OpenIntro Statistics.

Capítulo 3: Probabilidad

Metodología de análisis en Opinión Pública

Cátedra Olego
Facultad de Ciencias Sociales
Universidad de Buenos Aires

Definición de probabilidad

Procesos aleatorios

- Un proceso aleatorio es una situación en la que sabemos qué resultados podrían ocurrir, pero no sabemos qué resultado en particular ocurrirá.
- Ejemplos: lanzamientos de monedas, tiradas de dados, reproducción aleatoria de iTunes, si el mercado de valores sube o baja mañana, etc.
- Puede ser útil modelar un proceso como aleatorio incluso si no es realmente aleatorio.

MP3 Players › Stories › **iTunes: Just how random is random?**

iTunes: Just how random is random?

By David Braue on 08 March 2007

- **Introduction**
- **Say You, Say What?**
- **A role for labels?**
- **The new random**

Think that song has appeared in your playlists just a few too many times? David Braue puts the randomness of Apple's song shuffling to the test -- and finds some surprising results.

Quick -- think of a number between one and 20. Now think of another one, and another, and another.

Starting to repeat yourself? No surprise: in practice, many series of random numbers are far less random than you would think.

Computers have the same problem. Although all systems are able to pick random numbers, the method they use is often tied to specific other numbers -- for example, the time -- that means you could get a very similar series of 'random' numbers in different situations.

This tendency manifests itself in many ways. For anyone who uses their iPod heavily, you've probably noticed that your supposedly random 'shuffling' iPod seems to be particularly fond of the Bee Gees, Melissa Etheridge or Pavarotti. Look at a random playlist that iTunes generates for you, and you're likely to notice several songs from one or two artists, while other artists go completely unrepresented.

http://www.cnet.com.au/itunes-just-how-random-is-random-339274094.htm

Definición Axiomática de la Probabilidad

Hay varias interpretaciones posibles de la probabilidad pero estas están de acuerdo en las reglas matemáticas que debe seguir la probabilidad.

Las suposiciones para establecer los axiomas se pueden resumir de la siguiente manera: Sea $(\Omega, \mathrm{F}, \mathrm{P})$ un espacio de medida, con P(E) la probabilidad de algún Evento y $\mathrm{P}(\Omega) = 1$. Entonces $(\Omega, \mathrm{F}, \mathrm{P})$ es un espacio de probabilidad, con espacio muestral Ω, espacio de eventos F y medida de probabilidad P.

- primer axioma

$$\mathrm{P}(\mathrm{E}) \in \mathbb{R}, \mathrm{P}(\mathrm{E}) \geq 0 \qquad \forall \mathrm{E} \in \mathrm{F}$$

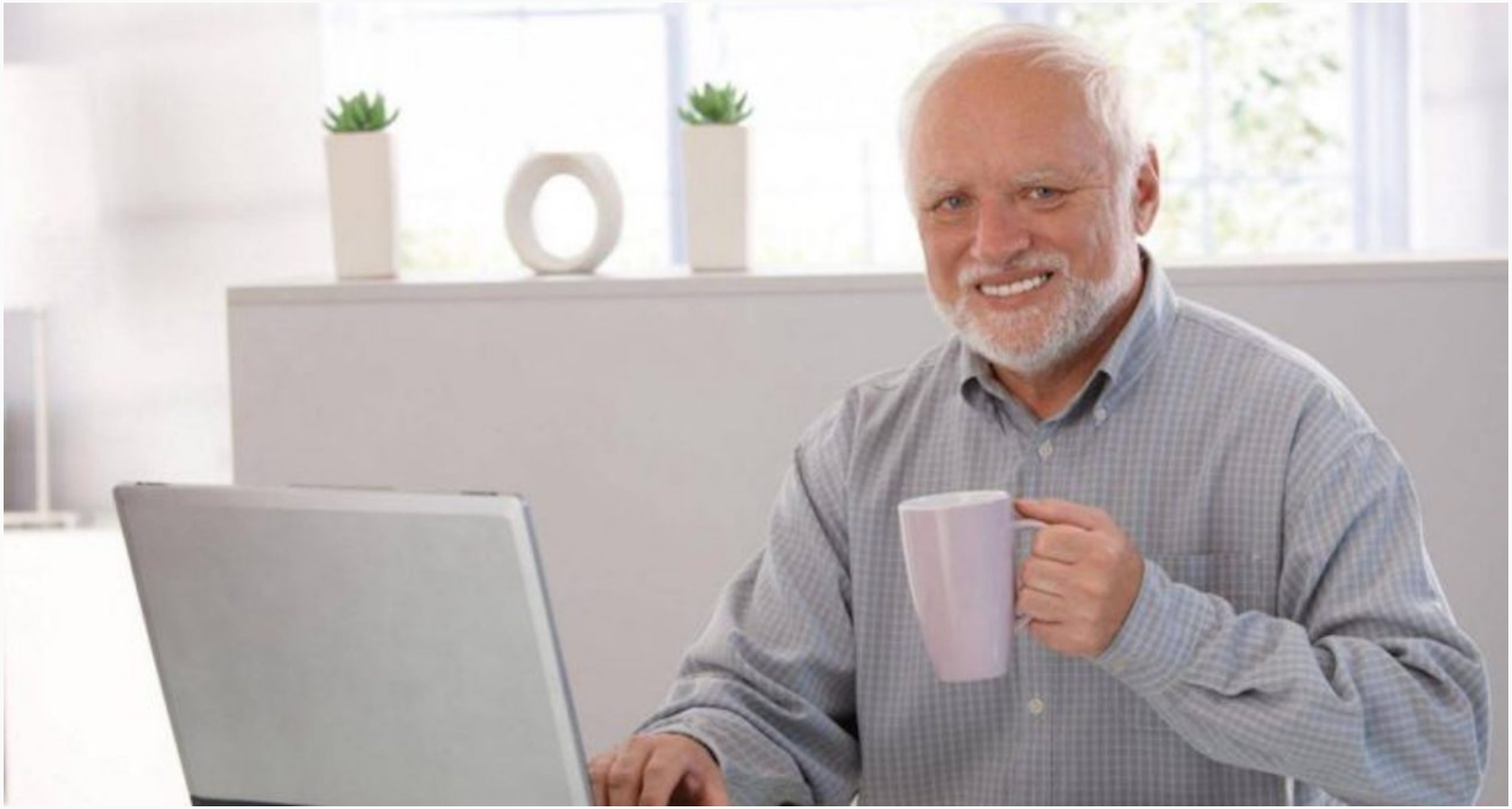
- Segundo axioma

$$\mathrm{P}(\Omega) = 1$$

- Tercer axioma

$$\mathrm{P}\left(\bigcup_{i=1}^{\infty} \mathrm{E}_i\right) = \sum_{i=1}^{\infty} \mathrm{P}(\mathrm{E}_i)$$

Definición Axiomática de la Probabilidad



Definición Axiomática de la Probabilidad

Hay varias interpretaciones posibles de la probabilidad pero estas están de acuerdo en las reglas matemáticas que debe seguir la probabilidad.

Las suposiciones para establecer los axiomas se pueden resumir de la siguiente manera: Sea $(\Omega, \mathrm{F}, \mathrm{P})$ un espacio de medida, con P(E) la probabilidad de algún Evento y $\mathrm{P}(\Omega) = 1$. Entonces $(\Omega, \mathrm{F}, \mathrm{P})$ es un espacio de probabilidad, con espacio muestral Ω, espacio de eventos F y medida de probabilidad P.

- primer axioma
  La probabilidad de un evento es un número real no negativo:

$$\mathrm{P}(\mathrm{E}) \in \mathbb{R}, \mathrm{P}(\mathrm{E}) \geq 0 \qquad \forall \mathrm{E} \in \mathrm{F}$$

- Segundo axioma
  Esta es la suposición de unidad de medida: que la probabilidad de que ocurra al menos uno de los eventos elementales en todo el espacio muestral es 1

$$\mathrm{P}(\Omega) = 1$$

- Tercer axioma
  Esta es la suposición de σ-aditividad. Cualquier secuencia contable de conjuntos disjuntos $\mathrm{E}_1, \mathrm{E}_2, \ldots$ satisface

$$\mathrm{P}\left(\bigcup \mathrm{E}_1 \bigcup \mathrm{E}_2 \bigcup \ldots\right) = \mathrm{P}(\mathrm{E}_1) + \mathrm{P}(\mathrm{E}_2) + \ldots$$

Interpretaciones de la Probabilidad

- Interpretación frecuentista:
  - La probabilidad de un resultado es la proporción de veces que ocurriría el resultado si observáramos el proceso aleatorio un número infinito de veces.
- Interpretación bayesiana:
  - Un bayesiano interpreta la probabilidad como un grado subjetivo de creencia: Para el mismo evento, dos personas separadas podrían tener diferentes puntos de vista y así asignar diferentes probabilidades.
  - Ampliamente popularizado por el avance revolucionario en tecnología y métodos computacionales durante los últimos veinte años.

Práctica

¿Cuál de los siguientes eventos le sorprendería más?

(a) exactamente 3 caras en 10 lanzamientos de moneda

(b) exactamente 3 caras en 100 lanzamientos de moneda

(c) exactamente 3 caras en 1000 lanzamientos de moneda

Práctica

¿Cuál de los siguientes eventos le sorprendería más?

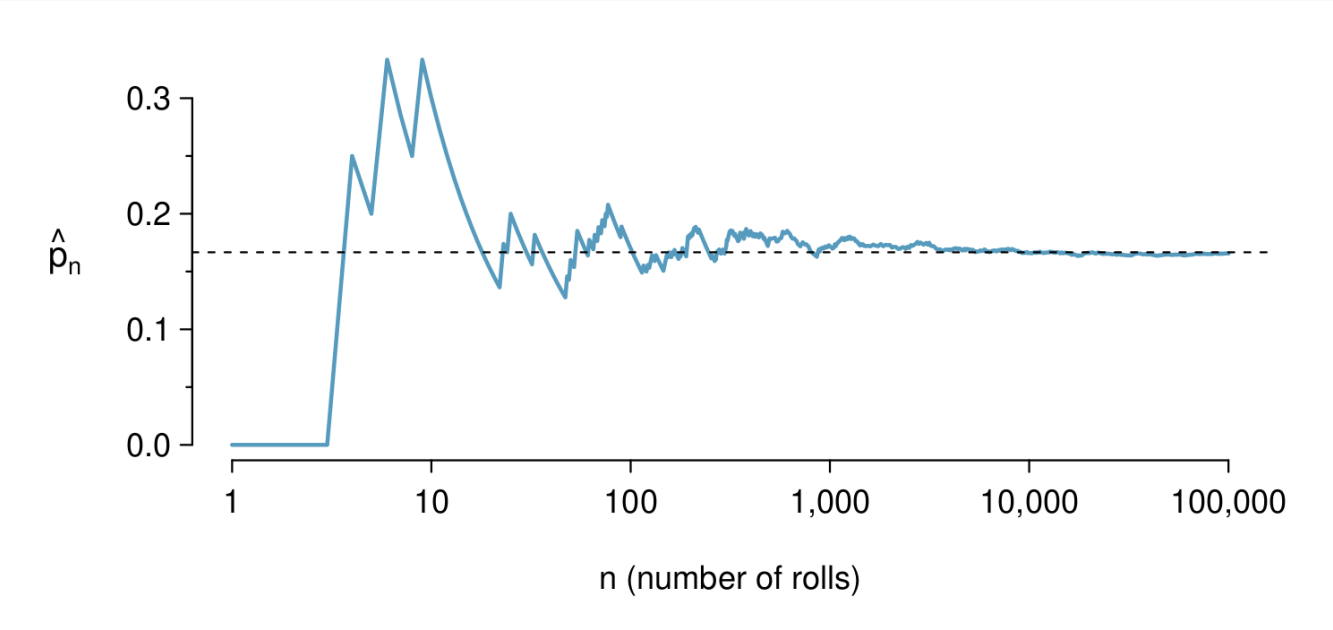
(a) exactamente 3 caras en 10 lanzamientos de moneda

(b) exactamente 3 caras en 100 lanzamientos de moneda

(c) exactamente 3 caras en 1000 lanzamientos de moneda

Ley de los grandes números

Ley de los grandes números establece que a medida que se recolectan más observaciones, la proporción de ocurrencias con un resultado particular, $\hat{p}_n$, converge a la probabilidad de ese resultado, p.



