Tema 5. Muestreo y distribuciones muestrales

En este tema:

- Muestreo y muestras aleatorias simples.
- Distribución de la media muestral:
  - Esperanza y varianza.
  - Distribución exacta en el caso normal.
  - Distribución aproximada en el caso general (teorema central del límite).
- Distribución de la varianza muestral:
  - Esperanza.
  - Distribución en el caso normal.

Muestreo

Motivación

- En muchos casos se desea obtener información estadística sobre poblaciones numerosas:
  - Situación laboral de las personas en edad de trabajar en España.
  - Fiabilidad de un modelo de automóvil en un año.
  - Precipitación anual en la Comunidad de Madrid.
- Puede ser imposible (por falta de recursos) obtener la información relativa a todos los individuos.
- Se estudia una muestra significativa de la población.
  - Un subconjunto de la población que permita obtener información fiable sobre el total de dicha población.

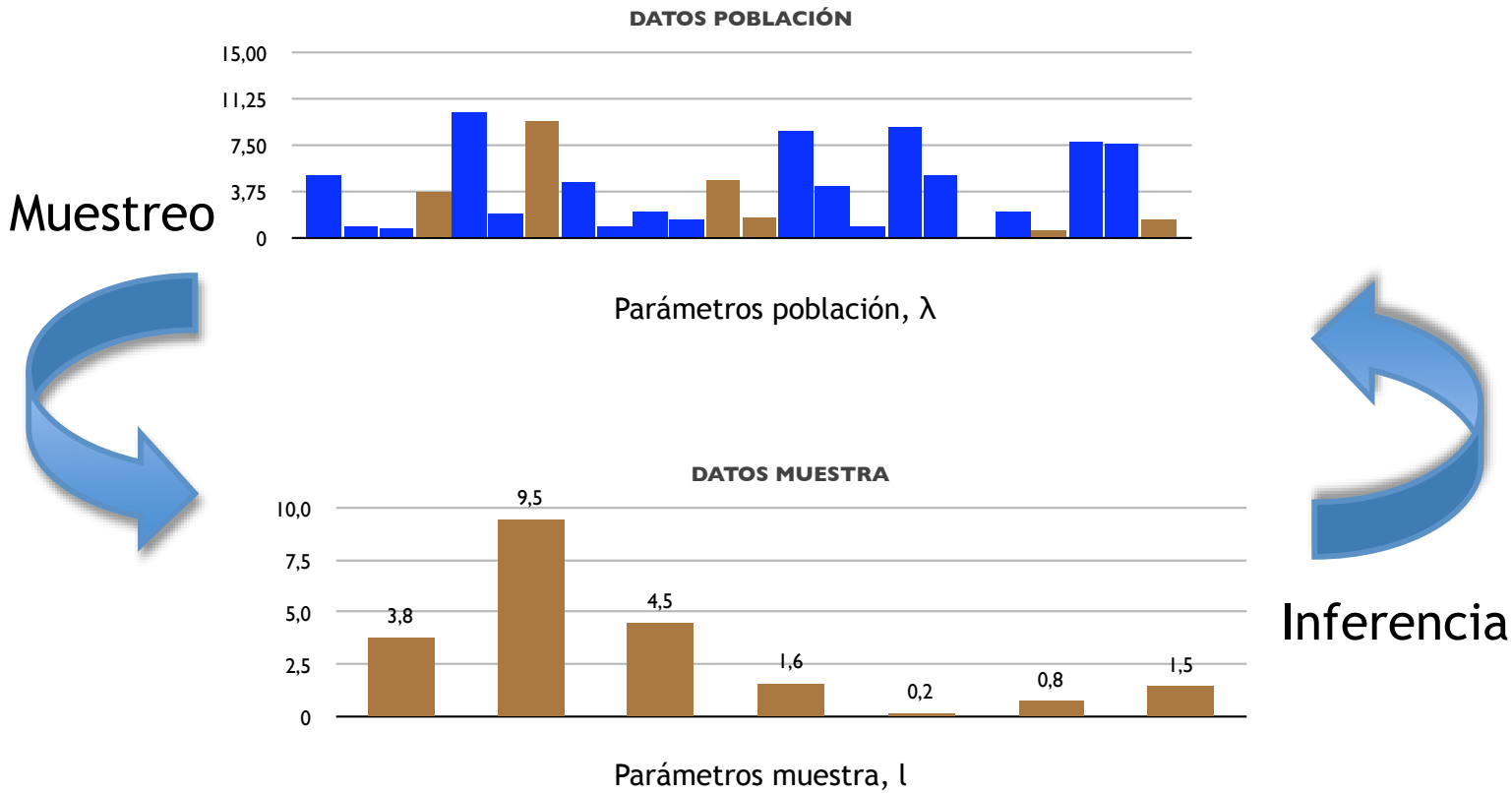
Muestras aleatorias simples

Cómo seleccionar una muestra

- Tamaño reducido.
- Ausencia de sesgos.
  - Conclusiones obtenidas de la muestra son válidas para la población.
- Facilidad en la definición de la muestra.
- Mejor alternativa: Muestras aleatorias simples
  - Cada miembro de la población tiene la misma probabilidad de pertenecer a la muestra.
  - La selección se realiza de manera independiente.
    - La selección de un individuo concreto no afecta a la probabilidad de seleccionar cualquiera de los otros.

Procedimiento de inferencia

- Partiendo de la distribución de la variable aleatoria en la muestra
- Obtener información sobre distribución de la variable en la población
- Valores de interés: cálculo de estadísticos para la media, varianza, proporciones



Ejemplo de muestreo e inferencia

Ejemplo

Consideremos el ejemplo de la figura anterior:

- Población compuesta por 24 individuos.
- Variable aleatoria de interés:
  - Tiempo para completar una consulta médica.
- Valores:

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

- Promedio de la población: $4,0$

Ejemplo de muestreo e inferencia

Muestra 1

- Muestra seleccionada en la figura, tamaño 7:

Muestra	3,8	9,5	4,8	1,6	0,2	0,8	1,5

- Estadístico de interés: promedio de la muestra $3,1$.
- Error (sesgo) relativo: $(4,0 - 3,1)/4,0 = 0,225$.

