable aleatoria dado el evento B es* $E\left[X^n\middle|B\right]=\int_{\mathbb{R}}x^n dF_{X|B}(x|B)$. *El n-ésimo momento central condicional de la variable aleatoria dado el evento B es*

$E\left[(X-E[X|B])^n|B\right]=\int_{\mathbb{R}}\left(x-E[X|B]\right)^n dF_{X|B}(x|B)$, *donde* $E[X|B]$ *es el primero momento condicional de X dado B.*

Sea, por ejemplo, una variable aleatoria Gaussiana con media cero y varianza σ^2, X. ¿Cuáles serán el valor esperado y la varianza condicionales de X dado que $X>0$? Por la definición 49 sabemos que

$$f_{X|X>0}(x|X>0)=\begin{cases}0 & x\le 0\\ \dfrac{f_X(x)}{\int_0^\infty f_X(s)ds} & x>0\end{cases}$$

donde $f_X(x)=\dfrac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2}$ y $\int_0^\infty f_X(s)ds=1/2$. Esto es,

$$f_{X|X>0}(x|X>0)=\frac{1}{\sigma}\sqrt{\frac{2}{\pi}}e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2},\ x>0$$

En consecuencia

$$E\left[X|X>0\right]=\frac{1}{\sigma}\sqrt{\frac{2}{\pi}}\int_0^\infty xe^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2}dx$$

Esta integral se puede calcular fácilmente haciendo el cambio de variable $y=(x/\sigma)^2/2$ de manera que $dy=xdx/\sigma^2$:

$$E\left[X|X>0\right]=\sigma\sqrt{\frac{2}{\pi}}\int_0^\infty e^{-y}dy=\sigma\sqrt{\frac{2}{\pi}}$$

ya que $\int_0^\infty e^{-y}dy=-e^{-y}\Big|_0^\infty=1$. De acuerdo con este resultado, podemos calcular

$$V\left[X|X>0\right]=\frac{1}{\sigma}\sqrt{\frac{2}{\pi}}\int_0^\infty\left(x-\sigma\sqrt{2/\pi}\right)^2 e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2}dx$$

$$V\left[X|X>0\right]=\frac{1}{\sigma}\sqrt{\frac{2}{\pi}}\int_0^\infty x^2e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2}dx-\sigma^2\frac{2}{\pi}$$

Como el integrando tiene simetría par con respecto a x=0, podemos integrar de -∞ a +∞ y dividir por 2:

$$V\left[X|X>0\right]=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^\infty x^2e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2}dx-\sigma^2\frac{2}{\pi}$$

donde podemos reconocer la varianza de X en el primer término, σ^2, de donde

$$V\left[X|X>0\right]=\sigma^2\left(1-\frac{2}{\pi}\right)$$

51. Memoria de una Distribución

Dada una variable aleatoria X con CDF $F_X(x)$, podemos considerar la distribución condicional de la cola de la distribución, $P[X > x+s \mid X > s]$, esto es, cuál es la probabilidad de que la variable sea mayor a x+s dado que ya sabemos que es mayor a s. Si esta probabilidad depende de s, se dice que la distribución tiene memoria. En otro caso, la distribución carece de memoria. La única distribución discreta sin memoria es la geométrica. La única distribución continua sin memoria es la exponencial.

Para ver la importancia de esta definición, consideremos por ejemplo tres modelos diferentes para el tiempo de transmisión de un paquete:

Modelo 1: T_1 es una v.a. uniformemente distribuida entre 0 y S: $f_1(t) = \begin{cases} \frac{1}{S} & 0 \le t \le S \\ 0 & \text{otro } t \end{cases}$

Modelo 2: T_2 es una v.a. exponencialmente distribuida con parámetro λ: $f_2(t) = \lambda e^{-\lambda t}, \quad t \ge 0$

