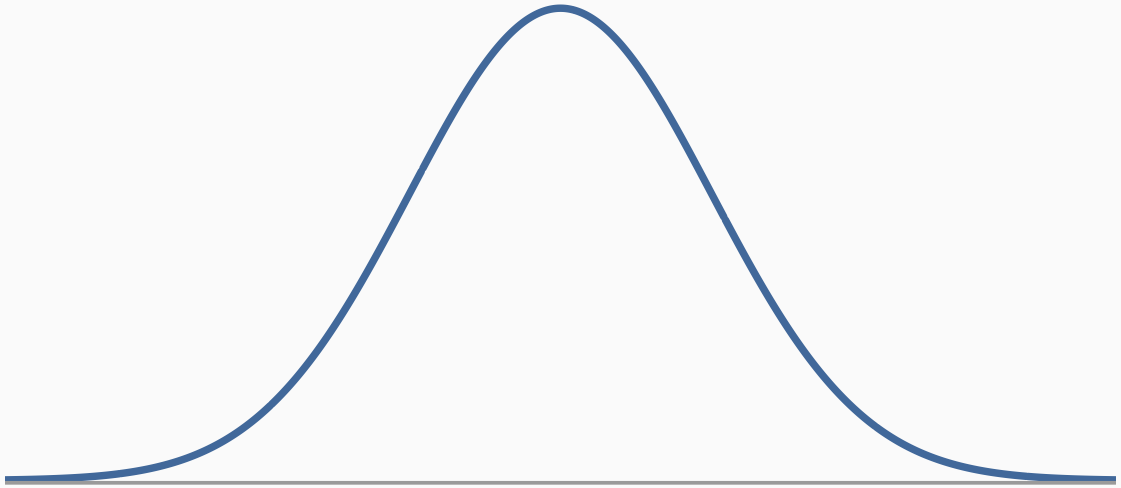
Metodología de análisis en Opinión Pública

Cátedra Olego
Facultad de Ciencias Sociales
Universidad de Buenos Aires

Distribución normal

Distribución normal

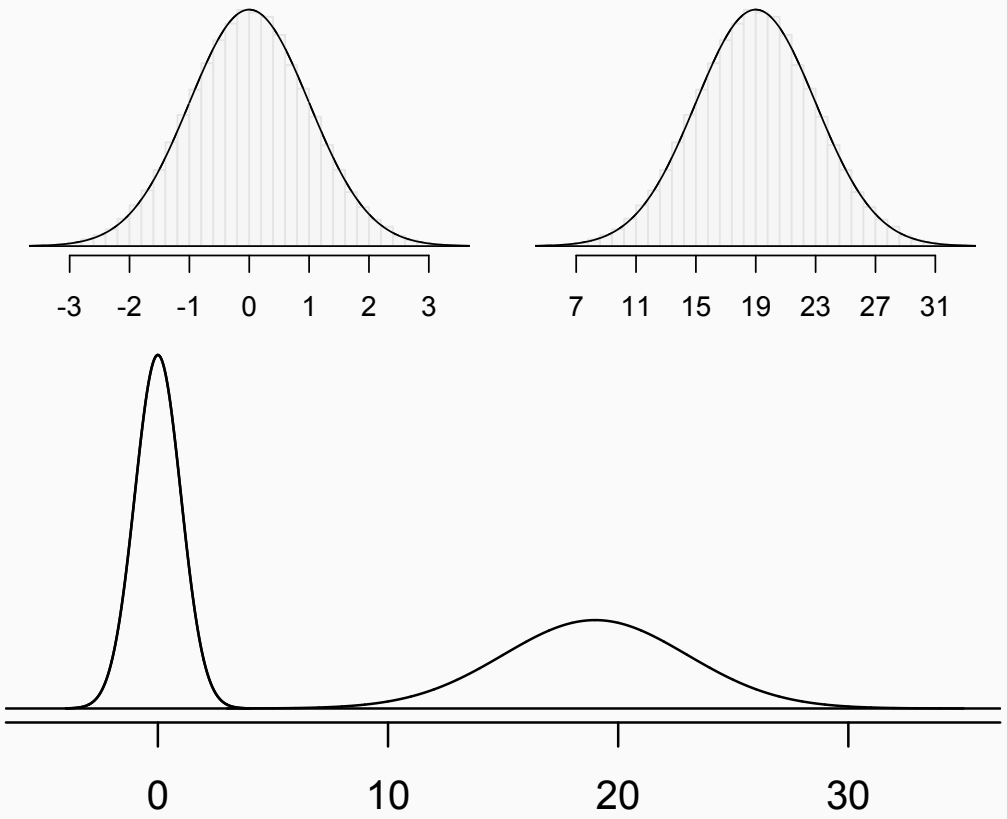
- Curva acampanada unimodal y simétrica
- Muchas variables son casi normales, pero ninguna es exactamente normal
- Denotado como $\mathrm{N}(\mu, \sigma)$ $\rightarrow$ Normal con media μ y desviación estándar σ



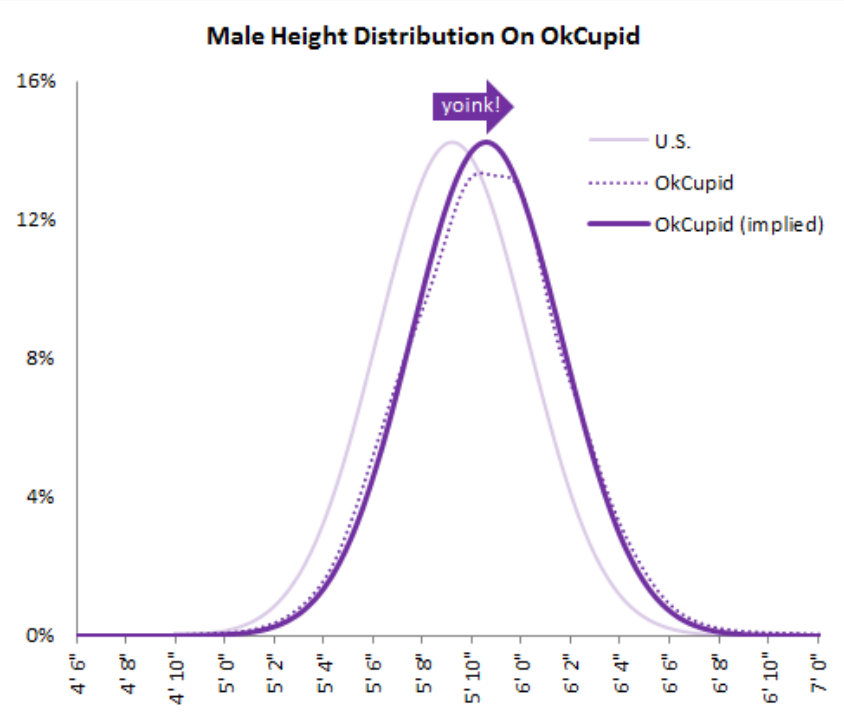
Distribuciones normales con diferentes parámetros

μ: media, σ: desviación estándar

$\mathrm{N}(\mu = 0, \sigma = 1)$ $\mathrm{N}(\mu = 19, \sigma = 4)$



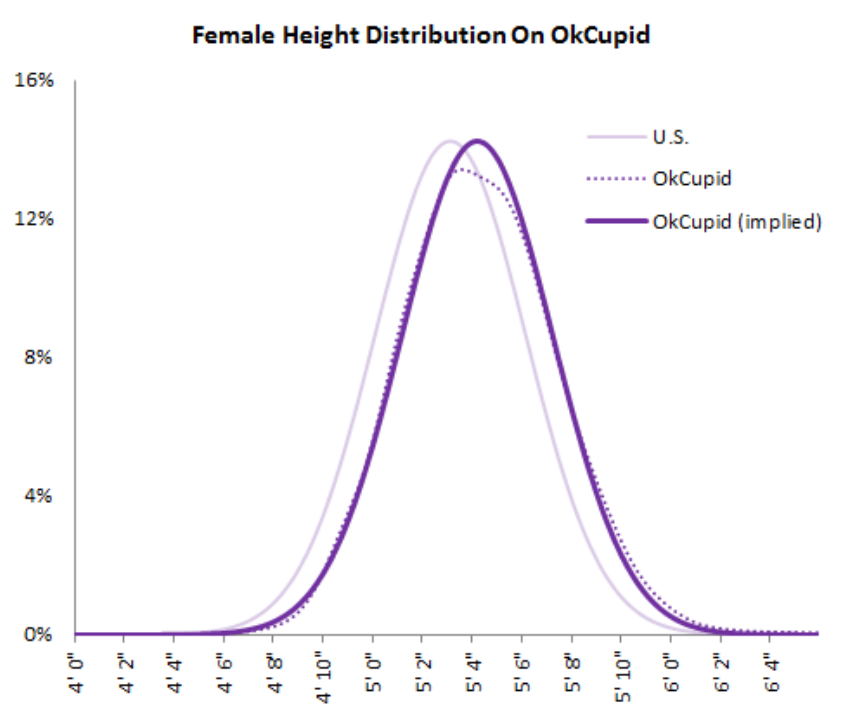
Alturas de los hombres



Las alturas masculinas en OkCupid casi siguen la distribución normal esperada, excepto que todo se desplaza a la derecha de donde debería estar. A casi todos los hombres les gusta agregar un par de pulgadas. También puede ver una vanidad más sutil en el trabajo: comenzando en aproximadamente 5' 8", la parte superior de la curva punteada se inclina aún más hacia la derecha. Esto significa que los hombres, a medida que se acercan a los seis pies, redondean un poco más de lo habitual, estirándose para ese codiciado punto de referencia psicológico.

http://blog.okcupid.com/index.php/the-biggest-lies-in-online-dating/

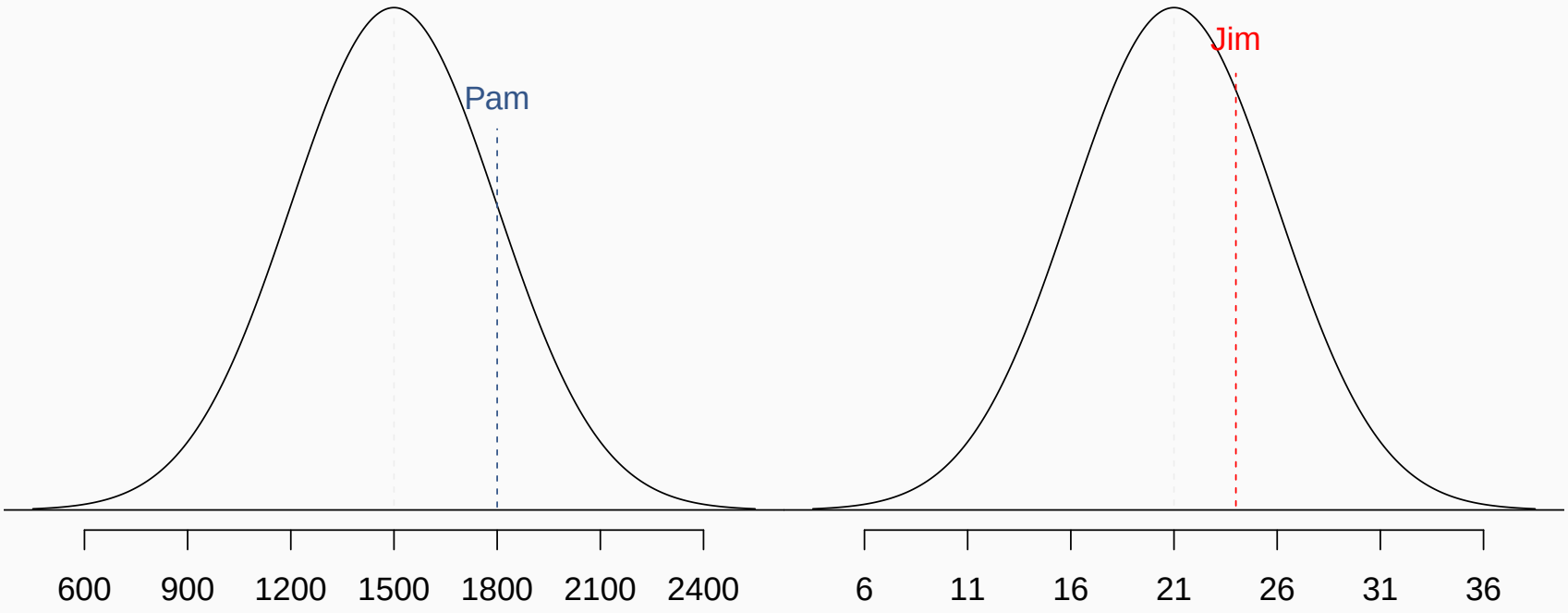
Alturas de las mujeres



"Cuando observamos los datos de las mujeres, nos sorprendió ver que la exageración de la estatura estaba tan extendida, aunque sin la inclinación hacia una estatura de referencia".

