Análisis Estadístico

Estimación, Ley de los Grandes Números y Teorema Central del Límite.

San Andres

21 de mayo de 2022

Introducción

Los avances en las tecnologías de información permiten que grandes cantidades de datos estén disponibles para ser analizados e interpretados:

- **Registros de transacciones:** de cajeros automáticos, comercios, transacciones con tarjetas de débito y crédito, cantidad de personas que entran a un local.
- **Estadísticas públicas:** gobierno abierto, series de variables económicas, estadísticas de salud, de justicia.
- **Opinión pública:** opiniones sobre políticos, sobre medidas del gobierno, sobre la marcha de la economía.
- **Mercados financieros:** series temporales del valor de acciones o de derivados financieros.
- **Internet:** visitas a páginas y la relación entre ellas, items comprados online.

Introducción

La Estadística nos enseña cómo organizar, analizar e interpretar datos para tomar decisiones y evaluar:

- **Descripciones del estado de situación:** tasa de desempleo, obesidad, inflación, apoyo a un candidato político.
- **Relaciones de causa-efecto, evaluaciones de impacto:** impacto en las ventas de una estrategia de marketing, impacto en la calidad educativa de una mejora en el salario docente.
- **Predicciones a futuro (Forecasting):**
  - *finanzas*: riesgo de una estrategia de inversión.
  - *marketing*: Market Basket Analysis: si un usuario compró pasajes, ¿le ofrecemos una valija?
  - *credit scoring*: riesgo de no pago de un préstamo bancario.
  - *economía*: inflación, desempleo, crecimiento del PBI.

Introducción

En el curso pasado estudiamos Teoría de Probabilidad, la rama de la matemática que nos da reglas sobre cómo operar con probabilidades y razonar coherentemente frente a la incertidumbre.

En un problema típico del curso pasado suponíamos que, de alguna manera, conocíamos la distribución que seguía cierta variable aleatoria de interés.

Introducción

Example

La cantidad de clientes que entran a un local a diario es una variable aleatoria X con distribución Poisson de parámetro 10.

- ¿Cuál es la probabilidad de que no entre ninguna persona al local?

$$P(X = 0).$$

- ¿Cual es el número esperado de clientes en un día?

$$E(X).$$

- ¿Cuál es el la probabilidad de que en un semana entren más de 80 personas al local?

$$P(X_L + X_M + \ldots + X_D > 80).$$

Introducción

Si fuéramos los dueños del local en cuestión dispondríamos de datos (¡números!), registros de cúanta gente entró al local cada día en algún lapso.

Podríamos **modelar** la cantidad de gente que entra al local cada día como una variable Poisson, ya que sabemos que la Poisson es útil para modelar procesos de número de llegadas.

¿Cómo hacemos para estimar, inteligentemente con los datos que tenemos, el parámetro de la Poisson?

Éste es exactamente el problema que trata de resolver la inferencia estadística.

Introducción

Teoría de Probabilidad: Asumimos que una variable aleatoria X tiene una cierta distribución F conocida y respondemos preguntas como probabilidades de eventos.

Teoría de Estadística: Tenemos una muestra $X_1 \dots X_n$ de variables aleatorias y queremos conocer cosas sobre la distribución con la que fue generada.

Inferencia estadística

El objetivo de la inferencia estadística es sacar conclusiones y/o tomar decisiones concernientes a algún parámetro desconocido basándose sólo en los datos de una muestra. A grandes rasgos hay tres tipos de inferencias. A partir del ingreso de c/persona en una muestra queremos

Estimación puntual	**estimar** la media del ingreso de la población.
Estimación por intervalos	proveer un intervalo que, con mucha probabilidad, cubra a la media del ingreso mensual de la población
Test de hipótesis	evaluar la evidencia a favor o en contra de la hipotesis de que el ingreso medio de la población es menor a $50000.

Estimación puntual

¿Qué queremos decir con **estimar**?

estimar $\rightarrow$ "adivinar"
(lógicamente, usando la información con la que contamos)

Problema de estimación puntual

Estimar *un parámetro* asociado a una variable aleatoria de interés.

