Cap. 5 : Distribuciones muestrales

Alexandre Blondin Massé

Departamento de Informática y Matematica
Université du Québec à Chicoutimi

18 de junio del 2015
Modelado de sistemas aleatorios
Ingeniería de sistemas, producción y ambiental

Tabla de contenidos

Teorema

Teorema

Sean X_1, X_2, ..., X_n ***variables aleatorias*** *que tienen la* ***misma distribución*** *de* ***media*** μ *y de* ***varianza*** σ^2*. Si* $n \to \infty$*, entonces la distribución de*

$$X_1 + X_2 + \ldots + X_n$$

es aproximadamente ***normal*** *de* ***media*** $n\mu$ *y de* ***varianza*** $n\sigma^2$*.*

Corolario

La variable aleatoria

$$\frac{X_1 + X_2 + \ldots + X_n - n\mu}{\sigma\sqrt{n}}$$

sigue ***aproximadamente*** *una* ***distribución normal estándar****.*

Ejemplo (1/2)

- Una compañía de seguros tiene 25000 **clientes**.
- El precio reclamado **anualmente** por un cliente es una variable aleatoria de **media** 320\$ y de **desviación estándar** 540\$.
- ¿Cuál es la probabilidad que la **cantidad total** reclamada por los clientes sea **mayor que** $8{,}3M\$$.

Ejemplo (1/2)

- Una compañía de seguros tiene 25000 **clientes**.
- El precio reclamado **anualmente** por un cliente es una variable aleatoria de **media** 320\$ y de **desviación estándar** 540\$.
- ¿Cuál es la probabilidad que la **cantidad total** reclamada por los clientes sea **mayor que** $8{,}3M\$$.
- Sea X_i la cantidad reclamada por el **cliente** i y $X = \sum_{i=1}^{n} X_i$.

Ejemplo (1/2)

- Una compañía de seguros tiene 25000 **clientes**.
- El precio reclamado **anualmente** por un cliente es una variable aleatoria de **media** 320\$ y de **desviación estándar** 540\$.
- ¿Cuál es la probabilidad que la **cantidad total** reclamada por los clientes sea **mayor que** $8{,}3M\$$.
- Sea X_i la cantidad reclamada por el **cliente** i y $X = \sum_{i=1}^{n} X_i$.
- Como n est bastante **grande**, se aplica el **teorema del límite central**.

Ejemplo (2/2)

- Aquí, $n = 25000$.

Ejemplo (2/2)

- Aquí, $n = 25000$.
- Entonces, por el **teorema del límite central**, tenemos que X es aproximadamente **normal** con parámetros

$$\begin{aligned} \mu &= 320 \times 25000 = 8 \times 10^6 \\ \sigma &= 540\sqrt{25000} \approx 8{,}5381 \times 10^4. \end{aligned}$$

Ejemplo (2/2)

- Aquí, $n = 25000$.
- Entonces, por el **teorema del límite central**, tenemos que X es aproximadamente **normal** con parámetros

$$
\begin{aligned}
\mu &= 320 \times 25000 = 8 \times 10^6 \\
\sigma &= 540\sqrt{25000} \approx 8{,}5381 \times 10^4.
\end{aligned}
$$

- Consecuamente,

$$P(X > 8{,}3 \times 10^6)$$

Ejemplo (2/2)

- Aquí, $n = 25000$.
- Entonces, por el **teorema del límite central**, tenemos que X es aproximadamente **normal** con parámetros

$$\begin{aligned} \mu &= 320 \times 25000 = 8 \times 10^6 \\ \sigma &= 540\sqrt{25000} \approx 8{,}5381 \times 10^4. \end{aligned}$$

- Consecuamente,

$$P(X > 8{,}3 \times 10^6) \approx P\left(Z > \frac{8{,}3 \times 10^6 - 8 \times 10^6}{8{,}5381 \times 10^4}\right)$$

Ejemplo (2/2)

- Aquí, $n = 25000$.
- Entonces, por el **teorema del límite central**, tenemos que X es aproximadamente **normal** con parámetros

$$\begin{aligned} \mu &= 320 \times 25000 = 8 \times 10^6 \\ \sigma &= 540\sqrt{25000} \approx 8{,}5381 \times 10^4. \end{aligned}$$

- Consecuamente,

$$\begin{aligned} P(X > 8{,}3 \times 10^6) &\approx P\left(Z > \frac{8{,}3 \times 10^6 - 8 \times 10^6}{8{,}5381 \times 10^4}\right) \\ &\approx P(Z > 3{,}51) \end{aligned}$$

Ejemplo (2/2)

- Aquí, $n = 25000$.
- Entonces, por el **teorema del límite central**, tenemos que X es aproximadamente **normal** con parámetros

$$\begin{aligned} \mu &= 320 \times 25000 = 8 \times 10^6 \\ \sigma &= 540\sqrt{25000} \approx 8{,}5381 \times 10^4. \end{aligned}$$

- Consecuamente,

$$\begin{aligned} P(X > 8{,}3 \times 10^6) &\approx P\left(Z > \frac{8{,}3 \times 10^6 - 8 \times 10^6}{8{,}5381 \times 10^4}\right) \\ &\approx P(Z > 3{,}51) \\ &\approx \mathbf{0{,}00023}. \end{aligned}$$

Aplicación (1/3)

- El **peso total** W en toneladas que un **puente puede soportar** tiene una distribución **normal** de **media** 400 y de **desviación estándar** 40.
- Suponemos que el **peso** en toneladas de un carro sigue una **distribución normal** de **media** 3 y de **desviación estándar** $0{,}3$.
- ¿Cuántos carros son necesarios para que el **riesgo** que haya un problema sea **mayor que** $10\,\%$?

Aplicación (1/3)

- El **peso total** W en toneladas que un **puente puede soportar** tiene una distribución **normal** de **media** 400 y de **desviación estándar** 40.
- Suponemos que el **peso** en toneladas de un carro sigue una **distribución normal** de **media** 3 y de **desviación estándar** $0{,}3$.
- ¿Cuántos carros son necesarios para que el **riesgo** que haya un problema sea **mayor que** $10\,\%$?
- Suponemos que hay n **carros** en el puente y sea W_i **el peso** del carro i.

