Estadística

Tema 5. Inferencia estadística

María Dolores Frías Domínguez
Jesús Fernández Fernández
Carmen María Sordo

Departamento de Matemática Aplicada y Ciencias de la Computación

TEMA 5: Inferencia Estadística

Muestreo:

- Tamaño y calidad de la muestra
- Muestreo aleatorio

Inferencia estadística:

- Estimación de una proporción
- Estimación de una media
- Estimación de una varianza

María Dolores Frías, Jesús Fernández y Carmen María Sordo

Muestreo

POBLACIÓN: todos los estudiantes de la Universidad de Cantabria

MUESTRA: alumnos de 1º de Grado de Ingeniería Química de la Universidad de Cantabria.

Necesidad del muestreo:

1. **Coste reducido**: la recogida y tratamiento de datos resulta más barato al trabajar con una pequeña parte de la población

2. Mayor **rapidez** en la evaluación del resultado final (ej. escrutinio de votos de las primeras mesas electorales).

3. **Imposibilidad** material por destrucción del objeto a estudio (ej. duración de bombillas, si se estudia toda la población no quedarían bombillas para vender).

Es importante elegir una muestra que represente bien a la población.

Muestreo Aleatorio

Todos los elementos tienen la misma probabilidad de ser incluidos en la muestra.

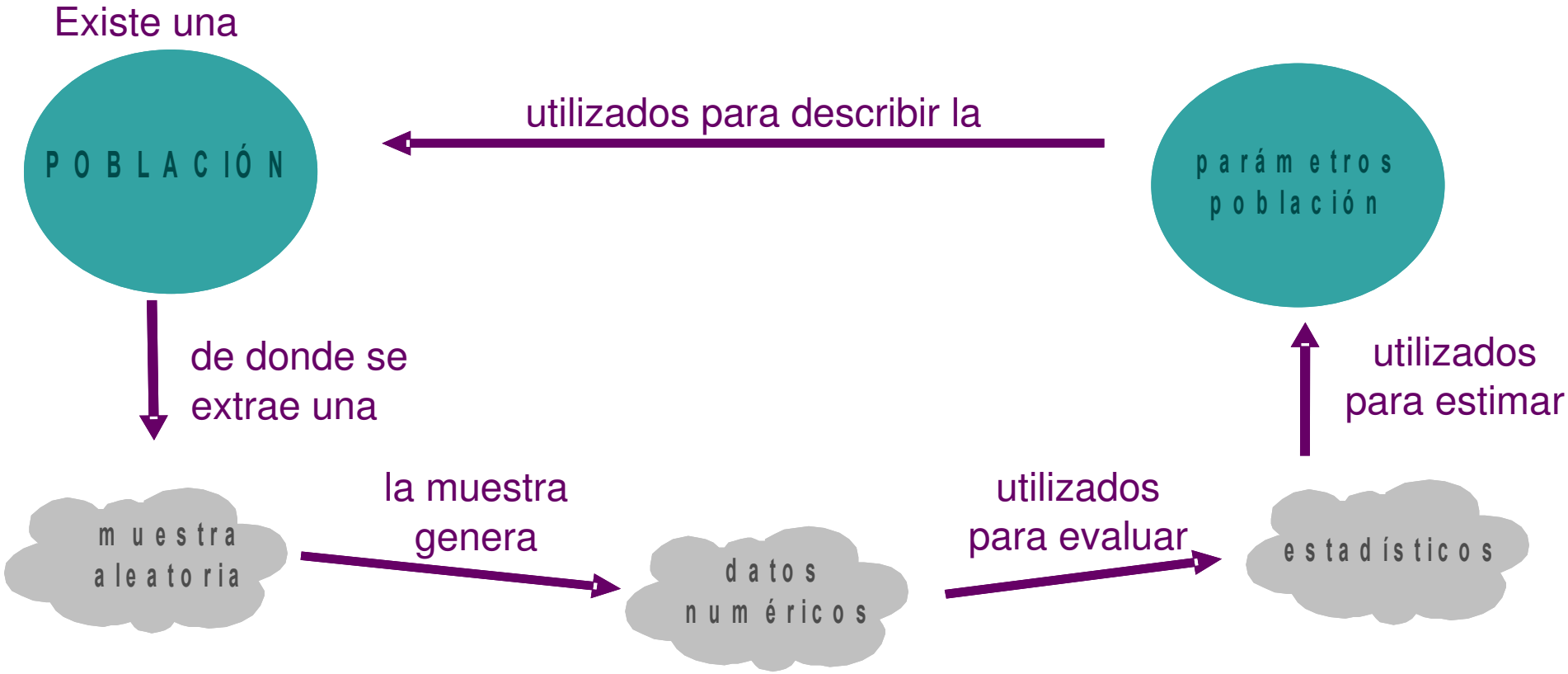
Sin reposición de los elementos: no se permite que un mismo individuo sea seleccionado más de una vez.

Con reposición: un elemento puede ser extraído varias veces.

Cuando la población es grande la diferencia entre ambos casos es mínima.

Inferencia Estadística

El problema que aparece con más frecuencia en la práctica es el de la estimación de parámetros de la población, que son desconocidos.