Estimación puntual

Example (Relación de dependencia)

Elegimos una persona al azar de la población de adultos argentions y le preguntamos si tiene o no un *trabajo en relación de dependencia*. Sea X la variable aleatoria que vale 1 si la persona está *en relación de dependencia* y 0 si no.

X es una variable aleatoria $Ber(p)$, donde p es la fracción de personas adultas en *relación de dependencia* en Argentina. Un parámetro de interés aquí podría ser

$$p = \frac{\#\text{Adultos argentinos en relación de dependencia}}{\#\text{Adultos Argentinos}}.$$

p **es un número fijo, ¡pero que desconocemos!**

Estimación puntual

Example (Número esperado de clientes)

Sea X la variable aleatoria que representa el número de clientes que entran cada día a un comercio.

Un parámetro de interés aquí es $E(X)$, el número de clientes esperado.

Estimación puntual

Example (Retorno esperado)

Sea X la variable aleatoria que representa el retorno diario de cierto activo financiero.

Un parámetro de interés aquí es $E(X)$, el retorno esperado.

También nos puede interesar estimar

- $Var(X)$ (si existe...), como una medida del riesgo del activo.
- $P(X > 0)$ que es la probabilidad de un retorno positivo.

Estimación puntual

Example (Distribución de la riqueza)

Elegimos una persona al azar de la población adulta de Buenos Aires. Sean

- X su ingreso neto mensual,
- F_X su función de distribución acumulada y
- F_X^{-1} su función cuantil.

Un parámetro de interés aquí es

$$q = F_X^{-1}(0{,}9)$$

(¿cúanto hay que ganar por mes para estar en el 10 % más rico?).

Nos podría interesar (ambiciosamente) toda la función F_X!

Estimación puntual

Example (Distribución de la riqueza)

Elegimos una persona al azar de la población adulta de Buenos Aires. Sean

- X su ingreso neto mensual,
- F_X su función de distribución acumulada y
- F_X^{-1} su función cuantil.

Un parámetro de interés aquí es

$$q = F_X^{-1}(0{,}9)$$

(¿cúanto hay que ganar por mes para estar en el 10 % más rico?).

Nos podría interesar (ambiciosamente) toda la función F_X!

Estimación puntual

Problema de estimación puntual

Dada una variable X, nos puede interesar estimar diversas cantidades asociadas a la variable:

- $E(X)$,
- $Var(X)$,
- $P(X > 0)$,
- y más.

En la práctica no conocemos completamente la *distribución* de X. En el ejemplo de la *relación de dependencia*, sabemos que X es Bernoulli pero nos falta el parámetro p que es justamente lo que queremos estimar.

Estimación puntual

Supongamos que para el ejemplo de la *relación de dependencia* obtuvimos las siguientes mediciones

$$x_1, \ldots, x_{10}$$

obtenidas todas que obtuvimos eligiendo 10 personas al *azar*.

1	2	3	4	5	6	7	8	9	10
1	1	1	0	1	0	0	0	1	1

Con estos datos, queremos construir una estimación del parámetro p. Por ejemplo, parece natural tomar

$$\frac{x_1 + x_2 + \cdots + x_{10}}{10} = 0{,}6$$

```
## En R:
set.seed(seed = 1)  # para que sea reproducible
muestra = rbinom(n = 10, size = 1, prob = 0.65)
mean(muestra)  # calculamos un posible estimador
```

Estimación puntual

Supongamos que para el ejemplo de la *relación de dependencia* obtuvimos las siguientes mediciones

$$x_1, \ldots, x_{10}$$

obtenidas todas que obtuvimos eligiendo 10 personas al *azar*.