http://blog.okcupid.com/index.php/the-biggest-lies-in-online-dating/

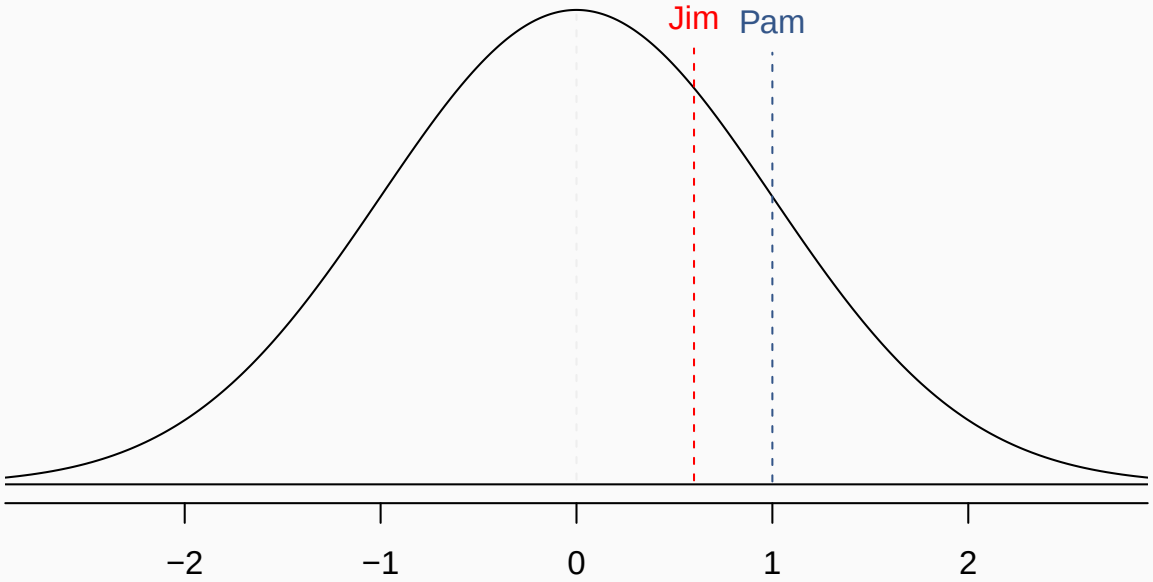
Los puntajes de SAT se distribuyen casi normalmente con una media de 1500 y una desviación estándar de 300. Los puntajes de ACT se distribuyen de manera casi normal con una media de 21 y una desviación estándar de 5. Un funcionario de admisiones de la universidad desea determinar cuál de los dos solicitantes obtuvo mejores puntajes en su prueba estandarizada con con respecto a los demás examinados: Pam, que obtuvo un 1800 en su SAT, o Jim, que obtuvo un 24 en su ACT?



Estandarización con puntuaciones Z

Dado que no podemos simplemente comparar estos dos puntajes brutos, en su lugar comparamos cuántas desviaciones estándar más allá de la media es cada observación.

- La puntuación de Pam es $\frac{1800-1500}{300} = 1$ desviación estándar por encima de la media.
- La puntuación de Jim es $\frac{24-21}{5} = 0,6$ desviaciones estándar por encima de la media.



Estandarización con puntuaciones Z (cont.)

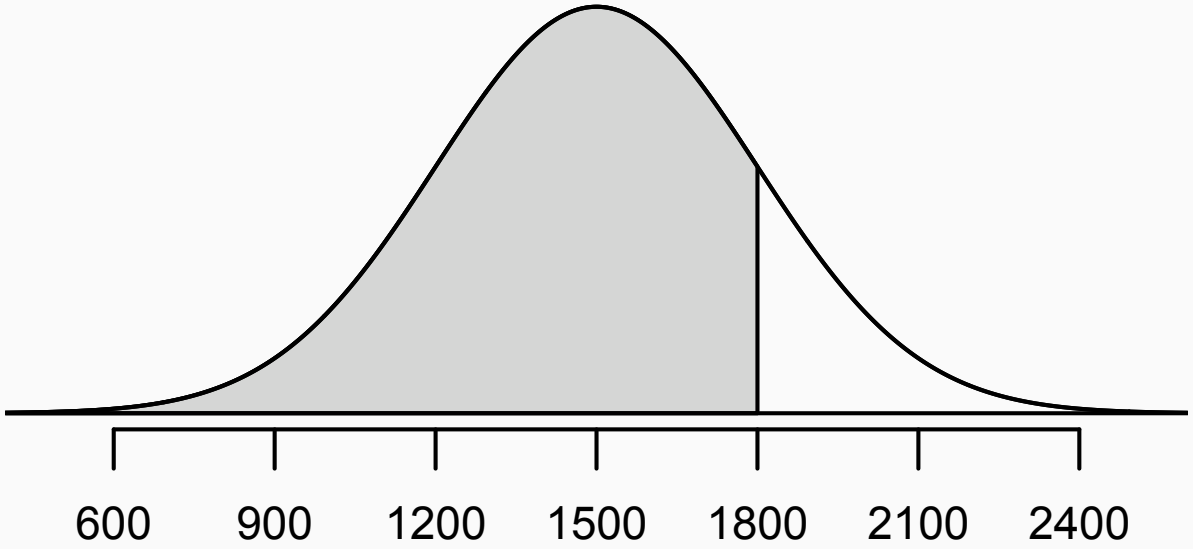
- Estos se denominan puntuaciones estandarizadas o puntuaciones Z.
- La puntuación Z de una observación es el número de desviaciones estándar por encima o por debajo de la media.

$$Z = \frac{\text{obs} - \text{media}}{\text{SD}}$$

- Los puntajes Z se definen para distribuciones de cualquier forma, pero solo cuando la distribución es normal podemos usar puntajes Z para calcular percentiles.
- Las observaciones que están a más de 2 sd. de la media ($|Z| > 2$) generalmente se consideran inusuales.

Percentiles

- Percentil es el porcentaje de observaciones que caen por debajo de un punto de datos determinado.
- Gráficamente, el percentil es el área debajo de la curva de distribución de probabilidad a la izquierda de esa observación.



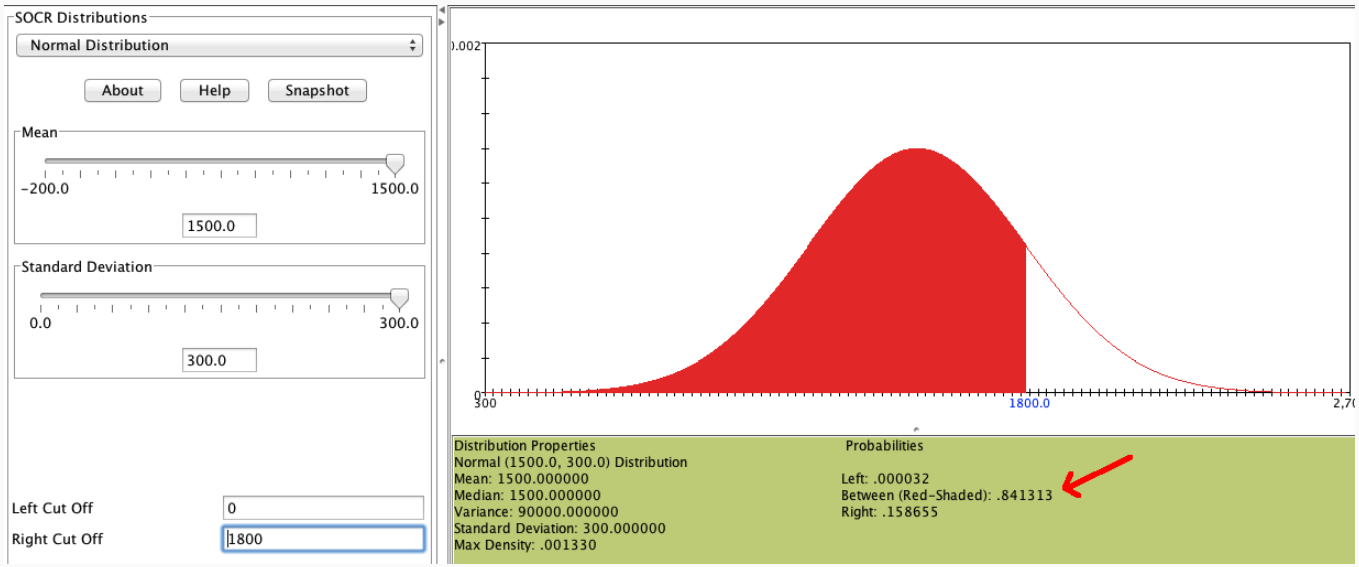
Calcular percentiles - usando computación

Hay muchas formas de calcular percentiles/áreas bajo la curva:

- R:

```
> pnorm(1800, media = 1500, sd = 300)
[1] 0.8413447
```

- subprograma: https://gallery.shinyapps.io/dist_calc/



Cálculo de percentiles: uso de tablas

	Second decimal place of Z									
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.0
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.53
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.57
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.61
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.65
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.68
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.72
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.75
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.78
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.81
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.83
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.86
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.88
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.90

Seis sigma

"El término proceso seis sigma proviene de la noción de que si uno tiene seis desviaciones estándar entre la media del proceso y el límite de especificación más cercano, como se muestra en el gráfico, prácticamente ningún elemento dejará de cumplir con las especificaciones".

$$6\sigma$$

http://en.wikipedia.org/wiki/Six_Sigma

Control de calidad

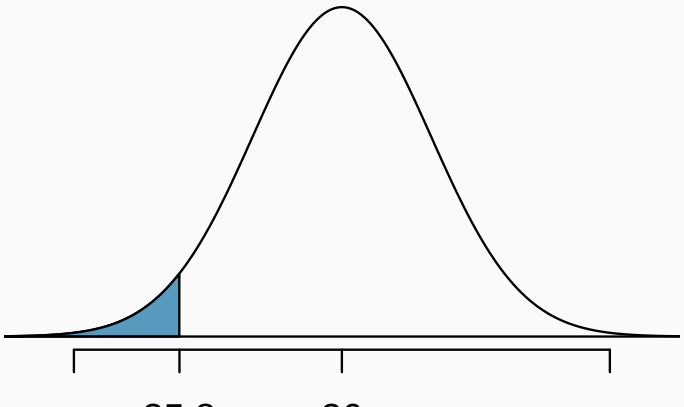
En la fábrica de salsa de tomate Heinz, se supone que las cantidades que entran en las botellas de salsa de tomate se distribuyen normalmente con una media de 36 oz. y desviación estándar 0.11 oz. Una vez cada 30 minutos se selecciona una botella de la línea de producción y se anota su contenido con precisión. Si la cantidad de ketchup en la botella es inferior a 35,8 oz. o más de 36.2 oz., entonces la botella no pasa la inspección de control de calidad. ¿Qué porcentaje de botellas tiene menos de 35.8 onzas de ketchup?