Aplicación (1/3)

- El **peso total** W en toneladas que un **puente puede soportar** tiene una distribución **normal** de **media** 400 y de **desviación estándar** 40.
- Suponemos que el **peso** en toneladas de un carro sigue una **distribución normal** de **media** 3 y de **desviación estándar** $0{,}3$.
- ¿Cuántos carros son necesarios para que el **riesgo** que haya un problema sea **mayor que** $10\,\%$?
- Suponemos que hay n **carros** en el puente y sea W_i **el peso** del carro i.
- Obtenemos

$$P_n = P(W_1 + W_2 + \cdots + W_n \geq W)$$

Aplicación (1/3)

- El **peso total** W en toneladas que un **puente puede soportar** tiene una distribución **normal** de **media** 400 y de **desviación estándar** 40.
- Suponemos que el **peso** en toneladas de un carro sigue una **distribución normal** de **media** 3 y de **desviación estándar** 0,3.
- ¿Cuántos carros son necesarios para que el **riesgo** que haya un problema sea **mayor que** 10 %?
- Suponemos que hay n **carros** en el puente y sea W_i **el peso** del carro i.
- Obtenemos

$$
\begin{aligned}
P_n &= P(W_1 + W_2 + \cdots + W_n \geq W) \\
&= P(W_1 + W_2 + \cdots + W_n - W \geq 0).
\end{aligned}
$$

Aplicación (2/3)

- Por el teorema del límite central, $\sum_{i=1}^{n} W_i$ es aproximadamente normal de **media** $3n$ y de varianza $(0{,}3)^2 n = \mathbf{0{,}09n}$.

Aplicación (2/3)

- Por el teorema del límite central, $\sum_{i=1}^{n} W_i$ es aproximadamente normal de **media** $3n$ y de varianza $(0{,}3)^2 n = \mathbf{0{,}09n}$.
- También, parece razonable que W y las W_i son **independientes**;

Aplicación (2/3)

- Por el teorema del límite central, $\sum_{i=1}^{n} W_i$ es aproximadamente normal de **media** $3n$ y de varianza $(0{,}3)^2 n = \mathbf{0{,}09n}$.
- También, parece razonable que W y las W_i son **independientes**;
- W sigue una **distribución normal** de tal manera que la variable aleatoria $X = \sum_{i=1}^{n} W_i - W$ es aproximadamente normal con

$$\begin{aligned} E[X] &= E\left[\sum_{i=1}^{n} W_i - W\right] = 3n - 400 \\ \mathrm{Var}(X) &= \mathrm{Var}\left(\sum_{i=1}^{n} W_i\right) + \mathrm{Var}(W) = 0{,}09n + 1600. \end{aligned}$$

Aplicación (3/3)

- Buscamos z que satisface $P(Z \geq z) = 0{,}1$;

Aplicación (3/3)

- Buscamos z que satisface $P(Z \geq z) = 0{,}1$;
- Leyendo una table de distribución normal, $z \approx 1{,}28$.

Aplicación (3/3)

- Buscamos z que satisface $P(Z \geq z) = 0{,}1$;
- Leyendo una table de distribución normal, $z \approx 1{,}28$.
- Entonces, buscamos n de tal manera que

$$\frac{400 - 3n}{\sqrt{0{,}09n + 1600}} \leq 1{,}28.$$

Aplicación (3/3)

- Buscamos z que satisface $P(Z \geq z) = 0{,}1$;
- Leyendo una table de distribución normal, $z \approx 1{,}28$.
- Entonces, buscamos n de tal manera que

$$\frac{400 - 3n}{\sqrt{0{,}09n + 1600}} \leq 1{,}28.$$

- Obtenemos

$$\frac{400 - 3n}{\sqrt{0{,}09n + 1600}} = 1{,}28$$

Aplicación (3/3)

- Buscamos z que satisface $P(Z \geq z) = 0{,}1$;
- Leyendo una table de distribución normal, $z \approx 1{,}28$.
- Entonces, buscamos n de tal manera que

$$\frac{400 - 3n}{\sqrt{0{,}09n + 1600}} \leq 1{,}28.$$

- Obtenemos

$$\frac{400 - 3n}{\sqrt{0{,}09n + 1600}} = 1{,}28 \quad \Rightarrow \quad 160000 - 2400n + 9n^2 = 1{,}28^2(0{,}09n + 1600)$$

Aplicación (3/3)

- Buscamos z que satisface $P(Z \geq z) = 0{,}1$;
- Leyendo una table de distribución normal, $z \approx 1{,}28$.
- Entonces, buscamos n de tal manera que

$$\frac{400 - 3n}{\sqrt{0{,}09n + 1600}} \leq 1{,}28.$$

- Obtenemos

$$\begin{aligned}\frac{400 - 3n}{\sqrt{0{,}09n + 1600}} = 1{,}28 \quad &\Rightarrow \quad 160000 - 2400n + 9n^2 = 1{,}28^2(0{,}09n + 1600) \\ &\Rightarrow \quad 9n^2 - 2400{,}147456n + 157378{,}56 = 0\end{aligned}$$

Aplicación (3/3)

- Buscamos z que satisface $P(Z \geq z) = 0{,}1$;
- Leyendo una table de distribución normal, $z \approx 1{,}28$.
- Entonces, buscamos n de tal manera que

$$\frac{400 - 3n}{\sqrt{0{,}09n + 1600}} \leq 1{,}28.$$

- Obtenemos

$$\begin{aligned}
\frac{400 - 3n}{\sqrt{0{,}09n + 1600}} = 1{,}28 \quad & \Rightarrow \quad 160000 - 2400n + 9n^2 = 1{,}28^2(0{,}09n + 1600) \\
& \Rightarrow \quad 9n^2 - 2400{,}147456n + 157378{,}56 = 0 \\
& \Rightarrow \quad n \approx \mathbf{116{,}2110} \text{ o } n \approx \mathbf{150{,}4721}.
\end{aligned}$$