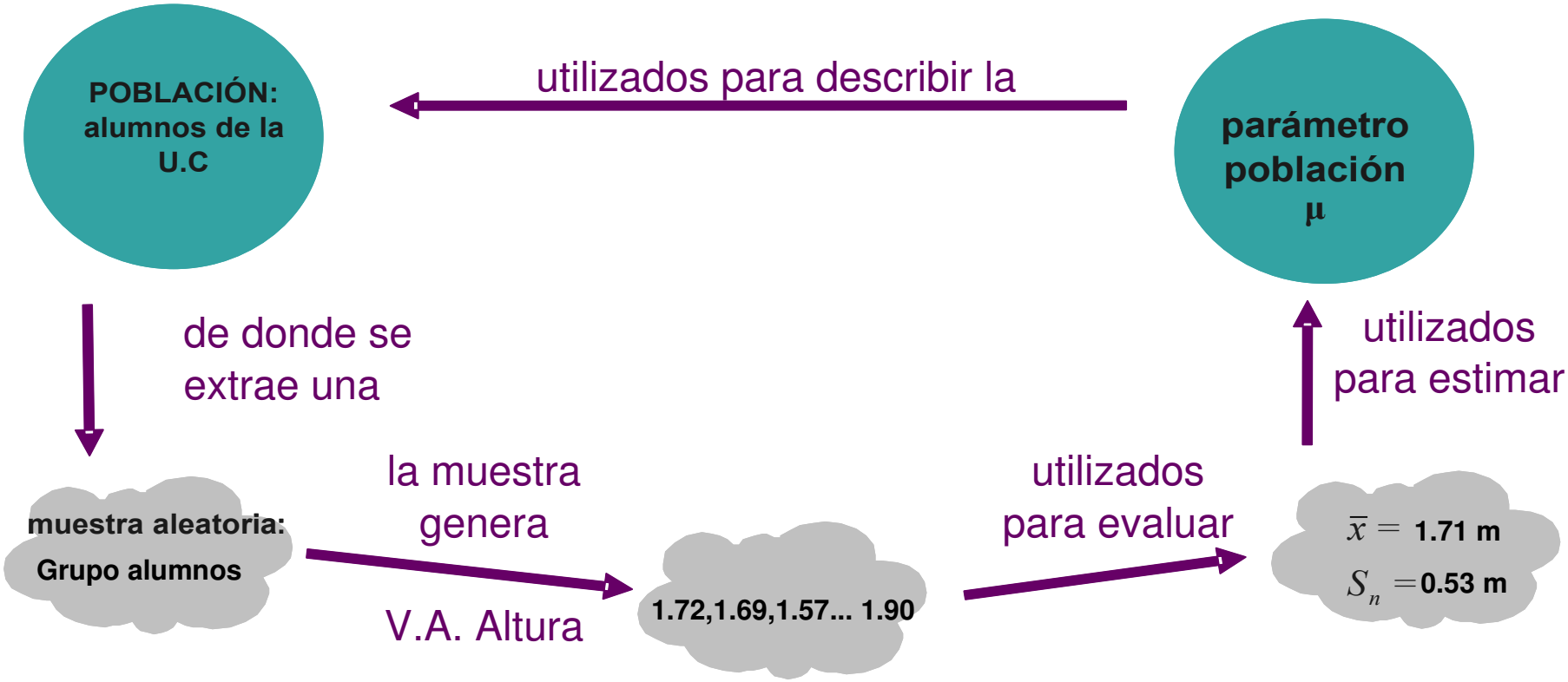


El objetivo es **deducir (inferir) características de una población mediante el estudio de una muestra**

María Dolores Frías, Jesús Fernández y Carmen María Sordo

Inferencia Estadística

Se desea conocer la altura de los alumnos de la Universidad de Cantabria.



Conjunto de métodos estadísticos que permiten deducir (inferir) como se distribuye la población en estudio a partir de la información que proporciona una muestra.

Inferencia Estadística

Estimación puntual: Obtener un pronóstico numérico único sobre un parámetro de la distribución.

Estimación por intervalos: Obtener un margen de variación para un parámetro de la distribución.

θ Población, parámetro

- proporción P
- media μ
- varianza σ^2

$\hat{\theta}$ Muestra, estimador parámetro

- proporción p
- media $\bar{x}$
- Varianza S_n^2

$$Objetivo: min|\theta - \hat{\theta}|$$

María Dolores Frías, Jesús Fernández y Carmen María Sordo

Estimación de una proporción

Dada una población con N individuos de los cuales M poseen cierta propiedad (e.g. mujeres) que no poseen los demás, la **proporción poblacional** se define como **P = M/N**

Si se elige una muestra de esa población de tamaño n, en la que aparecen m individuos con esa propiedad, entonces la **proporción muestral** se define como **p = m/n**

$$\theta \longrightarrow P$$

$$\hat{\theta} \longrightarrow p$$

La proporción poblacional (P) es constante mientras que cada muestra puede tener una proporción muestral (p) distinta.

La **proporción muestral** es una **variable aleatoria** por lo que es importante determinar su distribución.

Distribución de la proporción muestral

La distribución de la proporción muestral es la distribución de probabilidad de todos los valores posibles de la proporción muestral (p)

Muestreo con reemplazamiento o población infinita:

El número de individuos (m) que poseen la propiedad en la muestra es una variable aleatoria **binomial**.

La media y varianza de la proporción muestral serán:

$$E[p] = \frac{E[m]}{n} = \frac{n\,P}{n} = P$$

$$Var[p] = Var[m/n] = \frac{Var[m]}{n^2} = \frac{n\,P\,(1-P)}{n^2} = \frac{P\,(1-P)}{n}$$

Distribución de la proporción muestral

La distribución de la proporción muestral es la distribución de probabilidad de todos los valores posibles de la proporción muestral (p)

Muestreo sin reemplazamiento y población finita:

El número de individuos (m) que poseen la propiedad en la muestra es una variable aleatoria **hipergeométrica**.

La media y varianza de la proporción muestral serán:

$$E[p] = P$$

$$Var[p] = \frac{N-n}{N-1}\frac{P(1-P)}{n}$$

Distribución de la proporción muestral

- El valor medio de la función de probabilidad coincide con la proporción poblacional P.
- La varianza disminuye a medida que aumenta el tamaño de la muestra (n).
- La distribución de la proporción muestral p se aproxima a la distribución normal (μ=E[p] y σ^2=Var[p]) para n tendiendo a infinito.

$$F_{\hat{p}}(p) = P(\hat{p} \leq p) = P(m/n \leq p) = P(m \leq np) =$$
$$= F_{B(n,P)}(np) \approx F_{N(0,1)}\left(\frac{p-P}{\sqrt{\frac{P(1-P)}{n}}}\right)$$

si conociésemos P, podríamos calcular la probabilidad de que la proporción muestral sea menor que un cierto valor o que esté entre ciertos valores.

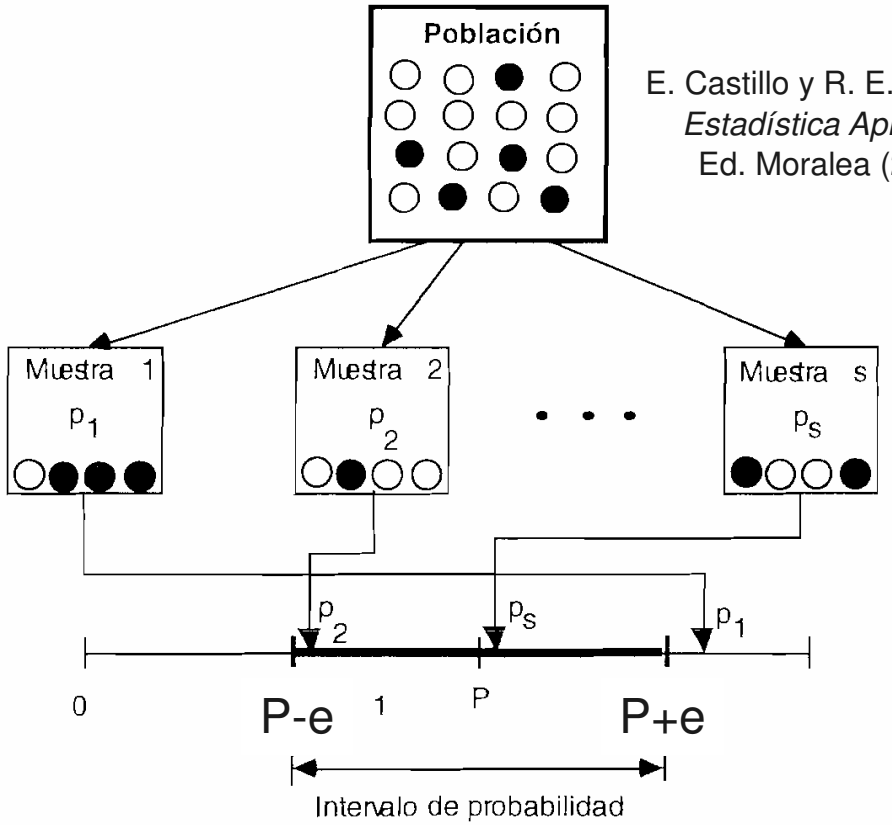
Intervalos de probabilidad de una proporción

Conocida la proporción poblacional, P, y la distribución de la proporción muestral, podemos obtener un intervalo donde la v.a. p tiene una probabilidad dada (1-α) de estar.

El 100(1-α)% de las muestras aleatorias dan un valor de la proporción muestral dentro de este intervalo.