Control de calidad

En la fábrica de salsa de tomate Heinz, se supone que las cantidades que entran en las botellas de salsa de tomate se distribuyen normalmente con una media de 36 oz. y desviación estándar 0.11 oz. Una vez cada 30 minutos se selecciona una botella de la línea de producción y se anota su contenido con precisión. Si la cantidad de ketchup en la botella es inferior a 35,8 oz. o más de 36.2 oz., entonces la botella no pasa la inspección de control de calidad. ¿Qué porcentaje de botellas tiene menos de 35.8 onzas de ketchup?

Sea X = cantidad de ketchup en una botella:

$$X \sim N(\mu = 36, \sigma = 0.11)$$



$$Z = \frac{35,8 - 36}{0,11} = -1,82$$

Encontrar la probabilidad exacta - usando R

```
> pnorm(-1.82, media = 0, sd = 1)
[1] 0,0344
```

Ó

```
> pnorm(35.8, media = 36, sd = 0.11)
[1] 0,0345
```

Práctica

¿Qué porcentaje de botellas pasan la inspección de control de calidad?

(a) 1,82%

(b) 3.44%

(c) 6.88%

(d) 93.12%

(e) 96.56%

Práctica

¿Qué porcentaje de botellas pasan la inspección de control de calidad?

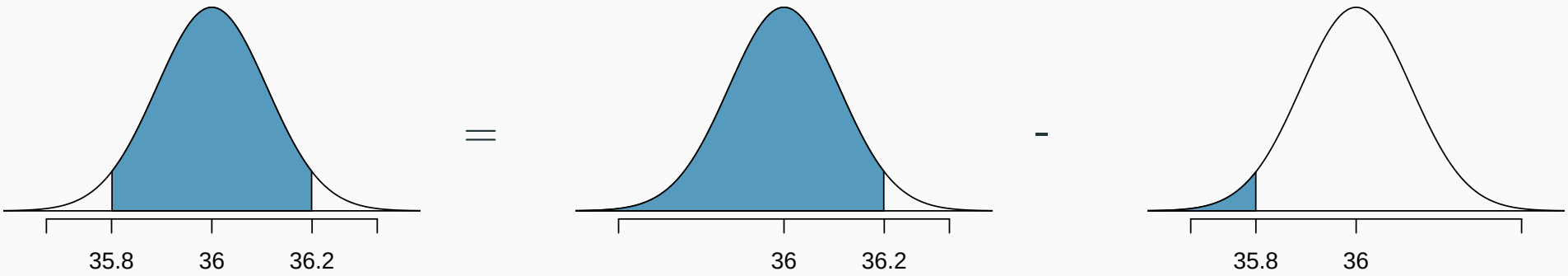
(a) 1,82%

(b) 3.44%

(c) 6.88%

(d) 93.12%

(e) 96.56%



Práctica

¿Qué porcentaje de botellas pasan la inspección de control de calidad?

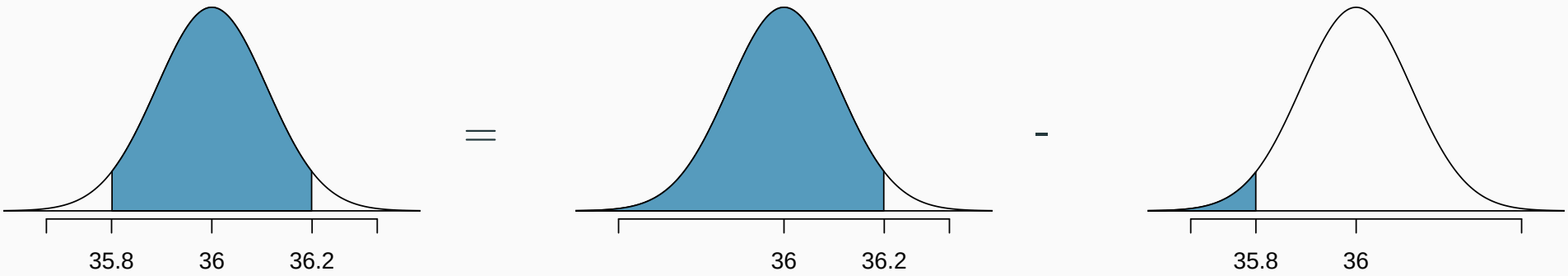
(a) 1,82%

(b) 3.44%

(c) 6.88%

(d) 93.12%

(e) 96.56%



$$Z_{35.8} = \frac{35.8 - 36}{0.11} = -1.82$$

$$Z_{36.2} = \frac{36.2 - 36}{0.11} = 1.82$$

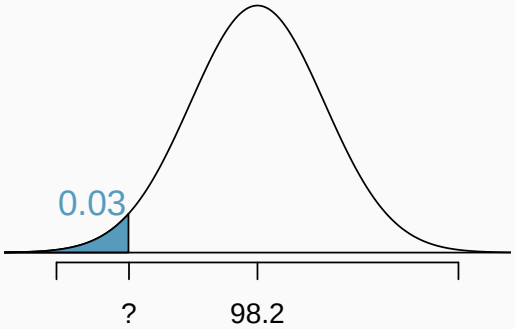
$$P(35,8 < X < 36,2) = P(-1,82 < Z < 1,82) = 0,9656 - 0,0344 = 0,9312$$

Encontrar puntos de corte

La temperatura corporal de los seres humanos sanos se distribuye casi normalmente con una media de 98,2°F y una desviación estándar de 0,73°F. ¿Cuál es el límite para el 3% más bajo de la temperatura del cuerpo humano?

Encontrar puntos de corte

La temperatura corporal de los seres humanos sanos se distribuye casi normalmente con una media de 98,2°F y una desviación estándar de 0,73°F. ¿Cuál es el límite para el 3% más bajo de la temperatura del cuerpo humano?



$$
\begin{aligned}
P(X < x) &= 0.03 \rightarrow P(Z < \text{-}1.88) = 0.03 \\
Z &= \frac{\text{obs} - \text{mean}}{SD} \rightarrow \frac{x - 98.2}{0.73} = -1.88 \\
x &= (-1,88 \times 0,73) + 98,2 = 96,8°F
\end{aligned}
$$

```
> qnorm(0.03)
[1] -1.880794
```

Mackowiak, Wasserman y Levine (1992), Una evaluación crítica de 98,6 grados F, el límite superior de la temperatura corporal normal y otros legados de Carl Reinhold August Wunderlick.

Práctica

La temperatura corporal de los seres humanos sanos se distribuye casi normalmente con una media de 98,2°F y una desviación estándar de 0,73°F. ¿Cuál es el límite para el 10% más alto de las temperaturas del cuerpo humano?

(a) 97.3°F

(b) 99.1°F

(c) 99.4°F

(d) 99.6°F

Práctica

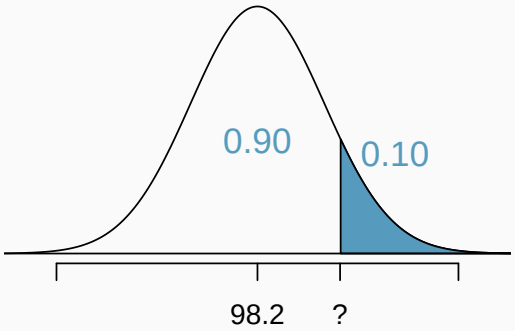
La temperatura corporal de los seres humanos sanos se distribuye casi normalmente con una media de 98,2°F y una desviación estándar de 0,73°F. ¿Cuál es el límite para el 10% más alto de las temperaturas del cuerpo humano?