Aplicación (3/3)

- Buscamos z que satisface $P(Z \geq z) = 0{,}1$;
- Leyendo una table de distribución normal, $z \approx 1{,}28$.
- Entonces, buscamos n de tal manera que

$$\frac{400 - 3n}{\sqrt{0{,}09n + 1600}} \leq 1{,}28.$$

- Obtenemos

$$\begin{aligned}
\frac{400 - 3n}{\sqrt{0{,}09n + 1600}} = 1{,}28 \quad & \Rightarrow \quad 160000 - 2400n + 9n^2 = 1{,}28^2(0{,}09n + 1600) \\
& \Rightarrow \quad 9n^2 - 2400{,}147456n + 157378{,}56 = 0 \\
& \Rightarrow \quad n \approx \mathbf{116{,}2110} \text{ o } n \approx \mathbf{150{,}4721}.
\end{aligned}$$

- Pero, hay que **rechazar** $n \approx 150{,}4721$ (falsa solución).

Aplicación (3/3)

- Buscamos z que satisface $P(Z \geq z) = 0{,}1$;
- Leyendo una table de distribución normal, $z \approx 1{,}28$.
- Entonces, buscamos n de tal manera que

$$\frac{400 - 3n}{\sqrt{0{,}09n + 1600}} \leq 1{,}28.$$

- Obtenemos

$$\begin{aligned}\frac{400 - 3n}{\sqrt{0{,}09n + 1600}} = 1{,}28 \quad &\Rightarrow \quad 160000 - 2400n + 9n^2 = 1{,}28^2(0{,}09n + 1600)\\ &\Rightarrow \quad 9n^2 - 2400{,}147456n + 157378{,}56 = 0\\ &\Rightarrow \quad n \approx \mathbf{116{,}2110} \text{ o } n \approx \mathbf{150{,}4721}.\end{aligned}$$

- Pero, hay que **rechazar** $n \approx 150{,}4721$ (falsa solución).
- Así, a partir de $n \geq 117$, el riesgo es **mayor que** $10\,\%$.

Aproximación de una variable binomial

- Sea X una variable aleatoria **binomial** de parámetros (n, p);
- Entonces
$$X = X_1 + X_2 + \ldots + X_n,$$
donde
$$X_i = \begin{cases} 1, & \text{si la prueba } i \text{ es un éxito;} \\ 0, & \text{de otra manera.} \end{cases}$$
- También,
$$E[X_i] = \mathbf{p} \quad \text{y} \quad \mathrm{Var}(X_i) = \mathbf{p(1-p)}.$$
- Por el teorema del límite central
$$\frac{X - np}{\sqrt{np(1-p)}}$$
es aproximadamente **normal estándar**.

Ejemplo (1/2)

- En un programa de estudios, el número **promedio** de estudiantes que son aceptados cada año es 150.
- En el pasado, cerca de 30 % de los estudiantes se inscriben de hecho al programa.
- Consecuamente, el director del programa acepta 450 **aplicaciones** por año.
- ¿Cuál es la probabilidad que **más de** 150 **estudiantes** decidan inscribirse en el programa?

Ejemplo (2/2)

- Sea X el número de estudiantes que se inscriben al programa.

Ejemplo (2/2)

- Sea X el número de estudiantes que se inscriben al programa.
- Asumiendo que los estudiantes se inscriben de manera **independentes**, tenemos que $X \sim \mathcal{B}(450; 0{,}3)$.

Ejemplo (2/2)

- Sea X el número de estudiantes que se inscriben al programa.
- Asumiendo que los estudiantes se inscriben de manera **independentes**, tenemos que $X \sim \mathcal{B}(450; 0{,}3)$.
- Entonces

$$P(X > 150) = \sum_{i=151}^{450} P(X = i) = \sum_{i=151}^{450} \binom{450}{i} (0{,}3)^i (0{,}7)^{450-i}.$$

Ejemplo (2/2)

- Sea X el número de estudiantes que se inscriben al programa.
- Asumiendo que los estudiantes se inscriben de manera **independentes**, tenemos que $X \sim \mathcal{B}(450; 0{,}3)$.
- Entonces

$$P(X > 150) = \sum_{i=151}^{450} P(X = i) = \sum_{i=151}^{450} \binom{450}{i} (0{,}3)^i (0{,}7)^{450-i}.$$

- Es **imposible** calcular esta suma a mano.

Ejemplo (2/2)

- Sea X el número de estudiantes que se inscriben al programa.
- Asumiendo que los estudiantes se inscriben de manera **independentes**, tenemos que $X \sim \mathcal{B}(450; 0{,}3)$.
- Entonces

$$P(X > 150) = \sum_{i=151}^{450} P(X = i) = \sum_{i=151}^{450} \binom{450}{i} (0{,}3)^i (0{,}7)^{450-i}.$$

- Es **imposible** calcular esta suma a mano.
- Se puede **aproximar** con una distribución **normal**:

$$P(X > 150) \approx P\left(Z \geq \frac{\mathbf{150{,}5} - 450 \cdot 0{,}3}{\sqrt{450 \cdot 0{,}3 \cdot 0{,}7}}\right) \approx P(Z \geq 1{,}59) \approx \mathbf{0{,}06}.$$

Corrección de continuidad

- La distribución binomial es **discreta** mientras que la distribución normal es **continua**;
- Sea X una variable aleatoria que sigue una **distribución binomial**;
- Tenemos que $P(X = i) = P(i - 0{,}5 < X < i + 0{,}5)$;
- Cuando se aproxima por una distribución normal, es mejor **corrigir agregando** o **restando** 0,5.
- En el ejemplo, preferimos

$$P(X > 150) = P(X > 150{,}5) \approx P\left(Z \geq \frac{\mathbf{150{,}5} - 450 \cdot 0{,}3}{\sqrt{450 \cdot 0{,}3 \cdot 0{,}7}}\right)$$

a

$$P(X > 150) \approx P\left(Z \geq \frac{\mathbf{150} - 450 \cdot 0{,}3}{\sqrt{450 \cdot 0{,}3 \cdot 0{,}7}}\right).$$

Tabla de contenidos

La media muestral

Definición

Sea una población de objetos. A cada objeto se asocia un ***valor númerico****. Esta población tiene una* ***media*** μ ***conocida*** *y una* ***varianza*** σ^2 ***non conocida****. Sea*

$$X_1, X_2, \ldots, X_n$$

una ***muestra*** *de esta población. Entonces la* ***media muestral*** *es*

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

La media muestral es una variable aleatoria

- Suponemos que una **muestra** X_1, X_2, ..., X_n se elige **al azar**;
- Entonces $\overline{X}$ es una **variable aleatoria**;
- ¿Cuál es la **distribución de** $\overline{X}$?