La probabilidad se puede ilustrar tirando un dado muchas veces. Sea p̂ n la proporción de resultados que son 1 después de los primeros n lanzamientos. A medida que aumenta el número de tiradas, $\hat{p}_n$ convergerá a la probabilidad de sacar un 1, $p = 1/6$. La figura muestra esta convergencia para 100 000 lanzamientos de dados. La tendencia de $\hat{p}_n$ a estabilizarse alrededor de p está descrita por la Ley de los Grandes Números.

Ley de los grandes números (cont.)

Al lanzar una moneda justa, si sale cara en cada uno de los primeros 10 lanzamientos, ¿cuál crees que es la probabilidad de que salga otra cara en el siguiente lanzamiento? 0,5, menos de 0,5 o más de 0,5?

$$\underline{\text{H}}\ \underline{\text{H}}\ \underline{\text{H}}\ \underline{\text{H}}\ \underline{\text{H}}\ \underline{\text{H}}\ \underline{\text{H}}\ \underline{\text{H}}\ \underline{\text{H}}\ \underline{\text{H}}\ \underline{?}$$

- La probabilidad sigue siendo 0,5, o todavía hay un 50% de posibilidades de que salga otra cara en el siguiente lanzamiento.

$$\text{P(H el } 11^{\text{th}} \text{ tirar)} = \text{P(T el } 11^{\text{th}} \text{ tirar)} = 0.5$$

- La moneda no es ”vencida” por cruz.
- El malentendido común de LLN es que se supone que los procesos aleatorios compensan lo que sucedió en el pasado; esto simplemente no es cierto y también se llama falacia del jugador (o ley de los promedios).

Resultados disjuntos y no disjuntos

Resultados disjuntos (mutuamente excluyentes): No pueden ocurrir al mismo tiempo.

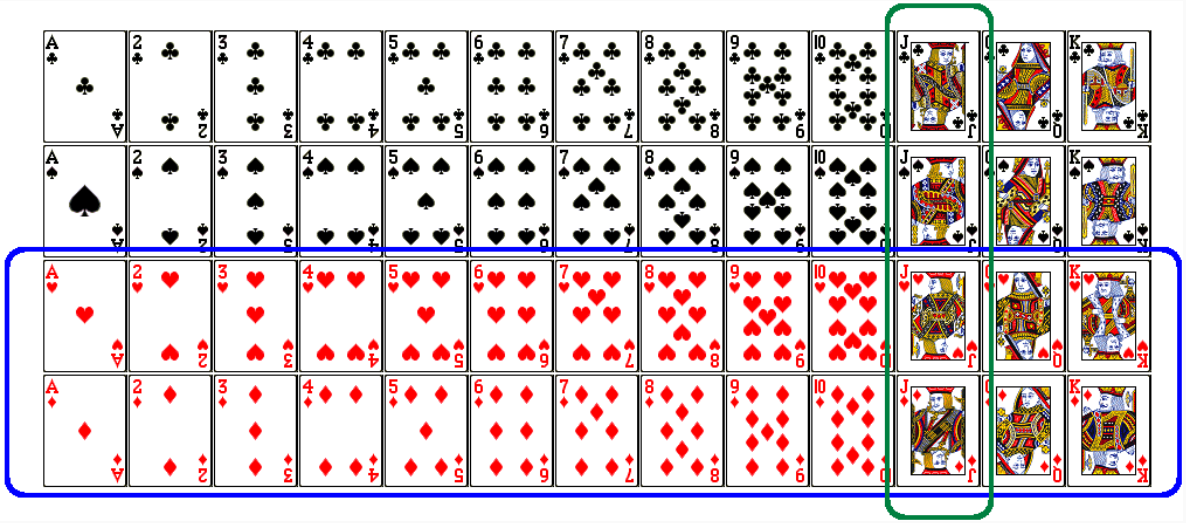
- El resultado de un lanzamiento de una sola moneda no puede ser cara y cruz.
- Un estudiante no puede reprobar y aprobar una clase.
- Una sola carta extraída de una baraja no puede ser un as y una reina.

Resultados no disjuntos: Pueden ocurrir al mismo tiempo.

- Un estudiante puede obtener una A en Estadísticas y una A en Econ en el mismo semestre.

Unión de eventos no disjuntos

¿Cuál es la probabilidad de sacar una jota o una carta roja de un mazo completo bien barajado?



$$\begin{aligned} \text{P(jota o roja)} &= \text{P(jota)} + \text{P(roja)} - \text{P(jota y roja)} \\ &= \frac{4}{52} + \frac{26}{52} - \frac{2}{52} = \frac{28}{52} \end{aligned}$$

Cifra de http://www.milefoot.com/math/discrete/counting/cardfreq.htm.

Práctica

¿Cuál es la probabilidad de que un estudiante seleccionado al azar piense que la marihuana debería legalizarse o esté de acuerdo con las opiniones políticas de sus padres?

	Compartir la política de los padres		
Legalizar MJ	No	Sí	Total
No	11	40	51
Sí	36	78	114
Total	47	118	165

(a) $\frac{40+36-78}{165}$

(b) $\frac{114+118-78}{165}$

(c) $\frac{78}{165}$

(d) $\frac{78}{188}$

(e) $\frac{11}{47}$

Práctica

¿Cuál es la probabilidad de que un estudiante seleccionado al azar piense que la marihuana debería legalizarse o esté de acuerdo con las opiniones políticas de sus padres?

	Compartir la política de los padres		
Legalizar MJ	No	Sí	Total
No	11	40	51
Sí	36	78	114
Total	47	118	165

(a) $\frac{40+36-78}{165}$

(b) $\frac{114+118-78}{165}$

(c) $\frac{78}{165}$

(d) $\frac{78}{188}$

(e) $\frac{11}{47}$

Resumen

Regla de adición general

$$P(A \text{ o } B) = P(A) + P(B) - P(A \text{ y } B)$$

Note: Para eventos disjuntos $P(A \text{ y } B) = 0$, entonces la fórmula anterior se simplifica a $P(A \text{ o } B) = P(A) + P(B)$.

Distribuciones de probabilidad

Una distribución de probabilidad enumera todos los eventos posibles y las probabilidades con las que ocurren.

- La distribución de probabilidad para el género de un niño:

Evento	Masculino	Femenino
Probabilidad	0.5	0.5

- Reglas para distribuciones de probabilidad:
  1. Los eventos listados deben ser disjuntos
  2. Cada probabilidad debe estar entre 0 y 1
  3. Las probabilidades deben sumar 1
- La distribución de probabilidad para los géneros de dos niños:

Evento	MM	FF	MF	FM
Probabilidad	0.25	0.25	0.25	0.25

Práctica

En una encuesta, el 52% de los encuestados dijeron que son demócratas. ¿Cuál es la probabilidad de que un encuestado seleccionado al azar de esta muestra sea republicano?

(a) 0.48

(b) más de 0.48

(c) menos de 0.48

(d) no se puede calcular usando solo la información dada

Si los únicos dos partidos políticos son republicano y demócrata, entonces (a) es posible. Sin embargo, también es posible que algunas personas no se afilien a un partido político o se afilien a un partido que no sean estos dos. Entonces (c) también es posible. Sin embargo, (b) definitivamente no es posible ya que daría como resultado que la probabilidad total de que el espacio muestral sea superior a 1.

Espacio muestral y complementos

Espacio muestral es la colección de todos los posibles resultados de un ensayo.

- Una pareja tiene un hijo, ¿cuál es el espacio muestral para el género de este hijo? $S = \{M, F\}$
- Una pareja tiene dos hijos, ¿cuál es el espacio muestral para el género de estos niños? $S = \{MM, FF, FM, MF\}$

Eventos complementarios son dos eventos mutuamente excluyentes cuyas probabilidades suman 1.

- Una pareja tiene un hijo. Si sabemos que el niño no es un niño, ¿cuál es el género de este niño? { ~~M~~, F } → Niño y niña son resultados complementarios.
- Una pareja tiene dos hijos, si sabemos que ambos no son niñas, ¿cuáles son las posibles combinaciones de género para estos niños? { MM, ~~FF~~, FM, MF }

Independencia

Dos procesos son independientes si conocer el resultado de uno no proporciona información útil sobre el resultado del otro.

- Saber que la moneda cayó en una cara en el primer lanzamiento no proporciona ninguna información útil para determinar en qué caerá la moneda en el segundo lanzamiento. → Los resultados de dos lanzamientos de una moneda son independientes.
- Saber que la primera carta extraída de una baraja es un as sí proporciona información útil para determinar la probabilidad de sacar un as en el segundo sorteo. → Los resultados de dos sorteos de una baraja de cartas (sin reemplazo) son dependientes.

Práctica

Entre el 9 y el 12 de enero de 2013, SurveyUSA entrevistó a una muestra aleatoria de 500 residentes de NC y les preguntó si creían que la posesión generalizada de armas protege a los ciudadanos respetuosos de la ley del crimen o hace que la sociedad sea más peligrosa. El 58% de todos los encuestados dijo que protege a los ciudadanos. El 67% de los encuestados blancos, el 28% de los encuestados negros y el 64% de los encuestados hispanos comparten esta opinión. ¿Cuál de los siguientes es cierto?

La opinión sobre la posesión de armas y el origen étnico racial es más probable

(a) complementario

(b) mutuamente excluyentes

(c) independiente

(d) dependiente

(e) disjunto

Práctica

Entre el 9 y el 12 de enero de 2013, SurveyUSA entrevistó a una muestra aleatoria de 500 residentes de NC y les preguntó si creían que la posesión generalizada de armas protege a los ciudadanos respetuosos de la ley del crimen o hace que la sociedad sea más peligrosa. El 58% de todos los encuestados dijo que protege a los ciudadanos. El 67% de los encuestados blancos, el 28% de los encuestados negros y el 64% de los encuestados hispanos comparten esta opinión. ¿Cuál de los siguientes es cierto?

La opinión sobre la posesión de armas y el origen étnico racial es más probable

(a) complementario

(b) mutuamente excluyentes

(c) independiente

(d) dependiente

(e) disjunto

Comprobación de la independencia

Si P(A ocurre, dado que B es verdadera) = P(A | B) = P(A), entonces A y B son independientes.

P(protege a los ciudadanos) = 0.58

P(residente de Carolina del Norte seleccionado al azar dice que la posesión de armas protege a los ciudadanos, dado que el residente es blanco) = P(protege a los ciudadanos | blancos) = 0,67

P(protege a los ciudadanos | Negro) = 0.28

P(protege a los ciudadanos | hispanos) = 0,64

P(protege a los ciudadanos) varía según la raza/origen étnico, por lo tanto, la opinión sobre la propiedad de armas y el origen étnico de la raza probablemente dependa.