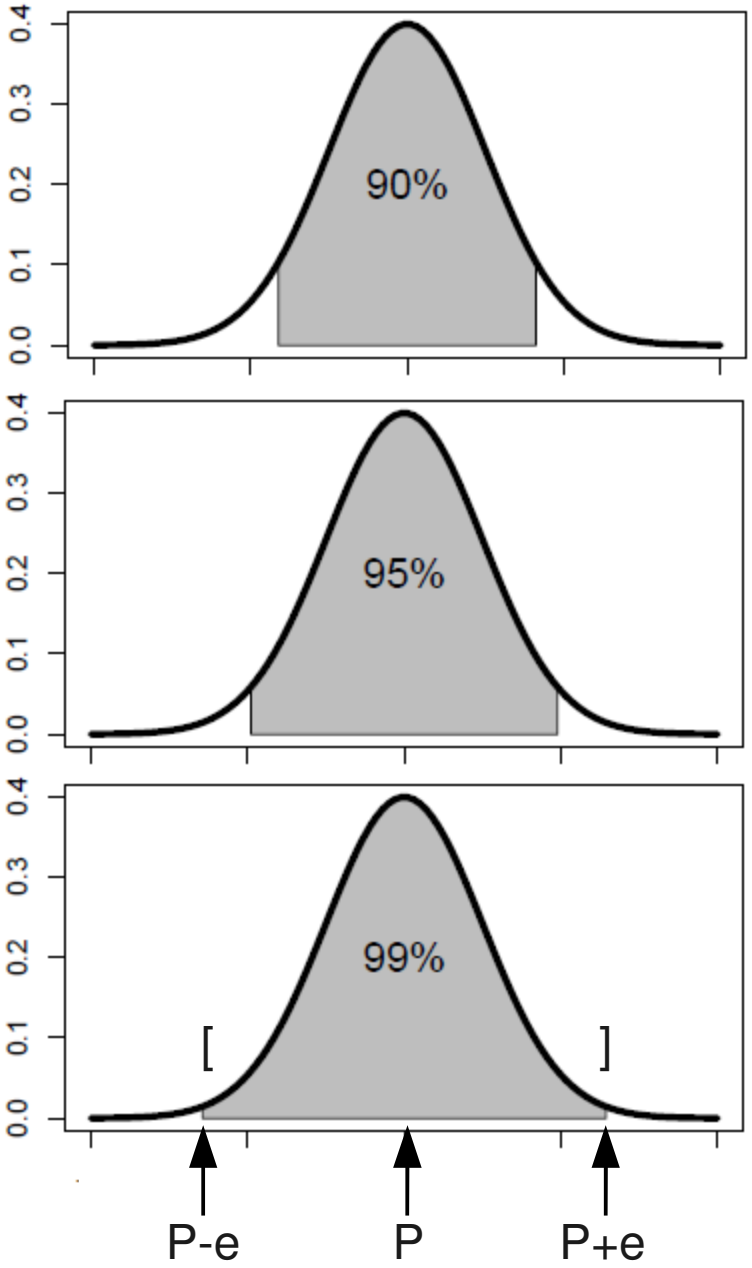
Ese intervalo se denomina **intervalo de probabilidad** de una proporción con confianza 1-α.

Los intervalos de probabilidad permanecen constantes para diferentes muestras.



E. Castillo y R. E. Pruneda
Estadística Aplicada.
Ed. Moralea (2001)

Intervalos de probabilidad de una proporción



1-α (nivel de confianza, valores típicos: 90% 95% 99%)

α es el nivel de significación

Existen infinitos intervalos 1-α. Nos quedaremos con un **intervalo simétrico** respecto al valor central P que, en el caso la distribución normal, es el de menor tamaño para una confianza dada.

$$P(P-e \leq p \leq P+e) = 1-\alpha$$

Intervalos de probabilidad de una proporción

Si el tamaño de la muestra es suficientemente grande, la variable aleatoria p tiende a la ley normal y los intervalos de probabilidad pueden obtenerse con las tablas de la ley normal.

$$\begin{aligned} P(P-e \leq p \leq P+e) = & \quad F_{N(0,1)}\left(\frac{P+e-P}{\sigma_p}\right) - F_{N(0,1)}\left(\frac{P-e-P}{\sigma_p}\right) = \\ = & \quad F_{N(0,1)}\left(\frac{e}{\sigma_p}\right) - \left(1 - F_{N(0,1)}\left(\frac{e}{\sigma_p}\right)\right) \\ = & \quad 2F_{N(0,1)}\left(\frac{e}{\sigma_p}\right) - 1 = 1 - \alpha \end{aligned}$$

$$\Rightarrow \quad F_{N(0,1)}\left(\frac{e}{\sigma_p}\right) = 1 - \frac{\alpha}{2} \quad \Rightarrow \quad e = z_{\alpha/2}\sigma_p$$

Donde $z_{\alpha/2} = F^{-1}_{N(0,1)}(1-\alpha/2)$

Intervalos de probabilidad de una proporción

$$P \pm e$$

Muestreo sin reemplazamiento y población finita

$$P \pm z_{\alpha/2}\sqrt{\frac{N-n}{N-1}\frac{P(1-P)}{n}}$$

$$nP > 5$$
$$n(1-P) > 5$$
$$n/N < 0.9$$

Condiciones de validez

Muestreo con reemplazamiento o población infinita

$$P \pm z_{\alpha/2}\sqrt{\frac{P(1-P)}{n}}$$

$$nP > 5$$
$$n(1-P) > 5$$

α	z_α
0.001	3.090
0.005	2.576
0.010	2.326
0.025	1.960
0.050	1.640
0.100	1.280

$$z_{\alpha/2} = F^{-1}_{N(0,1)}(1-\alpha/2)$$

Intervalos de probabilidad de una proporción

$$P \pm e$$

$$z_{\alpha/2} = F^{-1}_{N(0,1)}(1-\alpha/2)$$

Muestreo sin reemplazamiento y población finita

$$P \pm z_{\alpha/2}\sqrt{\frac{N-n}{N-1}\frac{P(1-P)}{n}}$$

$$nP > 5$$
$$n(1-P) > 5$$
$$n/N < 0.9$$

Condiciones de validez

Muestreo con reemplazamiento o población infinita

$$P \pm z_{\alpha/2}\sqrt{\frac{P(1-P)}{n}}$$

$$nP > 5$$
$$n(1-P) > 5$$

Ejemplo

La población de internados en un centro médico es de 1000 enfermos, de los cuales el 20% padecen afecciones cardiacas. Se elige una muestra de 50 enfermos del fichero de registro. Calcular el intervalo de probabilidad al 0.95 de p para el caso de muestreo sin y con reemplazamiento.

Intervalos de probabilidad de una proporción

Ejemplo

En este caso se tiene $nP = 10 > 5$, $n(1-P) = 40 > 5$ *y* $n/N = 0.05 < 0.9$ *por lo que es válida la aproximación normal. Además, por tratarse de un intervalo de probabilidad del* 95 %, z_α *vale* 1.96. *Por tanto, el intervalo de probabilidad pedido es*

$$\begin{aligned} & P \pm z_{\alpha/2}\sqrt{\tfrac{N-n}{N-1}\tfrac{P(1-P)}{n}} = \\ & = 0.2 \pm 1.96\sqrt{\tfrac{1000-50}{(1000-1)}\tfrac{0.2\times(1-0.2)}{50}} = \\ & = 0.2 \pm 0.108 \qquad \Rightarrow \qquad (0.092, 0.308) \end{aligned}$$

Para el caso de muestreo con reemplazamiento resulta:

$$\begin{aligned} & P \pm z_\alpha\sqrt{\tfrac{P(1-P)}{n}} = 0.2 \pm 1.96\sqrt{\tfrac{0.2\times(1-0.2)}{50}} = \\ & = 0.2 \pm 0.111 \qquad \Rightarrow \qquad (0.089, 0.311) \end{aligned}$$

Estimadores de una proporción

En la realidad, el problema más frecuente es el de la estimación de los parámetros de la población. Para ello se extrae de la población una muestra de tamaño n y conocida ésta se trata de estimar P.