1	2	3	4	5	6	7	8	9	10
1	1	1	0	1	0	0	0	1	1

Con estos datos, queremos construir una estimación del parámetro p. Por ejemplo, parece natural tomar

$$\frac{x_1 + x_2 + \cdots + x_{10}}{10} = 0{,}6$$

```
## En R:
set.seed(seed = 1)  # para que sea reproducible
muestra = rbinom(n = 10, size = 1, prob = 0.65)
mean(muestra)  # calculamos un posible estimador
```

Estimación puntual

El valor estimado

$$\frac{x_1 + x_2 + \cdots + x_{10}}{10} = 0{,}6.$$

es una **función de los datos**.

Nuestras mediciones $x_1, \ldots, x_{10}$ son **realizaciones** de variables aleatorias $X_1, \ldots, X_{10}$ donde

$$X_i = \begin{cases} 1 & \text{si la i-ésima persona trabaja en relación de dependencia} \\ 0 & \text{si la i-ésima persona no trabaja en relación de dependencia} \end{cases}$$

Estimación puntual

El valor estimado

$$\frac{x_1 + x_2 + \cdots + x_{10}}{10} = 0{,}6.$$

es una **función de los datos**.

Nuestras mediciones $x_1, \ldots, x_{10}$ son **realizaciones** de variables aleatorias $X_1, \ldots, X_{10}$ donde

$$X_i = \begin{cases} 1 & \text{si la i-ésima persona trabaja en relación de dependencia} \\ 0 & \text{si la i-ésima persona no trabaja en relación de dependencia} \end{cases}$$

Estimación puntual

El valor estimado

$$\frac{x_1 + x_2 + \cdots + x_{10}}{10} = 0{,}6.$$

es una **función de los datos**.

Nuestras mediciones $x_1, \ldots, x_{10}$ son **realizaciones** de variables aleatorias $X_1, \ldots, X_{10}$ donde

$$X_i = \begin{cases} 1 & \text{si la i-ésima persona trabaja en relación de dependencia} \\ 0 & \text{si la i-ésima persona no trabaja en relación de dependencia} \end{cases}$$

Media muestral

- Observemos que
$$\frac{x_1 + x_2 + \cdots + x_{10}}{10},$$
es un número fijo.
- Notemos también que si repetimos el experimento y son seleccionadas otras personas nuestro valor estimado podría cambiar.

Definition (Media muestral)

Dadas variables aleatorias $X_1, \ldots, X_n$ su media muestral es la **variable aleatoria** definida por

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n}X_1 + \cdots + \frac{1}{n}X_n\,.$$

Notación

Usamos **minúsculas** para **números fijos**, **mayúsculas** para **variables aleatorias**.

Media muestral

```
## En R:
> set.seed(1)
> muestras <- replicate(5, rbinom(n=10,
                                  size=1,
                                  prob=0.65))
> muestras
      [,1] [,2] [,3] [,4] [,5]
 [1,]    1    1    0    1    0
 [2,]    1    1    1    1    1
 [3,]    1    0    0    1    0
 [4,]    0    1    1    1    1
 [5,]    1    0    1    0    1
 [6,]    0    1    1    0    0
 [7,]    0    0    1    0    1
 [8,]    0    0    1    1    1
 [9,]    1    1    0    0    0
[10,]    1    0    1    1    0
> colMeans(muestras)
[1] 0.6 0.5 0.7 0.6 0.5
```

Estimadores

Nuestra estimación de p, la fracción de adultos argentinos que trabaja en relación de dependencia, es una **realización** de la media muestral

$$\frac{X_1 + \cdots + X_{10}}{10}.$$

La media muestral es un **estimador** de p.

Definition (Estimador)

Dadas variables aleatorias $X_1, \ldots, X_n$, para un parámetro cualquiera θ, un estimador es una función, digamos $\widehat{\theta}_n$, de $X_1, \ldots, X_n$.

Observación (muy importante)

- Un estimador es ¡una variable aleatoria!
- Un valor estimado es una realización de un estimador (número fijo).