Modelo 3: T_3 es una v.a. con distribución de Pareto, con parámetros a, b: $f_3(t) = \frac{a}{b}\left(\frac{b}{x}\right)^{a+1} \qquad t \ge b$

La pregunta que nos haremos es simple: Si el paquete que está en servicio empezó a ser transmitido hace τ segundos, ¿cuánto tiempo le falta para terminar su transmisión? Esta es otra variable aleatoria, el tiempo remanente, cuya distribución es la cola condicional de la distribución original. Esto es, si T es el tiempo de transmisión y el paquete lleva τ segundos siendo transmitido, lo que queremos saber es con qué probabilidad faltan más de t segundos para terminar la transmisión:

$$P[T > t+\tau \mid T > \tau] = \frac{P[T > t+\tau]}{P[T > \tau]}$$

En el modelo 1 tenemos $P[T>s]$ = máx(0,mín(1,1 – s/S)), de manera que $P[T>t+\tau|T>\tau]$ = máx(0,mín(1,1 – $t/(S-\tau)$)). Esto es, la distribución del tiempo remanente está uniformemente distribuida entre 0 y $S-\tau$, de manera que entre mayor tiempo hayamos esperado, menor tiempo nos falta por esperar. De hecho el tiempo promedio de espera se reduce de $S/2$ a $(S-\tau)/2$, de manera que el tiempo que falta se acerca a cero a medida que τ se acerca a S. Esta dependencia entre el tiempo que hemos esperado y el que nos falta por esperar se conoce como la **memoria de la distribución uniforme**, que es negativa porque entre más hemos esperado menos falta por esperar.

En el modelo 2 tenemos que $P[T>s] = e^{-\lambda s}$, de manera que $P[T>t+\tau|T>\tau] = e^{-\lambda(t+\tau)}/e^{-\lambda\tau} = e^{-\lambda t}$: El tiempo que falta por esperar sigue siendo una variable aleatoria exponencial con parámetro λ, independientemente de que ya hayamos esperado 1 milisegundo, 1 minuto ó 1 año. De hecho, originalmente el tiempo promedio de espera era $1/\lambda$ y, después de esperar τ segundos, el tiempo promedio del tiempo que falta por esperar sigue siendo $1/\lambda$, independientemente de τ. Esta propiedad tan particular se conoce como la **falta de memoria de la distribución exponencial**.

En el modelo 3 tenemos que $P[T>s] = (b/s)^a$, de manera que $P[T>t+\tau|T>\tau] = [\tau/(t+\tau)]^a$. Esto es, la probabilidad de que debamos esperar t segundos más dado que ya hemos esperado τ segundos

¡aumenta con τ!, de manera que tiende a uno a medida que τ tiende a infinito: entre mayor tiempo hayamos esperado, mayor tiempo nos falta por esperar. De hecho, suponiendo que a>1, el tiempo promedio que falta por esperar después de haber esperado τ segundos pasa de $ab/(a\text{-}1)$ a $\tau/(a\text{-}1)$, que tiende a infinito si τ tiende a infinito. Esta dependencia entre el tiempo que hemos esperado y el que nos falta por esperar se conoce como la **memoria de la distribución Pareto**. El hecho de que el tiempo remanente aumente con el tiempo que llevamos es un efecto de la **cola pesada** de la distribución, una de las **leyes de potencia** que tienen grandes implicaciones en la ingeniería de tráfico en redes de comunicaciones.

Las distribuciones exponencial y geométrica son las únicas distribuciones que no tienen memoria. En efecto, como necesitamos que $P[T>t+\tau|T>\tau] = P[T>t+\tau]/P[T>\tau] = P[T>t]$, será necesario que $P[T>t+\tau]$ se pueda expresar como el producto $P[T>t]P[T>\tau]$. En el caso continuo, esto exige que la cola de la distribución sea de la forma $P[T>t]=e^{-\lambda t}$. En el caso discreto, esto exige que la cola de la distribución sea $P[X>n]=p^n$.