Esperanza de la media muestral

Se puede calcular la **esperanza** de $\overline{X}$ sin conocer su distribución:

$$E[\overline{X}] \quad = \quad E\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right]$$

Esperanza de la media muestral

Se puede calcular la **esperanza** de $\overline{X}$ sin conocer su distribución:

$$
\begin{aligned}
E[\overline{X}] &= E\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right] \\
&= \frac{1}{n}\left(E[X_1] + E[X_2] + \cdots + E[X_n]\right)
\end{aligned}
$$

Esperanza de la media muestral

Se puede calcular la **esperanza** de $\overline{X}$ sin conocer su distribución:

$$\begin{aligned}
E[\overline{X}] &= E\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right] \\
&= \frac{1}{n}\left(E[X_1] + E[X_2] + \cdots + E[X_n]\right) \\
&= \frac{1}{n}(\mu + \mu + \cdots + \mu)
\end{aligned}$$

Esperanza de la media muestral

Se puede calcular la **esperanza** de $\overline{X}$ sin conocer su distribución:

$$
\begin{aligned}
E[\overline{X}] &= E\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right] \\
&= \frac{1}{n}\left(E[X_1] + E[X_2] + \cdots + E[X_n]\right) \\
&= \frac{1}{n}(\mu + \mu + \cdots + \mu) \\
&= \mu.
\end{aligned}
$$

Varianza de la media muestral

De manera similar, se calcula la **varianza**:

$$\mathrm{Var}[\overline{X}] \quad = \quad \mathrm{Var}\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right)$$

Varianza de la media muestral

De manera similar, se calcula la **varianza**:

$$\begin{aligned}
\mathrm{Var}[\overline{X}] &= \mathrm{Var}\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) \\
&= \frac{1}{n^2}\left(\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n)\right)
\end{aligned}$$

Varianza de la media muestral

De manera similar, se calcula la **varianza**:

$$\begin{aligned}
\mathrm{Var}[\overline{X}] &= \mathrm{Var}\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) \\
&= \frac{1}{n^2}\left(\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n)\right) \\
&= \frac{1}{n^2}\left(\sigma^2 + \sigma^2 + \cdots + \sigma^2\right)
\end{aligned}$$

Varianza de la media muestral

De manera similar, se calcula la **varianza**:

$$
\begin{aligned}
\mathrm{Var}[\overline{X}] &= \mathrm{Var}\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) \\
&= \frac{1}{n^2}\left(\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n)\right) \\
&= \frac{1}{n^2}\left(\sigma^2 + \sigma^2 + \cdots + \sigma^2\right) \\
&= \frac{1}{n^2} \cdot n\sigma^2
\end{aligned}
$$

Varianza de la media muestral

De manera similar, se calcula la **varianza**:

$$\begin{aligned}
\mathrm{Var}[\overline{X}] &= \mathrm{Var}\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) \\
&= \frac{1}{n^2}\left(\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n)\right) \\
&= \frac{1}{n^2}\left(\sigma^2 + \sigma^2 + \cdots + \sigma^2\right) \\
&= \frac{1}{n^2} \cdot n\sigma^2 \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

Distribución de $\overline{X}$

- La variable $\overline{X}$ tiene una esperanza de μ;
- Se puede **estimar** μ;
- ¿Cuál es la **calidad** de la estimación?

Distribución de $\overline{X}$

- La variable $\overline{X}$ tiene una esperanza de μ;
- Se puede **estimar** μ;
- ¿Cuál es la **calidad** de la estimación?
- Depiende de la **dispersión de** $\overline{X}$;

Distribución de $\overline{X}$

- La variable $\overline{X}$ tiene una esperanza de μ;
- Se puede **estimar** μ;
- ¿Cuál es la **calidad** de la estimación?
- Depiende de la **dispersión de** $\overline{X}$;
- Notamos que $\mathrm{Var}(\overline{X}) = \sigma^2/n$;

Distribución de $\overline{X}$

- La variable $\overline{X}$ tiene una esperanza de μ;
- Se puede **estimar** μ;
- ¿Cuál es la **calidad** de la estimación?
- Depiende de la **dispersión de $\overline{X}$**;
- Notamos que $\mathrm{Var}(\overline{X}) = \sigma^2/n$;
- Consecuamente, cuando $n \to \infty$, $\mathrm{Var}(\overline{X}) \to 0$;

Distribución de $\overline{X}$

- La variable $\overline{X}$ tiene una esperanza de μ;
- Se puede **estimar** μ;
- ¿Cuál es la **calidad** de la estimación?
- Depiende de la **dispersión de** $\overline{X}$;
- Notamos que $\mathrm{Var}(\overline{X}) = \sigma^2/n$;
- Consecuamente, cuando $n \to \infty$, $\mathrm{Var}(\overline{X}) \to 0$;
- En otras palabras, **mayor es la muestra**, mayor es la probabilidad que $\overline{X}$ sea una **estimación correcta para** μ.

Aplicación del teorema del límite central

Teorema

Sea X_1, X_2, ..., X_n una muestra de una población de ***media*** μ *y de* ***varianza*** σ^2. *Sea*

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}.$$

Entonces la variable aleatoria

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

sigue aproximadamente una distribución ***normal estándar***.

Nota

De manera general, para que la aproximación sea aceptable, es preferable que $n \geq 30$.