Determinación de la dependencia basada en datos de muestra

- Si las probabilidades condicionales calculadas en base a datos de muestra sugieren dependencia entre dos variables, el siguiente paso es realizar una prueba de hipótesis para determinar si la diferencia observada entre las probabilidades es probable o improbable que haya ocurrido por casualidad.
- Si la diferencia observada entre las probabilidades condicionales es grande, entonces hay evidencia más fuerte de que la diferencia es real.
- Si una muestra es grande, incluso una pequeña diferencia puede proporcionar una fuerte evidencia de una diferencia real.

Vimos que P(protege a los ciudadanos | blancos) = 0.67 y P(protege a los ciudadanos | hispanos) = 0.64. ¿Bajo qué condición estaría más convencido de una diferencia real entre las proporciones de blancos e hispanos que piensan que la posesión generalizada de armas protege a los ciudadanos? n = 500 o n = 50000

n = 50000

Regla del producto para eventos independientes

$$\mathrm{P(A\ y\ B) = P(A) \times P(B)}$$

O más generalmente, $\mathrm{P(A_1\ y\ \cdots\ y\ A_k) = P(A_1) \times \cdots \times P(A_k)}$

Si lanzas una moneda dos veces, ¿cuál es la probabilidad de obtener dos sellos seguidos?

$$\mathrm{P(T\ en\ el\ primer\ lanzamiento) \times P(T\ en\ el\ segundo\ lanzamiento)} = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

Práctica

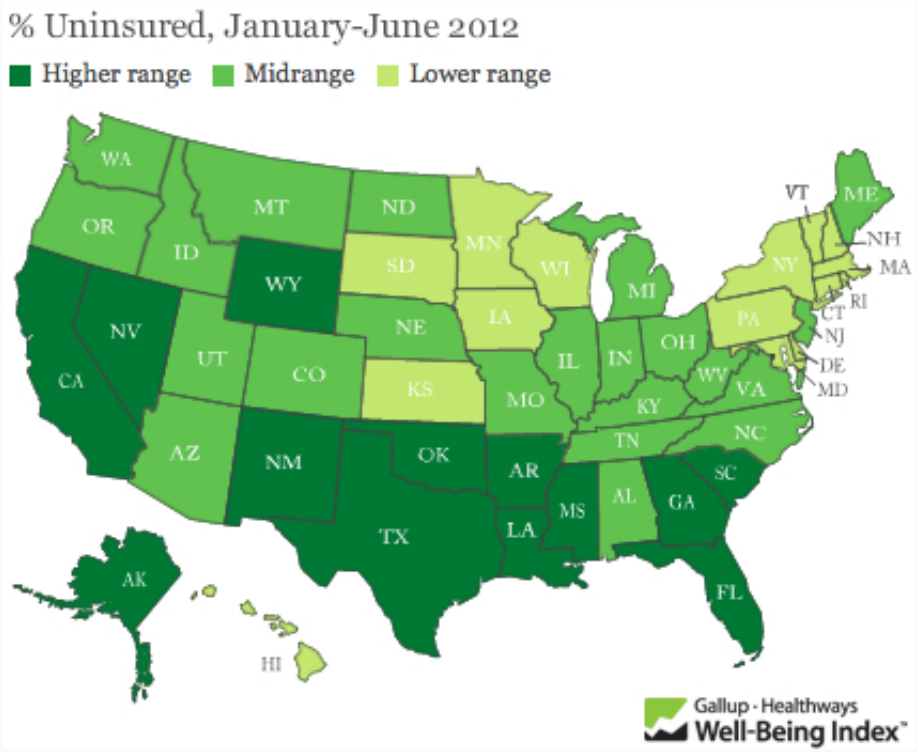
Una encuesta reciente de Gallup sugiere que el 25,5% de los tejanos no tienen seguro médico en junio de 2012. Suponiendo que la tasa de personas sin seguro se mantuvo constante, ¿cuál es la probabilidad de que dos tejanos seleccionados al azar no tengan seguro?

(a) 25.5^2

(b) 0.255^2

(c) 0.255×2

(d) $(1 - 0.255)^2$



http://www.gallup.com/poll/156851/uninsured-rate-stable-across-states-far-2012.aspx

Práctica

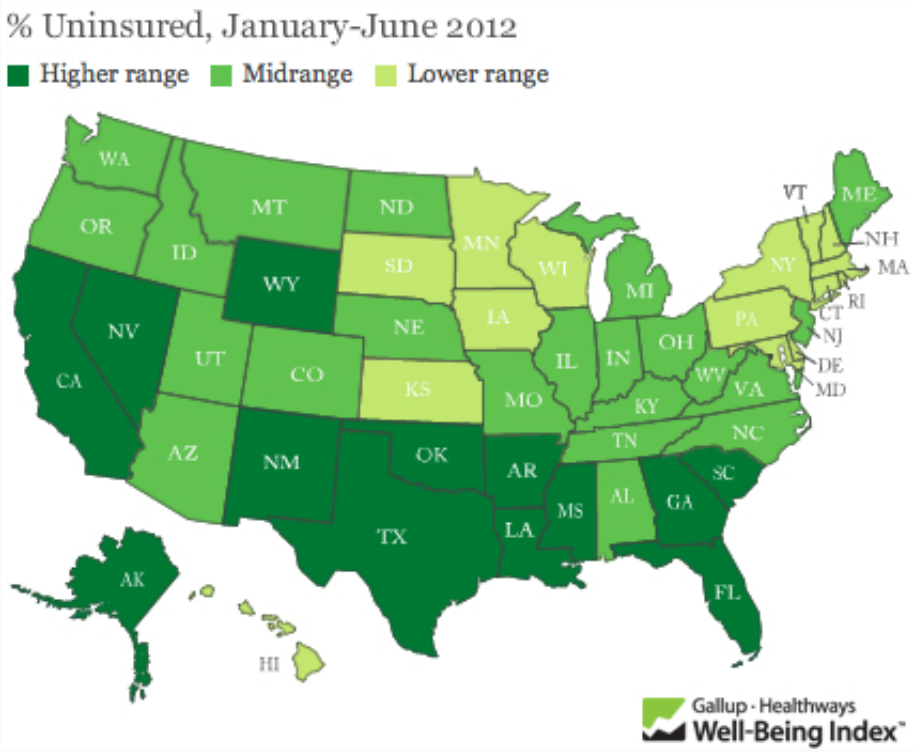
Una encuesta reciente de Gallup sugiere que el 25,5% de los tejanos no tienen seguro médico en junio de 2012. Suponiendo que la tasa de personas sin seguro se mantuvo constante, ¿cuál es la probabilidad de que dos tejanos seleccionados al azar no tengan seguro?

(a) 25.5^2

(b) 0.255^2

(c) 0.255×2

(d) $(1 - 0.255)^2$



http://www.gallup.com/poll/156851/uninsured-rate-stable-across-states-far-2012.aspx

Disjuntos vs. complementarios

¿La suma de probabilidades de dos eventos disjuntos siempre suman 1?

No necesariamente, puede haber más de 2 eventos en el espacio muestral, p. Fiesta de afiliación.

¿La suma de probabilidades de dos eventos complementarios siempre suman 1?

Sí, esa es la definición de complementario, p. Cabeza y cola.

Poniendo todo junto...

Si tuviéramos que seleccionar al azar a 5 tejanos, ¿cuál es la probabilidad de que al menos uno no tenga seguro?

- Si tuviéramos que seleccionar al azar 5 tejanos, el espacio de muestra para el número de tejanos que no tienen seguro sería:

$$S = \{0, 1, 2, 3, 4, 5\}$$

- Nos interesan los casos en los que al menos una persona no tiene seguro:

$$S = \{0, 1, 2, 3, 4, 5\}$$

- Entonces podemos dividir el espacio muestral en dos categorías:

$$S = \{0, \text{al menos uno}\}$$

Poniendo todo junto...

Como la probabilidad del espacio muestral debe sumar 1:

$$\begin{aligned}
\text{Prob(al menos 1 sinseguro)} &= 1 - \text{Prob(ninguno sinseguro)} \\
&= 1 - [(1 - 0.255)^5] \\
&= 1 - 0.745^5 \\
&= 1 - 0.23 \\
&= 0.77
\end{aligned}$$

Al menos 1

$$\text{P(al menos uno)} = 1 - \text{P(ninguno)}$$

Práctica

Alrededor del 20% de los estudiantes universitarios son vegetarianos o veganos. ¿Cuál es la probabilidad de que, entre una muestra aleatoria de 3 estudiantes universitarios, al menos uno sea vegetariano o vegano?

(a) $1 - 0,2 \times 3$

(b) $1 - 0.2^3$

(c) 0.8^3

(d) $1 - 0,8 \times 3$

(e) $1 - 0.8^3$

P(al menos 1 de verduras) =
$1 - \text{P(ninguna verduras)}$
$= 1 - (1 - 0.2)^3$
$= 1 - 0.8^3$
$= 1 - 0.512 = 0.488$

Práctica

Alrededor del 20% de los estudiantes universitarios son vegetarianos o veganos. ¿Cuál es la probabilidad de que, entre una muestra aleatoria de 3 estudiantes universitarios, al menos uno sea vegetariano o vegano?

(a) $1 - 0,2 \times 3$

(b) $1 - 0.2^3$

(c) 0.8^3

(d) $1 - 0,8 \times 3$

(e) $1 - 0.8^3$

P(al menos 1 de verduras) =
1 – P(ninguna verduras)
$= 1 - (1 - 0.2)^3$
$= 1 - 0.8^3$
$= 1 - 0.512 = 0.488$

Probabilidad condicional

Recaída

Los investigadores asignaron al azar a 72 usuarios crónicos de cocaína en tres grupos: desipramina (antidepresivo), litio (tratamiento estándar para la cocaína) y placebo. Los resultados del estudio se resumen a continuación.

	recaída	no recaída	total
desipramina	10	14	24
litio	18	6	24
placebo	20	4	24
total	48	24	72

http://www.oswego.edu/~srp/stats/2_way_tbl_1.htm

¿Cuál es la probabilidad de que un paciente no recaiga?

	recaída	no recaída	total
desipramina	10	14	24
litio	18	6	24
placebo	20	4	24
total	48	24	72

Probabilidad marginal

¿Cuál es la probabilidad de que un paciente recaiga?

	recaída	no recaída	total
desipramina	10	14	24
litio	18	6	24
placebo	20	4	24
total	48	24	72

Probabilidad marginal

¿Cuál es la probabilidad de que un paciente recaiga?

	recaída	no recaída	total
desipramina	10	14	24
litio	18	6	24
placebo	20	4	24
total	48	24	72

$$\text{P(recaída)} = \frac{48}{72} \approx 0.67$$

Probabilidad conjunta

¿Cuál es la probabilidad de que un paciente reciba el antidepresivo (desipramina) y recaiga?

	recaída	no recaída	total
desipramina	10	14	24
litio	18	6	24
placebo	20	4	24
total	48	24	72

Probabilidad conjunta

¿Cuál es la probabilidad de que un paciente reciba el antidepresivo (desipramina) y recaiga?

	recaída	no recaída	total
desipramina	10	14	24
litio	18	6	24
placebo	20	4	24
total	48	24	72

$$\text{P(recaída y desipramina)} = \frac{10}{72} \approx 0.14$$

Probabilidad condicional

La probabilidad condicional del resultado del interés A dada la condición B se calcula como

$$P(A|B) = \frac{P(A \text{ y } B)}{P(B)}$$

	recaída	no recaída	total
desipramina	10	14	24
litio	18	6	24
placebo	20	4	24
total	48	24	72

$$\begin{aligned} &P(\text{recada}|\text{desipramina}) \\ &= \frac{P(\text{recada y desipramina})}{P(\text{desipramina})} \\ &= \frac{10/72}{24/72} \\ &= \frac{10}{24} \\ &= 0,42 \end{aligned}$$

Probabilidad condicional (cont.)