52. Función Característica

Sea X una variable aleatoria. La función característica de X es una función de los reales en los complejos, definida de la siguiente manera

$$\phi_X(\omega) = E\left[e^{j\omega X}\right] = \int_{\mathbb{R}} e^{j\omega x} dF_X(x)$$

donde $j = \sqrt{(-1)}$.

Nótese que, con excepción del signo del exponente, la función característica de una variable aleatoria es la transformada de Fourier de su función de densidad de probabilidad o, tratándose de variables discretas, la transformada de Fourier en tiempo discreto de su función de distribución de probabilidad:

$$\phi_X(\omega) = \begin{cases} \int_{-\infty}^{\infty} f_X(x) e^{j\omega x} dx & continua \\ \sum_{k=-\infty}^{\infty} p_k e^{jk\omega} & discreta \end{cases}$$

Posiblemente recordemos cómo ésta era una herramienta muy útil en la soluciones de ecuaciones diferenciales (o de diferencia), así como en el cálculo de integrales (o sumas) de convolución en sistemas lineales. Pues bien, más adelante encontraremos este tipo de ecuaciones en el estudio de múltiples variables aleatorias y, en ese sentido, la función característica cumple un papel igualmente importante. En esas operaciones, si recuerda bien, se operaba en el dominio de la frecuencia y, posteriormente, se retornaba al dominio del tiempo. Ese retorno se consigue mediante la fórmula inversa de la función característica:

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_X(\omega) e^{-j\omega x} d\omega$$

$$p_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} \phi_X(\omega) e^{-jk\omega} d\omega$$

Los cambios de signo en el exponente complejo con respecto a la transformada de Fourier y la transformada inversa obedecen a una convención muy simple: en teoría de señales los vectores suelen ser columnas y en teoría de probabilidades los vectores suelen ser filas, lo cual exige cambiar el vector que se transpone y se conjuga al calcular el producto interno representado por la transformada.

Como un ejemplo de la utilidad de la función característica, nótese que si $\phi_X(\omega)$ es n veces diferenciable en el origen (ω=0), los n-ésimos momentos de X resultan fáciles de calcular:

$$E\left[X^n\right] = j^{-n}\frac{d^n}{d\omega^n}\phi_X(0)$$

Como se puede demostrar fácilmente al derivar la definición 52.

Por ejemplo, consideremos una variable aleatoria X exponencialmente distribuida con parámetro λ, cuya función característica es

$$\Phi_X(\omega) = \lambda\int_0^\infty e^{(j\omega-\lambda)x}dx = \frac{\lambda}{j\omega-\lambda}e^{(j\omega-\lambda)x}\Big|_0^\infty = \frac{\lambda}{\lambda-j\omega}$$

El cálculo de las estadísticas de primer y segundo orden resultan en $E[X] = \frac{1}{j}\frac{j\lambda}{(\lambda-j\omega)^2}\Big|_{\omega=0} = \frac{1}{\lambda}$

y $E[X^2] = \frac{1}{j^2}\frac{-2\lambda}{(\lambda-j\omega)^3}\Big|_{\omega=0} = \frac{2}{\lambda^2}$, con lo cual $V[X] = \frac{1}{\lambda^2}$.

Igualmente fácil es considerar una variable Y normalmente distribuida con parámetros μ y σ^2:

$$\Phi_Y(\omega) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2+j\omega y}dy = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{(y-\mu-j\omega\sigma^2)^2+(\omega^2\sigma^4-2j\omega\mu\sigma^2)}{2\sigma^2}}dy = \cdots$$

$$e^{-\frac{\omega^2\sigma^4-2j\omega\mu\sigma^2}{2\sigma^2}}\left[\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{(y-\mu-j\omega\sigma^2)^2}{2\sigma^2}}dy\right] = e^{j\omega\mu-\frac{\omega^2\sigma^2}{2}}$$