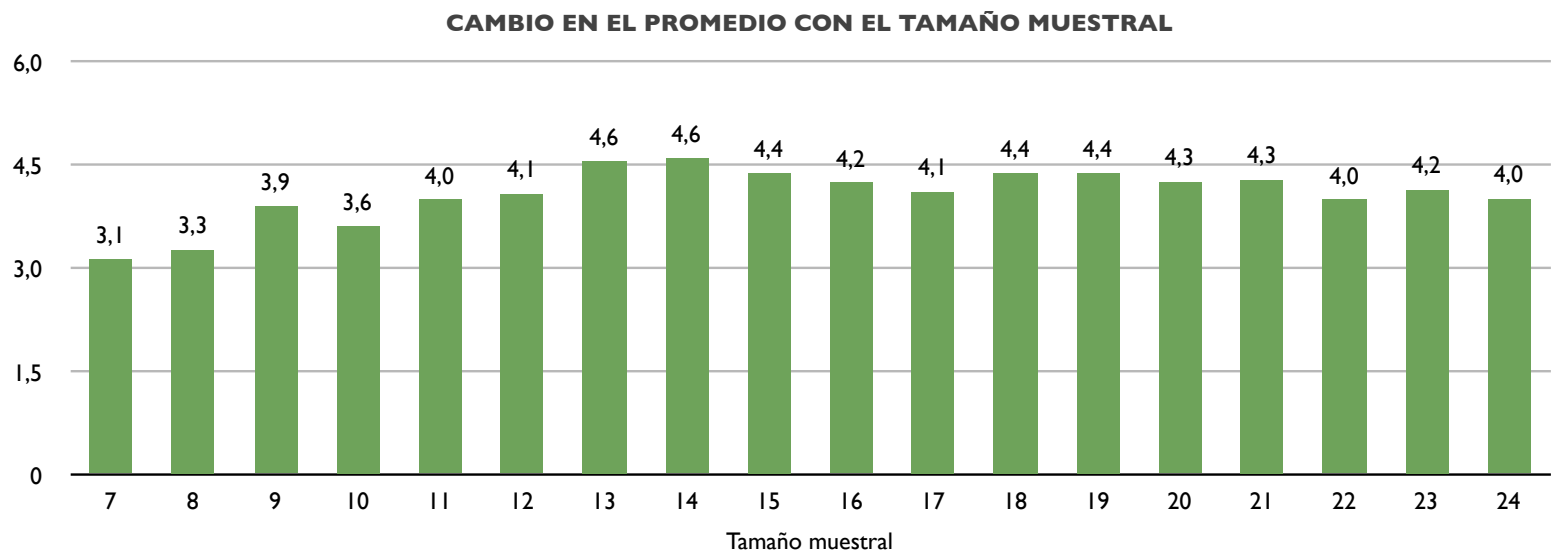
Cambios en el muestreo

- Selecciones alternativas de los elementos de la muestra.
- Aumento del tamaño de la muestra.

Ejemplo de muestreo

Cambios en el tamaño muestral

- Si a la muestra del ejemplo anterior le añadimos nuevos elementos, el promedio muestral cambia.
- Se aproxima al valor de la media poblacional



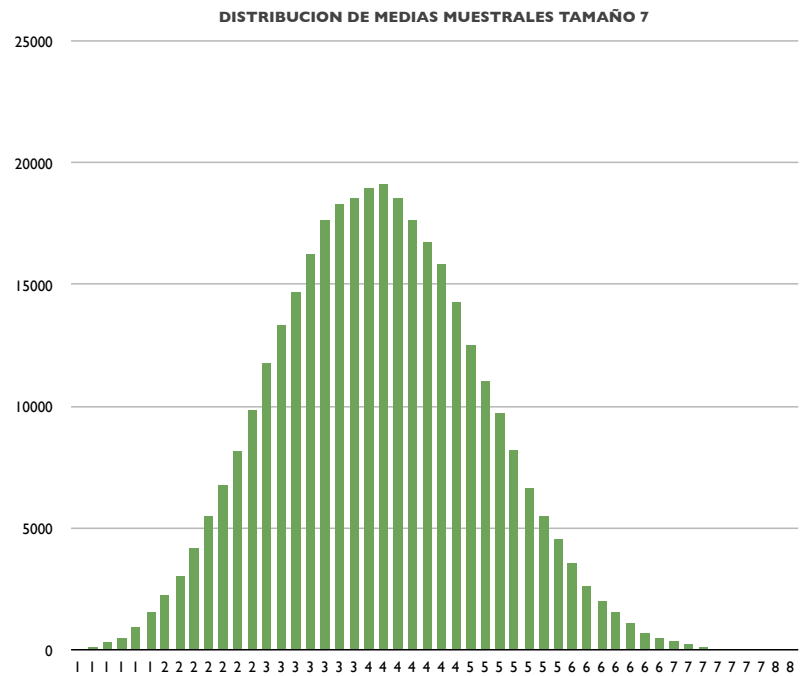
Ejemplo de muestreo

Selección de observaciones

- Si seleccionamos las primeras 7 observaciones: 5,1 1,0 0,9 3,8 18,2 2,1 9,5.
- Promedio de la muestra: $5,8$.