Si sabemos que un paciente recibió el antidepresivo (desipramina), ¿cuál es la probabilidad de que recaiga?

	recaída	no recaída	total
desipramina	10	14	24
litio	18	6	24
placebo	20	4	24
total	48	24	72

$$\text{P}(\text{recaída} \mid \text{desipramina}) = \frac{10}{24} \approx 0.42$$

$$\text{P}(\text{recaída} \mid \text{litio}) = \frac{18}{24} \approx 0.75$$
$$\text{P}(\text{recaída} \mid \text{placebo}) = \frac{20}{24} \approx 0,83$$

Probabilidad condicional (cont.)

Si sabemos que un paciente recayó, ¿cuál es la probabilidad de que haya recibido el antidepresivo (desipramina)?

	recaída	no recaída	total
desipramina	10	14	24
litio	18	6	24
placebo	20	4	24
total	48	24	72

$$\text{P}(\text{desipramina} \mid \text{recaída}) = \frac{10}{48} \approx 0.21$$

$$\text{P}(\text{litio} \mid \text{recaída}) = \frac{18}{48} \approx 0.375$$
$$\text{P}(\text{placebo} \mid \text{recaída}) = \frac{20}{48} \approx 0,42$$

Regla general de multiplicación

- Anteriormente vimos que si dos eventos son independientes, su probabilidad conjunta es simplemente el producto de sus probabilidades. Si no se cree que los eventos sean independientes, la probabilidad conjunta se calcula de forma ligeramente diferente.
- Si A y B representan dos resultados o eventos, entonces

$$P(A \text{ y } B) = P(A|B) \times P(B)$$

  Tenga en cuenta que esta fórmula es simplemente la fórmula de probabilidad condicional, reorganizada.
- Es útil pensar en A como el resultado de interés y en B como la condición.

Independencia y probabilidades condicionales

Considere la siguiente distribución (hipotética) de género y especialidad de los estudiantes en una clase de introducción a la estadística:

	social ciencia	no social ciencia	total
Mujer	30	20	50
hombre	30	20	50
total	60	40	100

- La probabilidad de que un estudiante seleccionado al azar se especialice en ciencias sociales es $\frac{60}{100} = 0.6$.
- La probabilidad de que un estudiante seleccionado al azar se especialice en ciencias sociales dado que es mujer es $\frac{30}{50} = 0.6$.
- Dado que P(SS|M) también es igual a 0,6, la concentración de los estudiantes de esta clase no depende de su género: $P(SS \mid F) = P(SS)$.

Independencia y probabilidades condicionales (cont.)

Genéricamente, si P(A|B) = P(A) entonces se dice que los eventos A y B son independientes.

- Conceptualmente: Dar B no nos dice nada acerca de A.
- Matemáticamente: Sabemos que si los eventos A y B son independientes, P(A y B) = P(A) × P(B). Después,

$$\mathrm{P(A|B)} = \frac{\mathrm{P(A\ y\ B)}}{\mathrm{P(B)}} = \frac{\mathrm{P(A)} \times \mathrm{P(B)}}{\mathrm{P(B)}} = \mathrm{P(A)}$$

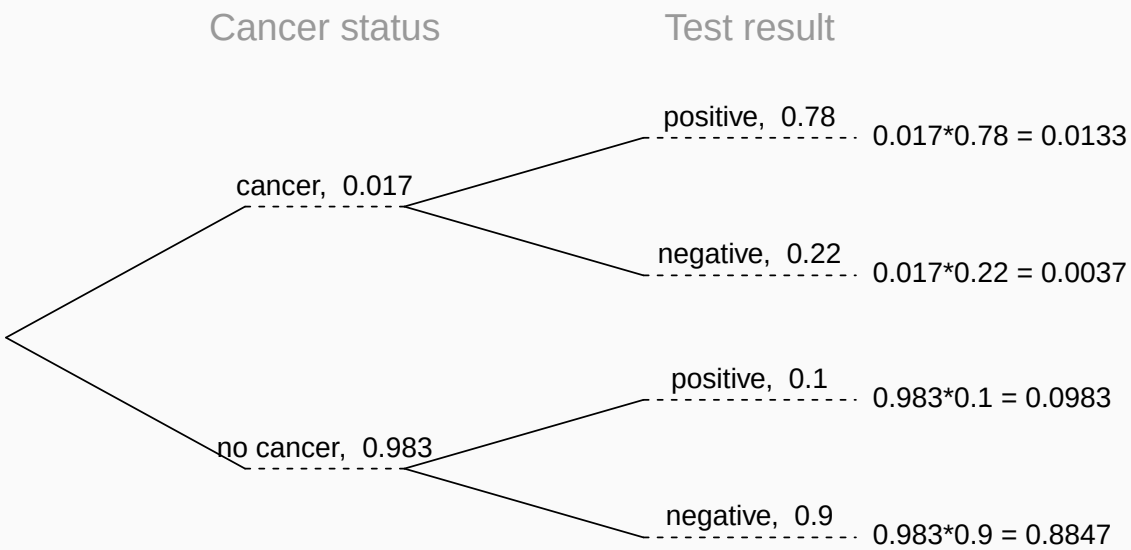
Exámenes de detección de cáncer de mama

- La Sociedad Americana del Cáncer estima que alrededor del 1,7% de las mujeres tienen cáncer de mama.
  http://www.cancer.org/cancer/cancerbasics/cancer-prevalence
- Susan G. Komen For The Cure Foundation afirma que la mamografía identifica correctamente alrededor del 78% de las mujeres que realmente tienen cáncer de mama.
  http:
  //ww5.komen.org/BreastCancer/AccuracyofMammograms.html
- Un artículo publicado en 2003 sugiere que hasta el 10% de todas las mamografías dan falsos positivos en pacientes que no tienen cáncer.
  http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1360940

Note: Estos porcentajes son aproximados y muy difíciles de estimar.

Invertir probabilidades

Cuando una paciente se somete a un examen de detección de cáncer de mama, hay dos afirmaciones contrapuestas: la paciente tenía cáncer y la paciente no tiene cáncer. Si una mamografía arroja un resultado positivo, ¿cuál es la probabilidad de que el paciente realmente tenga cáncer?



$$
\begin{aligned}
&\mathrm{P}(\mathrm{C}|+) \\
&= \frac{\mathrm{P}(\mathrm{C}\ \mathrm{y}\ +)}{\mathrm{P}(+)} \\
&= \frac{0.0133}{0.0133 + 0.0983} \\
&= 0.12
\end{aligned}
$$

Note: Los diagramas de árbol son útiles para invertir probabilidades: nos dan P(+|C) y nos piden P(C|+).

Práctica

Supongamos que una mujer que se hace la prueba una vez y obtiene un resultado positivo quiere volver a hacerse la prueba. En la segunda prueba, ¿cuál deberíamos suponer que es la probabilidad de que esta mujer específica tenga cáncer?

(a) 0.017

(b) 0.12

(c) 0.0133

(d) 0.88

Práctica

Supongamos que una mujer que se hace la prueba una vez y obtiene un resultado positivo quiere volver a hacerse la prueba. En la segunda prueba, ¿cuál deberíamos suponer que es la probabilidad de que esta mujer específica tenga cáncer?

(a) 0.017

(b) 0.12

(c) 0.0133

(d) 0.88

Práctica

¿Cuál es la probabilidad de que esta mujer tenga cáncer si esta segunda mamografía también dio un resultado positivo?

(a) 0.0936

(b) 0.088

(c) 0.48

(d) 0.52

Práctica

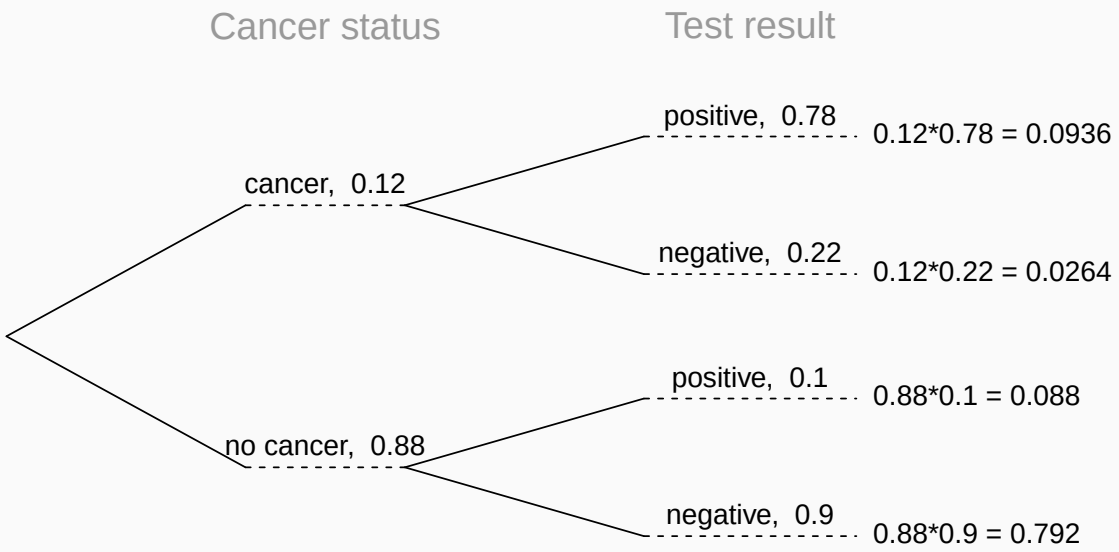
¿Cuál es la probabilidad de que esta mujer tenga cáncer si esta segunda mamografía también dio un resultado positivo?

(a) 0.0936

(b) 0.088

(c) 0.48

(d) 0.52



Práctica

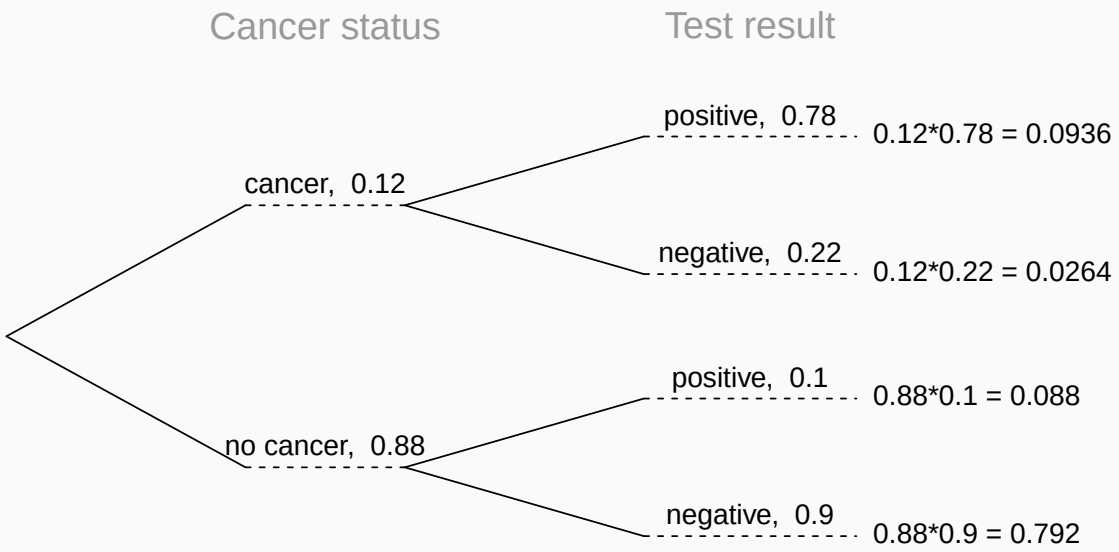
¿Cuál es la probabilidad de que esta mujer tenga cáncer si esta segunda mamografía también dio un resultado positivo?