Estimadores

En el ejemplo que veníamos trabajando de la *relación de dependencia*:

$$\text{Estimador:} \quad \frac{X_1 + \cdots + X_{10}}{10}$$

$$\text{Valor estimado :} \quad \frac{x_1 + \cdots + x_{10}}{10} = 0{,}6$$

Otro estimador (¿qué opina del estimador?),

$$\text{Estimador :} \quad \frac{X_1 + X_2}{2}$$

$$\text{Valor estimado :} \quad \frac{x_1 + x_2}{2} = \frac{1+1}{2} = 1$$

Cualquier función de $X_1, \ldots, X_n$ es un estimador. Más adelante veremos más parámetros y estimadores. Por ahora, nos enfocamos en la media muestral.

Estimadores

En el ejemplo que veníamos trabajando de la *relación de dependencia*:

$$\text{Estimador:} \quad \frac{X_1 + \cdots + X_{10}}{10}$$

$$\text{Valor estimado :} \quad \frac{x_1 + \cdots + x_{10}}{10} = 0{,}6$$

Otro estimador (¿qué opina del estimador?),

$$\text{Estimador :} \quad \frac{X_1 + X_2}{2}$$

$$\text{Valor estimado :} \quad \frac{x_1 + x_2}{2} = \frac{1+1}{2} = 1$$

Cualquier función de $X_1, \ldots, X_n$ es un estimador. Más adelante veremos más parámetros y estimadores. Por ahora, nos enfocamos en la media muestral.

Distribución muestral

La media muestra es una variable aleatoria

Dadas variables aleatorias $X_1, \ldots, X_n$, la media muestral

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

es una variable aleatoria y por lo tanto tiene su propia distribución.

Distribucion muestral o distribucion del estimador

Distribución de la media muestral

La distribución de un estimador se suele llamar **distribución muestral**.

Distribución muestral

Veamos que la media muestral es una variable aleatoria.

```
## Definimos los parametros
n_muestra <- 20
parametro <- 0.7
## Generamos 1000 muestras
muestras <- replicate(n = 1000,
                      rbinom(n = n_muestra,
                             size = 1,
                             prob = parametro))
## Calculamos la media de cada muestra
medias <- colMeans(muestras)
## Graficamos el histograma
hist(medias, freq = TRUE,
     main = paste("Histograma, n=", n_muestra),
     xlab = "Medias muestrales",
     col = "steelblue", lwd = 3, density = 45,
     xlim = c(0, 1))
```

Distribución muestral

¿Qué pasa si cambiamos el parámetro p y mantenemos fijo el tamaño de muestra?

¿Qué pasa si fijamos el parámetro p y aumentamos el tamaño de la muestra?

Propiedades de estimadores

Nos interesa entender qué propiedades tienen distintos estimadores de distintos parámetros para responder, entre otras, las siguientes preguntas:

- ¿Qué propiedades tiene la **distribución muestral** del estimador?
- ¿Cómo medimos si un estimador es **mejor** que otro? ¿Cúal es la manera **óptima** de estimar un parámetro?
- ¿Podemos garantizar que, con mucha probabilidad, el estimador va a estar **cerca** del valor verdadero (fijo y desconocido) que queremos estimar?
- ¿Cuántas (y qué tipo de) observaciones necesitamos para garantizar cierto **margen de error** con probabilidad alta?

Para la mayoría de los problemas, sólo vamos a poder dar respuestas aproximadas a estas preguntas, basadas en resultados **asintóticos** (con un número de observaciones que tiende a infinito).

Comenzaremos con un caso particular e importante: **estimar una esperanza usando la media muestral**.

Propiedades de estimadores

Nos interesa entender qué propiedades tienen distintos estimadores de distintos parámetros para responder, entre otras, las siguientes preguntas:

- ¿Qué propiedades tiene la **distribución muestral** del estimador?
- ¿Cómo medimos si un estimador es **mejor** que otro? ¿Cúal es la manera **óptima** de estimar un parámetro?
- ¿Podemos garantizar que, con mucha probabilidad, el estimador va a estar **cerca** del valor verdadero (fijo y desconocido) que queremos estimar?
- ¿Cuántas (y qué tipo de) observaciones necesitamos para garantizar cierto **margen de error** con probabilidad alta?