(a) 97.3°F

(b) 99.1°F

(c) 99.4°F

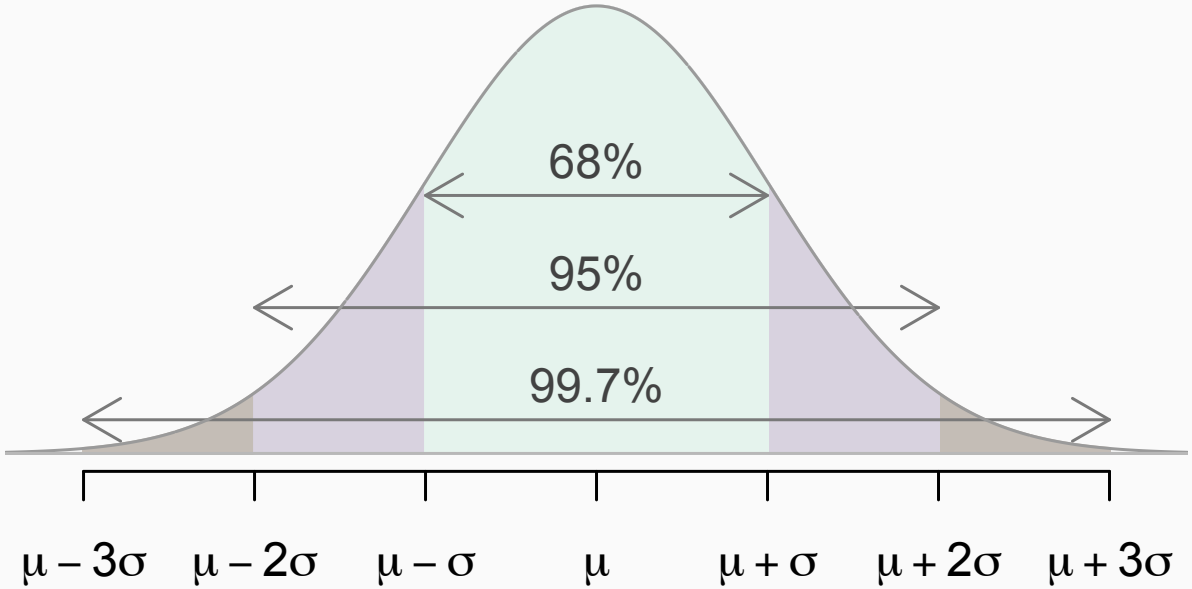
(d) 99.6°F



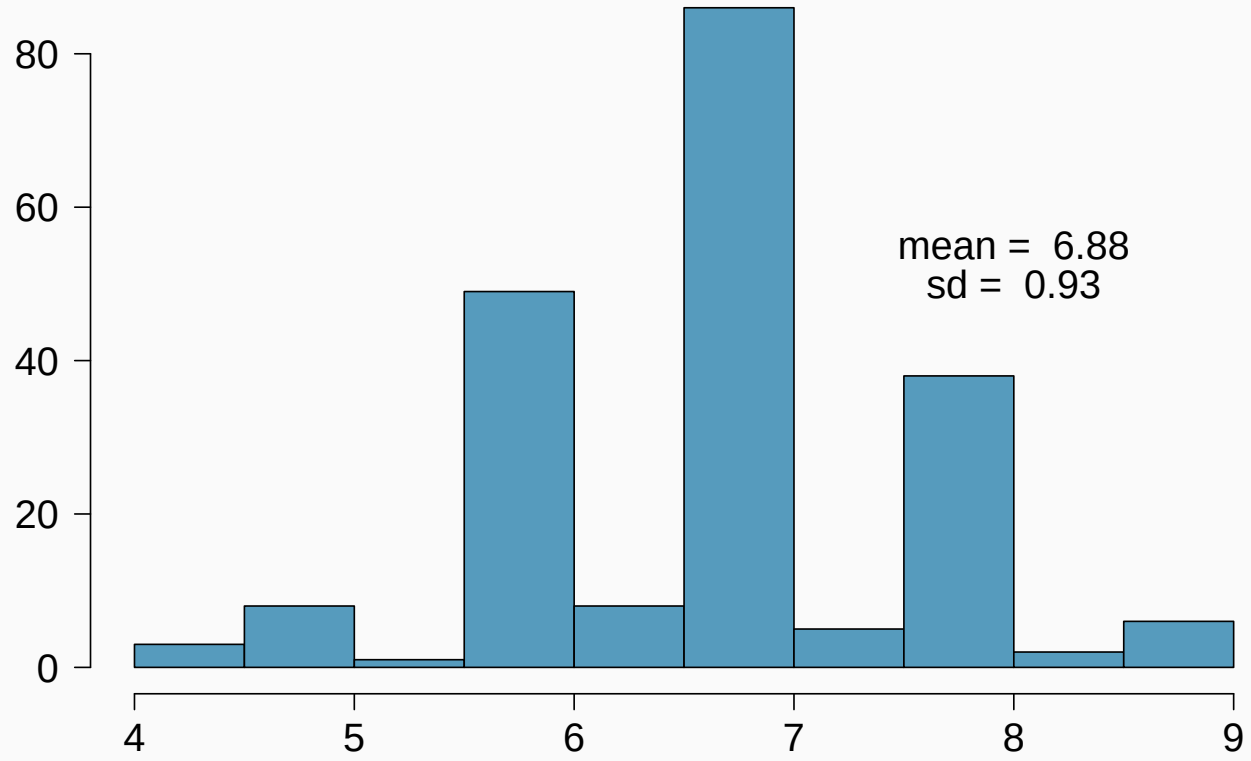
$$
\begin{aligned}
P(X > x) &= 0.10 \rightarrow P(Z < 1.28) = 0.90 \\
Z &= \frac{\text{obs} - \text{mean}}{SD} \rightarrow \frac{x - 98.2}{0.73} = 1.28 \\
x &= (1,28 \times 0,73) + 98,2 = 99,1
\end{aligned}
$$

Regla 68-95-99.7

- Para datos casi normalmente distribuidos,
  - alrededor del 68% cae dentro de 1 SD de la media,
  - alrededor del 95% cae dentro de 2 SD de la media,
  - alrededor del 99,7% cae dentro de 3 SD de la media.
- Es posible que las observaciones se alejen 4, 5 o más desviaciones estándar de la media, pero estos casos son muy raros si los datos son casi normales.



Número de horas de sueño en las noches de escuela



- Media = 6,88 horas, SD = 0,92 horas

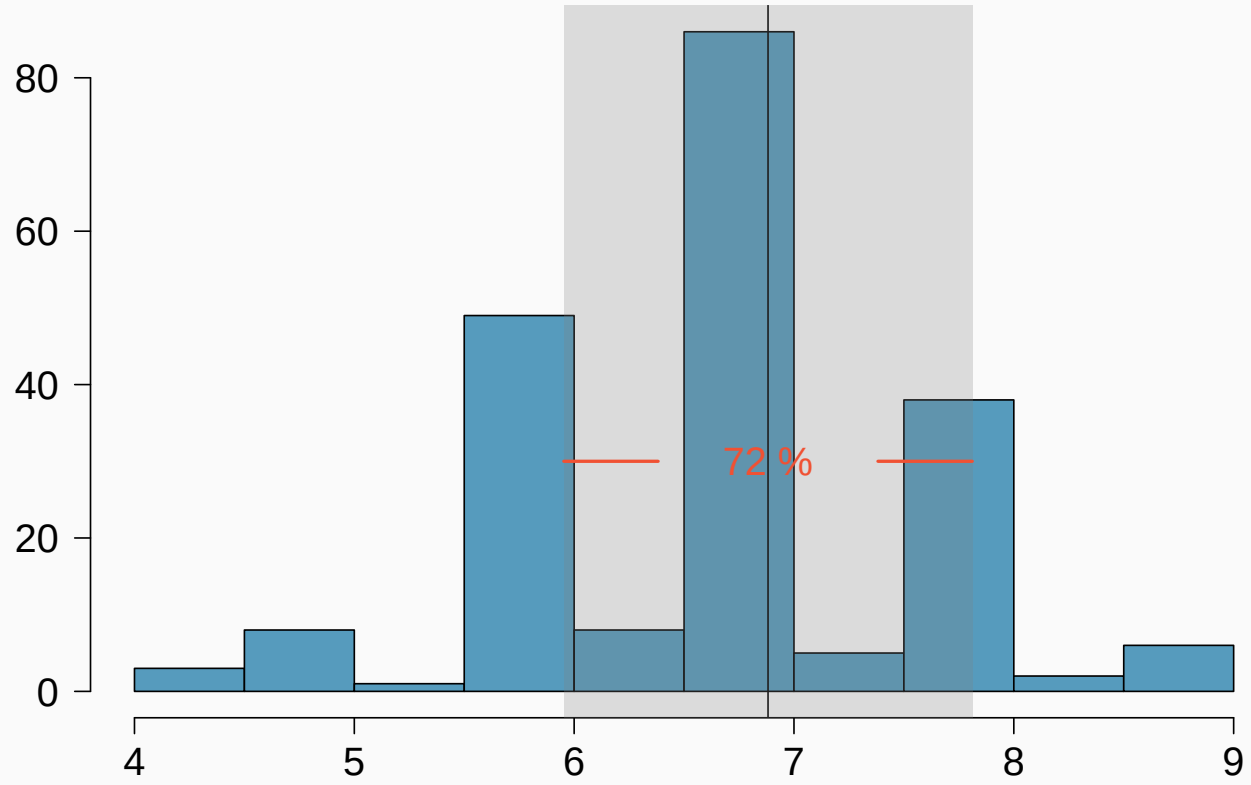
  72% de los datos están dentro de 1 SD de la media:

  6.88 ± 0.93

  92% de los datos están dentro de 2 SD de la media:

  $6.88 \pm 2 \times 0.93$

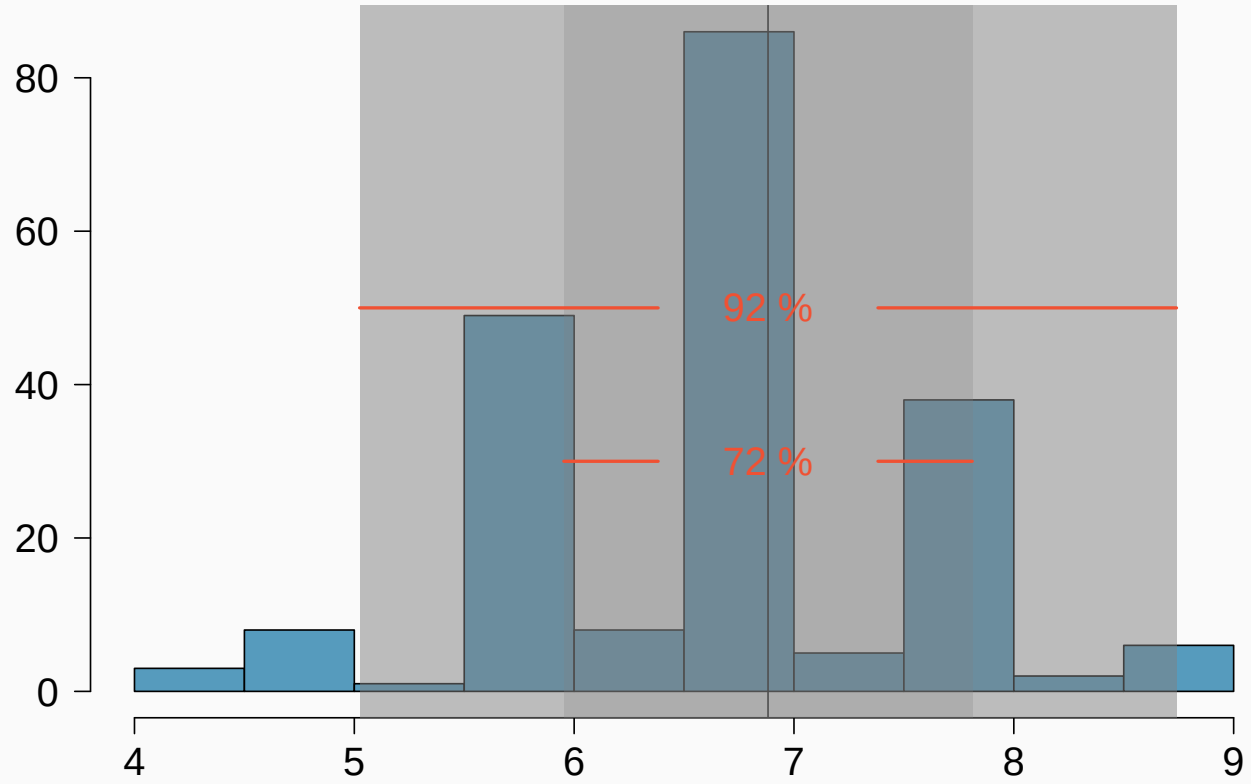
Número de horas de sueño en las noches de escuela



- Media = 6,88 horas, SD = 0,92 horas
- 72% de los datos están dentro de 1 SD de la media: 6.88 ± 0.93

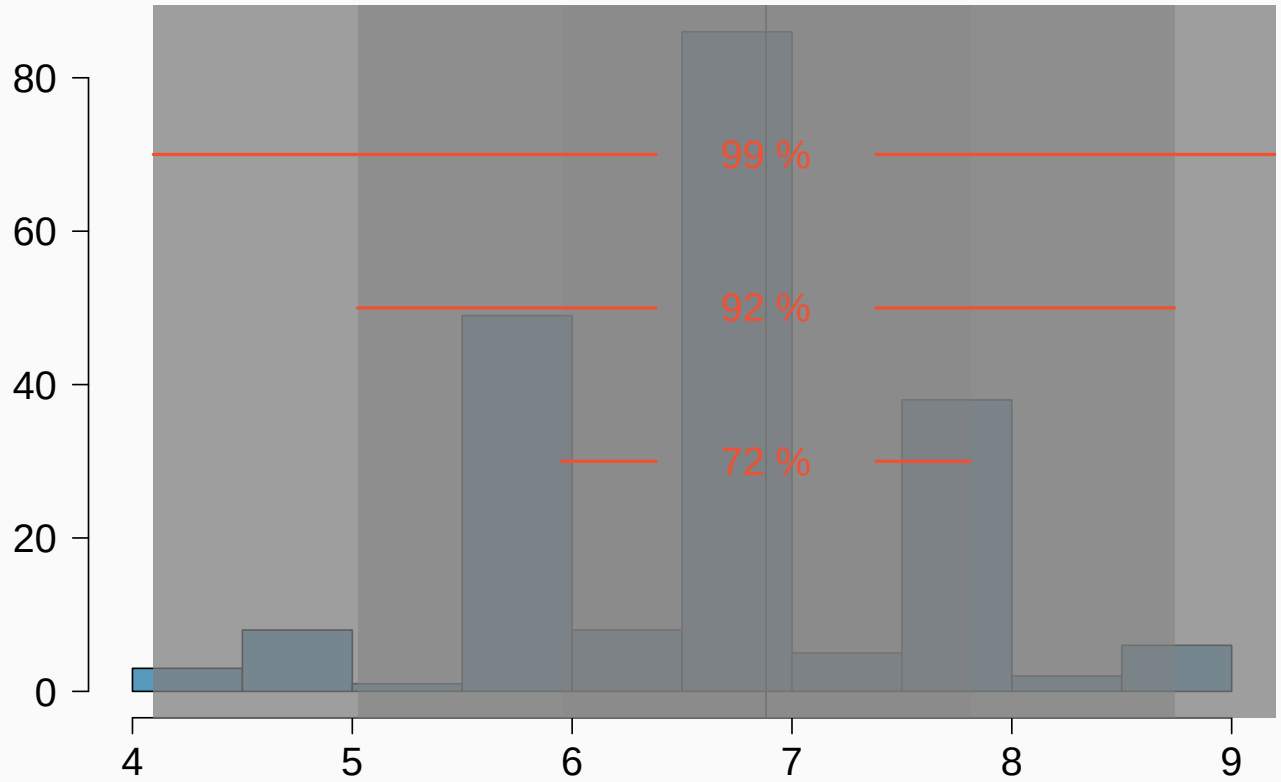
  92% de los datos están dentro de 2 SD de la media: $6.88 \pm 2 \times 0.93$