(a) 0.0936

(b) 0.088

(c) 0.48

(d) 0.52



$$P(C|+) = \frac{P(C \text{ y } +)}{P(+)} = \frac{0.0936}{0.0936 + 0.088} = 0.52$$

Teorema de Bayes

- La fórmula de probabilidad condicional que hemos visto hasta ahora es un caso especial del Teorema de Bayes, que es aplicable incluso cuando los eventos tienen más de dos resultados.
- Teorema de Bayes:

$$\begin{aligned}&\mathrm{P}(\text{resultado } \mathrm{A}_1 \text{ de variable 1} \mid \text{resultado B de variable 2})\\ &= \frac{\mathrm{P}(\mathrm{B}|\mathrm{A}_1)\mathrm{P}(\mathrm{A}_1)}{\mathrm{P}(\mathrm{B}|\mathrm{A}_1)\mathrm{P}(\mathrm{A}_1) + \mathrm{P}(\mathrm{B}|\mathrm{A}_2)\mathrm{P}(\mathrm{A}_2) + \cdots + \mathrm{P}(\mathrm{B}|\mathrm{A}_k)\mathrm{P}(\mathrm{A}_k)}\end{aligned}$$

donde $\mathrm{A}_2, \cdots, \mathrm{A}_k$ representan todos los demás resultados posibles de la variable 1.

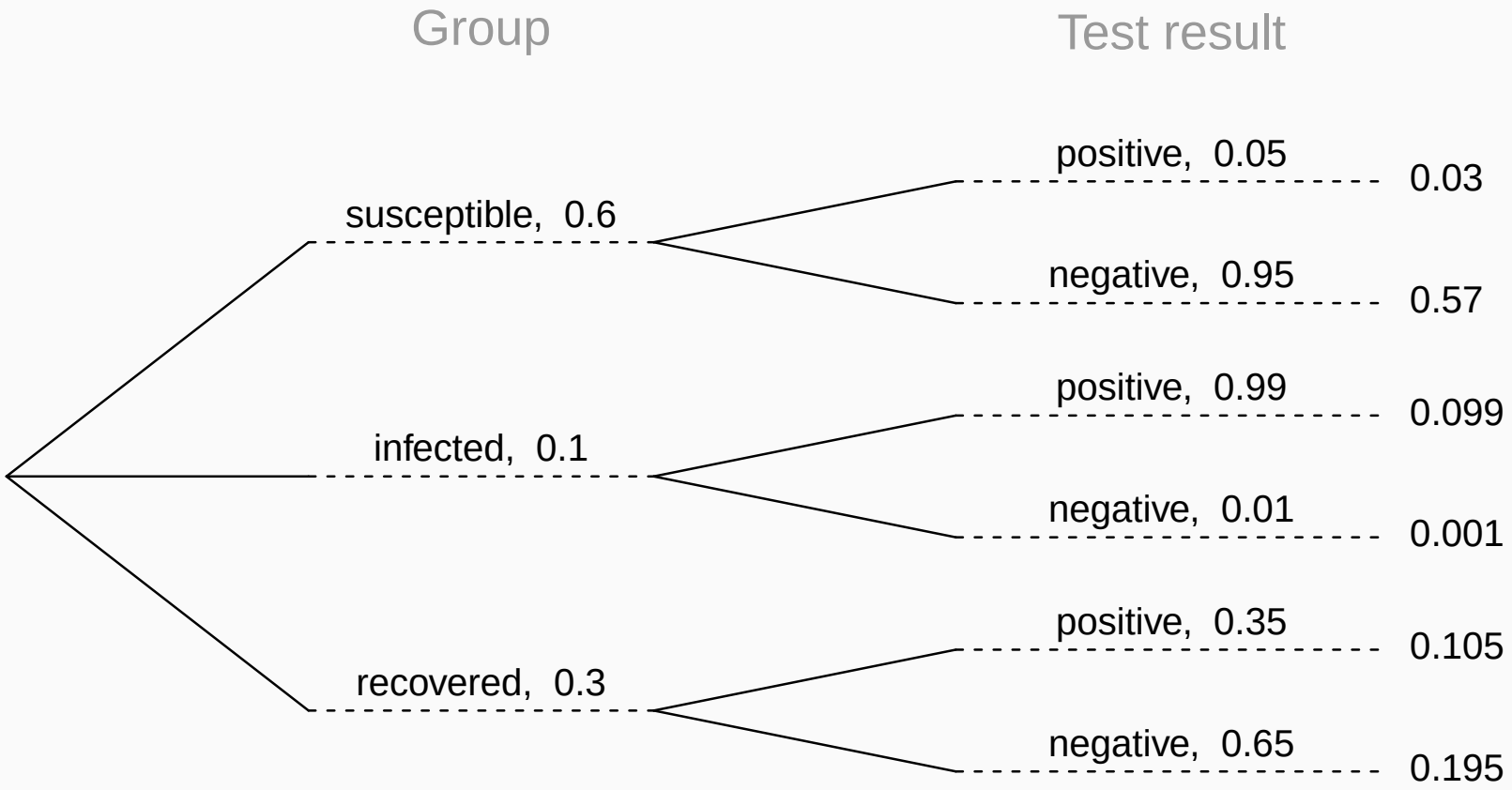
Actividad de la aplicación: Invertir probabilidades

Un modelo epidemiológico común para la propagación de enfermedades es el modelo SIR, donde la población se divide en tres grupos: susceptible, infectado y recuperado. Este es un modelo razonable para enfermedades como la varicela, donde una sola infección suele proporcionar inmunidad a infecciones posteriores. A veces, estas enfermedades también pueden ser difíciles de detectar.

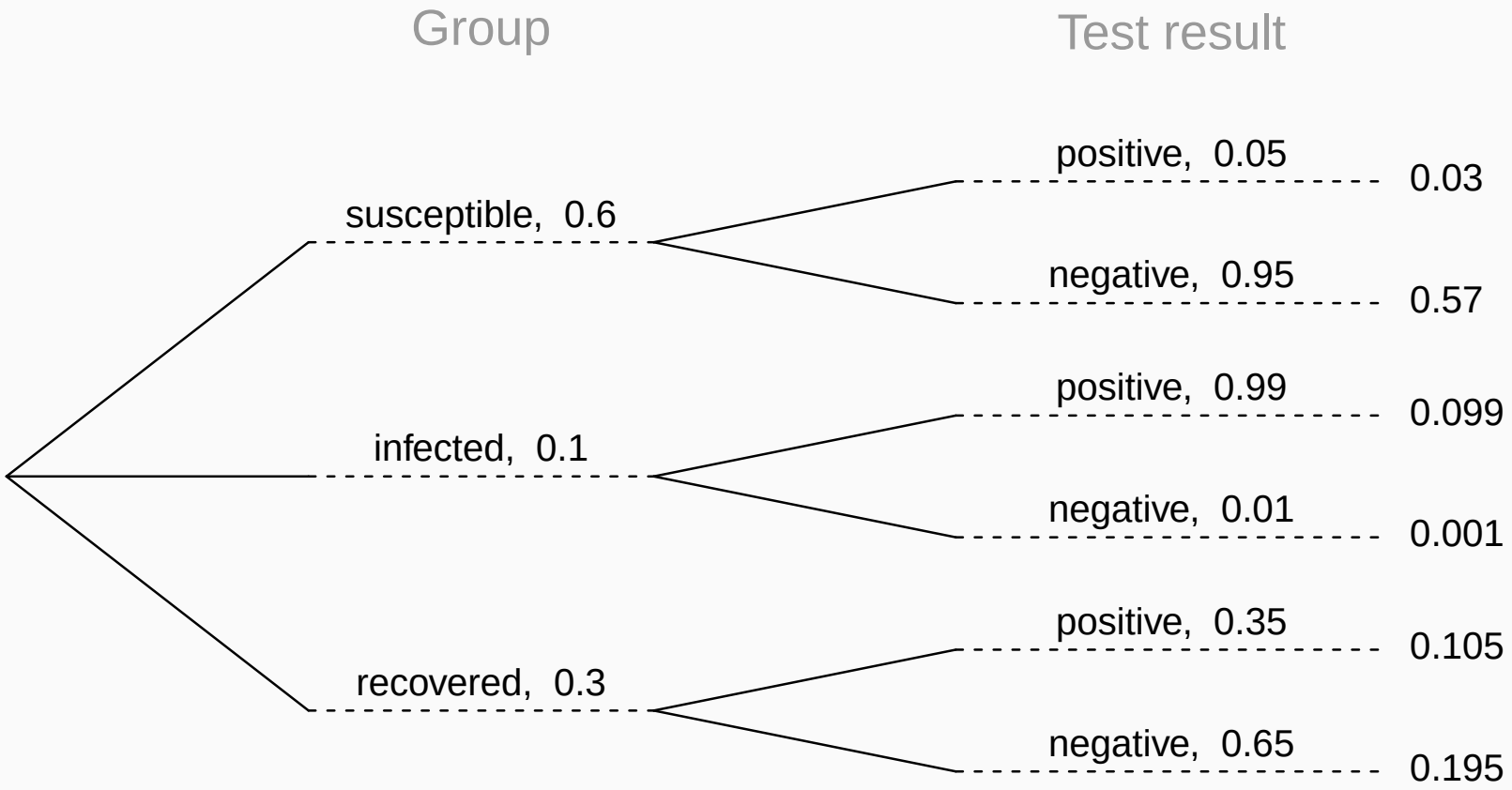
Imagina una población en medio de una epidemia donde el 60% de la población se considera susceptible, el 10% está infectado y el 30% está recuperado. La única prueba de la enfermedad es precisa el 95% de las veces para individuos susceptibles, el 99% para individuos infectados, pero el 65% para individuos recuperados. (Nota: en este caso, preciso significa devolver un resultado negativo para individuos susceptibles y recuperados y un resultado positivo para individuos infectados).

Dibuje un árbol de probabilidad para reflejar la información dada anteriormente. Si el individuo ha dado positivo, ¿cuál es la probabilidad de que realmente esté infectado?

Actividad de la aplicación: Inversión de probabilidades (cont.)



Actividad de la aplicación: Inversión de probabilidades (cont.)



$$\mathrm{P}(\mathrm{inf}|+) = \frac{\mathrm{P}(\mathrm{inf\ y\ }+)}{\mathrm{P}(+)} = \frac{0.099}{0.03 + 0.099 + 0.105} \approx 0.423$$

Muestreo de una pequeña población

Muestreo con reemplazo

Al muestrear con reemplazo, vuelve a colocar lo que acaba de dibujar.