Para la mayoría de los problemas, sólo vamos a poder dar respuestas aproximadas a estas preguntas, basadas en resultados **asintóticos** (con un número de observaciones que tiende a infinito).

Comenzaremos con un caso particular e importante: **estimar una esperanza usando la media muestral**.

Propiedades de estimadores

Necesitamos estudiar dos resultados fundacionales de la teoría de la probabilidad:

- la **Ley de los Grandes Números**
- y el **Teorema Central del Límite**.

Estos teoremas caracterizan el comportamiento **asintótico** (cuando $n \to \infty$) de la media muestral en dos sentidos.

Veremos más adelante que estos dos sentidos se corresponden con dos manera distintas de definir convergencia de sucesiones de variables aleatorias.

La media muestral

Definition (Muestra aleatoria)

Si $X_1, \ldots, X_n$ son independientes e idénticamente distribuidas (i.i.d) diremos que son una *muestra aleatoria* de tamaño n.

Estimar una esperanza usando la media muestral

Nos interesa estimar $E(X)$ para cierta variable aleatoria X. Hasta nuevo aviso, vamos a suponer que tenemos una muestra aleatoria $X_1, \ldots, X_n$ con la misma distribución que X. Consideramos el estimador de $E(X)$ dado por

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

La media muestral

Definition (Muestra aleatoria)

Si $X_1, \ldots, X_n$ son independientes e idénticamente distribuidas (i.i.d) diremos que son una *muestra aleatoria* de tamaño n.

Estimar una esperanza usando la media muestral

Nos interesa estimar $E(X)$ para cierta variable aleatoria X. Hasta nuevo aviso, vamos a suponer que tenemos una muestra aleatoria $X_1, \ldots, X_n$ con la misma distribución que X. Consideramos el estimador de $E(X)$ dado por

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

Propiedades de la media muestral

En la siguiente proposición, formalizamos algo que ya habíamos notado en las simulaciones con R.

Proposición

Si existe $E(X)$, *entonces*

$$E\left(\overline{X}_n\right) = E(X).$$

Si existe $Var(X)$, *entonces*

$$Var(\overline{X}_n) = \frac{Var(X)}{n}.$$

Idea

La distribución de $\overline{X}_n$ está centrada en $E(X)$ y su varianza disminuye (¿tiene sentido?) a medida que aumenta el tamaño de la muestra.

Propiedades de la media muestral

En la siguiente proposición, formalizamos algo que ya habíamos notado en las simulaciones con R.

Proposición

Si existe $E(X)$, *entonces*

$$E\left(\overline{X}_n\right) = E(X).$$

Si existe $Var(X)$, *entonces*

$$Var(\overline{X}_n) = \frac{Var(X)}{n}.$$

Idea

La distribución de $\overline{X}_n$ está centrada en $E(X)$ y su varianza disminuye (¿tiene sentido?) a medida que aumenta el tamaño de la muestra.

Propiedades de la media muestral

Demostración.

- Usamos las propiedades de la esperanza

$$E\left(\overline{X}_n\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right)$$
$$= \frac{1}{n}\sum_{i=1}^{n} E\left(X_i\right) = \frac{1}{n}nE(X) = E(X)$$

- Con las propiedades de la varianza (y la independencia)

$$Var\left(\overline{X}_n\right) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}Var\left(\sum_{i=1}^{n} X_i\right)$$
$$= \frac{1}{n^2}\sum_{i=1}^{n} Var\left(X_i\right) = \frac{1}{n^2}nVar(X) = \frac{Var(X)}{n}.$$

Desigualdades de Markov y Tchebysheff

Desigualdad de Markov –> Es poco probable que una variable aleatoria (X), tome valores muy alejados de su media

Proposición (Desigualdades de Markov y Tchebysheff)

Markov *Si* $E(X)$ *existe, para todo* $M > 0$

$$P(|X| \geq M) \leq \frac{E(|X|)}{M}.$$

Tchebysheff *Si* $E(X)$ *y* $Var(X)$ *existen, para todo* $M > 0$

$$P(|X - E(X)| \geq M) \leq \frac{Var(X)}{M^2}.$$

Desigualdad de Tchebycheff—> La probabilidad de que la distancia de una v.a. (X) a su valor esperado, esta acotada

Desigualdades de Markov y Tchebysheff

Idea

Markov	Es poco probable que una variable aleatoria sea "mucho" más grande que su esperanza.
Tchebysheff	Es poco probable que una variable aleatoria esté "muy" lejos de su esperanza.