Número de horas de sueño en las noches de escuela



- Media = 6,88 horas, SD = 0,92 horas
- 72% de los datos están dentro de 1 SD de la media: 6.88 ± 0.93
- 92% de los datos están dentro de 1 SD de la media: $6.88 \pm 2 \times 0.93$

Número de horas de sueño en las noches de escuela



- Media = 6,88 horas, SD = 0,92 horas
- 72% de los datos están dentro de 1 SD de la media: 6.88 ± 0.93
- 92% de los datos están dentro de 2 SD de la media: $6.88 \pm 2 \times 0.93$

Práctica

¿Cuál de los siguientes es falso?

(a) La mayoría de las puntuaciones Z en una distribución sesgada a la derecha son negativas.

(b) En distribuciones sesgadas, la puntuación Z de la media puede ser diferente de 0.

(c) Para una distribución normal, IQR es menor que $2 \times \text{SD}$.

(d) Los puntajes Z son útiles para determinar cuán inusual es un punto de datos en comparación con el resto de los datos en la distribución.

Práctica

¿Cuál de los siguientes es falso?

(a) La mayoría de las puntuaciones Z en una distribución sesgada a la derecha son negativas.

(b) En distribuciones sesgadas, la puntuación Z de la media puede ser diferente de 0.

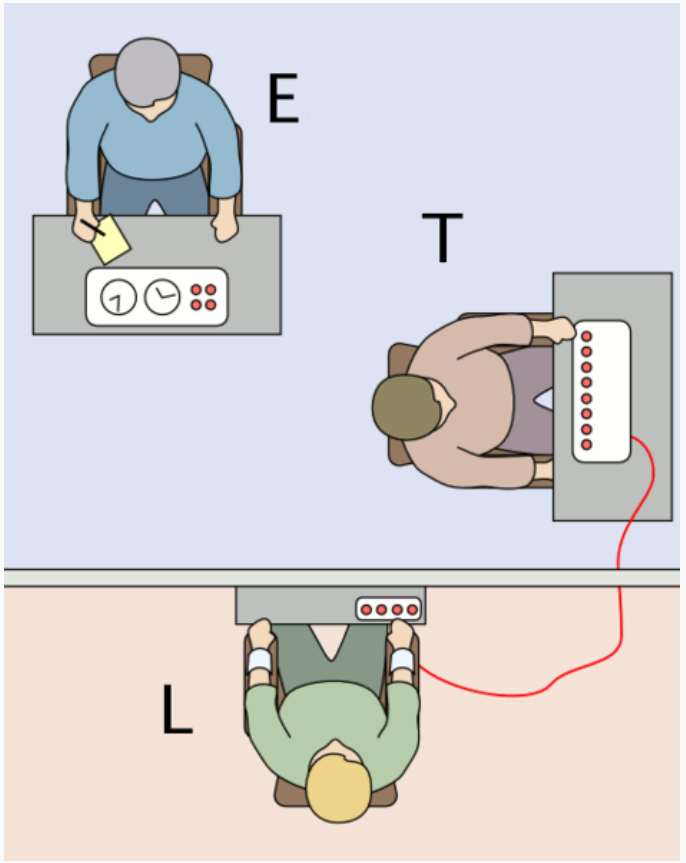
(c) Para una distribución normal, IQR es menor que $2 \times \text{SD}$.

(d) Los puntajes Z son útiles para determinar cuán inusual es un punto de datos en comparación con el resto de los datos en la distribución.

Distribución de Bernoulli

Experimento de Milgram

- Stanley Milgram, un psicólogo de la Universidad de Yale, realizó una serie de experimentos sobre la obediencia a la autoridad a partir de 1963.
- El experimentador (E) le ordena al maestro (T), el sujeto del experimento, que dé severas descargas eléctricas a un alumno (L) cada vez que el alumno responde una pregunta incorrectamente.
- El alumno es en realidad un actor, y las descargas eléctricas no son reales, pero se reproduce un sonido pregrabado cada vez que el maestro administra una descarga eléctrica.



http://en.wikipedia.org/wiki/File:Milgram_Experiment_v2.png

Experimento de Milgram (continuación)

- Estos experimentos midieron la voluntad de los participantes del estudio de obedecer a una figura de autoridad que les instruía para realizar actos que entraban en conflicto con su conciencia personal.
- Milgram descubrió que alrededor del 65% de las personas obedecerían a la autoridad y darían tales descargas.
- A lo largo de los años, investigaciones adicionales sugirieron que este número es aproximadamente constante en todas las comunidades y el tiempo.

Variables aleatorias de Bernouilli

Distribución de Bernouilli

Si p representa la probabilidad de éxito, (1 – p) representa la probabilidad de fracaso y x representa el resultado observado

$$\text{P(exito en el ensayo)} = p^x(1-p)^{1-x}$$

Distribución de Bernouilli describe la probabilidad dado un experimento con solo dos alternativas posibles, de encontrar un éxito.

Variables aleatorias de Bernouilli

- Cada persona en el experimento de Milgram puede considerarse como un ensayo.
- Una persona es etiquetada como éxito si se niega a administrar una descarga severa, y como fracaso si administra dicha descarga.
- Dado que solo el 35% de las personas se negaron a administrar una descarga, la probabilidad de éxito es $p = 0,35$.
- Cuando un ensayo individual tiene sólo dos resultados posibles, se denomina variable aleatoria de Bernoulli.

Distribución geométrica

Probabilidades geométricas

Si p representa la probabilidad de éxito, (1 – p) representa la probabilidad de fracaso y n representa el número de intentos independientes

$$\text{P(exito en el } n^{\text{th}} \text{ ensayo)} = (1-p)^{n-1}p$$

Distribución geométrica describe el tiempo de espera hasta el éxito de independientes e idénticamente distribuidas (iid) variables aleatorias de Bernouilli.

- independencia: los resultados de los ensayos no se afectan entre sí
- idéntico: la probabilidad de éxito es la misma para cada ensayo

Distribución geométrica

Dr. Smith quiere repetir los experimentos de Milgram, pero solo quiere tomar muestras de personas hasta que encuentre a alguien que no le cause un shock severo. ¿Cuál es la probabilidad de que se detenga después de la primera persona?

$$\text{P}(1^{\text{st}} \text{ persona se niega}) = 0.35$$

Distribución geométrica

Dr. Smith quiere repetir los experimentos de Milgram, pero solo quiere tomar muestras de personas hasta que encuentre a alguien que no le cause un shock severo. ¿Cuál es la probabilidad de que se detenga después de la primera persona?

$$\text{P}(1^{\text{st}} \text{ persona se niega}) = 0.35$$

... la tercera persona?

$$\text{P}(1^{\text{st}} \text{ y } 2^{\text{nd}} \text{ shock, } 3^{\text{rd}} \text{ se niega}) = \underline{\text{S}}_{0.65} \times \underline{\text{S}}_{0.65} \times \underline{\text{R}}_{0.35} = 0.65^2 \times 0.35 \approx 0.15$$

Distribución geométrica

Dr. Smith quiere repetir los experimentos de Milgram, pero solo quiere tomar muestras de personas hasta que encuentre a alguien que no le cause un shock severo. ¿Cuál es la probabilidad de que se detenga después de la primera persona?

$$\text{P}(1^{\text{st}} \text{ persona se niega}) = 0.35$$

... la tercera persona?

$$\text{P}(1^{\text{st}} \text{ y } 2^{\text{nd}} \text{ shock, } 3^{\text{rd}} \text{ se niega}) = \underline{\text{S}}_{0.65} \times \underline{\text{S}}_{0.65} \times \underline{\text{R}}_{0.35} = 0.65^2 \times 0.35 \approx 0.15$$

... ¿la décima persona?

Distribución geométrica

Dr. Smith quiere repetir los experimentos de Milgram, pero solo quiere tomar muestras de personas hasta que encuentre a alguien que no le cause un shock severo. ¿Cuál es la probabilidad de que se detenga después de la primera persona?

$$\text{P}(1^{\text{st}} \text{ persona se niega}) = 0.35$$

... la tercera persona?

$$\text{P}(1^{\text{st}} \text{ y } 2^{\text{nd}} \text{ shock, } 3^{\text{rd}} \text{ se niega}) = \underset{0.65}{\text{S}} \times \underset{0.65}{\text{S}} \times \underset{0.35}{\text{R}} = 0.65^2 \times 0.35 \approx 0.15$$

... ¿la décima persona?

$$\text{P}(9 \text{ shock, } 10^{\text{th}} \text{ se niega}) = \underbrace{\underset{0.65}{\text{S}} \times \cdots \times \underset{0.65}{\text{S}}}_{9 \text{ de estos}} \times \underset{0.35}{\text{R}} = 0.65^9 \times 0.35 \approx 0.0072$$

¿Podemos calcular la probabilidad de sacar un 6 por primera vez en la tirada 6^{th} de un dado usando la distribución geométrica? Tenga en cuenta que lo que fue un éxito (sacar un 6) y lo que fue un fracaso (no sacar un 6) están claramente definidos y uno u otro debe suceder para cada intento.

(a) no, al tirar un dado hay más de 2 resultados posibles

(b) sí, por qué no

¿Podemos calcular la probabilidad de sacar un 6 por primera vez en la tirada 6^{th} de un dado usando la distribución geométrica? Tenga en cuenta que lo que fue un éxito (sacar un 6) y lo que fue un fracaso (no sacar un 6) están claramente definidos y uno u otro debe suceder para cada intento.

(a) no, al tirar un dado hay más de 2 resultados posibles

(b) sí, por qué no

$$\text{P(6 on the } 6^{\text{th}} \text{ roll)} = \left(\frac{5}{6}\right)^5 \left(\frac{1}{6}\right) \approx 0,067$$

Valor esperado

¿A cuántas personas se espera que el Dr. Smith examine antes de encontrar al primero que se niegue a administrar la descarga?

Valor esperado

¿A cuántas personas se espera que el Dr. Smith examine antes de encontrar al primero que se niegue a administrar la descarga?

El valor esperado, o la media, de una distribución geométrica se define como $\frac{1}{\mathrm{p}}$.

$$\mu = \frac{1}{\mathrm{p}} = \frac{1}{0,35} = 2,86$$

Valor esperado

¿A cuántas personas se espera que el Dr. Smith examine antes de encontrar al primero que se niegue a administrar la descarga?

El valor esperado, o la media, de una distribución geométrica se define como $\frac{1}{\mathrm{p}}$.

$$\mu = \frac{1}{\mathrm{p}} = \frac{1}{0,35} = 2,86$$

Se espera que evalúe a 2,86 personas antes de encontrar a la primera que se niegue a administrar la descarga.