- **Estimación puntual**: Se estima el valor de la proporción de la población (P) con el valor del parámetro de la muestra.

$$p \longrightarrow P$$

No da información alguna de la precisión de la estimación.

- **Intervalo de confianza**: Determina entre que valores (a, b] se encuentra la proporción de la población P con cierta probabilidad o certeza (1-α).

$$P(a \leq P \leq b) = 1 - \alpha$$

Complementa la estimación puntual precisando la exactitud de la estimación.

Intervalos de Confianza de una Proporción

Se dice que el intervalo (a,b] es un **intervalo de confianza** para P al nivel (1-α) si se verifica:

$$P(a \leq P \leq b) = 1 - \alpha$$

Partiendo del intervalo de probabilidad (1- α):

$$P(P - e \leq p \leq P + e) = 1 - \alpha$$

Esta expresión se puede escribir como:

$$P(p - e \leq P \leq p + e) = 1 - \alpha$$

Por lo que el intervalo [p-e, p+e] tiene una probabilidad asociada de (1- α) de contener al parámetro P.

Intervalos de confianza de una proporción

$$p \pm e$$

$$z_{\alpha/2} = F^{-1}_{N(0,1)}(1-\alpha/2)$$

Muestreo sin reemplazamiento y población finita

$$p \pm z_{\alpha/2}\sqrt{\frac{N-n}{N-1}\frac{p(1-p)}{n}}$$

$$n(p-e) > 5$$

$$n(1-p-e) > 5$$

$$n/N < 0.9$$

Condiciones de validez

Muestreo con reemplazamiento o población infinita

$$p \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

$$n(p-e) > 5$$

$$n(1-p-e) > 5$$

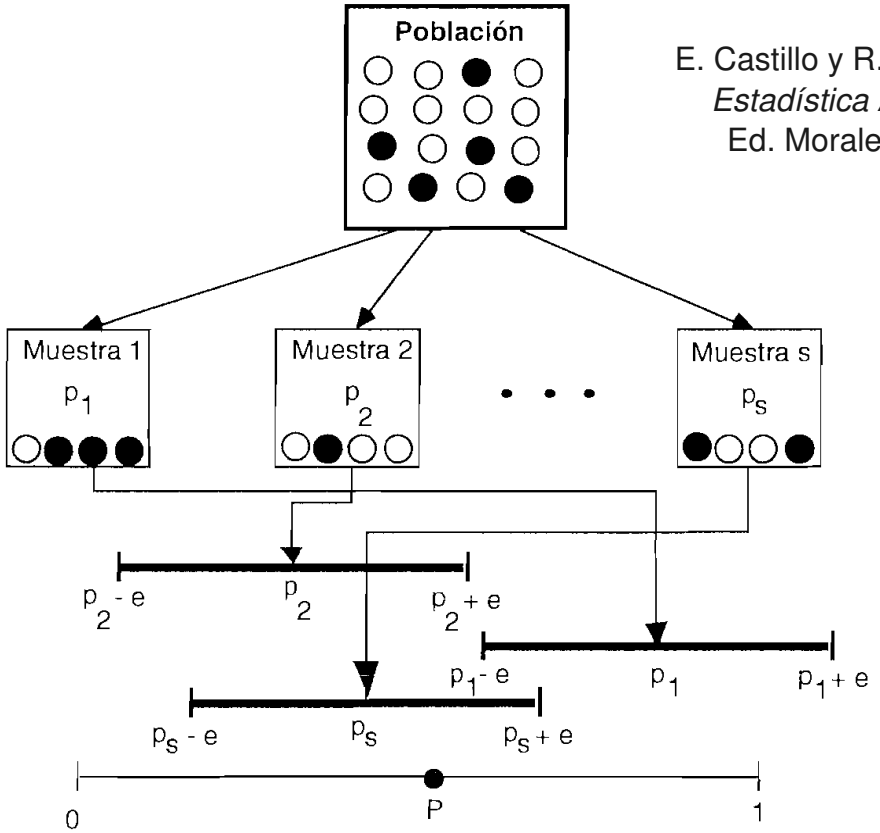
Ejemplo

En una muestra aleatoria de 50 rocas tomadas de una mina se observa que 20 de ellas son ricas en contenido mineral. Estimar puntualmente la proporción de rocas con alto contenido mineral en la mina. Calcular un intervalo de confianza 0.95 de esta proporción.

$$p = 20/50 = 0.4$$

$$0.4 \pm 1.96\sqrt{\frac{0.4 \times 0.6}{50}} = 0.40 \pm 0.14 \Rightarrow IC_{95\%} : (0.26, 0.54)$$

Intervalos de Confianza de una Proporción



E. Castillo y R. E. Pruneda
Estadística Aplicada.
Ed. Moralea (2001)

- Los intervalos de confianza sí cambian con las muestras.
- El 100(1- α)% de las muestras dan intervalos de confianza que contienen a la proporción poblacional.

Tamaño de muestra para estimar proporción

En la práctica el experimentador se plantea con qué error y nivel de confianza desea estimar la proporción y se calcula el tamaño de la muestra necesario.
Es decir, se conocen e y 1-α y se busca calcular n.

Muestreo sin reemplazamiento y población finita

$$e = z_{\alpha/2}\sqrt{\frac{N-n}{N-1}\,\frac{p(1-p)}{n}}$$

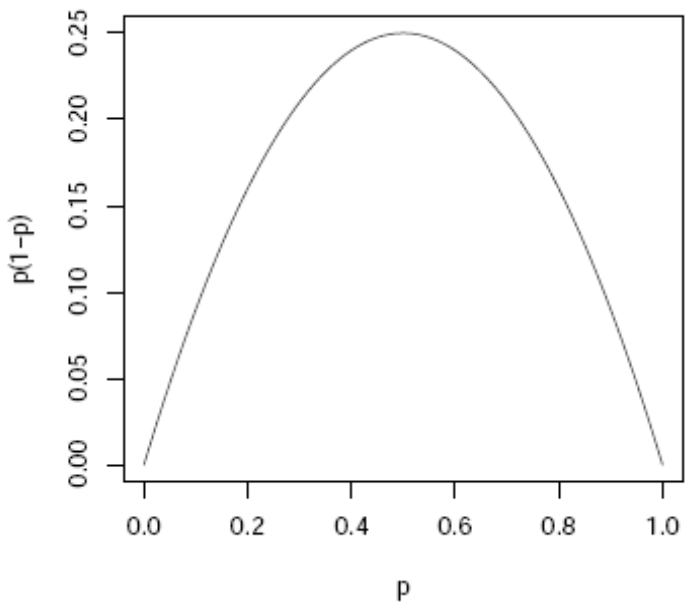
$$n = \left\lceil N\left(1 + \frac{e^2(N-1)}{z^2_{\alpha/2}p(1-p)}\right)^{-1}\right\rceil$$

Muestreo con reemplazamiento o población infinita

$$e = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

$$n = \left\lceil \frac{z^2_{\alpha/2}\,p\,(1-p)}{e^2}\right\rceil$$

Tamaño de muestra para estimar proporción



El cálculo de n implica el conocimiento previo de la proporción poblacional (que, como mucho, se podrá estimar DESPUÉS de tomar la muestra)

Si no se tiene idea del rango de valores de P(1-P), se puede usar el valor ¼ que es la cota superior de P(1-P).