(este truco de completar el cuadrado para obtener una integral conocida es muy típico al trabajar con la distribución normal). De estos resultados se encuentran fácilmente los momentos de la distribución Gaussiana:

$$E[Y] = \frac{1}{j}\frac{d}{d\omega}e^{j\omega\mu-\frac{\omega^2\sigma^2}{2}}\Big|_{\omega=0} = \mu \qquad E[Y^2] = -\frac{d^2}{d\omega^2}e^{j\omega\mu-\frac{\omega^2\sigma^2}{2}}\Big|_{\omega=0} = \mu^2+\sigma^2$$

de donde $V[Y]=\sigma^2$.

Por último, veamos el caso de una variable de Poisson, U, con intensidad ρ:

$$\Phi_U(j\omega) = \sum_{k=0}^{\infty} e^{j\omega k}\frac{\rho^k}{k!}e^{-\rho} = e^{-\rho}\sum_{k=0}^{\infty}\frac{(\rho e^{j\omega})^k}{k!} = e^{-\rho}\exp\left(\rho e^{j\omega}\right) = \exp\left(\rho\left(e^{j\omega}-1\right)\right)$$

$$E[U]=\frac{1}{j}\frac{d}{d\omega}\exp\left(\rho\left(e^{j\omega}-1\right)\right)\Big|_{\omega=0}=\rho e^{j\omega}\exp\left(\rho\left(e^{j\omega}-1\right)\right)\Big|_{\omega=0}=\rho$$

$$E[U^2]=-\frac{d^2}{d\omega^2}\exp\left(\rho\left(e^{j\omega}-1\right)\right)\Big|_{\omega=0}=\rho e^{j\omega}\exp\left(\rho\left(e^{j\omega}-1\right)\right)\left(1+\rho e^{j\omega}\right)\Big|_{\omega=0}=\rho(1+\rho)$$

$$V[U]=E[U^2]-E^2[U]=\rho$$

Cuando se trata de variables continuas o de variables discretas, la función característica (transformada de Fourier) se puede generalizar a la función generadora de momentos (transformada de Laplace) o a la función generadora de probabilidad (transformada Z). En este capítulo las mencionaremos brevemente aunque su utilidad principal se verá en los próximos capítulos al considerar distribuciones conjuntas de múltiples variables aleatorias.

53. Función Generadora de Momentos

Sea X una variable aleatoria continua. La función generadora de momentos de X es una función de los complejos en los complejos definida de la siguiente manera

$$M_X(s)=E\left[e^{sX}\right]=\int_{\mathbb{R}}e^{sx}dF_X(x)=\int_{\mathbb{R}}e^{sx}f_X(x)dx$$

Cuando s=jω, obtenemos la función característica de la distribución.

El nombre de Función generadora de momentos es fácil de comprender:

$$\frac{d^n}{ds^n}M_X(0)=\frac{d^n}{ds^n}E\left[e^{sX}\right]\Big|_{s=0}=E\left[\frac{d^n}{ds^n}e^{sX}\right]\Big|_{s=0}=E\left[X^n e^{sX}\right]\Big|_{s=0}=E\left[X^n\right]$$

de manera que el *n*-ésimo momento de *X* es la *n*-ésima derivada de la función generadora de momentos, evaluada en *s*=0.

Consideremos, por ejemplo, la función generadora de momentos de algunas variables conocidas:

1. Exponencial(λ): $M_X(s)=E\left[e^{sX}\right]=\lambda\int_0^\infty e^{-(\lambda-s)x}dx=\frac{\lambda}{\lambda-s},\quad s<\lambda$
2. Gaussiana(μ,σ): $M_X(s)=\exp(s(\mu+\sigma^2 s/2))$ -de la manera que se calculó $\Phi_X(\omega)$ en 52-
3. Pareto(a,b): $M_X(s)=E\left[e^{sX}\right]=ab^a\int_b^\infty e^{sx}x^{-a-1}dx=a(-bs)^a\Gamma(-a,-bs),\quad s<0$

   donde $\Gamma(u,v)$ es la función gamma incompleta inferior, $\Gamma(u,v)=\int_v^\infty e^{-x}x^{u-1}dx$
4. Uniforme(a,b): $M_X(s)=E\left[e^{sX}\right]=\frac{1}{b-a}\int_a^b e^{sx}dx=\frac{e^{sb}-e^{sa}}{s(b-a)}$