Ejemplo (1/2)

- Se mide la **distancia** entre la Tierra y una estrella.
- Como es díficil tener una medida precisa, se toman **varias medidas**.
- Suponemos que las medidas son **independentes**, que la **media** es d y que la **desviación estándar** es 2 **años-luz**.
- ¿Cuántas medidas se deben tomar para estar seguro a $95\,\%$ que la estimación es correcta con un error de $\pm 0{,}5$ **años-luz**?

Ejemplo (1/2)

- Se mide la **distancia** entre la Tierra y una estrella.
- Como es díficil tener una medida precisa, se toman **varias medidas**.
- Suponemos que las medidas son **independentes**, que la **media** es d y que la **desviación estándar** es 2 **años-luz**.
- ¿Cuántas medidas se deben tomar para estar seguro a $95\,\%$ que la estimación es correcta con un error de $\pm 0{,}5$ **años-luz**?
- Sea n el número de medidas tomadas.

Ejemplo (1/2)

- Se mide la **distancia** entre la Tierra y una estrella.
- Como es díficil tener una medida precisa, se toman **varias medidas**.
- Suponemos que las medidas son **independentes**, que la **media** es d y que la **desviación estándar** es 2 **años-luz**.
- ¿Cuántas medidas se deben tomar para estar seguro a $95\,\%$ que la estimación es correcta con un error de $\pm 0{,}5$ **años-luz**?
- Sea n el número de medidas tomadas.
- Entonces $\overline{X}$ es aproximadamente normal de **media** d y de **desviación estándar** $2/\sqrt{n}$.

Ejemplo (2/2)

- Consecuamente, la probabilidad que $\overline{X}$ sea dentro del intervalo de error es

$$P(-0{,}5 < \overline{X} - d < 0{,}5)$$

Ejemplo (2/2)

- Consecuamente, la probabilidad que $\overline{X}$ sea dentro del intervalo de error es

$$P(-0{,}5 < \overline{X} - d < 0{,}5) \quad = \quad P\left(\frac{-0{,}5}{2/\sqrt{n}} < \frac{\overline{X} - d}{2/\sqrt{n}} < \frac{0{,}5}{2/\sqrt{n}}\right)$$

Ejemplo (2/2)

- Consecuamente, la probabilidad que $\overline{X}$ sea dentro del intervalo de error es

$$
\begin{aligned}
P(-0{,}5 < \overline{X} - d < 0{,}5) &= P\left(\frac{-0{,}5}{2/\sqrt{n}} < \frac{\overline{X} - d}{2/\sqrt{n}} < \frac{0{,}5}{2/\sqrt{n}}\right) \\
&\approx P(-\sqrt{n}/4 < Z < \sqrt{n}/4)
\end{aligned}
$$

Ejemplo (2/2)

- Consecuamente, la probabilidad que $\overline{X}$ sea dentro del intervalo de error es

$$
\begin{aligned}
P(-0{,}5 < \overline{X} - d < 0{,}5) &= P\left(\frac{-0{,}5}{2/\sqrt{n}} < \frac{\overline{X} - d}{2/\sqrt{n}} < \frac{0{,}5}{2/\sqrt{n}}\right) \\
&\approx P(-\sqrt{n}/4 < Z < \sqrt{n}/4) \\
&= P(Z < \sqrt{n}/4) - P(Z < -\sqrt{n}/4)
\end{aligned}
$$

Ejemplo (2/2)

- Consecuamente, la probabilidad que $\overline{X}$ sea dentro del intervalo de error es

$$
\begin{aligned}
P(-0{,}5 < \overline{X} - d < 0{,}5) &= P\left(\frac{-0{,}5}{2/\sqrt{n}} < \frac{\overline{X} - d}{2/\sqrt{n}} < \frac{0{,}5}{2/\sqrt{n}}\right) \\
&\approx P(-\sqrt{n}/4 < Z < \sqrt{n}/4) \\
&= P(Z < \sqrt{n}/4) - P(Z < -\sqrt{n}/4) \\
&= 2P(Z < \sqrt{n}/4) - 1.
\end{aligned}
$$

Ejemplo (2/2)

- Consecuamente, la probabilidad que $\overline{X}$ sea dentro del intervalo de error es

$$\begin{aligned} P(-0{,}5 < \overline{X} - d < 0{,}5) &= P\left(\frac{-0{,}5}{2/\sqrt{n}} < \frac{\overline{X} - d}{2/\sqrt{n}} < \frac{0{,}5}{2/\sqrt{n}}\right) \\ &\approx P(-\sqrt{n}/4 < Z < \sqrt{n}/4) \\ &= P(Z < \sqrt{n}/4) - P(Z < -\sqrt{n}/4) \\ &= 2P(Z < \sqrt{n}/4) - 1. \end{aligned}$$

- Buscamos n que satisface $2P(Z < \sqrt{n}/4) - 1 \geq 0{,}95$, es decir
$$P(Z < \sqrt{n}/4) \geq 0{,}975.$$

Ejemplo (2/2)

- Consecuamente, la probabilidad que $\overline{X}$ sea dentro del intervalo de error es

$$\begin{aligned} P(-0{,}5 < \overline{X} - d < 0{,}5) &= P\left(\frac{-0{,}5}{2/\sqrt{n}} < \frac{\overline{X} - d}{2/\sqrt{n}} < \frac{0{,}5}{2/\sqrt{n}}\right) \\ &\approx P(-\sqrt{n}/4 < Z < \sqrt{n}/4) \\ &= P(Z < \sqrt{n}/4) - P(Z < -\sqrt{n}/4) \\ &= 2P(Z < \sqrt{n}/4) - 1. \end{aligned}$$

- Buscamos n que satisface $2P(Z < \sqrt{n}/4) - 1 \geq 0{,}95$, es decir

$$P(Z < \sqrt{n}/4) \geq 0{,}975.$$

- Entonces, $\sqrt{n}/4 \geq \mathbf{1{,}96}$ de tal manera que $n \geq 62$.