Muestreo sin reemplazamiento y población finita

$$n = \left\lceil \frac{z^2_{\alpha/2} N}{4e^2(N-1) + z^2_{\alpha/2}} \right\rceil$$

Muestreo con reemplazamiento o población infinita

$$n = \left\lceil \frac{z^2_{\alpha/2}}{4e^2} \right\rceil$$

Ejercicio

Se quiere estimar la proporción de zurdos en una población con una confianza del 95% y una precisión de 0.01.

1. ¿Cuál debe ser el tamaño de la muestra escogida?

2. Mediante un muestreo previo se estima que $p \approx 0.1$, ¿qué tamaño debe tener la muestra si para calcularlo se utiliza la estimación de p obtenida?

Estimación de una media

$$\theta \longrightarrow \mu$$

$$\hat{\theta} \longrightarrow \overline{x}$$

Dada una población con N individuos que poseen cierta propiedad (altura), esa propiedad o variable tendrá su media poblacional μ, aún cuando su valor numérico se desconozca.

Si se elige una muestra aleatoria de esa población de tamaño n, se puede observar dicha variable y obtener la media muestral

La media muestral es una variable aleatoria ya que cada muestra tiene un valor distinto, por lo que tiene interés estudiar su función de probabilidad y en especial su media y su varianza.

La distribución de la media muestral es la distribución de probabilidad de todos los valores posibles de la media muestral.

Distribución de la media muestral

Muestreo sin reemplazamiento y población finita

$$E[\bar{x}] = \mu$$

$$Var[\bar{x}] = \frac{\sigma^2(N-n)}{n(N-1)}$$

Muestreo con reemplazamiento o población infinita

$$E[\bar{x}] = \mu$$

$$Var[\bar{x}] = \frac{\sigma^2}{n}$$

- El valor esperado de la media muestral coincide con la media poblacional.
- La varianza de la media muestral disminuye a medida que aumenta el tamaño de la muestra (n).
- La función de distribución converge a la normal para n tendiendo a infinito (teorema del límite central).

$$F_{\overline{X}}(\overline{x}) = P(\overline{X} \leq \overline{x}) = F_{N(\mu,\sigma^2/n)}(\overline{x}) = F_{N(0,1)}\left(\frac{\overline{x}-\mu}{\sqrt{\sigma^2/n}}\right)$$

Intervalos de probabilidad de una media

Se denomina **intervalo de probabilidad** de una **media** a aquel intervalo para el cual se sabe con una confianza 1-α que la media muestral se encuentra en dicho intervalo.

El intervalo (a,b] es un intervalo para la media muestral con probabilidad 1-α si se verifica:

$$P(a \leq \bar{x} \leq b) = 1 - \alpha$$

Al igual que para proporciones, para la media el intervalo de especial interés es el simétrico respecto de la media de la población.

$$P(\mu - e \leq \bar{x} \leq \mu + e) = 1 - \alpha \longrightarrow [\mu - e, \mu + e]$$

Intervalos de probabilidad de una media

- **Varianza** de la población **conocida** y n grande (**n≥30**):

  La distribución muestral se puede aproximar por una normal, (Teor. central del límite)

  Muestreo sin reemplazamiento y población finita

  $$\mu \pm z_{\alpha/2}\sigma\sqrt{\frac{N-n}{n(N-1)}}$$

  Muestreo con reemplazamiento o población infinita

  $$\mu \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

  $$z_{\alpha/2} = F^{-1}_{N(0,1)}(1-\alpha/2)$$

- **Varianza** de la población **desconocida** y **n** es **pequeña**.

  No se puede emplear σ^2/n, en su lugar se toma S^2/n. Nótese que S^2 es una variable a aleatoria (depende de la muestra) por lo que $\frac{\overline{X}-\mu}{\sqrt{S^2/n}}$ ya no sigue una distribución normal tipificada.

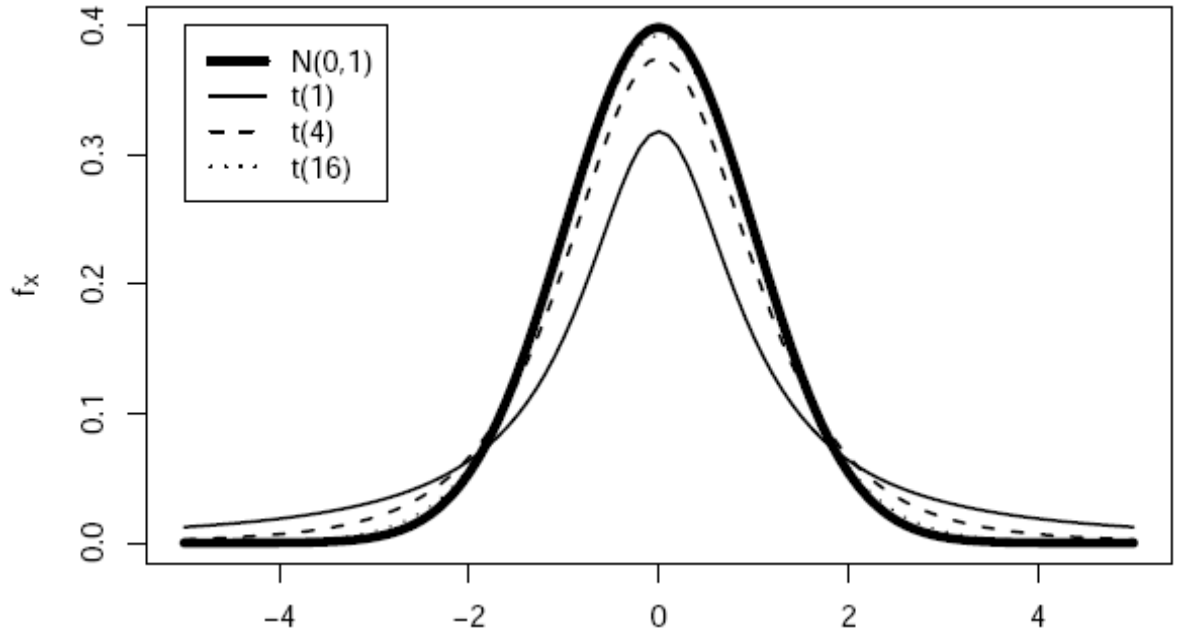
  En este caso, si la distribución de partida es normal, se considera el estadístico t que se distribuye según una t de Student con n-1 grados de libertad.

  $$t = \frac{\bar{x}-\mu}{\frac{S}{\sqrt{n}}} \longrightarrow \mu \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \qquad t_{n-1,\alpha/2} = F^{-1}_{t_{n-1}}(1-\alpha/2)$$

t de Student, t(n)

Distribución contínua con forma de campana, simétrica y unimodal. El parámetro n se denomina “grados de libertad”.