54. Función Generadora de Probabilidad

Sea X una variable aleatoria discreta que toma valores en los enteros no negativos, con p_k=Prob[X=k], k=0,1,2,... La función generadora de probabilidad de X es una función de los complejos en los complejos definida de la siguiente manera

$$G_X(z) = E\left[z^X\right] = \sum_{k=0}^{\infty} p_k z^k$$

Cuando $z=e^{j\omega}$, obtenemos la función característica de la distribución.

El nombre de Función generadora de probabilidad es fácil de comprender. Considérese una variable aleatoria discreta que toma valores en los enteros, de manera que p_k = P[X=k] para k=0,1,2,….

$$\frac{d^n}{dz^n} G_X(0) = \frac{d^n}{dz^n} \sum_{k=0}^{\infty} p_k z^k \bigg|_{z=0} = \sum_{k=0}^{\infty} p_k \frac{d^n}{dz^n} z^k \bigg|_{z=0} = \sum_{k=n}^{\infty} \frac{k!}{(k-n)!} p_k z^{k-n} \bigg|_{z=0} = n!\, p_n$$

$$\Rightarrow \quad p_n = \frac{1}{n!} \frac{d^n}{dz^n} G_X(0)$$

de manera que la distribución de probabilidad se puede obtener inmediatamente de la función generadora de probabilidad mediante diferenciación y normalización.

También ocurren cosas interesantes cuando evaluamos en z=1:

$$G_X(1) = \sum_{k=0}^{\infty} p_k = 1 \qquad \frac{d}{dz} G_X(1) = \sum_{k=1}^{\infty} k p_k = E[X] \qquad \frac{d^2}{dz^2} G_X(1) = \sum_{k=2}^{\infty} k(k-1) p_k = E[X(X-1)]$$

$$\frac{d^n}{dz^n} G_X(1) = \sum_{k=n}^{\infty} k(k-1)\cdots(k-n+1) p_k = E\left[X(X-1)\cdots(X-n+1)\right]$$

lo que indica que la función generadora de probabilidad también es, de algún modo, una función generadora de momentos.

Consideremos, por ejemplo, la función generadora de probabilidad de algunas variables conocidas:

1. Bernoulli(p): $G_X(z)=(1-p)+pz = 1 + p(z-1)$
2. Geométrica(p): $G_X(z) = \sum_{k=0}^{\infty} z^k p^k (1-p) = (1-p)\sum_{k=0}^{\infty} (zp)^k = (1-p)/(1-zp)$, $|zp|<1$
3. Binomial(n,p): $G_X(z) = \sum_{k=0}^{n} z^k \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=0}^{n} \binom{n}{k} (zp)^k (1-p)^{n-k} = \left(1+p(z-1)\right)^n$
4. Poisson(λ): $G_X(z) = \sum_{k=0}^{\infty} z^k \frac{\rho^k}{k!} e^{-\rho} = e^{-\rho} \sum_{k=0}^{\infty} \frac{(\rho z)^k}{k!} = e^{-\rho} e^{\rho z} = \exp\left(\rho(z-1)\right)$

La utilidad de las funciones característica, generadora de momentos y generadora de probabilidades se podrá apreciar mejor al considerar variables aleatorias conjuntamente distribuidas, que es el tema del siguiente capítulo.

