



Tema 5. Muestreo y distribuciones muestrales

Contenidos

- Muestreo y muestras aleatorias simples
- La distribución de la media en el muestreo
- La distribución de la varianza muestral

Lecturas recomendadas:

- Capítulo 7 del libro de Newbold, Carlson y Thorne (2009).
- Capítulo 7 del libro de Peña (2001).
- Capítulos 19 a 21 del libro de Peña y Romo (1997).



Tema 5. Muestreo y distribuciones muestrales

Objetivos de aprendizaje

- Saber qué es una muestra aleatoria simple
- Conocer la distribución de la media muestral
 - Su media y su varianza
 - Su distribución en el caso normal
 - Su distribución aproximada en el caso general (teorema central del límite)
- Conocer la distribución de la varianza muestral
 - Su media
 - Su distribución en el caso normal

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** representativa de la población
 - Un subconjunto de la población que permita obtener información fiable sobre el total de dicha población



Muestras aleatorias simples

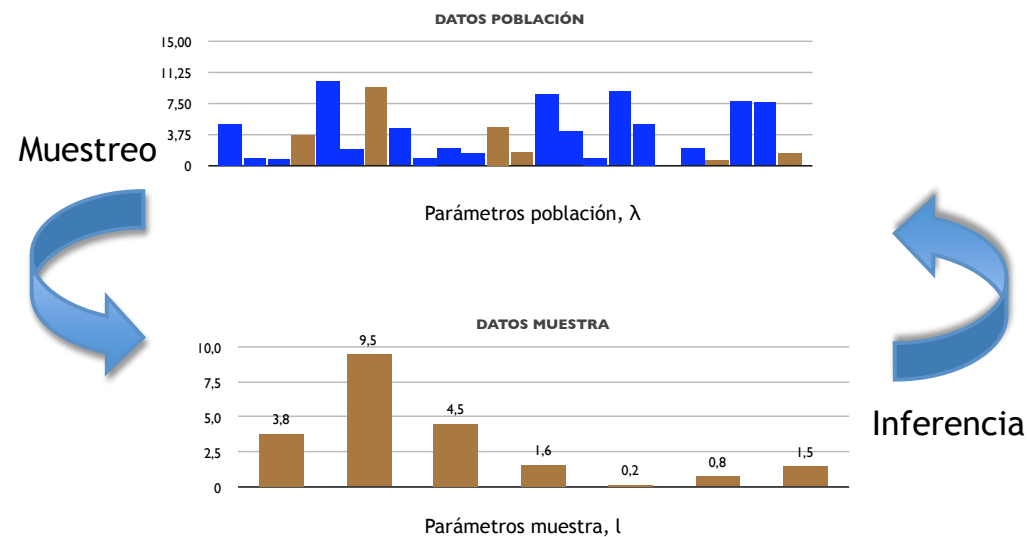
Cómo seleccionar una muestra

- 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
 - En la práctica, selección basada en números aleatorios