Eje de simetría en la recta X=0, por lo que su mediana = 0



n =1, 4, 16, ∞

Cuando n tiende a infinito la distribución t(n) tiende a la N(0,1)

t de Student, t(n)

Los cuantiles de la distribución t(n) aparecen en muchas fórmulas de inferencia estadística y se aproximan mediante tablas o mediante programas de ordenador.

n=10

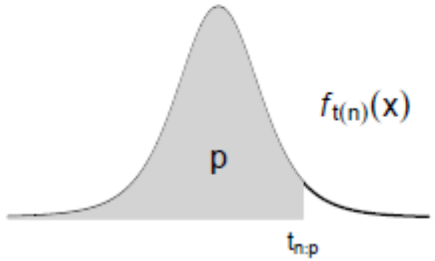
1-α=0.95

	Valores de p					
n	0.75	0.90	0.95	0.975	0.99	0.995
1	1.0000	3.0777	6.3138	12.7062	31.8205	63.6567
2	0.8165	1.8856	2.9200	4.3027	6.9646	9.9248
3	0.7649	1.6377	2.3534	3.1824	4.5407	5.8409
4	0.7407	1.5332	2.1318	2.7764	3.7469	4.6041
5	0.7267	1.4759	2.0150	2.5706	3.3649	4.0321
6	0.7176	1.4398	1.9432	2.4469	3.1427	3.7074
7	0.7111	1.4149	1.8946	2.3646	2.9980	3.4995
8	0.7064	1.3968	1.8595	2.3060	2.8965	3.3554
9	0.7027	1.3830	1.8331	2.2622	2.8214	3.2498
10	0.6998	1.3722	1.8125	2.2281	2.7638	3.1693
11	0.6974	1.3634	1.7959	2.2010	2.7181	3.1058
80	0.6776	1.2922	1.6641	1.9901	2.3739	2.6387
90	0.6772	1.2910	1.6620	1.9867	2.3685	2.6316
100	0.6770	1.2901	1.6602	1.9840	2.3642	2.6259
110	0.6767	1.2893	1.6588	1.9818	2.3607	2.6213
120	0.6765	1.2886	1.6577	1.9799	2.3578	2.6174

Tabla 5.4: Valores de $t_{n;p} \equiv F^{-1}_{t(n)}(p)$.

R tip

```
> # t_{n-1,α/2}
> qt(0.975, 9)
[1] 2.2622
```



Ejercicio

En un instituto se sabe que la estatura de los alumnos se ajusta a una $N(165,8^2)$ en cm. Calcular la probabilidad de que la altura media de 64 alumnos, elegidos al azar, esté entre 163 y 167 cm.

Estimadores de una media

Como ya se ha mencionado antes, en la realidad, el problema más frecuente es el de la estimación de los parámetros de la población. Para ello se extrae de la población una muestra de tamaño n y conocida ésta se trata de estimar μ.

- **Estimación puntual**: La media muestral es un buen estimador de la media de la población.

$$\overline{x} \longrightarrow \mu$$

No da información alguna de la precisión de la estimación.

- **Intervalo de confianza**: Determina entre que valores (a, b] se encuentra la media de la población μ con cierta probabilidad o certeza (1-α).

$$P(a \leq \mu \leq b) = 1 - \alpha$$

Complementa la estimación puntual precisando la exactitud de la estimación.

Intervalos de confianza de una media

Se dice que el intervalo (a,b] es un **intervalo de confianza** para **μ** al nivel (1-α) si se verifica:

$$P(a \leq \mu \leq b) = 1 - \alpha$$

Usando la hipótesis de normalidad y de la misma manera que se hizo para las proporciones:

$$P(\bar{x} - e \leq \mu \leq \bar{x} + e) = 1 - \alpha \longrightarrow e = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Si la aproximación normal no es válida (n pequeña y σ desconocida), al igual que se hizo con el intervalo de probabilidad, es necesario considerar el valor de la cuasivarianza muestral S^2 y calcular la variable t, que se distribuye según una t de Student.

$$t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}} \longrightarrow e = t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$$

Intervalos de confianza de una media

$$\bar{x} \pm e$$

- **Varianza** de la población **conocida** y n grande (**n≥30**):

Muestreo sin reemplazamiento y población finita

$$e = z_{\alpha/2}\sigma\sqrt{\frac{N-n}{n(N-1)}}$$

Muestreo con reemplazamiento o población infinita

$$e = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$z_{\alpha/2} = F^{-1}_{N(0,1)}(1-\alpha/2)$$

- **Varianza** de la población **desconocida** y **n** es **pequeña**.

$$e = t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$$

$$t_{n-1,\alpha/2} = F^{-1}_{t_{n-1}}(1-\alpha/2)$$

Tamaño de muestra para estimar media

Al igual que con la proporción, en la realidad el problema que se plantea se centra en estimar el tamaño de muestra necesario para estimar una media con un error y nivel de confianza dados.

Es decir, se conocen e y 1-α y se busca calcular n.

$$e = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \longrightarrow n = \left\lceil z^2_{\alpha/2}\frac{\sigma^2}{e^2} \right\rceil$$

Si la aproximación Normal no es válida, este cálculo se complica ya que n aparece implícitamente en $t_{n-1,\alpha/2}$. Además, S^2 también es desconocido hasta tomar la muestra.

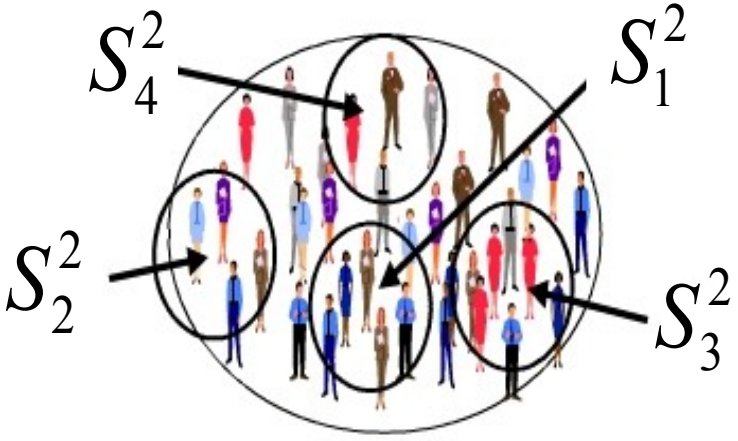
Ejercicio

Si la vida en horas de una bombilla eléctrica de 75 watios se distribuye de forma normal con una desviación típica de 5 horas y elegimos una m.a.s. de 30 bombillas cuya vida media es de 1014 horas, se pide:

1. Construir un intervalo de confianza para la vida media de las bombillas con un nivel de significación del 0.05.

2. Si queremos tener un nivel de confianza del 95% de que el error en la estimación de la vida media fuera menor de una hora, ¿Qué tamaño de la muestra elegiríamos?

Estimación de una varianza



$$\theta \longrightarrow \sigma^2$$

$$\hat{\theta} \longrightarrow S^2, \quad S_n^2$$

La varianza poblacional (σ^2) es constante mientras que cada muestra puede tener una varianza o cuasi-varianza muestral (S_n^2, S^2) distinta.

S_n^2 y S^2 son variables aleatorias por lo que es importante determinar su distribución

La distribución de la varianza (cuasi-varianza) muestral es la distribución de probabilidad de todos los valores posibles de la varianza (cuasi-varianza) muestral.

Distribución de la varianza muestral

$$E[S_n^2] = \frac{n-1}{n}\sigma^2$$

$$Var[S_n^2] = \frac{(n-1)^2}{n^3}(\mu_4 - \frac{n-3}{n-1}\sigma^4)$$

$$E[S^2] = \sigma^2$$

$$Var[S^2] = \frac{1}{n}(\mu_4 - \frac{n-3}{n-1}\sigma^4)$$

μ_4 es el momento de orden cuatro respecto de la media:

- El valor esperado de la varianza muestral **no coincide** con la varianza poblacional (se dice que es un estimador sesgado)
- El valor medio de las cuasi-varianzas muestrales si coincide con el de la varianza de la población (estimador centrado).