Tabla de contenidos

Definición

Definición

Sean Z_1, Z_2, ..., Z_n variables aleatorias ***normales estándares independientes****. Entonces la variable*

$$X = \sum_{i=1}^{n} Z_i^2 = Z_1^2 + Z_2^2 + \cdots + Z_n^2$$

sigue una ***distribución de χ^2 con n grados de libertad*** *y se escribe*

$$X \sim \chi_n^2.$$

Densidad de χ^2

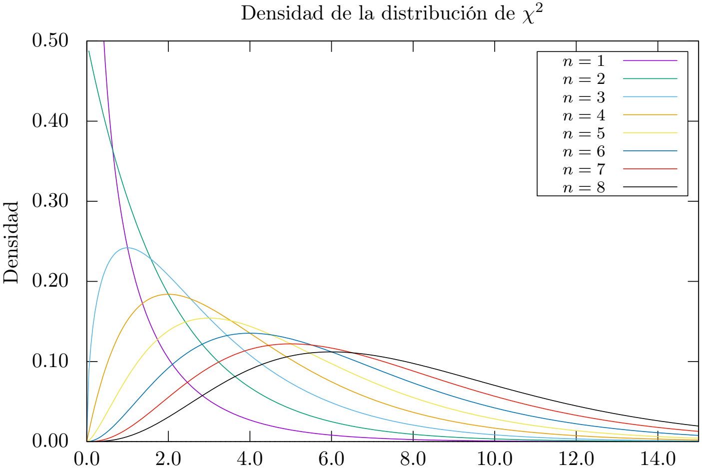

Tabla de los valores de la FDA de χ^2 con n grados de libertad

Sea $X \sim \chi^2_n$. Cada entrada de la tabla corresponde a la probabilidad $P(X \geq x)$. Por ejemplo, para $n = 6$, tenemos

$$P(X \geq 1.237) = 0.975,$$

que se obtiene leyend la entrada a la intersección de la fila $n = 6$ y de la columna 0.975.

n	0.99	0.975	0.95	0.05	0.025	0.01
1	0.000	0.001	0.004	3.841	5.024	6.635
2	0.020	0.051	0.103	5.991	7.378	9.210
3	0.115	0.216	0.352	7.815	9.348	11.345
4	0.297	0.484	0.711	9.488	11.143	13.277
5	0.554	0.831	1.145	11.070	12.833	15.086
6	0.872	1.237	1.635	12.592	14.449	16.812
7	1.239	1.690	2.167	14.067	16.013	18.475
8	1.646	2.180	2.733	15.507	17.535	20.090
9	2.088	2.700	3.325	16.919	19.023	21.666
10	2.558	3.247	3.940	18.307	20.483	23.209
11	3.053	3.816	4.575	19.675	21.920	24.725
12	3.571	4.404	5.226	21.026	23.337	26.217
13	4.107	5.009	5.892	22.362	24.736	27.688
14	4.660	5.629	6.571	23.685	26.119	29.141
15	5.229	6.262	7.261	24.996	27.488	30.578
16	5.812	6.908	7.962	26.296	28.845	32.000
17	6.408	7.564	8.672	27.587	30.191	33.409
18	7.015	8.231	9.390	28.869	31.526	34.805
19	7.633	8.907	10.117	30.144	32.852	36.191
20	8.260	9.591	10.851	31.410	34.170	37.566
21	8.897	10.283	11.591	32.671	35.479	38.932
22	9.542	10.982	12.338	33.924	36.781	40.289
23	10.196	11.689	13.091	35.172	38.076	41.638
24	10.856	12.401	13.848	36.415	39.364	42.980
25	11.524	13.120	14.611	37.652	40.646	44.314
26	12.198	13.844	15.379	38.885	41.923	45.642
27	12.879	14.573	16.151	40.113	43.195	46.963
28	13.565	15.308	16.928	41.337	44.461	48.278
29	14.256	16.047	17.708	42.557	45.722	49.588
30	14.953	16.791	18.493	43.773	46.979	50.892

Distribución de Student

Definición

Sea Z una variable aleatoria ***normal estándar*** *y χ^2_n una variable aleatoria que tiene una* ***distribución de*** χ^2 *con* n ***grados de libertad****. Suponemos que Z y χ^2_n sean* ***independentes****. Entonces la variable*

$$T_n = \frac{Z}{\sqrt{\chi^2_n / n}}$$

sigue una ***distribución de Student*** *o una* t***-distribución*** *con* n ***grados de libertad****.*

Densidad de Student

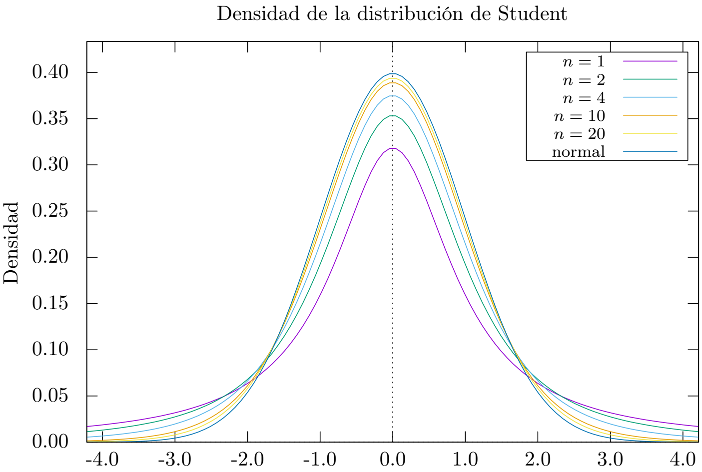

Tabla de Student

Tabla de los valores de la FDA de una distribución de Student con n grados de libertad

Sea $X \sim T_n$. Cada entrada de la tabla corresponde a la probabilidad $P(X \leq x)$. Por ejemplo, para $n = 6$, tenemos

$$P(X \leq 1.943) = 0.95,$$

que se obtiene leyendo la entrada a la intersección de la fila $n = 6$ y de la columna 0.95.

n	0.90	0.95	0.975	0.99	0.995
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750

Propiedades de las dos distribuciones

- Sean X_1 y X_2 dos variables **independentes** que siguen una distribución χ^2, con grados de libertad n_1 **y** n_2;
- Entonces $X_1 + X_2$ sigue una distribución χ^2 con grados de libertad $n_1 + n_2$;
- La distribución χ^2 permite estimar la **varianza** de una muestra;
- La distribución de Student permite realizar **pruebas de hipótesis**.