Procedimiento de inferencia

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: $(4,0 - 3,1)/4,0 = 0,225$

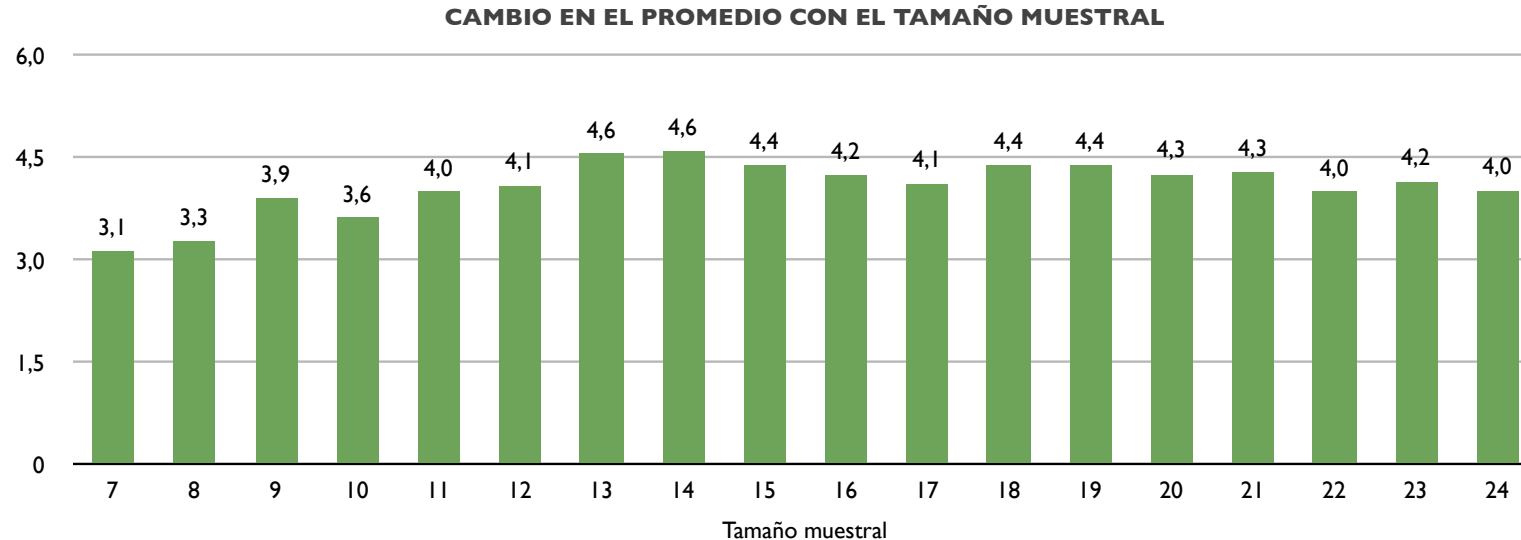
Cambios en el muestreo

- 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

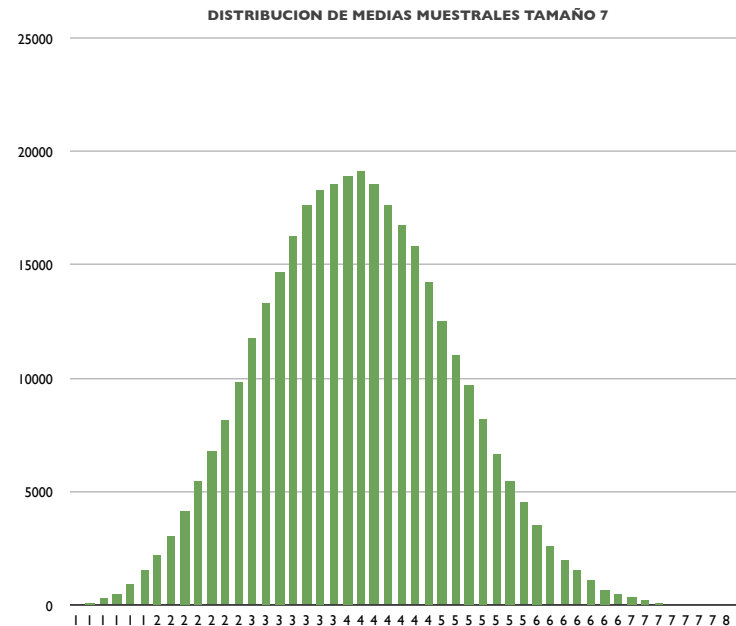


Ejemplo de muestreo

- Si seleccionamos las primeras 7 observaciones obtenemos un promedio de la muestra igual a 5,8:

Muestra 5,1 1,0 0,9 3,8 18,2 2,1 9,5

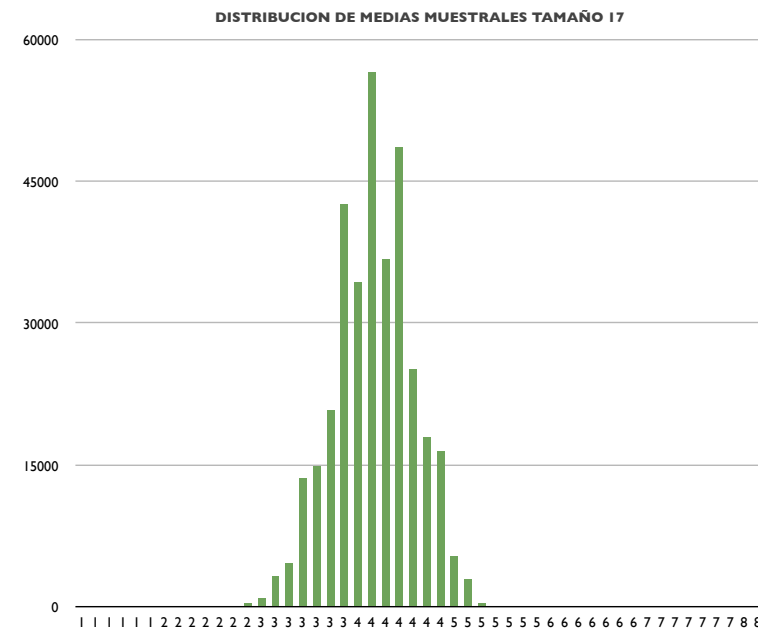
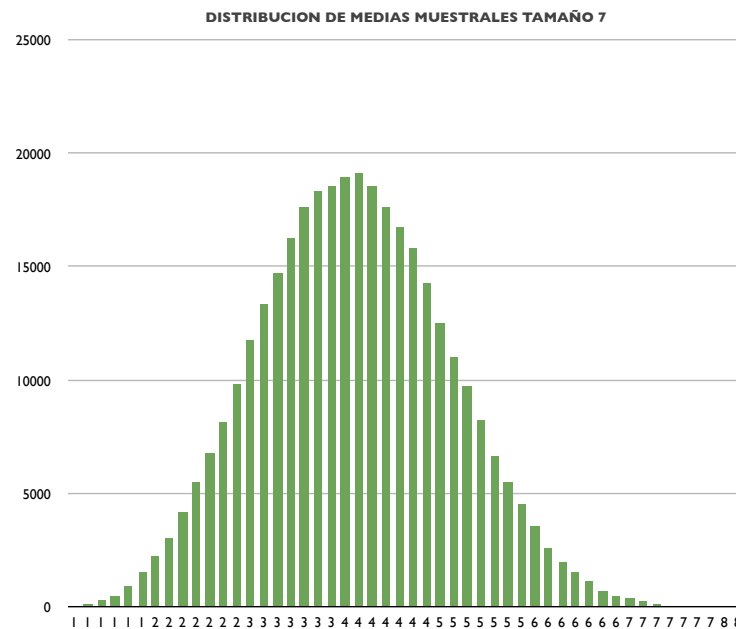
- Si consideramos todas las selecciones posibles de 7 observaciones (346,104 posibilidades):