- Imagina que tienes una bolsa con 5 fichas rojas, 3 azules y 2 naranjas. ¿Cuál es la probabilidad de que la primera ficha que saques sea azul?

$$5 \text{ 🔴} , 3 \text{ 🔵} , 2 \text{ 🟠}$$

$$\text{Prob}(1^{\text{st}} \text{ chip B}) = \frac{3}{5+3+2} = \frac{3}{10} = 0,3$$

- Suponga que de hecho sacó una ficha azul en el primer sorteo. Si saca con reemplazo, ¿cuál es la probabilidad de sacar una ficha azul en el segundo sorteo?

$$1^{\text{st}} \text{ dibujar: } 5 \text{ 🔴} , 3 \text{ 🔵} , 2 \text{ 🟠}$$
$$2^{\text{nd}} \text{ dibuja: } 5 \text{ 🔴} , 3 \text{ 🔵} , 2 \text{ 🟠}$$

$$\text{Prob}(2^{\text{nd}} \text{ chip B} | 1^{\text{st}} \text{ chip B}) = \frac{3}{10} = 0.3$$

Muestreo con reemplazo (cont.)

- Supongamos que en realidad sacaste una ficha naranja en el primer sorteo. Si saca con reemplazo, ¿cuál es la probabilidad de sacar una ficha azul en el segundo sorteo?

1^{st} dibujar: 5 🔴 , 3 🔵 , 2 🟠
2^{nd} dibuja: 5 🔴 , 3 🔵 , 2 🟠

$$\text{Prob}(2^{\text{nd}} \text{ chip B} | 1^{\text{st}} \text{ chip O}) = \frac{3}{10} = 0.3$$

- Si saca con reemplazo, ¿cuál es la probabilidad de sacar dos fichas azules seguidas?

1^{st} dibujar: 5 🔴 , 3 🔵 , 2 🟠
2^{nd} dibuja: 5 🔴 , 3 🔵 , 2 🟠

$$\text{Prob}(1^{\text{st}} \text{ chip B}) \cdot \text{Prob}(2^{\text{nd}} \text{ chip B} | 1^{\text{st}} \text{ chip B}) = 0,3 \times 0,3$$
$$= 0,3^2 = 0,09$$

Muestreo con reemplazo (cont.)

- Cuando se extrae con reemplazo, la probabilidad de que la segunda ficha sea azul no depende del color de la primera ya que todo lo que extraigamos en la primera extracción se devuelve a la bolsa.

$$\text{Prob(B|B)} = \text{Prob(B|O)}$$

- Además, esta probabilidad es igual a la probabilidad de sacar una ficha azul en el primer sorteo, ya que la composición de la bolsa nunca cambia cuando se muestrea con reemplazo.

$$\text{Prob(B|B)} = \text{Prob(B)}$$

- Al dibujar con reemplazo, los sorteos son independientes.

Muestreo sin reemplazo

Cuando dibujas sin reemplazo no vuelves a colocar lo que acabas de dibujar.

- Suponga que sacó una ficha azul en el primer sorteo. Si saca sin reemplazo, ¿cuál es la probabilidad de sacar una ficha azul en el segundo sorteo?

1^{st} dibujar: 5 🔴 , 3 🔵 , 2 🟠
2^{nd} dibuja: 5 🔴 , 2 🔵 , 2 🟠

$$\text{Prob}(2^{\text{nd}} \text{ chip B} | 1^{\text{st}} \text{ chip B}) = \frac{2}{9} = 0,22$$

- Si saca sin reemplazo, ¿cuál es la probabilidad de sacar dos fichas azules seguidas?

1^{st} dibujar: 5 🔴 , 3 🔵 , 2 🟠
2^{nd} dibuja: 5 🔴 , 2 🔵 , 2 🟠

$$\text{Prob}(1^{\text{st}} \text{ chip B}) \cdot \text{Prob}(2^{\text{nd}} \text{ chip B} | 1^{\text{st}} \text{ chip B}) = 0,3 \times 0,22$$
$$= 0.066$$

Muestreo sin reemplazo (cont.)

- Cuando se extrae sin reemplazo, la probabilidad de que la segunda ficha sea azul dado que la primera fue azul no es igual a la probabilidad de sacar una ficha azul en el primer sorteo, ya que la composición de la bolsa cambia con el resultado del primer sorteo.

$$\text{Prob(B|B)} \neq \text{Prob(B)}$$

Muestreo sin reemplazo (cont.)

- Cuando se extrae sin reemplazo, la probabilidad de que la segunda ficha sea azul dado que la primera fue azul no es igual a la probabilidad de sacar una ficha azul en el primer sorteo, ya que la composición de la bolsa cambia con el resultado del primer sorteo.

$$\text{Prob(B|B)} \neq \text{Prob(B)}$$

- Al dibujar sin reemplazo, los sorteos no son independientes.

Muestreo sin reemplazo (cont.)

- Cuando se extrae sin reemplazo, la probabilidad de que la segunda ficha sea azul dado que la primera fue azul no es igual a la probabilidad de sacar una ficha azul en el primer sorteo, ya que la composición de la bolsa cambia con el resultado del primer sorteo.

$$\text{Prob(B|B)} \neq \text{Prob(B)}$$

- Al dibujar sin reemplazo, los sorteos no son independientes.
- Es especialmente importante tomar nota de esto cuando los tamaños de muestra son pequeños. Si estuviéramos tratando con, digamos, 10,000 fichas en una bolsa (gigante), sacar una ficha de cualquier color no tendría un impacto tan grande en las probabilidades en el segundo sorteo.

Práctica

En la mayoría de los juegos de cartas, las cartas se reparten sin reemplazo. ¿Cuál es la probabilidad de recibir un as y luego un 3? Elija la respuesta más cercana.

(a) 0.0045

(b) 0.0059

(c) 0.0060

(d) 0.1553

Práctica

En la mayoría de los juegos de cartas, las cartas se reparten sin reemplazo. ¿Cuál es la probabilidad de recibir un as y luego un 3? Elija la respuesta más cercana.

(a) 0.0045

(b) 0.0059

(c) 0.0060

(d) 0.1553

$$\text{P(as entonces 3)} = \frac{4}{52} \times \frac{4}{51} \approx 0,0060$$

Variables aleatorias

Variables aleatorias

- Una variable aleatoria es una cantidad numérica cuyo valor depende del resultado de un evento aleatorio
  - Usamos una letra mayúscula, como X, para denotar una variable aleatoria
  - Los valores de una variable aleatoria se denotan con una letra minúscula, en este caso x
  - Por ejemplo, P(X = x)
- Hay dos tipos de variables aleatorias:
  - Variables aleatorias discretas a menudo toman solo valores enteros
    - Ejemplo: número de horas de crédito, diferencia en el número de horas de crédito este período frente al último
  - Variables aleatorias continuas toman valores reales (decimales)
    - Ejemplo: Costo de los libros este término, Diferencia en el costo de los libros este término vs el último

Expectativa

- A menudo nos interesa el resultado promedio de una variable aleatoria.
- A esto lo llamamos el valor esperado (media), y es un promedio ponderado de los posibles resultados

$$\mu = \mathrm{E(X)} = \sum_{\mathrm{i=1}}^{\mathrm{k}} \mathrm{x_i}\ \mathrm{P(X = x_i)}$$

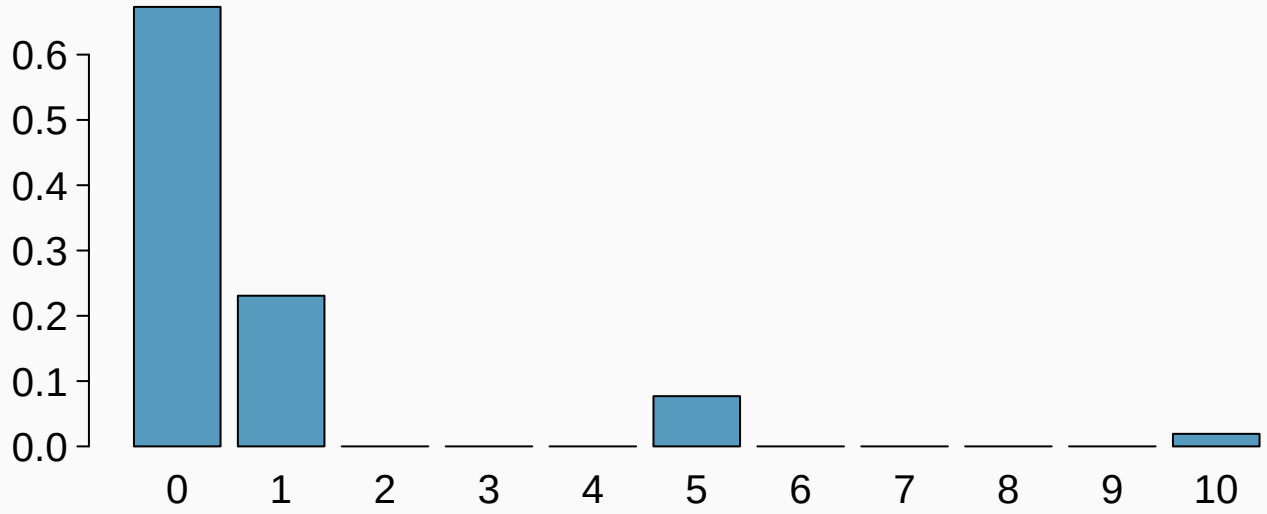
Valor esperado de una variable aleatoria discreta

En un juego de cartas ganas \$1 si sacas un corazón, \$5 si sacas un as (incluido el as de corazones), \$10 si sacas el rey de picas y nada por cualquier otra carta que saques . Escriba el modelo de probabilidad para sus ganancias y calcule su ganancia esperada.

Evento	X	P(X)	X P(X)
Corazón (no as)	1	$\frac{12}{52}$	$\frac{12}{52}$
As	5	$\frac{4}{52}$	$\frac{20}{52}$
Rey de picas	10	$\frac{1}{52}$	$\frac{10}{52}$
Todo lo demás	0	$\frac{35}{52}$	0
Total			$E(X) = \frac{42}{52} \approx 0.81$

Valor esperado de una variable aleatoria discreta (cont.)

A continuación se muestra una representación visual de la distribución de probabilidad de las ganancias de este juego:



Variabilidad

A menudo también nos interesa la variabilidad en los valores de una variable aleatoria.

$$\sigma^2 = \mathrm{Var}(X) = \sum_{i=1}^{k} (x_i - E(X))^2 P(X = x_i)$$

$$\sigma = \mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}$$

Variabilidad de una variable aleatoria discreta

Para el ejemplo del juego de cartas anterior, ¿cuánto esperaría que varíen las ganancias de un juego a otro?

X	P(X)	X P(X)	$(X - E(X))^2$	P(X) $(X - E(X))^2$
1	$\frac{12}{52}$	$1 \times \frac{12}{52} = \frac{12}{52}$	$(1 - 0.81)^2 = 0.0361$	$\frac{12}{52} \times 0,0361 = 0,0083$
5	$\frac{4}{52}$	$5 \times \frac{4}{52} = \frac{20}{52}$	$(5 - 0.81)^2 = 17.5561$	$\frac{4}{52} \times 17,5561 = 1,3505$
10	$\frac{1}{52}$	$10 \times \frac{1}{52} = \frac{10}{52}$	$(10 - 0.81)^2 = 84.4561$	$\frac{1}{52} \times 84,0889 = 1,6242$
0	$\frac{35}{52}$	$0 \times \frac{35}{52} = 0$	$(0 - 0.81)^2 = 0.6561$	$\frac{35}{52}$ veces$0.6561 = 0.4416$
		$E(X) = 0.81$		$V(X) = 3.4246$ $SD(X) = \sqrt{3.4246} = 1.85$

Combinaciones lineales

- Una combinación lineal de variables aleatorias X y Y viene dada por

$$aX + bY$$

  donde a y b son algunos números fijos.