  [Esta es la razón de haber introducido este estadístico en estadística descriptiva!]

- Las varianzas de la varianza y la cuasi-varianza muestral tienden a cero cuando n tiende a infinito.

Intervalos de probabilidad de una varianza

Se denomina **intervalo de probabilidad** de una **varianza** a aquel intervalo para el cual se sabe con una confianza 1-α que la varianza muestral se encuentra en dicho intervalo.

$$P(a \leq S_n^2 \leq b) = 1 - \alpha$$

Para el caso de la varianza y cuasi-varianza muestrales, no existe una distribución a la que converjan todos los casos posibles de distribución poblacional.

La distribución de la varianza o cuasi-varianza muestral depende en alto grado de cual sea la distribución poblacional de partida.

Para simplificar vamos a considerar en lo que sigue sólo el caso de población normal.

Intervalos de probabilidad de una varianza

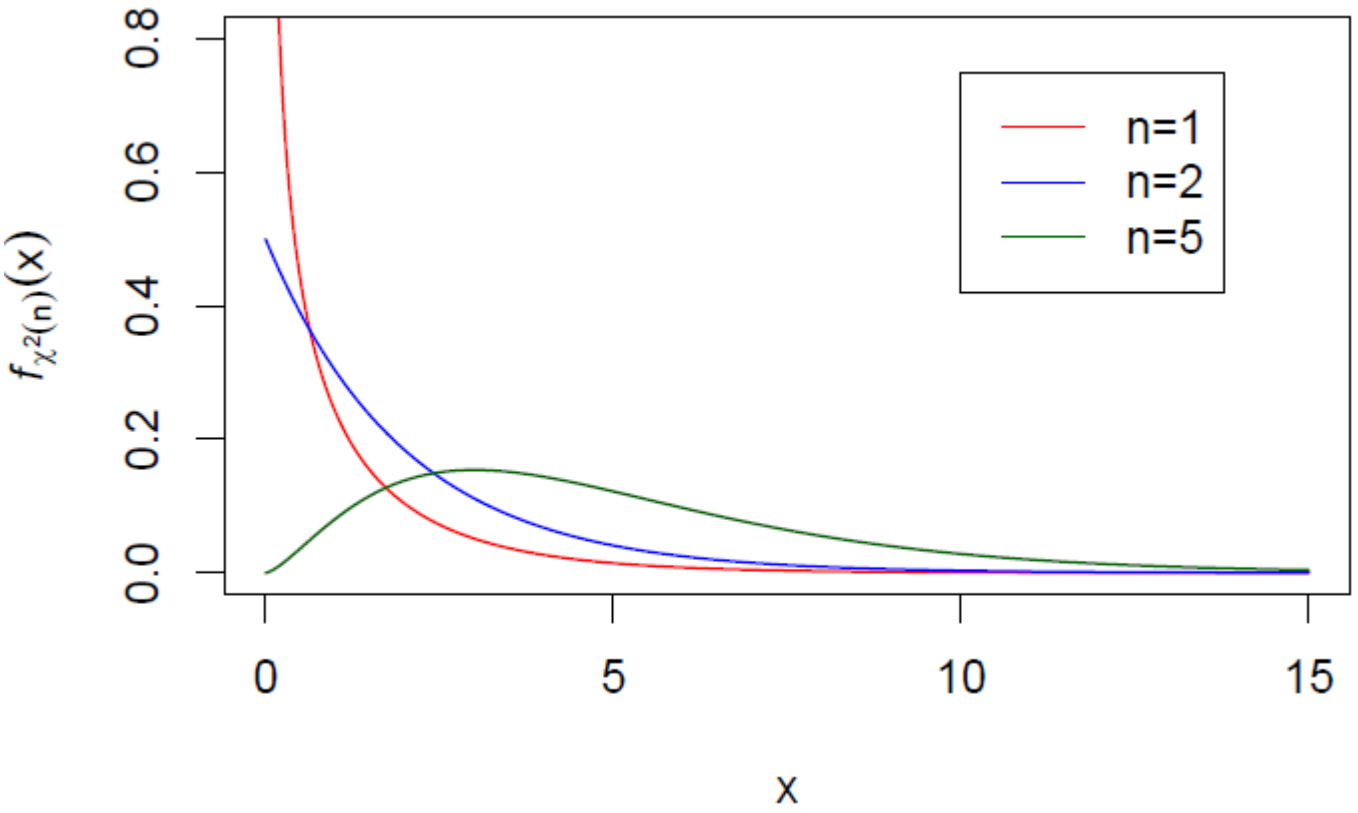
Si asumimos que la población sigue una distribución $N(\mu, \sigma^2)$, entonces la variable aleatoria

$$\frac{n\,S_n^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Distribución Chi-cuadrado, $\chi^2(n)$

Es una distribución contínua, que sólo toma valores positivos. Por lo tanto, NO es simétrica.

Es la distribución que sigue la suma de n variables aleatorias independientes N(0,1) elevadas al cuadrado. El parámetro n se denomina grados de libertad.



$$E[\chi^2(n)] = n$$

$$\text{var}[\chi^2(n)] = 2n$$

María Dolores Frías, Jesús Fernández y Carmen María Sordo

Intervalos de probabilidad de una varianza

Si asumimos que la población sigue una distribución $N(\mu, \sigma^2)$, entonces la variable aleatoria

$$\frac{n\,S_n^2}{\sigma^2} \sim \chi^2_{n-1}$$

Por tanto:

$$1-\alpha = P(a \leq S_n^2 \leq b) = P\left(\frac{na}{\sigma^2} \leq \frac{nS_n^2}{\sigma^2} \leq \frac{nb}{\sigma^2}\right) =$$

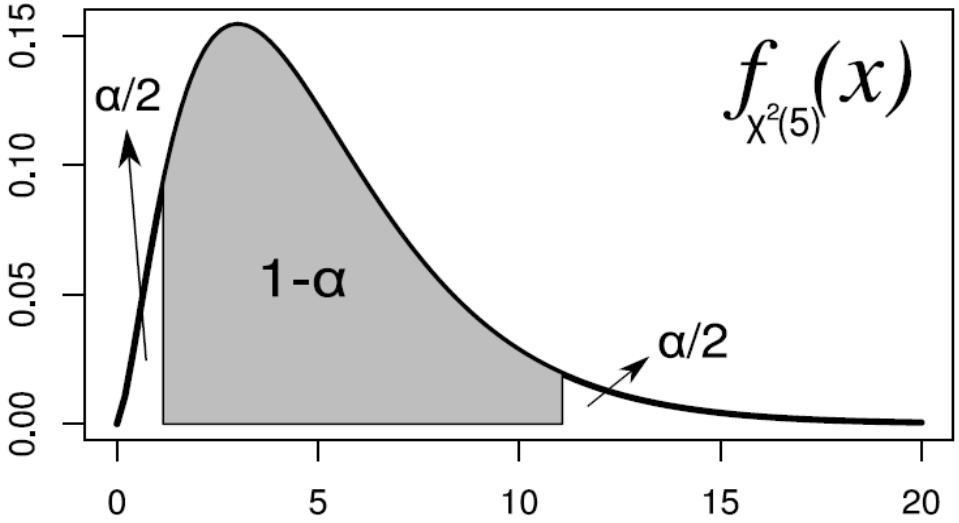
$$= F_{\chi^2(n-1)}\left(\frac{nb}{\sigma^2}\right) - F_{\chi^2(n-1)}\left(\frac{na}{\sigma^2}\right)$$

Pero hay infinitos valores de a y b que cumplen esta relación para una confianza dada.