Distribuciones en el muestreo

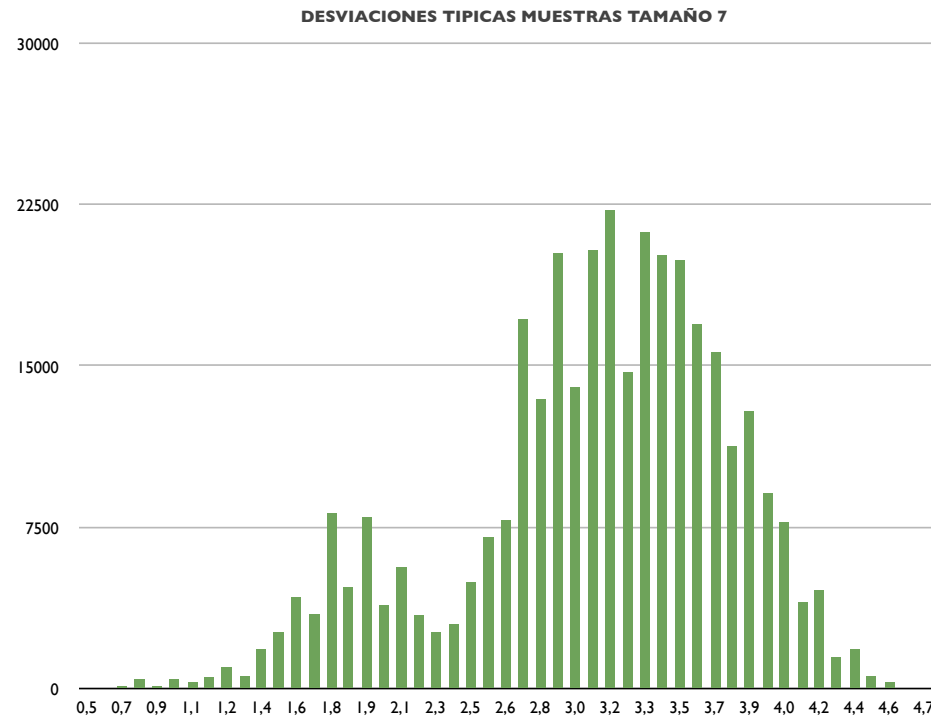
Distribución de la media muestral

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



Distribuciones en el muestreo

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





Distribuciones en el muestreo - Conclusiones

- 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
 - Variabilidad de estadísticos muestrales disminuye con el tamaño de la muestra



La distribución de la media muestral

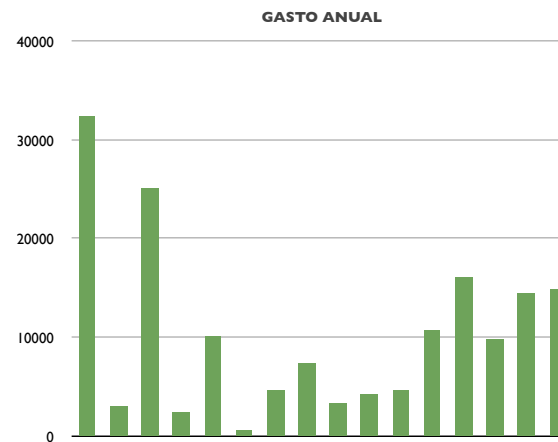
El problema de interés

- 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

- Valor de interés: estimación de la media nacional (media de la variable aleatoria)
 - A partir de los datos disponibles en la muestra
- ¿Qué estadístico de la muestra se parece al promedio nacional (media de la población)?
- **El valor esperado de la media de la muestra es la media de la población**

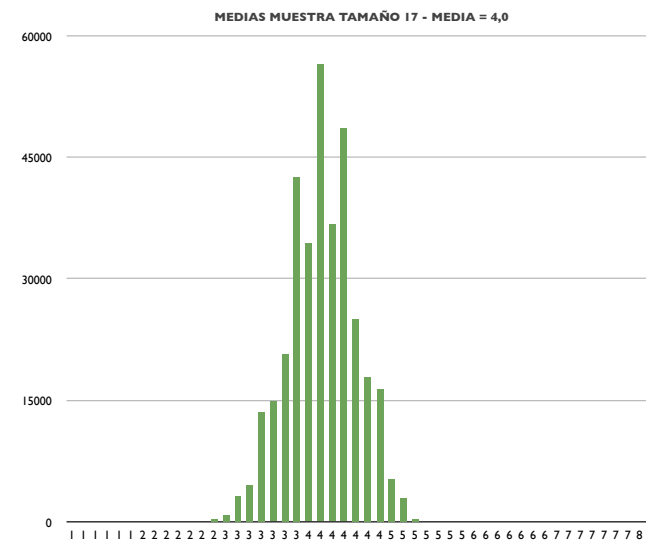
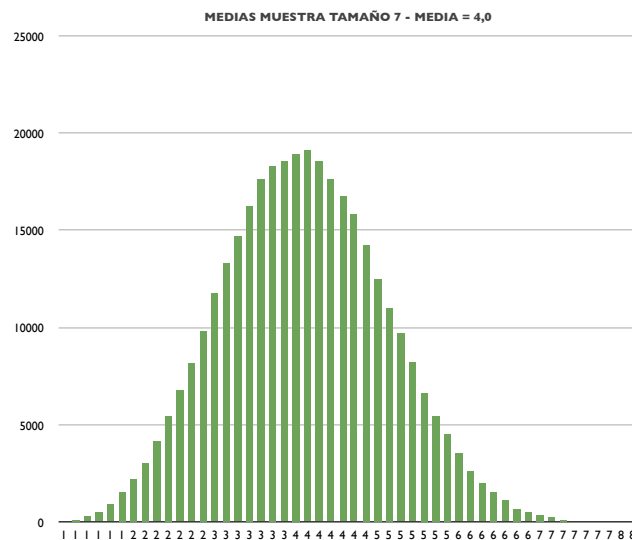
$$E \left[\frac{1}{n} \sum_{i=1}^n x_i \right] = E[X]$$