Valor esperado

¿A cuántas personas se espera que el Dr. Smith examine antes de encontrar al primero que se niegue a administrar la descarga?

El valor esperado, o la media, de una distribución geométrica se define como $\frac{1}{\mathrm{p}}$.

$$\mu = \frac{1}{\mathrm{p}} = \frac{1}{0,35} = 2,86$$

Se espera que evalúe a 2,86 personas antes de encontrar a la primera que se niegue a administrar la descarga.

Pero, ¿cómo puede evaluar a un número no entero de personas?

Media y desviación estándar de la distribución geométrica

$$\mu = \frac{1}{p} \qquad \sigma = \sqrt{\frac{1-p}{p^2}}$$

Media y desviación estándar de la distribución geométrica

$$\mu = \frac{1}{p} \qquad \sigma = \sqrt{\frac{1-p}{p^2}}$$

- Volviendo al experimento del Dr. Smith:

$$\sigma = \sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{1-0.35}{0.35^2}} = 2.3$$

- Se espera que el Dr. Smith analice a 2,86 personas antes de encontrar a la primera que se niegue a administrar la descarga, más o menos 2,3 personas.
- Estos valores solo tienen sentido en el contexto de repetir el experimento muchas veces.

Distribución binomial

Distribución binomial (cont.)

Probabilidades binomiales

Si p representa la probabilidad de éxito, (1 – p) representa la probabilidad de fracaso, n representa el número de ensayos independientes y k representa el número de éxitos

$$\text{P(k éxitos en n ensayos)} = \binom{n}{k} p^k \, (1 - p)^{(nk)}$$

Supongamos que seleccionamos al azar a cuatro personas para que participen en este experimento. ¿Cuál es la probabilidad de que exactamente 1 de ellos se niegue a administrar la descarga?

Supongamos que seleccionamos al azar a cuatro personas para que participen en este experimento. ¿Cuál es la probabilidad de que exactamente 1 de ellos se niegue a administrar la descarga?

Llamemos a estas personas Allen (A), Brittany (B), Caroline (C) y Damian (D). Cada uno de los cuatro escenarios a continuación satisfará la condición de "exactamente 1 de ellos se niega a administrar la descarga":

$$\text{Escenario 1: } \underbrace{0.35}_{\text{(A) rechazar}} \times \underbrace{0.65}_{\text{(B) shock}} \times \underbrace{0.65}_{\text{(C) choque}} \times \underbrace{0.65}_{\text{(D) choque}} = 0.0961$$

$$\text{Escenario 2: } \underbrace{0.65}_{\text{(A) shock}} \times \underbrace{0.35}_{\text{(B) residuos}} \times \underbrace{0.65}_{\text{(C) choque}} \times \underbrace{0.65}_{\text{(D) choque}} = 0.0961$$

$$\text{Escenario 3: } \underbrace{0.65}_{\text{(A) choque}} \times \underbrace{0.65}_{\text{(B) choque}} \times \underbrace{0.35}_{\text{(C) rechazar}} \times \underbrace{0.65}_{\text{(D) shock}} = 0.0961$$

$$\text{Escenario 4: } \underbrace{0.65}_{\text{(A) choque}} \times \underbrace{0.65}_{\text{(B) choque}} \times \underbrace{0.65}_{\text{(C) shock}} \times \underbrace{0.35}_{\text{(D) rechazar}} = 0.0961$$

La probabilidad de que exactamente 1 de 4 personas se nieguen a administrar la descarga es la suma de todas estas probabilidades.

Distribución binomial

La pregunta de la diapositiva anterior pedía la probabilidad de un número dado de éxitos, k, en un número dado de intentos, n, (k = 1 éxito en n = 4 intentos) , y calculamos esta probabilidad como

$$\#\text{ de escenarios} \times \text{P(único escenario)}$$

- # de escenarios: hay una manera menos tediosa de resolver esto, llegaremos a eso en breve...
- P(escenario único) = $p^k\ (1-p)^{(nk)}$
  pequeña probabilidad de éxito elevada a la potencia del número de éxitos, probabilidad de fracaso elevada a la potencia del número de fallos

La distribución binomial describe la probabilidad de tener exactamente k éxitos en n ensayos independientes de Bernouilli con probabilidad de éxito p.

Contando el # de escenarios

Anteriormente escribimos todos los escenarios posibles que se ajustan a la condición de exactamente una persona que se niega a administrar la descarga. Si n fuera mayor y/o k fuera diferente a 1, por ejemplo, n = 9 y k = 2:

RRSSSSSSS
SRRSSSSSS
SSRRSSSSS
...
SSRSSRSSS
...
SSSSSSSRR

escribir todos los escenarios posibles sería increíblemente tedioso y propenso a errores.

Calcular el # de escenarios

Binomio

La función de combinaciones sin repetición es útil para calcular el número de formas de elegir k éxitos en n intentos.

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

- k = 1, n = 4: $\binom{4}{1} = \frac{4!}{1!(4-1)!} = \frac{4\times3\times2\times1}{1\times(3\times2\times1)} = 4$
- k = 2, n = 9: $\binom{9}{2} = \frac{9!}{2!(9-2)!} = \frac{9\times8\times7!}{2\times1\times7!} = \frac{72}{2} = 36$

Note: También puede usar R para estos cálculos:

```
> elegir(9,2)
[1] 36
```

Propiedades de la función de selección

¿Cuál de las siguientes es falsa?

(a) Hay n formas de obtener 1 éxito en n intentos, $\binom{n}{1} = n$.

(b) Solo hay 1 forma de obtener n éxitos en n intentos, $\binom{n}{n} = 1$.

(c) Solo hay 1 forma de obtener n fallas en n pruebas, $\binom{n}{0} = 1$.

(d) Hay $n - 1$ formas de obtener $n - 1$ éxitos en n intentos, $\binom{n}{n-1} = n - 1$.

Propiedades de la función de selección

¿Cuál de las siguientes es falsa?

(a) Hay n formas de obtener 1 éxito en n intentos, $\binom{n}{1} = n$.

(b) Solo hay 1 forma de obtener n éxitos en n intentos, $\binom{n}{n} = 1$.

(c) Solo hay 1 forma de obtener n fallas en n pruebas, $\binom{n}{0} = 1$.

(d) Hay $n - 1$ formas de obtener $n - 1$ éxitos en n intentos, $\binom{n}{n-1} = n - 1$.

¿Cuál de las siguientes no es una condición que debe cumplirse para que se aplique la distribución binomial?

(a) los ensayos deben ser independientes

(b) el número de intentos, n, debe ser fijo

(c) el resultado de cada ensayo debe clasificarse como éxito o fracaso

(d) el número de éxitos deseados, k, debe ser mayor que el número de intentos

(e) la probabilidad de éxito, p, debe ser la misma para cada ensayo

¿Cuál de las siguientes no es una condición que debe cumplirse para que se aplique la distribución binomial?

(a) los ensayos deben ser independientes

(b) el número de intentos, n, debe ser fijo

(c) el resultado de cada ensayo debe clasificarse como éxito o fracaso

(d) el número de éxitos deseados, k, debe ser mayor que el número de intentos

(e) la probabilidad de éxito, p, debe ser la misma para cada ensayo

Una encuesta de Gallup de 2012 sugiere que el 26,2% de los estadounidenses son obesos. Entre una muestra aleatoria de 10 estadounidenses, ¿cuál es la probabilidad de que exactamente 8 sean obesos?

(a) bastante alto

(b) bastante bajo

Gallup: http://www.gallup.com/poll/160061/obesity-rate-stable-2012.aspx, 23 de enero de 2013.

Una encuesta de Gallup de 2012 sugiere que el 26,2% de los estadounidenses son obesos. Entre una muestra aleatoria de 10 estadounidenses, ¿cuál es la probabilidad de que exactamente 8 sean obesos?

(a) bastante alto

(b) bastante bajo

Gallup: http://www.gallup.com/poll/160061/obesity-rate-stable-2012.aspx, 23 de enero de 2013.

Una encuesta de Gallup de 2012 sugiere que el 26,2% de los estadounidenses son obesos. Entre una muestra aleatoria de 10 estadounidenses, ¿cuál es la probabilidad de que exactamente 8 sean obesos?

(a) $0.262^8 \times 0.738^2$

(b) $\binom{8}{10} \times 0.262^8 \times 0.738^2$

(c) $\binom{10}{8} \times 0,262^8 \times 0,738^2$

(d) $\binom{10}{8} \times 0.262^2 \times 0.738^8$

Una encuesta de Gallup de 2012 sugiere que el 26,2% de los estadounidenses son obesos. Entre una muestra aleatoria de 10 estadounidenses, ¿cuál es la probabilidad de que exactamente 8 sean obesos?

(a) $0.262^8 \times 0.738^2$

(b) $\binom{8}{10} \times 0.262^8 \times 0.738^2$

(c) $\binom{10}{8} \times 0,262^8 \times 0,738^2 \ = 45 \times 0,262^8 \times 0,738^2 = 0,0005$

(d) $\binom{10}{8} \times 0.262^2 \times 0.738^8$

El problema del cumpleaños

¿Cuál es la probabilidad de que 2 personas elegidas al azar compartan cumpleaños?

Bastante bajo, $\frac{1}{365} \approx 0.0027$.

El problema del cumpleaños

¿Cuál es la probabilidad de que 2 personas elegidas al azar compartan cumpleaños?

Bastante bajo, $\frac{1}{365} \approx 0.0027$.

¿Cuál es la probabilidad de que al menos 2 personas de 366 compartan cumpleaños?

El problema del cumpleaños

¿Cuál es la probabilidad de que 2 personas elegidas al azar compartan cumpleaños?

Bastante bajo, $\frac{1}{365} \approx 0.0027$.

¿Cuál es la probabilidad de que al menos 2 personas de 366 compartan cumpleaños?

¡Exactamente 1! (Excluyendo la posibilidad de un cumpleaños en año bisiesto).

El problema del cumpleaños (cont.)

¿Cuál es la probabilidad de que al menos 2 personas (1 coincidencia) de 121 personas compartan cumpleaños?

Algo complicado de calcular, pero podemos pensarlo como el complemento de la probabilidad de que no haya coincidencias en 121 personas.

$$
\begin{aligned}
\text{P(sin coincidencias)} &= 1 \times \left(1 - \frac{1}{365}\right) \times \left(1 - \frac{2}{365}\right) \times \cdots \times \left(1 - \frac{120}{365}\right) \\
&= \frac{365 \times 364 \times \cdots \times 245}{365^{121}} \\
&= \frac{365!}{365^{121} \times (365 - 121)!} \\
&= \frac{121! \times \binom{365}{121}}{365^{121}} \approx 0 \\
\text{P(al menos 1 coincidencia)} &\approx 1
\end{aligned}
$$

Valor esperado

Una encuesta de Gallup de 2012 sugiere que el 26,2% de los estadounidenses son obesos.

Entre una muestra aleatoria de 100 estadounidenses, ¿cuántos esperaría que fueran obesos?

- Bastante fácil, $100 \times 0.262 = 26.2$.
- O más formalmente, $\mu = \text{np} = 100 \times 0.262 = 26.2$.
- Pero esto no significa que en cada muestra aleatoria de 100 personas, exactamente 26,2 serán obesas. De hecho, eso ni siquiera es posible. En algunas muestras este valor será menor, y en otras más. ¿Cuánto esperaríamos que varíe este valor?