- El valor promedio de una combinación lineal de variables aleatorias viene dado por

$$E(aX + bY) = a \times E(X) + b \times E(Y)$$

Calcular la expectativa de una combinación lineal

En promedio, toma 10 minutos para cada problema de tarea de estadística y 15 minutos para cada problema de tarea de química. Esta semana tienes asignados 5 problemas de estadística y 4 de química. ¿Cuál es el tiempo total que espera dedicar a la tarea de estadística y química durante la semana?

$$\begin{aligned} \text{E}(\text{S} + \text{S} + \text{S} + \text{S} + \text{S} + \text{C} + \text{C} + \text{C} + \text{C}) &= 5 \times \text{E}(\text{S}) + 4 \times \text{E}(\text{C}) \\ &= 5 \times 10 + 4 \times 15 \\ &= 50 + 60 \\ &= 110 \text{ min} \end{aligned}$$

Combinaciones lineales

- La variabilidad de una combinación lineal de dos variables aleatorias independientes se calcula como

$$V(aX + bY) = a^2 \times V(X) + b^2 \times V(Y)$$

- La desviación estándar de la combinación lineal es la raíz cuadrada de la varianza.

Combinaciones lineales

- La variabilidad de una combinación lineal de dos variables aleatorias independientes se calcula como

$$V(aX + bY) = a^2 \times V(X) + b^2 \times V(Y)$$

- La desviación estándar de la combinación lineal es la raíz cuadrada de la varianza.

Note: Si las variables aleatorias no son independientes, el cálculo de la varianza se vuelve un poco más complicado y está más allá del alcance de este curso.

Calcular la varianza de una combinación lineal

La desviación estándar del tiempo que toma para cada problema de tarea de estadística es de 1,5 minutos, y es de 2 minutos para cada problema de química. ¿Cuál es la desviación estándar del tiempo que espera dedicar a la tarea de estadística y física durante la semana si tiene asignados 5 problemas de estadística y 4 de química? Suponga que el tiempo que toma completar cada problema es independiente de otro.

$$
\begin{aligned}
\mathrm{V}(\mathrm{S}+\mathrm{S}+\mathrm{S}+\mathrm{S}+\mathrm{S}+\mathrm{C}+\mathrm{C}+\mathrm{C}+\mathrm{C}) &= \mathrm{V}(\mathrm{S})+\mathrm{V}(\mathrm{S})+\mathrm{V}(\mathrm{S})+\mathrm{V}(\mathrm{S})+\mathrm{V}(\mathrm{S})+\mathrm{V}(\mathrm{C})+\mathrm{V}(\mathrm{C})+\mathrm{V}(\mathrm{C})+\mathrm{V}(\mathrm{C}) \\
&= 5\times\mathrm{V}(\mathrm{S})+4\times\mathrm{V}(\mathrm{C}) \\
&= 5\times 1.5^2+4\times 2^2 \\
&= 27.25
\end{aligned}
$$

Práctica

Jugar un juego de casino cuesta \$5. Si la primera carta que sacas es roja, entonces puedes sacar una segunda carta (sin reemplazo). Si la segunda carta es el as de tréboles, ganas \$500. Si no, no ganas nada, es decir, pierdes tus \$5. ¿Cuáles son sus ganancias/pérdidas esperadas al jugar este juego? Recuerde: ganancia/pérdida = ganancias - costo.

(a) Una ganancia de 5¢

(b) Una pérdida de 10¢

(c) Una pérdida de 25¢

(d) Una pérdida de 30¢

Evento	Ganancia	Beneficio: X	P(X)	X × P(X)
Rojo, A♣	500	500 - 5 = 495	$\frac{26}{52} \times \frac{1}{51} = 0.0098$	$495 \times 0.0098 = 4.851$
Otro	0	0 - 5 = -5	$1 - 0.0098 = 0.9902$	$-5 \times 0.9902 = -4.951$
				E(X) = −0.1

Práctica

Jugar un juego de casino cuesta \$5. Si la primera carta que sacas es roja, entonces puedes sacar una segunda carta (sin reemplazo). Si la segunda carta es el as de tréboles, ganas \$500. Si no, no ganas nada, es decir, pierdes tus \$5. ¿Cuáles son sus ganancias/pérdidas esperadas al jugar este juego? Recuerde: ganancia/pérdida = ganancias - costo.

(a) Una ganancia de 5¢

(b) Una pérdida de 10¢

(c) Una pérdida de 25¢

(d) Una pérdida de 30¢

Evento	Ganancia	Beneficio: X	P(X)	X × P(X)
Rojo, A♣	500	500 - 5 = 495	$\frac{26}{52} \times \frac{1}{51} = 0.0098$	$495 \times 0.0098 = 4.851$
Otro	0	0 - 5 = -5	$1 - 0.0098 = 0.9902$	$-5 \times 0.9902 = -4.951$
				E(X) = −0.1

Juego justo

Un juego justo se define como un juego que cuesta tanto como el pago esperado, es decir, la ganancia esperada es 0.

¿Crees que los juegos de casino en Las Vegas cuestan más o menos que los pagos esperados?

Si esos juegos cuestan menos de los pagos esperados, significaría que los casinos estarían perdiendo dinero en promedio y, por lo tanto, no podrían pagar todo esto:

Imagen de Moyan_Brenn en Flickr http://www.flickr.com/photos/aigle_dore/5951714693.

Simplificar variables aleatorias

Las variables aleatorias no funcionan como las variables algebraicas normales:

$$X + X \neq 2X$$

$$\begin{aligned} E(X + X) &= E(X) + E(X) \\ &= 2E(X) \end{aligned} \qquad \begin{aligned} \text{Var}(X + X) &= \text{Var}(X) + \text{Var}(X) \text{ (suponiendo independencia)} \\ &= 2\ \text{Var}(X) \end{aligned}$$

$$E(2X) = 2E(X) \qquad \begin{aligned} \text{Var}(2X) &= 2^2\ \text{Var}(X) \\ &= 4\ \text{Var}(X) \end{aligned}$$

$E(X + X) = E(2X)$, pero $\text{Var}(X + X) \neq \text{Var}(2X)$.

¿Sumar o multiplicar?

Una empresa tiene 5 Lincoln Town Cars en su flota. Los datos históricos muestran que el costo de mantenimiento anual de cada automóvil es en promedio de \$2,154 con una desviación estándar de \$132. ¿Cuál es la media y la desviación estándar del costo total anual de mantenimiento de esta flota?

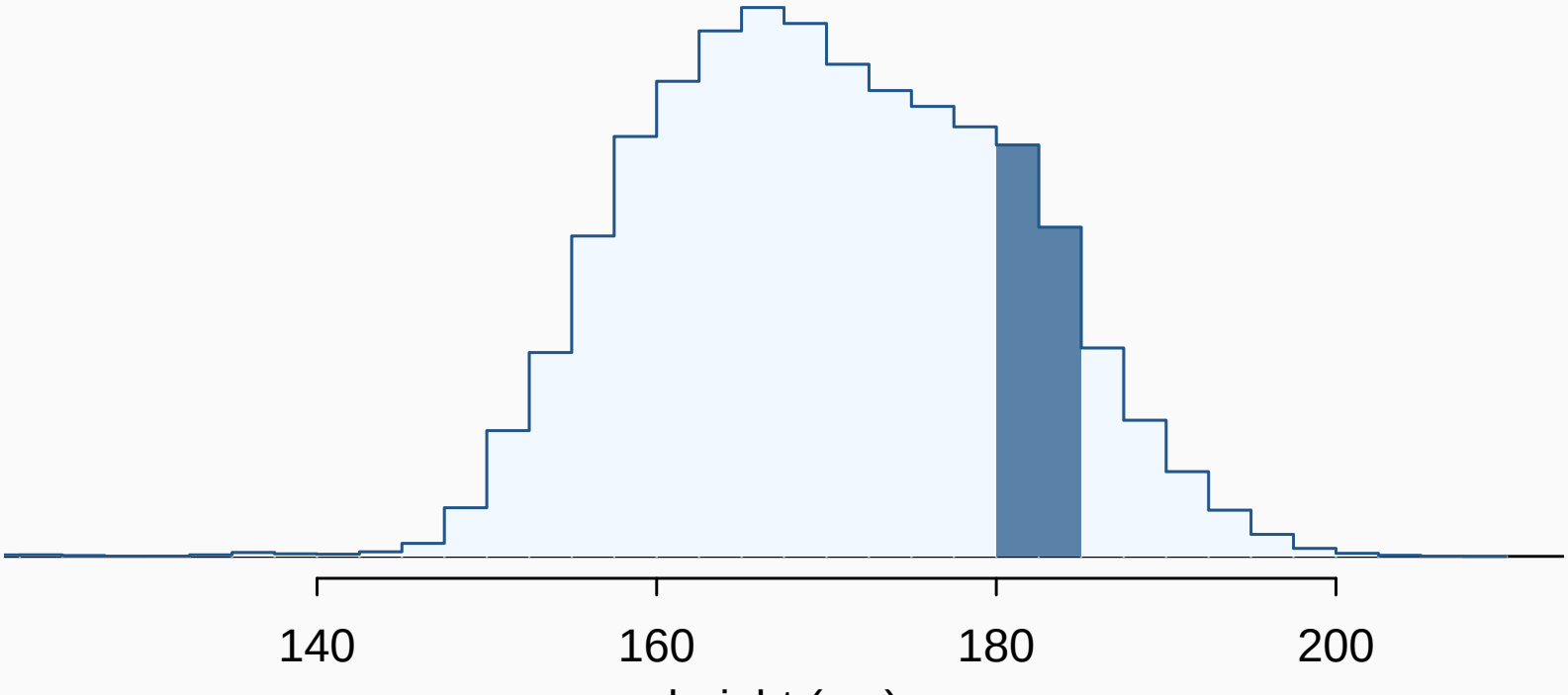
Tenga en cuenta que tenemos 5 autos cada uno con el costo de mantenimiento anual dado ($X_1 + X_2 + X_3 + X_4 + X_5$), no un auto que tenga 5 veces el costo de mantenimiento anual dado (5X).

$$
\begin{aligned}
E(X_1 + X_2 + X_3 + X_4 + X_5) &= E(X_1) + E(X_2) + E(X_3) + E(X_4) + E(X_5) \\
&= 5 \times E(X) = 5 \times 2,154 = \$10,770 \\
Var(X_1 + X_2 + X_3 + X_4 + X_5) &= Var(X_1) + Var(X_2) + Var(X_3) + Var(X_4) + Var(X_5) \\
&= 5 \times V(X) = 5 \times 132^2 = \$87,120 \\
DE(X_1 + X_2 + X_3 + X_4 + X_5) &= \sqrt{87,120} = 295.16
\end{aligned}
$$