Desigualdades de Markov y Tchebysheff

Idea

Sea X el ingreso mensual de un individuo elegido al azar de la población Argentina. Si tomamos $\varepsilon = 2E(X)$, la desigualdad de Markov (notar que $X \geq 0$) nos dice que

$$P\left(X \geq 2E(X)\right) \leq 1/2,$$

es decir, que es imposible que más de la mitad de la población tenga un ingreso de al menos el doble del ingreso promedio.

Desigualdades de Markov y Tchebysheff

Nota: Epsilon lo puedo pensar como un numero muy chico

Corollary

$$P\left(|\overline{X}_n - E(X)| < \varepsilon\right) \geq 1 - \frac{Var(X)}{n}\frac{1}{\varepsilon^2}$$

Interpretación frecuentista

Si repetimos muchas veces el experimento de tomar una muestra de tamaño n y calcular la correspondiente media muestral, **al menos** en una fracción de

$$1 - \frac{Var(X)}{n}\frac{1}{\varepsilon^2}$$

de los experimentos, nuestro valor estimado está a una distancia menor a ε del valor verdadero que queremos estimar.

¡Esto no dice nada sobre ninguna realización de la media muestral en particular!

Desigualdades de Markov y Tchebysheff

Ej1/ aplicacion tchebycheff:
Acotar la probabilidad de que en una muestra tamano n =250, con distribucion bernoulli, la media muestral diste en menos de 0.1 del valor verdadero de p, sabiendo p=0.4
P(|X - 0.4|) <= 0.1) >= 1 - (p*(1-p))/250
P(|X - 0.4|) <= 0.1) >= 0.9
por tchebycheff se que la probabilidad es mas grande o igual que 0.9

Example (Relación de dependencia)

Queremos estimar la proporción de la población adulta de Argentina que trabaja en relación de dependencia, digamos p. ¿Qué tamaño de muestra necesitamos para garantizar que con probabilidad al menos 0,95, el error de estimar p con $\overline{X}_n$ sea menor que 0,1?

Desigualdades de Markov y Tchebysheff

Example (Relación de dependencia)

Recordar que $X \sim Ber(p)$. Luego $E(X) = p$. Queremos encontrar n tal que

$$P\left(\left|\overline{X}_n - p\right| < 0{,}1\right) \geq 0{,}95,$$

equivalentemente

$$P\left(\left|\overline{X}_n - p\right| \geq 0{,}1\right) \leq 0{,}05,$$

Por Tchebysheff

$$P\left(\left|\overline{X}_n - p\right| \geq 0{,}1\right) \leq \frac{Var(X)}{n}\frac{1}{0{,}1^2} = \frac{p(1-p)}{n}\frac{1}{0{,}1^2}.$$

Pero no conocemos p. Sin embargo, podemos usar que $p(1-p) \leq 0{,}25$ para cualquier $p \in [0,1]$. Entonces

$$P\left(\left|\overline{X}_n - p\right| \geq 0{,}1\right) \leq \frac{0{,}25}{n}\frac{1}{0{,}1^2}.$$

Desigualdades de Markov y Tchebysheff

Example (Relación de dependencia)

$$P\left(\left|\overline{X}_n - p\right| \geq 0{,}1\right) \leq \frac{0{,}25}{n}\frac{1}{0{,}1^2}.$$

Queremos que

$$\frac{0{,}25}{n}\frac{1}{0{,}1^2} \leq 0{,}05,$$

que equivale a

$$\frac{0{,}25}{0{,}05}\frac{1}{0{,}1^2} \leq n,$$

o sea

$$n \geq 500.$$

Ley de los grandes números

Theorem

Sea $X_1, \ldots, X_n$ una muestra aleatoria tal que $E(X) = \mu$. Entonces para todo $\varepsilon > 0$

$$P\left(|\overline{X}_n - \mu| \geq \varepsilon\right) \underset{n \to \infty}{\to} 0$$

Interpretación frecuentista

Si n es grande con probabilidad muy alta la media muestral estará 'cerca' de $E(X)$.
Esta es claramente una propiedad deseable de $\overline{X}_n$ como estimador de $E(X)$.