Media y desviación estándar de la distribución binomial

$$\mu = \mathrm{np} \qquad \sigma = \sqrt{\mathrm{np}(1-\mathrm{p})}$$

Media y desviación estándar de la distribución binomial

$$\mu = \text{np} \qquad \sigma = \sqrt{\text{np}(1-\text{p})}$$

- Volviendo a la tasa de obesidad:

$$\sigma = \sqrt{\text{np}(1-\text{p})} = \sqrt{100 \times 0,262 \times 0,738} \approx 4,4$$

- Esperaríamos que 26,2 de cada 100 estadounidenses seleccionados al azar fueran obesos, con una desviación estándar de 4,4.

Note: La media y la desviación estándar de un binomio pueden no ser siempre números enteros, y eso está bien, estos valores representan lo que esperaríamos ver en promedio.

Observaciones inusuales

Usando la noción de que las observaciones que están a más de 2 desviaciones estándar de la media se consideran inusuales y la media y la desviación estándar que acabamos de calcular, podemos calcular un rango para el número plausible de estadounidenses obesos en muestras aleatorias de 100.

$$26,2 \pm (2 \times 4,4) = (17,4, 35)$$

Una encuesta de Gallup de agosto de 2012 sugiere que el 13% de los estadounidenses piensa que la educación en el hogar proporciona una educación excelente para los niños. ¿Se consideraría inusual una muestra aleatoria de 1000 estadounidenses donde solo 100 comparten esta opinión?

(a) No

(b) Si

	Excellent	Good	Only fair	Poor	Total excellent/ good
	%	%	%	%	%
Independent private school	31	47	13	2	78
Parochial or church-related schools	21	48	18	5	69
Charter schools	17	43	23	5	60
Home schooling	13	33	30	14	46
Public schools	5	32	42	19	37

Gallup, Aug. 9-12, 2012

Una encuesta de Gallup de agosto de 2012 sugiere que el 13% de los estadounidenses piensa que la educación en el hogar proporciona una educación excelente para los niños. ¿Se consideraría inusual una muestra aleatoria de 1000 estadounidenses donde solo 100 comparten esta opinión?

(a) No

(b) Si

$$\mu = \text{np} = 1000 \times 0,13 = 130$$

$$\sigma = \sqrt{\text{np}(1-\text{p})} = \sqrt{1000 \times 0,13 \times 0,87} \approx 10,6$$

http://www.gallup.com/poll/156974/private-schools-top-marks-educating-children.aspx

Una encuesta de Gallup de agosto de 2012 sugiere que el 13% de los estadounidenses piensa que la educación en el hogar proporciona una educación excelente para los niños. ¿Se consideraría inusual una muestra aleatoria de 1000 estadounidenses donde solo 100 comparten esta opinión?

(a) No

(b) Si

$$\mu = \text{np} = 1000 \times 0,13 = 130$$

$$\sigma = \sqrt{\text{np}(1-\text{p})} = \sqrt{1000 \times 0,13 \times 0,87} \approx 10,6$$

Método 1: Rango de observaciones usuales:
$130 \pm 2 \times 10.6 = (108.8, 151.2)$
100 está fuera de este rango, por lo que se consideraría inusual.

http://www.gallup.com/poll/156974/private-schools-top-marks-educating-children.aspx

Una encuesta de Gallup de agosto de 2012 sugiere que el 13% de los estadounidenses piensa que la educación en el hogar proporciona una educación excelente para los niños. ¿Se consideraría inusual una muestra aleatoria de 1000 estadounidenses donde solo 100 comparten esta opinión?

(a) No

(b) Si

$$\mu = \text{np} = 1000 \times 0,13 = 130$$

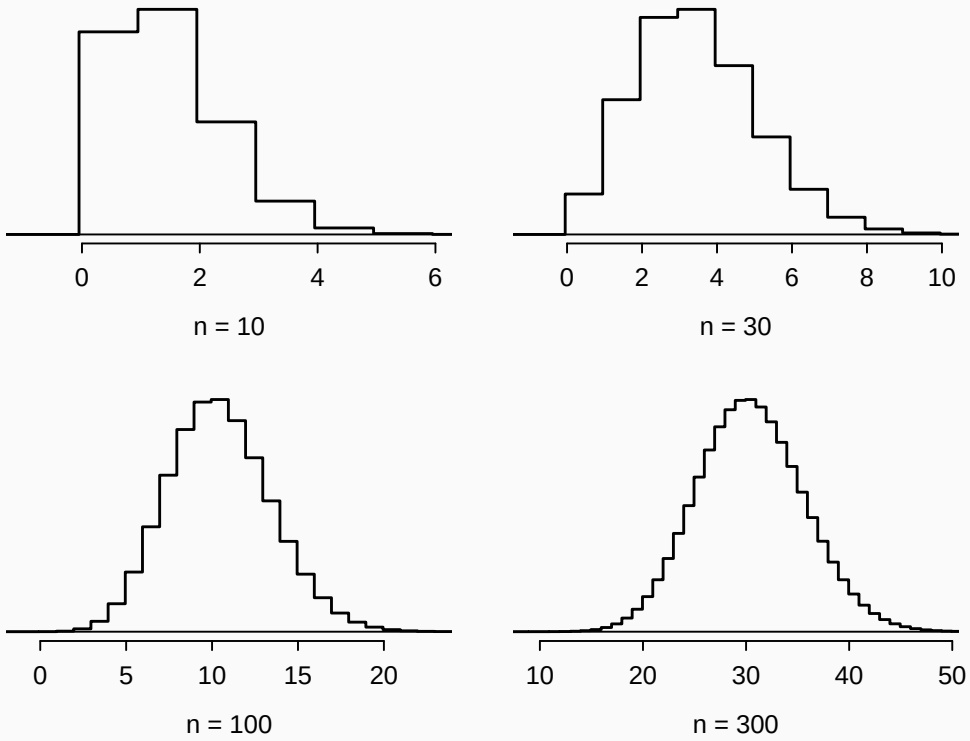
$$\sigma = \sqrt{\text{np}(1-\text{p})} = \sqrt{1000 \times 0,13 \times 0,87} \approx 10,6$$

Método 1: Rango de observaciones usuales:
$130 \pm 2 \times 10.6 = (108.8, 151.2)$
100 está fuera de este rango, por lo que se consideraría inusual.

Método 2: Puntuación Z de la observación:
$\text{Z} = \frac{\text{x–mean}}{\text{SD}} = \frac{100-130}{10.6} = -2.83$
100 es más de 2 SD por debajo de la media, por lo que se consideraría inusual.

Distribuciones del número de éxitos

Histogramas huecos de muestras del modelo binomial donde p = 0.10 y n = 10, 30, 100 y 300. ¿Qué sucede a medida que aumenta n?



¿Qué tan grande es lo suficientemente grande?

El tamaño de la muestra se considera lo suficientemente grande si el número esperado de éxitos y fracasos es al menos 10.

$$np \geq 10 \qquad \text{y} \qquad n(1-p) \geq 10$$

Abajo hay cuatro pares de parámetros de distribución Binomial. ¿Qué distribución se puede aproximar mediante la distribución normal?

(a) $\mathrm{n} = 100, \mathrm{p} = 0,95$

(b) $\mathrm{n} = 25, \mathrm{p} = 0,45$

(c) $\mathrm{n} = 150, \mathrm{p} = 0,05$

(d) $\mathrm{n} = 500, \mathrm{p} = 0.015$

Abajo hay cuatro pares de parámetros de distribución Binomial. ¿Qué distribución se puede aproximar mediante la distribución normal?

(a) $n = 100, p = 0,95$

(b) $n = 25, p = 0,45 \rightarrow 25 \times 0,45 = 11,25; 25 \times 0,55 = 13,75$

(c) $n = 150, p = 0,05$

(d) $n = 500, p = 0.015$

Un análisis de los usuarios de Facebook

Un estudio reciente encontró que “Los usuarios de Facebook obtienen más de lo que dan”. Por ejemplo:

- El 40% de los usuarios de Facebook de nuestra muestra hizo una solicitud de amistad, pero el 63% recibió al menos una solicitud
- Los usuarios de nuestra muestra presionaron el botón ”Me gusta” junto al contenido de sus amigos un promedio de 14 veces, pero su contenido ”me gustó” un promedio de 20 veces
- Los usuarios enviaron 9 mensajes personales, pero recibieron 12
- El 12% de los usuarios etiquetaron a un amigo en una foto, pero el 35% fueron ellos mismos etiquetados en una foto

¿Alguna idea de cómo se puede explicar este patrón?

http://www.pewinternet.org/Reports/2012/Facebook-users/Summary.aspx

Un análisis de los usuarios de Facebook

Un estudio reciente encontró que “Los usuarios de Facebook obtienen más de lo que dan”. Por ejemplo:

- El 40% de los usuarios de Facebook de nuestra muestra hizo una solicitud de amistad, pero el 63% recibió al menos una solicitud
- Los usuarios de nuestra muestra presionaron el botón ”Me gusta” junto al contenido de sus amigos un promedio de 14 veces, pero su contenido ”me gustó” un promedio de 20 veces
- Los usuarios enviaron 9 mensajes personales, pero recibieron 12
- El 12% de los usuarios etiquetaron a un amigo en una foto, pero el 35% fueron ellos mismos etiquetados en una foto

¿Alguna idea de cómo se puede explicar este patrón?

Los usuarios avanzados aportan mucho más contenido que el usuario típico.

http://www.pewinternet.org/Reports/2012/Facebook-users/Summary.aspx

Este estudio también encontró que aproximadamente el 25% de los usuarios de Facebook se consideran usuarios avanzados. El mismo estudio encontró que el usuario promedio de Facebook tiene 245 amigos. ¿Cuál es la probabilidad de que el usuario promedio de Facebook con 245 amigos tenga 70 o más amigos que se considerarían usuarios avanzados? Tenga en cuenta cualquier suposición que deba hacer.

Nos dan que $n = 245, p = 0.25$, y nos piden la probabilidad $P(K \geq 70)$. Para proceder, necesitamos independencia, que asumiremos pero que podríamos verificar si tuviéramos acceso a más datos de Facebook.

$$\begin{aligned} P(X \geq 70) &= P(K = 70 \text{ o } K = 71 \text{ o } K = 72 \text{ o } \cdots \text{ o } K = 245) \\ &= P(K = 70) + P(K = 71) + P(K = 72) + \cdots + P(K = 245) \end{aligned}$$

Esto parece mucho trabajo...