IV. Vectores Aleatorios Bi-dimensionales

55. Campo-σ de Borel en el Plano, $\mathcal{B}(\mathbb{R}^2)$

El campo-σ de Borel del plano real, $\mathcal{B}(\mathbb{R}^2)$, es el mínimo campo-σ que contiene a todos los subconjuntos de la forma

$$A(x,y) = \{(a,b)\in\mathbb{R}^2 : a \leq x,\ b \leq y\},\ \forall\ (x,y)\in\mathbb{R}^2.$$

Los subconjuntos de $\mathbb{R}^2$ que pertenecen a $\mathcal{B}(\mathbb{R}^2)$ se denominan "conjuntos de Borel".

Ya vimos cómo los conjuntos de la forma $(-\infty, x]$ resultaron fundamentales para pasar de espacios de probabilidad arbitrarios $(\Omega, \mathcal{F}, \mathbf{P})$ a espacios más fáciles de manejar definidos por una variable aleatoria, $(\mathbb{R}, \mathcal{B}(\mathbb{R}), F_X(x))$. De la misma manera, al trabajar con dos resultados numéricos diferentes, X_1 y X_2, asociados con un mismo experimento aleatorio $(\Omega, \mathcal{F}, \mathbf{P})$, podemos evitar el tener que describir explícitamente el espacio de probabilidad original si construimos un nuevo espacio implícito para el resultado (X_1,X_2) donde el espacio muestral está dado por el plano cartesiano $\mathbb{R}^2$, el campo-sigma de eventos es el campo-sigma de Borel de $\mathbb{R}^2$, y donde la medida de probabilidad está dada por una función de $\mathbb{R}^2$ en $\mathbb{R}$, $F_{X,Y}(x,y)=\mathbf{P}[\{\omega\in\Omega : X_1(\omega) \leq x_1\}\cap\{\omega\in\Omega : X_2(\omega) \leq x_2\},\ (x,y)\in\mathbb{R}^2]$. En este caso hablaremos del vector aleatorio (X_1, X_2), definido en 56, y de su función de distribución acumulativa conjunta $F_{X,Y}(x,y)$, definida en 57. Efectivamente, como es imposible pensar en asignar probabilidades a cada elemento del conjunto potencia de $\mathbb{R}^2$, nos limitamos a los conjuntos que pertenezcan a $\mathcal{B}(\mathbb{R}^2)$. Nuevamente, el campo sigma $\mathcal{B}(\mathbb{R}^2)$ incluye puntos, líneas, curvas, polígonos, círculos y otros subconjuntos razonables de $\mathbb{R}^2$. Limitándonos a uniones numerables de este tipo de eventos en $\mathbb{R}^2$, podemos construir el espacio de probabilidad $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2), F_{X,Y}(\cdot\,,\cdot))$, sobre el cual podremos aplicar toda la lógica booleana sin llegar a inconsistencias.

La siguiente figura muestra un conjunto elemental de Borel, $A(8,8)$, que es el producto cartesiano de los intervalos $\{x \leq 8\} \times \{y \leq 8\}$. Con complementos, uniones contables e intersecciones contables de estos conjuntos básicos podemos generar casi cualquier subconjunto imaginable de $\mathbb{R}^2$.

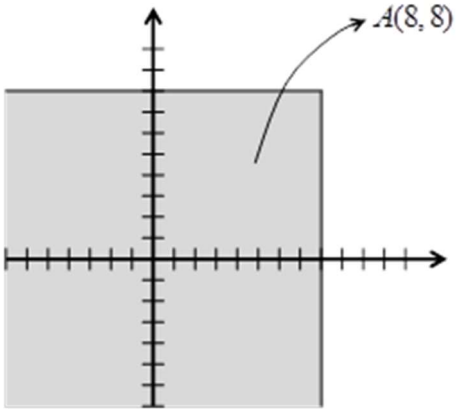


Figura 70. $A(8,8) = \{(a,b)\in\mathbb{R}^2 : a \leq 8,\ b \leq 8\}$