Tabla de contenidos

La varianza muestral

Definición

Sean X_1, X_2, ..., X_n variables aleatorias de una distribución de ***media*** *μ y de* ***varianza*** *σ^2. Sea $\overline{X}$ la* ***media muestral****.*
Entonces

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

se llama ***varianza muestral*** *y*

$$S = \sqrt{S^2}$$

se llama ***desviación estándar muestral****.*

Esperanza de la varianza muestral (1/2)

- Sabemos del **capítulo 1** que

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2,$$

para cualquier **vector de números** $x = (x_1, x_2, \ldots, x_n)$.

- Entonces,

$$(n-1)\mathbf{S^2} = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2.$$

- Calculamos la **esperanza** de los dos lados de esta igualdad.

Esperanza de la varianza muestral (2/2)

$(n-1)E[S^2]$

Esperanza de la varianza muestral (2/2)

$$(n-1)E[S^2] = E[(n-1)S^2]$$

Esperanza de la varianza muestral (2/2)

$$\begin{aligned}(n-1)E[S^2] &= E[(n-1)S^2] \\ &= E\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right]\end{aligned}$$

Esperanza de la varianza muestral (2/2)

$$
\begin{aligned}
(n-1)E[S^2] &= E[(n-1)S^2] \\
&= E\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right] \\
&= E\left[\sum_{i=1}^{n} X_i^2\right] - nE[\overline{X}^2]
\end{aligned}
$$

Esperanza de la varianza muestral (2/2)

$$
\begin{aligned}
(n-1)E[S^2] &= E[(n-1)S^2] \\
&= E\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right] \\
&= E\left[\sum_{i=1}^{n} X_i^2\right] - nE[\overline{X}^2] \\
&= nE[X_1^2] - nE[\overline{X}^2]
\end{aligned}
$$

Esperanza de la varianza muestral (2/2)

$$\begin{aligned}
(n-1)E[S^2] &= E[(n-1)S^2] \\
&= E\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right] \\
&= E\left[\sum_{i=1}^{n} X_i^2\right] - nE[\overline{X}^2] \\
&= nE[X_1^2] - nE[\overline{X}^2] \\
&= n(\mathrm{Var}(X_1) + E[X_1]^2) - n(\mathrm{Var}(\overline{X}) + E[\overline{X}]^2)
\end{aligned}$$

Esperanza de la varianza muestral (2/2)

$$
\begin{aligned}
(n-1)E[S^2] &= E[(n-1)S^2] \\
&= E\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right] \\
&= E\left[\sum_{i=1}^{n} X_i^2\right] - nE[\overline{X}^2] \\
&= nE[X_1^2] - nE[\overline{X}^2] \\
&= n(\mathrm{Var}(X_1) + E[X_1]^2) - n(\mathrm{Var}(\overline{X}) + E[\overline{X}]^2) \\
&= n\sigma^2 + n\mu^2 - n\sigma^2/n - n\mu^2
\end{aligned}
$$

Esperanza de la varianza muestral (2/2)

$$\begin{aligned}
(n-1)E[S^2] &= E[(n-1)S^2] \\
&= E\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right] \\
&= E\left[\sum_{i=1}^{n} X_i^2\right] - nE[\overline{X}^2] \\
&= nE[X_1^2] - nE[\overline{X}^2] \\
&= n(\mathrm{Var}(X_1) + E[X_1]^2) - n(\mathrm{Var}(\overline{X}) + E[\overline{X}]^2) \\
&= n\sigma^2 + n\mu^2 - n\sigma^2/n - n\mu^2 \\
&= n\sigma^2 - \sigma^2
\end{aligned}$$

Esperanza de la varianza muestral (2/2)

$$\begin{aligned}
(n-1)E[S^2] &= E[(n-1)S^2] \\
&= E\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right] \\
&= E\left[\sum_{i=1}^{n} X_i^2\right] - nE[\overline{X}^2] \\
&= nE[X_1^2] - nE[\overline{X}^2] \\
&= n(\mathrm{Var}(X_1) + E[X_1]^2) - n(\mathrm{Var}(\overline{X}) + E[\overline{X}]^2) \\
&= n\sigma^2 + n\mu^2 - n\sigma^2/n - n\mu^2 \\
&= n\sigma^2 - \sigma^2 \\
&= (n-1)\sigma^2,
\end{aligned}$$

de tal manera que $E[S^2] = \sigma^2$.

Tabla de contenidos

Población normal

- Sea una muestra X_1, X_2, ..., X_n de una población normal de **media** μ y de **varianza** σ^2;
- La **media muestral** es
$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n};$$
- La **varianza muestral** es
$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1};$$
- ¿Cuáles son las distribuciones de $\overline{X}$ y S^2?

Distribución de $\overline{X}$

- Como $X_i \sim \mathcal{N}(\mu, \sigma^2)$, tenemos que $\overline{X}$ sigue una **distribución normal** de media
$$E[\overline{X}| = \mu$$
y de varianza
$$\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}.$$
- Consecuamente:

Teorema

La variable
$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$
sigue una distribución ***normal estándar****.*

Distribución de S^2

Teorema

*Si X_1, X_2, ..., X_n es una muestra de una **población normal** de **media** μ y de **varianza** σ^2, entonces*

$$\frac{(n-1)S^2}{\sigma^2}$$

*sigue una distribución **de** χ^2 con $n-1$ **grados de libertad**.*

Independencia de $\overline{X}$ y S^2

Teorema

Si X_1, X_2, …, X_n es una muestra de una ***población normal*** *de* ***media*** *μ y de* ***varianza*** *σ^2, entonces las* ***variables aleatorias*** *$\overline{X}$ y S^2 son* ***independientes****.*

Corolario

La variable aleatoria

$$T = \sqrt{n}\,\frac{\overline{X} - \mu}{S}$$

sigue una ***distribución de Student*** *con $n-1$* ***grados de libertad*** *y se escribe $T \sim t_{n-1}$.*

Ejemplo

- El tiempo de uso de un procesor para procesar una tarea sigue una distribución **normal** de **media** 20 **nanosecondes** y de **desviación estándar** 3 **nanosecondes**.
- Se recoge una muestra de 15 **tareas**.
- ¿Cuál es la probabilidad que la varianza de la muestra sea **mayor que** 12?