Notemos que la conclusión es más débil que la que obtuvimos antes usando Tchebysheff, pero también que las hipótesis son más débiles ya que no requerimos que exista $Var(X)$.

Ley de los grandes números

```
medias <- list()
grilla_n <- seq(10, 1000, 10)
for (n in grilla_n) {
  muestras <- replicate(1000,
                        rpois(n = n, lambda = 1))
  medias_n <- colMeans(muestras)
  medias <- cbind(medias, medias_n)
}
## Graficamos
matplot(grilla_n, t(medias),
        pch = 10, col = "steelblue", lwd = 1,
        ylab = "Media muestral",
        xlab = "Tamano de muestra")
abline(h=1, lwd = 3, col = "darkorange")
```

Ley de los Grandes Números

Theorem (Ley (débil) de los grandes números)

Sea $X_1, \ldots, X_n$ una muestra aleatoria tal que $E(X) = \mu$. Entonces para todo $\varepsilon > 0$

$$P\left(|\overline{X}_n - \mu| \geq \varepsilon\right) \underset{n\to\infty}{\longrightarrow} 0$$

Demostración.

La demostración está fuera del alcanza del curso. Pero podemos probar que el resultado vale si asumimos que $Var(X)$ es finita. En este caso, por Tchebysheff, para todo $\varepsilon > 0$ vale que

$$P\left(|\overline{X}_n - \mu| \geq \varepsilon\right) \leq \frac{Var(X)}{n}\frac{1}{\varepsilon^2} \underset{n\to\infty}{\longrightarrow} 0\,.$$

□

Teorema Central del Límite

Recordemos lo que habíamos visto en las simulaciones con R:

> *para tamaños de muestra grande, si $X_1, \ldots, X_n$ es una muestra aleatoria con la misma distribución que X, la distribución de $\overline{X}_n$ se concentra alrededor de $E(X)$ y toma una forma similar a la campana normal, aun cuando la distribución de X es muy distinta de la normal.*

Theorem (Teorema central de límite)

Sea $X_1, \ldots, X_n$ una muestra aleatoria tal que $E(X) = \mu$ y $Var(X) = \sigma^2$, entonces, para todo $z \in \mathbb{R}$

$$P\left(\frac{\sqrt{n}\,(\overline{X}_n - \mu)}{\sigma} \leq z\right) \underset{n\to\infty}{\longrightarrow} \Phi(z),$$

donde $\Phi(z) = P(Z \leq z)$ para $Z \sim N(0,1)$.

Teorema Central del Límite

Idea

Para tamaños de muestra grande, la distribución de $\overline{X}_n$ es aproximadamente normal.

$$\frac{\sqrt{n}\,(\overline{X}_n - \mu)}{\sigma} \approx N(0,1)$$

y

$$\overline{X}_n \approx N\left(\mu, \frac{\sigma^2}{n}\right) .$$

El TCL nos va a permitir (más adelante en la materia) *cuantificar la incertidumbre* que tenemos sobre nuestra estimación de μ usando la media muestral.

Observación (importante)

El teorema se refiere a la distribución de $\overline{X}_n$, **no** a la de $X_1, \ldots, X_n$.

Teorema Central del Límite

Idea

El TCL nos dice que, una versión debidamente re-escalada de $\overline{X}_n$ **converge en cierto sentido** a una distribución normal $N(0,1)$.
Pensemos por qué es razonable este re-escalamiento. Sabemos que

$$E(\overline{X}_n) = \mu \quad Var(\overline{X}_n) = \frac{\sigma^2}{n}.$$

Luego

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} = \frac{\overline{X}_n - E(\overline{X}_n)}{SD(\overline{X}_n)}.$$

El cambio de escala es consiste en

- restar la esperanza y
- dividir por el desvío!