Cambios para diferentes selecciones

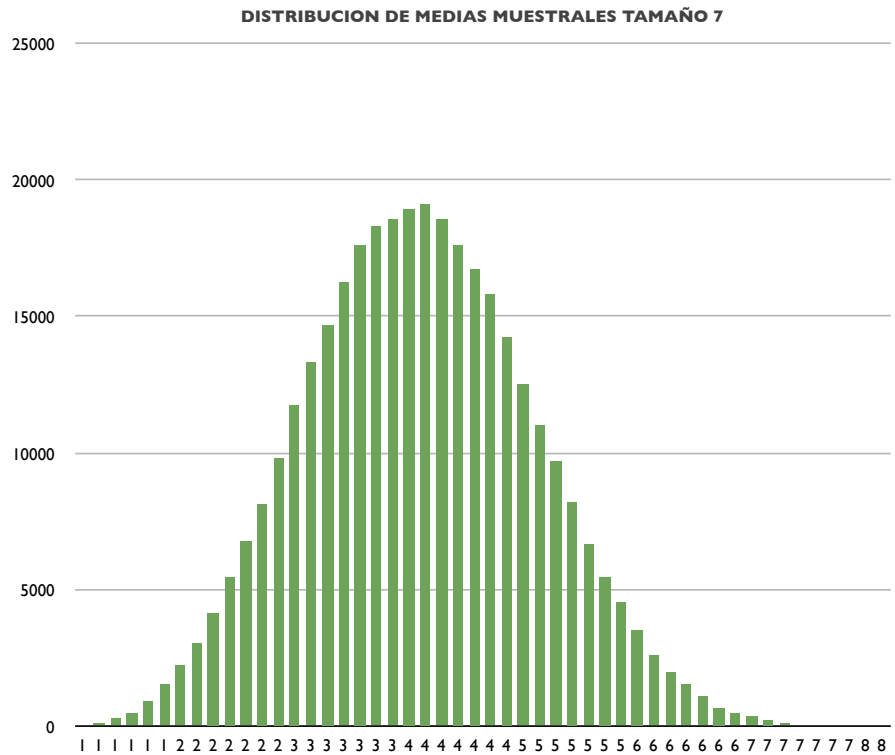
- Todas las selecciones posibles de 7 observaciones (346.104 posibilidades):

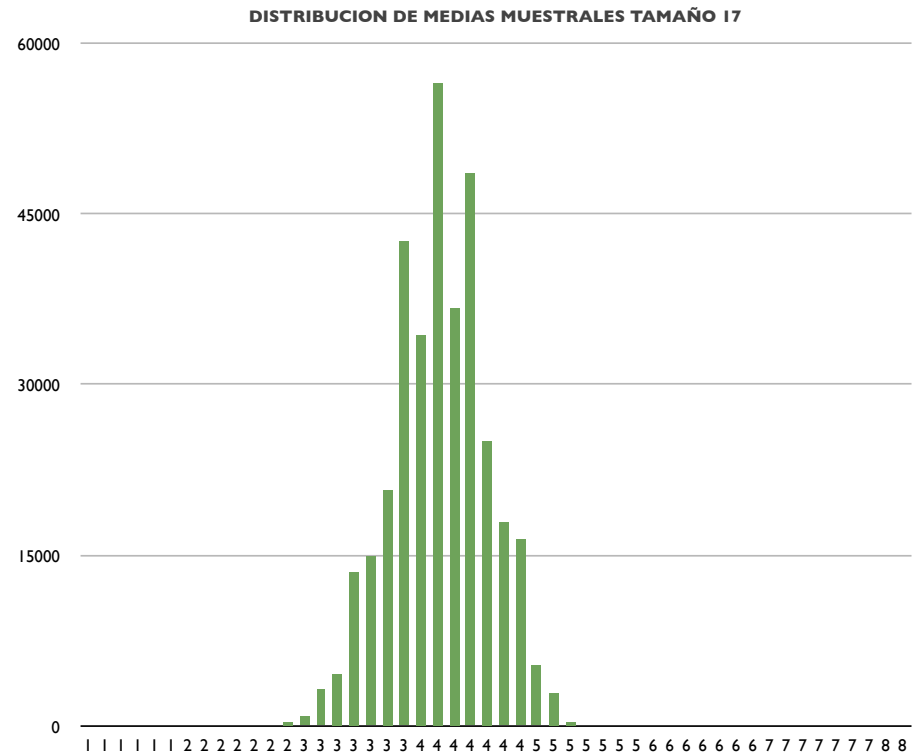


Distribuciones en el muestreo

Distribución de la media muestral

- Para todas las muestras de tamaño 7 y 17 obtenemos:

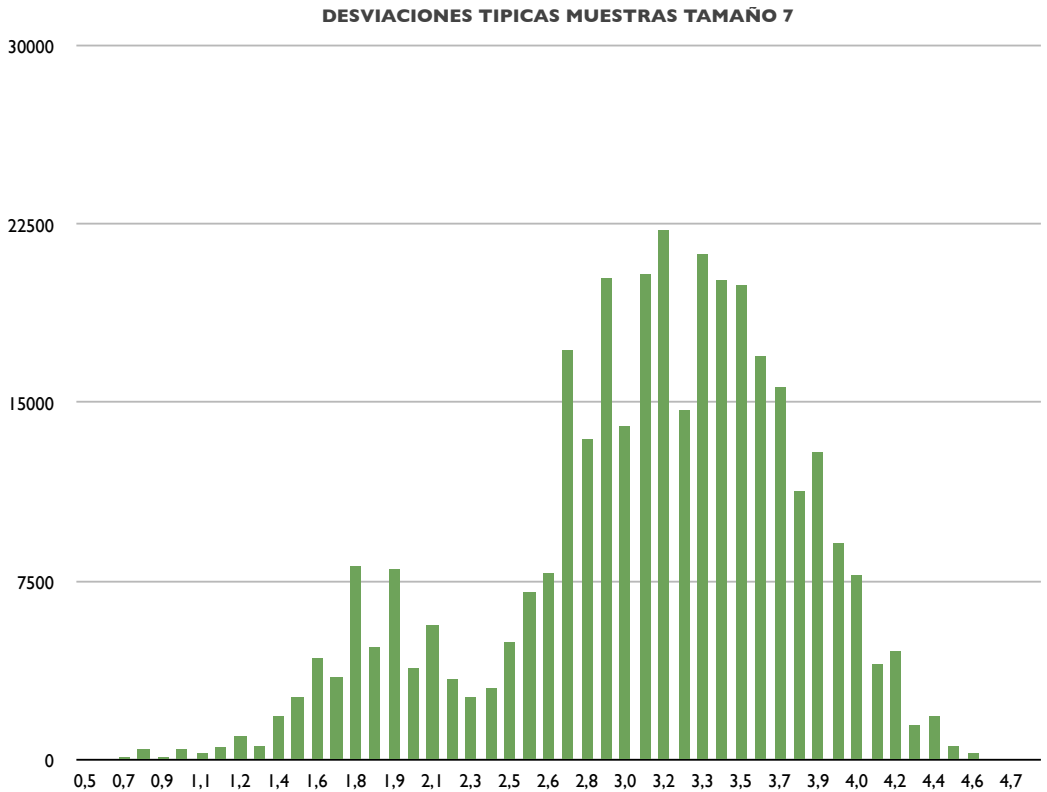




Distribuciones en el muestreo

Distribución de la varianza muestral

- Se obtienen resultados similares para otros estadísticos
- Para la desviación típica de muestras de tamaño 7 obtenemos:



Distribuciones en el muestreo

Conclusiones

- Una muestra aleatoria simple de tamaño n de una v.a. X es un conjunto de v.a. independientes, todas con la misma distribución que X:

$$\{X_i\}_{i=1}^{n} \text{ i.i.d.}$$

- El valor del promedio muestral es una variable aleatoria (los estadísticos son variables aleatorias).
  - Depende de la selección (aleatoria) de los individuos en la muestra:
- Distribución muestral del estadístico: distribución de probabilidad del valor de interés para todas las muestras del mismo tamaño.
- La distribución muestral cambia con el tamaño de la muestra.
  - La variabilidad de los estadísticos muestrales disminuye con el tamaño de la muestra.