Ejemplo

- El tiempo de uso de un procesor para procesar una tarea sigue una distribución **normal** de **media** 20 **nanosecondes** y de **desviación estándar** 3 **nanosecondes**.
- Se recoge una muestra de 15 **tareas**.
- ¿Cuál es la probabilidad que la varianza de la muestra sea **mayor que** 12?
- Tenemos $n = 15$ y $\sigma^2 = 9$.
- Entonces

$$P(S^2 > 12) = P\left(\frac{14S^2}{9} > \frac{14}{9} \cdot 12\right) = P(\chi^2_{14} > 18{,}67) \approx \mathbf{0{,}1779}.$$

Calculadores

- A veces, las tablas son **incompletas**, tenemos que utilizar **ordenadores** para calcular las probabilidades;
- Por ejemplo, $P(\chi 2_{14} > 18{,}67)$ y $P(\chi 2_{14} \leq 18{,}67)$ no se encuentran en la tabla de χ^2.
- Con Scipy, se obtiene fácilmente

```
>>> from scipy import stats
>>> 1 - stats.chi2.cdf(18.67, 14)
0.17794580609499322
```

- También existen **online calculadores**:

https://www.easycalculation.com/statistics/statistics.php

Población finita

- Sea una **población de N objetos** y $p \in [0,1]$ la **proporción** de elementos que tienen una **característica dada**.
- Una **muestra aleatoria de n objetos** es un conjunto de objetos $\{X_1, X_2, \ldots, X_n\}$ eligidos con **probabilidad uniforme**, es decir que cada de los $\binom{N}{n}$ subconjuntos tienen la misma probabilidad.
- Claramente, X_1, X_2, …, X_n no son **independentes**, porque siguen una **distribución hipergeométrica** (sin reemplazo);
- Pero cuando **N es grande**, la dependencia vuelve **casi nulla** y tiende hacia una **distribución binomial**.
- Entonces, se puede utilizar la **aproximación normal**.

Ejemplo (1/2)

- Suponemos que $45\,\%$ de una población prefiere un candidato para las próximas elecciones.
- Sea una muestra de 200 personas elegidas al azar.
- Calculen
  1. (a) la **esperanza** y la **varianza** del número de personas de la muestra que prefieren el candidato.
  2. (b) la **probabilidad** que más que la **mitad** de las personas prefieren el candidato.

Ejemplo (1/2)

- Suponemos que $45\,\%$ de una población prefiere un candidato para las próximas elecciones.
- Sea una muestra de 200 personas elegidas al azar.
- Calculen
  1. (a) la **esperanza** y la **varianza** del número de personas de la muestra que prefieren el candidato.
  2. (b) la **probabilidad** que más que la **mitad** de las personas prefieren el candidato.
- Tenemos $n = 200$ y $p = 0{,}45$. Entonces,

$$\overline{X} \quad = \quad np = 90,$$

Ejemplo (1/2)

- Suponemos que $45\,\%$ de una población prefiere un candidato para las próximas elecciones.
- Sea una muestra de 200 personas elegidas al azar.
- Calculen
  1. (a) la **esperanza** y la **varianza** del número de personas de la muestra que prefieren el candidato.
  2. (b) la **probabilidad** que más que la **mitad** de las personas prefieren el candidato.
- Tenemos $n = 200$ y $p = 0{,}45$. Entonces,

$$\begin{aligned} \overline{X} &= np = 90, \\ S &= \sqrt{np(1-p)} = \sqrt{200 \cdot 0{,}45 \cdot 0{,}55} \approx 7{,}0356. \end{aligned}$$

Ejemplo (2/2)

- X es binomial con parámetros $n = 200$ y $p = 0{,}45$.

Ejemplo (2/2)

- X es binomial con parámetros $n = 200$ y $p = 0{,}45$.
- Se puede aproximar con la distribución normal:

$$P(X \geq 101)$$

Ejemplo (2/2)

- X es binomial con parámetros $n = 200$ y $p = 0{,}45$.
- Se puede aproximar con la distribución normal:

$$P(X \geq 101) \quad = \quad P(X \geq 100{,}5)$$

Ejemplo (2/2)

- X es binomial con parámetros $n = 200$ y $p = 0{,}45$.
- Se puede aproximar con la distribución normal:

$$\begin{aligned} P(X \geq 101) &= P(X \geq 100{,}5) \\ &\approx P\left(\frac{X - 90}{7{,}0356} \geq \frac{100{,}5 - 90}{7{,}0356}\right) \end{aligned}$$

Ejemplo (2/2)

- X es binomial con parámetros $n = 200$ y $p = 0{,}45$.
- Se puede aproximar con la distribución normal:

$$
\begin{aligned}
P(X \geq 101) &= P(X \geq 100{,}5) \\
&\approx P\left(\frac{X - 90}{7{,}0356} \geq \frac{100{,}5 - 90}{7{,}0356}\right) \\
&\approx P(Z \geq 1{,}4924)
\end{aligned}
$$

Ejemplo (2/2)

- X es binomial con parámetros $n = 200$ y $p = 0{,}45$.
- Se puede aproximar con la distribución normal:

$$
\begin{aligned}
P(X \geq 101) &= P(X \geq 100{,}5) \\
&\approx P\left(\frac{X-90}{7{,}0356} \geq \frac{100{,}5-90}{7{,}0356}\right) \\
&\approx P(Z \geq 1{,}4924) \\
&\approx \mathbf{0{,}0678}.
\end{aligned}
$$