- 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 $\bar{x}$ (una medida del error) vale

$$V[\bar{x}] = V \left[\frac{1}{n} \sum_{i=1}^n x_i \right] = \frac{1}{n} \sigma^2$$

- En el ejemplo anterior, $V[\bar{x}] = 76,458,643$ y $s[\bar{x}] = 8,744$ 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 solo 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

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



Distribución de la media muestral

- Queremos obtener una medida del error de estimación
- Utilizando el resultado

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

- Pero habitualmente no conocemos σ^2
 - Aproximamos este valor con el correspondiente a la muestra (razonable si n es grande)
- De las tablas de la normal construimos un intervalo que nos proporciona una indicación del error
- El intervalo se selecciona de manera que $P(-\beta \leq Z \leq \beta) = \alpha$ para el nivel de error (confianza) α deseado



Distribución de la media muestral - Ejemplo

- Suponemos una distribución normal de la variable gasto anual de hogares
 - Escogemos un nivel de confianza de 0,95
 - De las tablas de la normal estándar sabemos que para $Z \sim N(0, 1)$

$$P(-1,96 \leq Z \leq 1,96) = 0,95$$

- De los datos muestrales, la media muestral vale $\bar{x} = 10,353$ y la desviación típica muestral vale $s = 8,744$
- Por el resultado anterior sobre la distribución de la media muestral,

$$P(-\beta \leq \frac{\bar{x} - E[X]}{s} \leq \beta) = P(-1,96 \leq \frac{10353 - E[X]}{8744} \leq 1,96) = 0,95$$

$$[10353 - 1,96 \times 8744, 10353 + 1,96 \times 8744] = [-6785, 27491]$$



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

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

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

conforme $n \rightarrow \infty$

- La distribución de la media muestral se parece a una distribución normal para 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.
- Repetimos el estudio que hemos realizado para la media muestral
- Partimos de que la varianza muestral es una variable aleatoria
- Queremos relacionar sus momentos con los de la población
- Y si es posible, identificar su distribución



La distribución de la varianza muestral

Esperanza de la varianza muestral

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

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

- El valor esperado de la varianza muestral no es la varianza de la población
- Definamos la varianza muestral como

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$



La distribución de la varianza muestral

Esperanza de la varianza muestral

- Con esta definición, tenemos $E[s^2] = \sigma^2$
 - El valor esperado de s^2 coincide con el valor deseado (varianza de la población)
 - s^2 es un estimador insesgado de σ^2

Distribución de la varianza muestral

- Nos gustaría tener información adicional sobre la varianza muestral y su distribución
 - La distribución de la varianza muestral no es simétrica: tiene asimetría positiva.

La distribución de la varianza muestral

Distribución de la varianza muestral

- Si la variable X tiene una distribución normal
 - La distribución de $(n - 1)s^2/\sigma^2$ es una χ^2 (**chi-cuadrado**) con $n - 1$ **grados de libertad** (χ^2_{n-1})