La distribución de la media muestral

El problema de interés

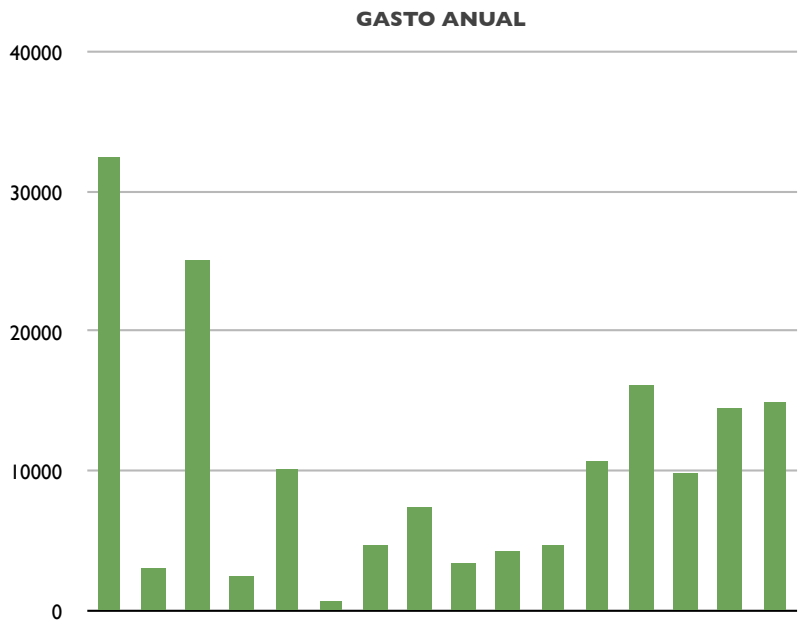
- La media poblacional es un parámetro de gran interés en muchas situaciones prácticas.
- Por ejemplo, queremos conocer el promedio de:
  - los ingresos familiares en España el año 2007.
  - la proporción de préstamos morosos el último mes.
  - el precio de compra de viviendas en la Comunidad de Madrid el pasado mes.
- A partir de una muestra (reducida) de valores queremos calcular:
  - Una buena aproximación al valor correcto (inevitablemente con error).
  - Y una estimación del error en la aproximación.

La distribución de la media muestral

Ejemplo

- Información sobre el gasto familiar en España
- Disponemos de los datos siguientes (gasto anual por hogar, EPF)

Gasto	32545,76	3140,24	25205,64	2474,28	10242,34	721,16
	4855,80	7449,74	3466,50	4400,80	4740,00	10830,00
	16240,88	9840,12	14534,96	14960,00		



La distribución de la media muestral

Algunas respuestas

- Muestra de tamaño n, $\{X_i\}_{i=1}^n$, de una variable aleatoria X (gasto de hogares).
- Queremos estimar la media nacional (esperanza de X) a partir de la muestra.
- Se define el estadístico media muestral como