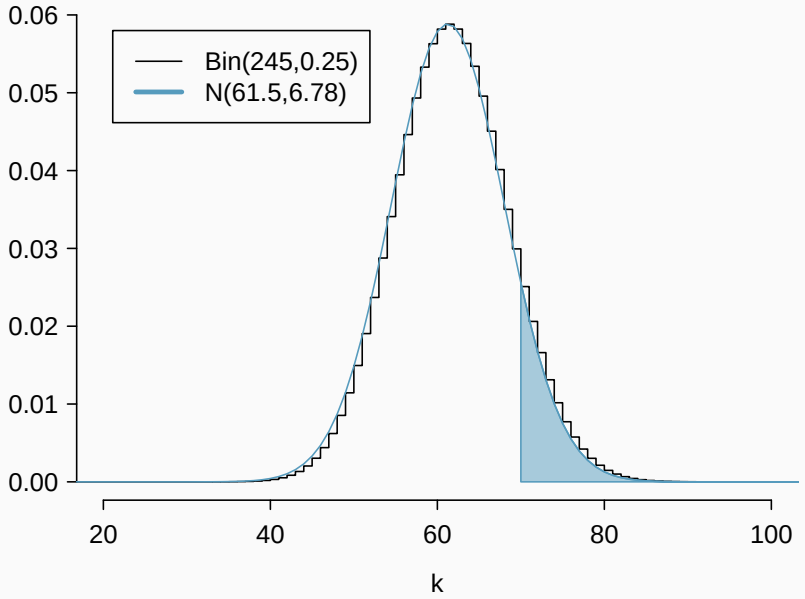
Aproximación normal al binomio

Cuando el tamaño de la muestra es lo suficientemente grande, la distribución binomial con los parámetros n y p puede aproximarse mediante el modelo normal con los parámetros $\mu = \text{np}$ y $\sigma = \sqrt{\text{np}(1-\text{p})}$

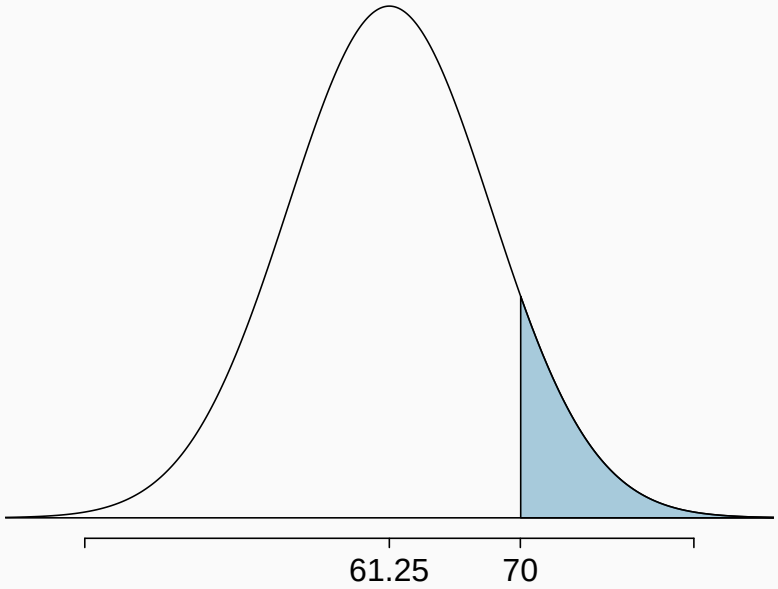
- En el caso de los usuarios avanzados de Facebook, $\text{n} = 245$ y $\text{p} = 0,25$.

$$\mu = 245 \times 0,25 = 61,25 \qquad \sigma = \sqrt{245 \times 0,25 \times 0,75} = 6,78$$

- $\text{Bin}(\text{n} = 245, \text{p} = 0,25) \approx \text{N}(\mu = 61,25, \sigma = 6,78)$.



¿Cuál es la probabilidad de que el usuario promedio de Facebook con 245 amigos tenga 70 o más amigos que serían considerados usuarios avanzados?



$$Z = \frac{\text{obs} - \text{mean}}{\text{SD}} = \frac{70 - 61.25}{6.78} = 1.29$$

$$P(Z > 1.29) = 1 - 0.9015 = 0.0985$$

```
> pnorm(1.29)
[1] 0.9014747
```

La aprox. normal se descompone en pequeños intervalos

- La aproximación normal a la distribución binomial tiende a funcionar mal cuando se estima la probabilidad de un pequeño rango de recuentos, incluso cuando se cumplen las condiciones.
- Esta aproximación para intervalos de valores generalmente se mejora si los valores de corte se extienden en 0.5 en ambas direcciones.
- La sugerencia de agregar área adicional al aplicar la aproximación normal suele ser útil cuando se examina un rango de observaciones. Si bien también es posible aplicar esta corrección al calcular un área de cola, el beneficio de la modificación generalmente desaparece ya que el intervalo total suele ser bastante amplio.

Distribución binomial negativa

Distribución binomial negativa

$$P(k^{th}\text{exito en la prueba } n^{th}) = \binom{n-1}{k-1} p^k \, (1-p)^{nk}$$

donde p es la probabilidad de que un ensayo individual sea un éxito. Se supone que todos los ensayos son independientes.

- La distribución binomial negativa describe la probabilidad de observar el k^{th} éxito en la n^{th} prueba.
- Las siguientes cuatro condiciones son útiles para identificar un caso binomial negativo:
  1. Los ensayos son independientes.
  2. Cada resultado de la prueba se puede clasificar como un éxito o un fracaso.
  3. La probabilidad de éxito (p) es la misma para cada prueba.
  4. La última prueba debe ser un éxito.

  Tenga en cuenta que las tres primeras condiciones son comunes a la distribución binomial.

A un estudiante universitario que trabaja en un laboratorio de psicología se le pide que reclute 10 parejas para participar en un estudio. Decide pararse fuera del centro de estudiantes y preguntar a cada 5^{th} persona que sale del edificio si tiene una relación y, de ser así, si le gustaría participar en el estudio con su pareja. Supongamos que la probabilidad de encontrar a esa persona es del 10%. ¿Cuál es la probabilidad de que necesite preguntarle a 30 personas antes de alcanzar su meta?

A un estudiante universitario que trabaja en un laboratorio de psicología se le pide que reclute 10 parejas para participar en un estudio. Decide pararse fuera del centro de estudiantes y preguntar a cada 5^{th} persona que sale del edificio si tiene una relación y, de ser así, si le gustaría participar en el estudio con su pareja. Supongamos que la probabilidad de encontrar a esa persona es del 10%. ¿Cuál es la probabilidad de que necesite preguntarle a 30 personas antes de alcanzar su meta?

Dado: p = 0.10, k = 10, n = 30. Se nos pide encontrar la probabilidad de éxito de 10^{th} en la prueba de 30^{th}, por lo tanto, usamos la distribución binomial negativa.

A un estudiante universitario que trabaja en un laboratorio de psicología se le pide que reclute 10 parejas para participar en un estudio. Decide pararse fuera del centro de estudiantes y preguntar a cada 5^{th} persona que sale del edificio si tiene una relación y, de ser así, si le gustaría participar en el estudio con su pareja. Supongamos que la probabilidad de encontrar a esa persona es del 10%. ¿Cuál es la probabilidad de que necesite preguntarle a 30 personas antes de alcanzar su meta?

Dado: p = 0.10, k = 10, n = 30. Se nos pide encontrar la probabilidad de éxito de 10^{th} en la prueba de 30^{th}, por lo tanto, usamos la distribución binomial negativa.

$$\begin{aligned} \text{P}(10^{\text{th}} \text{ éxito en la prueba de } 30^{\text{th}}) &= \binom{29}{9} \times 0.10^{10} \times 0.90^{20} \\ &= 10.015.005 \times 0,10^{10} \times 0,90^{20} \\ &= 0.00012 \end{aligned}$$

Binomio vs. binomio negativo

¿En qué se diferencia la distribución binomial negativa de la distribución binomial?

- En el caso binomial, normalmente tenemos un número fijo de intentos y en su lugar consideramos el número de éxitos.
- En el caso binomial negativo, examinamos cuántos intentos se necesitan para observar un número fijo de éxitos y requerimos que la última observación sea un éxito.

Práctica

¿Cuál de los siguientes describe un caso en el que usaríamos la distribución binomial negativa para calcular la probabilidad deseada?

(a) Probabilidad de que un niño de 5 años mida más de 42 pulgadas.

(b) Probabilidad de que 3 de 10 lanzamientos de softbol sean exitosos.

(c) Probabilidad de recibir una mano de color en el póquer.

(d) Probabilidad de fallar 8 tiros antes del primer golpe.

(e) Probabilidad de golpear la pelota por 3rd vez en el 8^{th} intento.

Práctica

¿Cuál de los siguientes describe un caso en el que usaríamos la distribución binomial negativa para calcular la probabilidad deseada?

(a) Probabilidad de que un niño de 5 años mida más de 42 pulgadas.

(b) Probabilidad de que 3 de 10 lanzamientos de softbol sean exitosos.

(c) Probabilidad de recibir una mano de color en el póquer.

(d) Probabilidad de fallar 8 tiros antes del primer golpe.

(e) Probabilidad de golpear la pelota por 3rd vez en el 8^{th} intento.

Distribución de Poisson

Distribución de Poisson

$$\text{P(observar k eventos raros)} = \frac{\lambda^{\text{k}} \text{e}^{-\lambda}}{\text{k!}}$$

donde k puede tomar un valor de 0, 1, 2, etc., y k! representa k-factorial.
La letra e ≈ 2.718 es la base del logaritmo natural.
La media y la desviación estándar de esta distribución son λ y $\sqrt{\lambda}$, respectivamente.

- La distribución de Poisson suele ser útil para estimar el número de eventos raros en una población grande durante una unidad de tiempo corta para una población fija si los individuos dentro de la población son independientes.
- La tasa para una distribución de Poisson es el número promedio de ocurrencias en una población mayoritariamente fija por unidad de tiempo, y normalmente se denota por λ.
- Usando la tasa, podemos describir la probabilidad de observar exactamente k eventos raros en una sola unidad de tiempo.

Suponga que en una región rural de un país en desarrollo ocurren fallas en el suministro eléctrico siguiendo una distribución de Poisson con un promedio de 2 fallas cada semana. Calcula la probabilidad de que en una semana determinada la electricidad falle solo una vez.

Dado $\lambda = 2$.

$$\begin{aligned} \text{P(solo 1 falla en una semana)} &= \frac{2^1 \times \mathrm{e}^{-2}}{1!} \\ &= \frac{2 \times \mathrm{e}^{-2}}{1} \\ &= 0.27 \end{aligned}$$

Suponga que en una región rural de un país en desarrollo ocurren fallas en el suministro eléctrico siguiendo una distribución de Poisson con un promedio de 2 fallas cada semana. Calcula la probabilidad de que en un dia determinado la electricidad falle tres veces.

Nos dan la tasa de falla semanal, pero para responder a esta pregunta primero debemos calcular la tasa promedio de falla en un día determinado: $\lambda_{\text{i}} = \frac{2}{7} = 0.2857$. Tenga en cuenta que estamos asumiendo que la probabilidad de corte de energía es la misma en cualquier día de la semana, es decir, asumimos independencia.

$$\begin{aligned} \text{P(3 fallas en un dia dado)} &= \frac{0.2857^3 \times \text{e}^{-0.2857}}{3!} \\ &= \frac{0.2857^3 \times \text{e}^{-0.2857}}{6} \\ &= 0.0029 \end{aligned}$$

¿Es Poisson?

- Una variable aleatoria puede seguir una distribución de Poisson si el evento que se considera es raro, la población es grande y los eventos ocurren independientemente unos de otros
- Sin embargo, podemos pensar en situaciones en las que los eventos no son realmente independientes. Por ejemplo, si estamos interesados en la probabilidad de un cierto número de bodas durante un verano, debemos tener en cuenta que los fines de semana son más populares para las bodas.
- En este caso, un modelo de Poisson a veces puede seguir siendo razonable si permitimos que tenga una tasa diferente para tiempos diferentes; podríamos modelar la tasa como más alta los fines de semana que entre semana.
- La idea de modelar tasas para una distribución de Poisson contra una segunda variable (día de la semana) forma el base de algunos métodos más avanzados llamados modelos lineales generalizados. Hay más allá del alcance de este curso, pero discutiremos una base de modelos lineales en los capítulos 7 y 8.

Práctica

¿Una variable aleatoria que sigue a cuál de las siguientes distribuciones puede tomar valores que no sean números enteros positivos?

(a) veneno

(b) Binomial negativo

(c) Binomio

(d) Normal

(e) geométrico

Práctica

¿Una variable aleatoria que sigue a cuál de las siguientes distribuciones puede tomar valores que no sean números enteros positivos?

(a) veneno

(b) Binomial negativo

(c) Binomio

(d) Normal

(e) geométrico