Distribuciones continuas

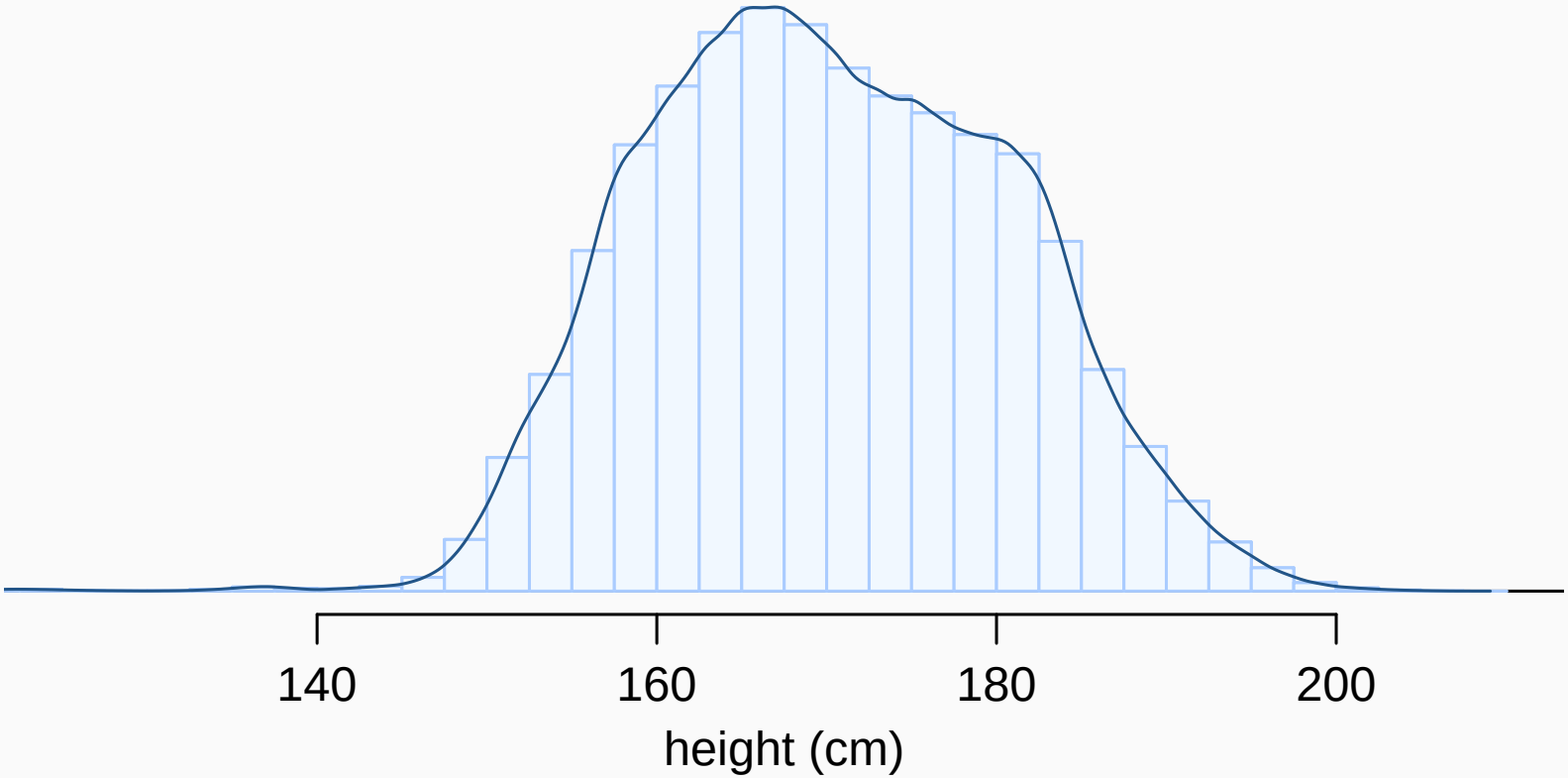
Distribuciones continuas

- A continuación se muestra un histograma de la distribución de alturas de los adultos estadounidenses.
- La proporción de datos que cae en los contenedores sombreados da la probabilidad de que un adulto estadounidense muestreado aleatoriamente mida entre 180 cm y 185 cm (alrededor de 5'11" a 6'1").

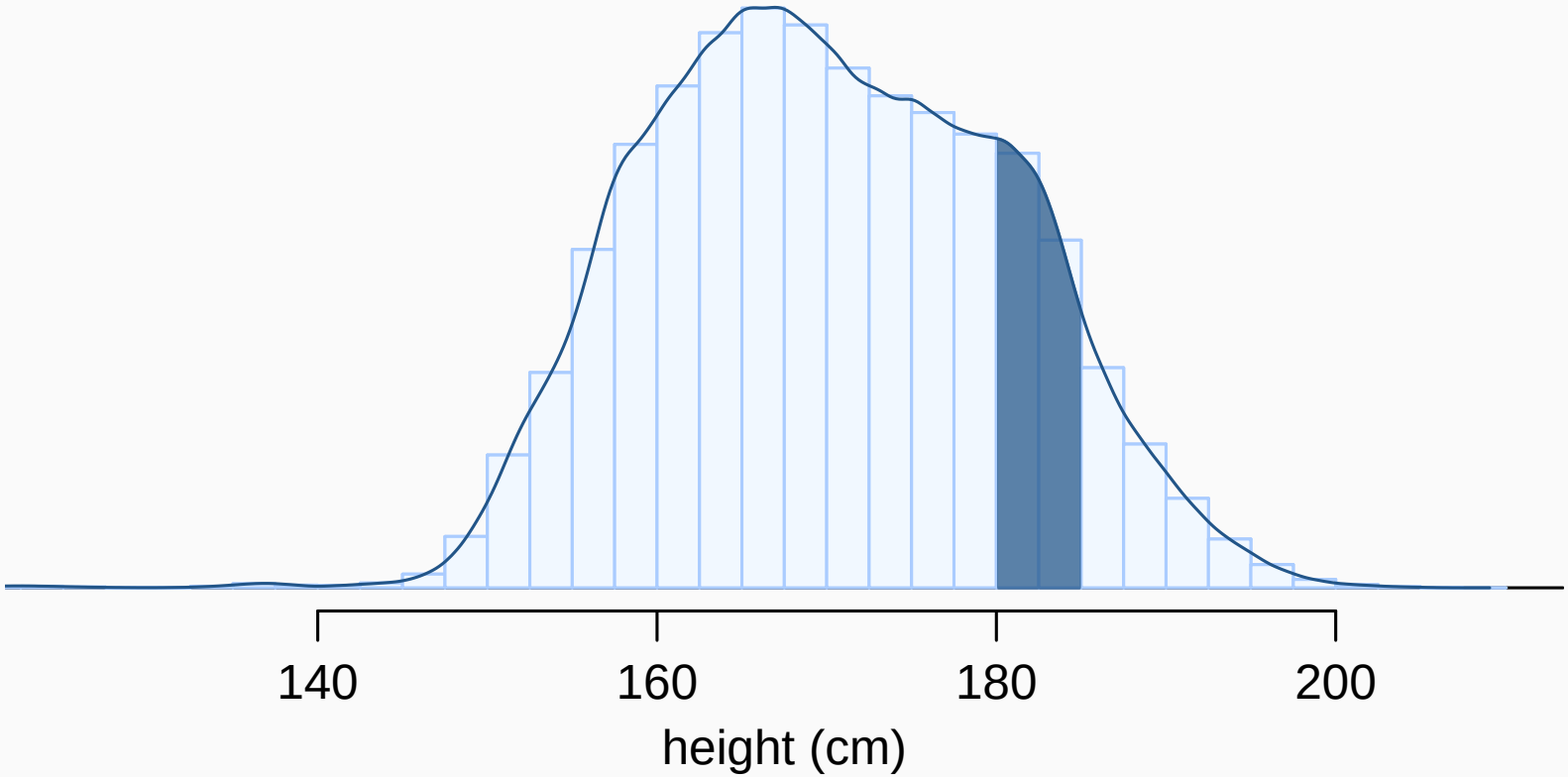


De histogramas a distribuciones continuas

Dado que la altura es una variable numérica continua, su función de densidad de probabilidad es una curva suave.

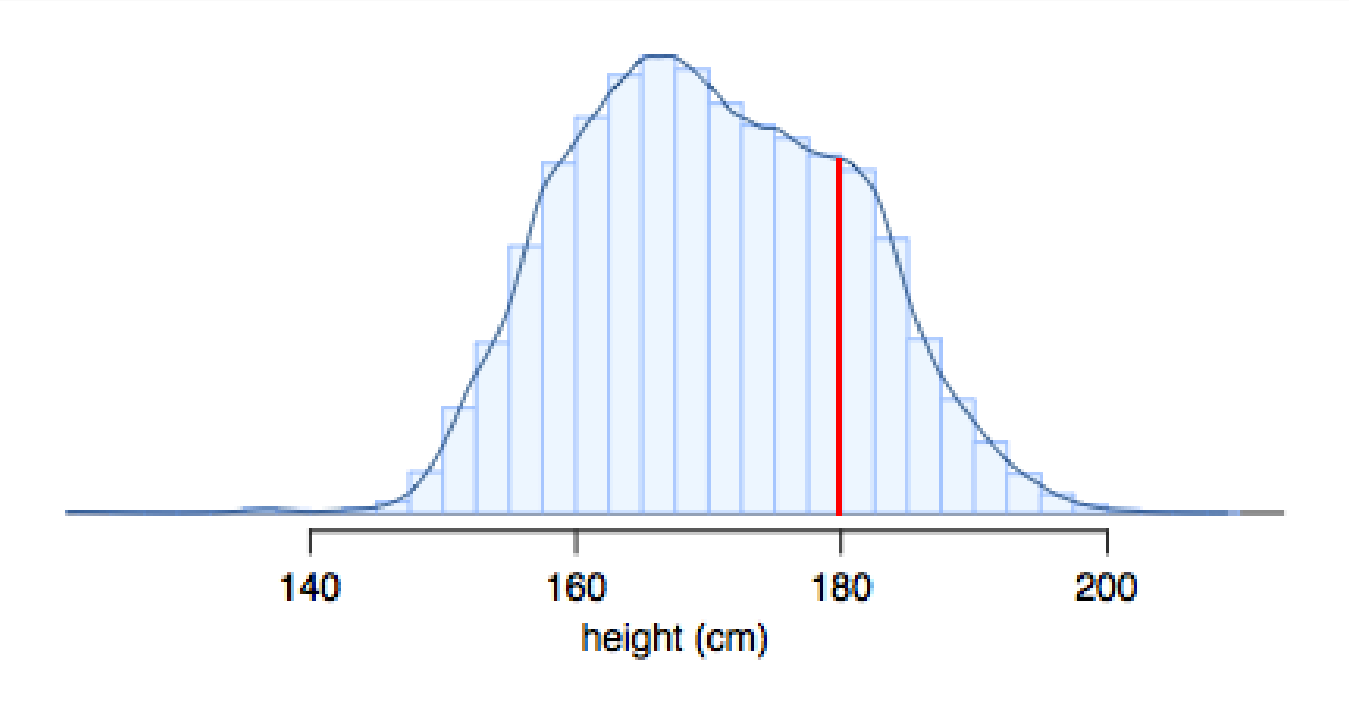


Por lo tanto, la probabilidad de que un adulto estadounidense muestreado aleatoriamente mida entre 180 cm y 185 cm también se puede estimar como el área sombreada bajo la curva.



Por definición...

Dado que las probabilidades continuas se estiman como ”el área bajo la curva”, la probabilidad de que una persona tenga exactamente 180 cm (o cualquier valor exacto) se define como 0.



OpenIntro Statistics

Capítulo 3 del libro OpenIntro Statistics.
Las diapositivas están basadas en las slides desarrolladas por Mine Çetinkaya-Rundel de OpenIntro y el contenido del libro.