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

LA MEDIA MUESTRAL ES UNA VARIABLE ALEATORIA

- El valor esperado de la media muestral es la media de la población

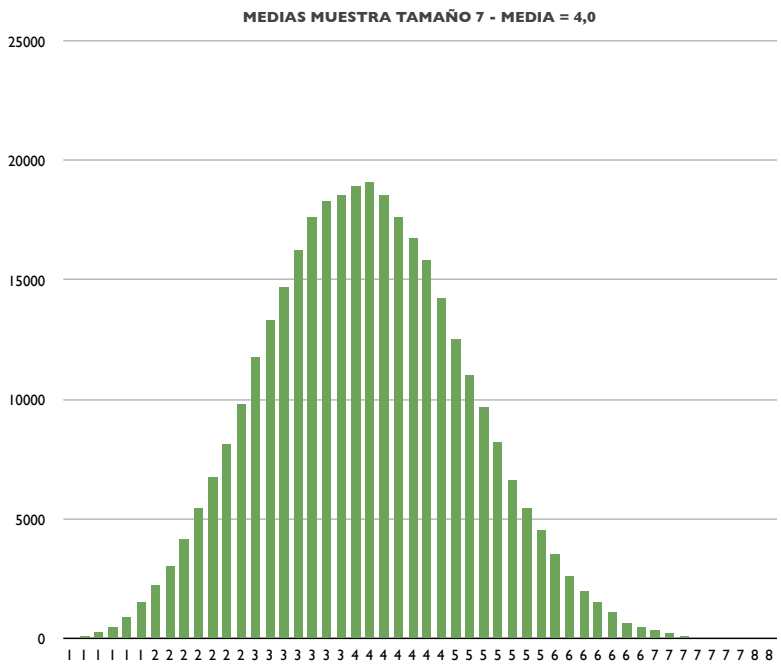
$$\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \mathrm{E}[X]$$

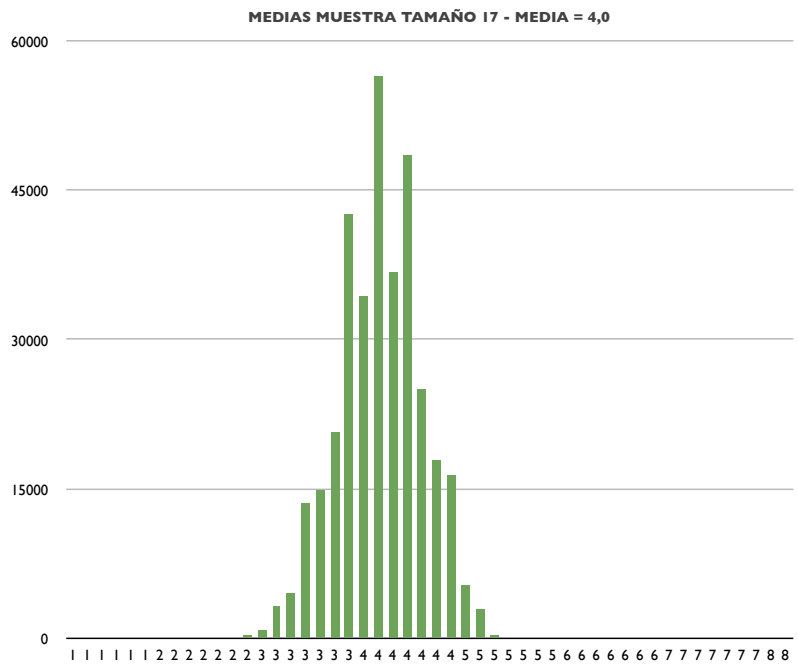
- Estimamos la media de la población a partir de la media de la muestra.
  - En nuestro ejemplo: 10353,01 euros

La distribución de la media muestral

Más datos de la distribución

- Media de una muestra en general diferente de la media de la población
- ¿Podemos conocer la magnitud del error que estamos cometiendo?
  - Depende de la distribución de la media muestral
  - En particular, de su variabilidad (desviación respecto de la media)
  - ¿En cual de los casos siguientes tenemos menos error?





Distribución de la media muestral

La variabilidad de la media muestral

- La varianza de la media muestral, $\overline{X}$, (una medida del error) es

$$\text{Var}[\overline{X}] \stackrel{def.}{=} \text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sigma^2$$

- En el ejemplo anterior, $\text{Var}[\bar{x}] = 76.458.643$ y $\sigma[\bar{x}] = 8.744$ euros.
- El valor de la varianza decrece si n aumenta.
- Podemos reducir el error aumentando el tamaño de la muestra.
  - La reducción en el error es lenta.
  - Para reducir el error (medido por la desviación típica) a la mitad debemos aumentar el tamaño de la muestra 4 veces.