Si $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, entonces la distribución $\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \sim \mathcal{N}(0,1)$ es **exacta**.

Teorema Central del Límite

¿Qué tan grande debe ser el n?

- Si la distribución de X es simétrica, es decir, si existe algún punto c tal que el histograma o la función de densidad de X es simétrica respecto de c, típicamente con $n > 20$ la aproximación normal del TCL es adecuada.
- Si la distribución de X no es simétrica, no hay una "receta", mientras más grande sea el n, mejor.

Teorema Central del Límite

```
n_muestra <- 50
n_rep <- 1000
parametro <- 0.5  # asimetrica 0.95 o 0.05
esperanza <- parametro
desvio <- sqrt(parametro*(1-parametro))
escalar <- FALSE
muestras <- replicate(n=n_rep,
                      rbinom(n=n_muestra, size=1,
                             prob=parametro))
medias <- colMeans(muestras)
par(mfrow=c(1, 2))
hist(medias, col="steelblue", angle=45, density=20,
     freq=FALSE, main="Histograma de medias",
     xlab="Medias", xlim=c(0, 1))
hist(sqrt(n_muestra)*(medias-esperanza)/desvio,
     col="steelblue", angle=45, density=20,
     freq=FALSE, main="Histograma de medias",
     xlab="Medias")#, xlim=c(0, 1))
curve(dnorm(x), add=TRUE, col="darkorange",
      lwd=3, lty=2)
```

Teorema Central del Límite

Example

- Vas a entrevistar a $n = 100$ personas elegidas al azar de toda la población de Argentina y les vas a preguntar si están a favor o en contra de una reforma previsional.
- Suponiendo que el 40 % de las personas está en contra, ¿cuál es la probabilidad de que, en tu encuesta, la proporción de personas en contra de la reforma previsional esté entre 0,38 y 0,42?

P(0.38<=X(raya)<=0.42)=?—>aplicando TCL N(0.4, (0.4*0.6)/100)
V(X) = P* 1-P

P(0.38<=X(raya)<=0.42)= pnorm(q = 0.42,mean = 0.4,sd = sqrt((0.4*(1-.4))/100)) - pnorm(q = 0.38,mean = 0.4,sd = sqrt((0.4*(1-.4)))/100)
P(0.38<=X(raya)<=0.42)=0.658

Aca no es lo mismo >= que >, dado que no es continuo lo que estamos contando

Teorema Central del Límite

Example

El gobierno de la Ciudad de Buenos Aires está interesado en estimar la proporción de fumadores que habita la Ciudad. Para ello, el Ministerio de Salud decidió encuestar a n habitantes al azar y preguntarles si son o no fumadores. ¿Qué valor debe tener n para que esta proporción no difiera de la real en más de 0,01 con una probabilidad mayor o igual que 0,95?

Teorema Central del Límite

Hasta ahora nuestra gran hipótesis es que obtuvimos una **muestra aleatoria** $X_1, \ldots, X_n$. Es decir, suponemos que $X_1, \ldots, X_n$ son **independientes** e **idénticamente distribuidas**. Es razonable (al menos aproximadamente) suponer que podemos conseguir esto

- En el último ejemplo, de la reforma previsional, si elegimos a los entrevistados sacando bolillas de un bolillero sin reposición y además suponemos que todos responden: sí, aproximadamente, porque la población de Argentina es muy grande.
- Si $X_1, \ldots, X_n$ son los retornos de un activo financiero en n días consecutivos de trading: no, no van a ser idénticamente distribuidas, ¡la varianza no va a ser constante!
- Si $X_1, \ldots, X_n$ son ventas mensuales de cerveza en n meses consecutivos: no, no van a ser independientes, ¡esperamos una dependencia temporal y estacionalidad!