Distribución de la media muestral

La distribución de la media muestral

- El valor de la varianza de la media muestral sólo nos dice si el error puede ser grande o pequeño.
- Para obtener una respuesta más precisa deberíamos conocer la distribución de la media muestral.
- Si la variable X tiene una distribución normal, entonces

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mathrm{E}[X], \sqrt{\sigma^2/n})$$

(por ser combinación lineal de v.a. indep. con dist. normal, ver Tema 4)

Luego,

$$\frac{\overline{X} - \mathrm{E}[X]}{\sqrt{\sigma^2/n}} \sim N(0,1)$$

Distribución de la media muestral

El teorema central del límite

- Distribución de la media muestral si X no es normal
- Si cumple ciertas condiciones: Teorema central del límite (ver Tema 3)

Dada una muestra aleatoria simple $\{X_i\}_{i=1}^n$ de tamaño n obtenida de una variable aleatoria X con media $\mathrm{E}[X]$ y varianza σ^2 finitas, se cumple que

$$\frac{\frac{1}{n}\sum_{i=1}^n X_i - \mathrm{E}[X]}{\sqrt{\sigma^2/n}} \to N(0,1)$$

cuando $n \to \infty$.

- La distribución de la media muestral se parece a una distribución normal para muestras grandes.
- Aplicación similar al caso anterior cuando se tienen muestras grandes.

La distribución de la varianza muestral

La varianza muestral

- En muchos casos es importante conocer el valor de la varianza de la población.
  - Para aplicar el teorema central del límite.
  - Para estimar riesgos en inversiones (el riesgo depende de la varianza).
  - Para estimar desigualdades en ingresos, rentas, etc.
- Se define el estadístico varianza muestral como

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

LA VARIANZA MUESTRAL ES UNA VARIABLE ALEATORIA

- Queremos relacionar su media y su varianza con las de la población.
- Y si es posible, identificar su distribución.

La distribución de la varianza muestral

La esperanza de la varianza muestral

- Si $\bar{X}$ denota la media muestral, se tiene que

$$\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\right]=\frac{n-1}{n}\sigma^2$$

- El valor esperado de esa v.a. no es la varianza de la población.
- OJO: la varianza de la media muestral tampoco es la varianza de la población.
- Por eso definimos la varianza muestral como

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

  dividiendo entre $n-1$ y no entre n.

- Con esta definición, tenemos $\mathrm{E}[S^2]=\sigma^2$.
  - El valor muestral de S^2 (s^2) se puede emplear como una aproximación de la varianza de la población.

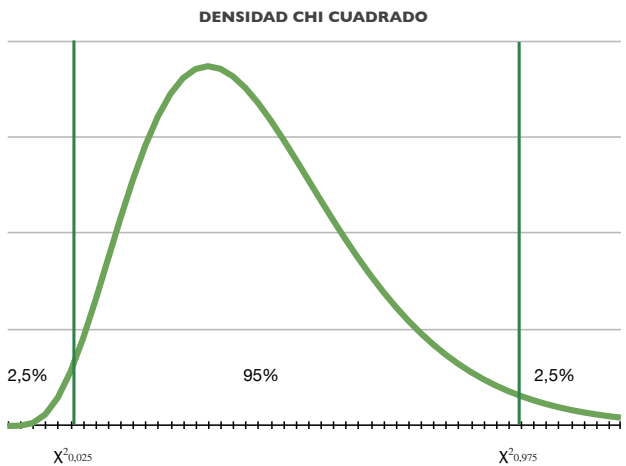
La distribución de la varianza muestral

Distribución de la varianza muestral

- Nos gustaría tener más información sobre la distr. de la varianza muestral.
  - La varianza muestral es no negativa.
  - La distribución de la varianza muestral es asimétrica a la derecha.
- Además, si la variable X tiene distribución normal (ver Tema 3)

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

distribución χ^2 (chi-cuadrado) con $n-1$ grados de libertad.



Función de distribución en las tablas