Intervalos de probabilidad de una varianza

$$F_{\chi^2(n-1)}\left(\frac{na}{\sigma^2}\right) = \frac{\alpha}{2} \quad \Rightarrow \quad a = \frac{\sigma^2}{n} F^{-1}_{\chi^2(n-1)}\left(\frac{\alpha}{2}\right)$$

$$F_{\chi^2(n-1)}\left(\frac{nb}{\sigma^2}\right) = 1 - \frac{\alpha}{2} \quad \Rightarrow \quad b = \frac{\sigma^2}{n} F^{-1}_{\chi^2(n-1)}\left(1 - \frac{\alpha}{2}\right)$$



Al no ser simétrica la distribución, este NO es el intervalo más pequeño para una confianza dada. Sólamente es el más sencillo de calcular.

$$P\left(\frac{\chi^2_{(n-1,\alpha/2)}\sigma^2}{n} \leq S_n^2 \leq \frac{\chi^2_{(n-1,1-\alpha/2)}\sigma^2}{n}\right) = 1 - \alpha$$

$$\chi^2_{(n-1,\alpha/2)} = F^{-1}_{\chi^2(n-1)}(\alpha/2) \qquad \chi^2_{(n-1,1-\alpha/2)} = F^{-1}_{\chi^2(n-1)}(1-\alpha/2)$$

Intervalos de probabilidad de una varianza

	Valores de p									
n	0.005	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990	0.995
1	0.0_439	0.0_316	0.0_398	0.0_239	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
31	14.458	15.655	17.539	19.281	21.434	41.422	44.985	48.232	52.191	55.003
32	15.134	16.362	18.291	20.072	22.271	42.585	46.194	49.480	53.486	56.328
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

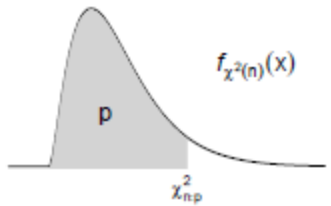


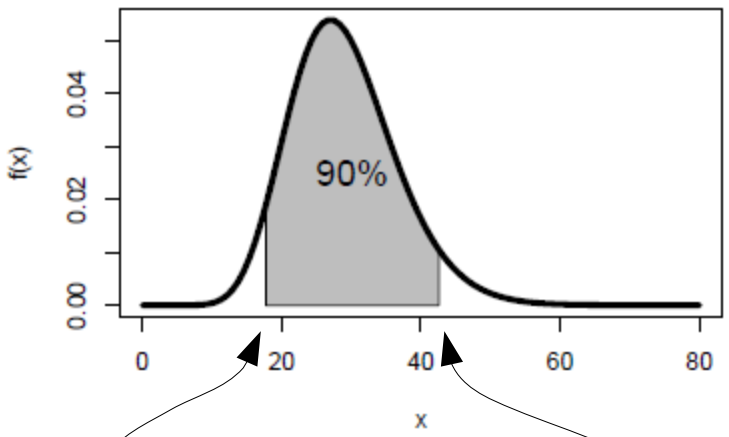
Tabla 5.5: Valores de $\chi^2_{n;p} \equiv F^{-1}_{\chi^2(n)}(p)$. Los subíndices indican el número de repeticiones de un digito. Por ejemplo, $F^{-1}_{\chi^2(1)}(0.005) = 0.0_439 = 0.000039$.

María Dolores Frías, Jesús Fernández y Carmen María Sordo

Intervalos de probabilidad de una varianza

n=30

$1-\alpha = 0.9 \quad \alpha/2 = 0.05$



$F^{-1}_{\chi^2_{(29)}}(0.05) = 17.7084$

$F^{-1}_{\chi^2_{(29)}}(0.95) = 42.5570$

	Valores de p									
n	0.005	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990	0.995
1	0.0_439	0.0_316	0.0_398	0.0_239	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11		1	16.473	34.382	37.652			6.928
26	11.160	12	χ^2	9	17.292	35.563	38.885	χ^2		8.290
27	11.808	12	$_{n-1,\alpha/2}$	1	18.114	36.741	40.113	$_{n-1,1-\alpha/2}$		9.645
28	12.461	13		8	18.939	37.916	41			0.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
31	14.458	15.655	17.539	19.281	21.434	41.422	44.985	48.232	52.191	55.003
32	15.134	16.362	18.291	20.072	22.271	42.585	46.194	49.480	53.486	56.328
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

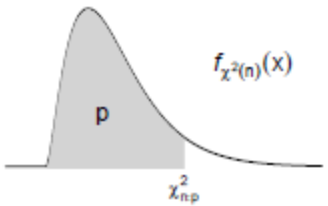


Tabla 5.5: Valores de $\chi^2_{n;p} \equiv F^{-1}_{\chi^2(n)}(p)$. Los subíndices indican el número de repeticiones de un digito. Por ejemplo, $F^{-1}_{\chi^2(1)}(0.005) = 0.0_439 = 0.000039$.

R tip

```
> # χ²_{n-1,α/2}
> qchisq(0.05,29)
> # χ²_{n-1,1-α/2}
> qchisq(0.95,29)
```

María Dolores Frías, Jesús Fernández y Carmen María Sordo

Intervalos de probabilidad de una cuasi-varianza

Para la cuasi-varianza el intervalo de probabilidad se calcularía de la misma manera:

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

sigue una distribución Chi-cuadrado con n-1 grados de libertad,

El intervalo de probabilidad vendría dado de la forma:

$$P\left(\chi^2_{n-1,\frac{\alpha}{2}}\frac{\sigma^2}{n-1} \leq S^2 \leq \chi^2_{n-1,1-\frac{\alpha}{2}}\frac{\sigma^2}{n-1}\right) = 1-\alpha$$

Estimadores de una varianza

En la realidad, el problema más frecuente es el de la estimación de los parámetros de la población. Para ello se extrae de la población una muestra de tamaño n y conocida ésta se trata de estimar σ^2.

- **Estimación puntual**: La varianza y cuasi-varianza muestral son buenos estimadores de la varianza de la población. La cuasivarianza tiene la ventaja de se un estimador centrado de σ^2.

$$S^2 \longrightarrow \sigma^2$$

No da información alguna de la precisión de la estimación.

- **Intervalo de confianza**: Determina entre que valores (a, b] se encuentra la varianza de la población con cierta probabilidad o certeza (1-α).

$$P(a \leq \sigma^2 \leq b) = 1 - \alpha$$

Complementa la estimación puntual precisando la exactitud de la estimación.

Intervalos de confianza de una varianza

De la misma manera que se hizo para el intervalo de probabilidad (1- α):

$$P\left(\frac{nS_n^2}{\chi^2_{(n-1,1-\alpha/2)}} \leq \sigma^2 \leq \frac{nS_n^2}{\chi^2_{(n-1,\alpha/2)}}\right) = 1-\alpha$$

es un **intervalo de confianza para la varianza poblacional** si la población de partida es normal. Por la definición de la cuasi-varianza muestral, este intervalo también se puede escribir como:

$$\left(\frac{(n-1)S^2}{\chi^2_{(n-1,1-\alpha/2)}}, \frac{(n-1)S^2}{\chi^2_{(n-1,\alpha/2)}}\right)$$

Ejercicio

Se sabe que el peso por bloque de un cierto preparado de hormigón se distribuye de forma normal. Con el objeto de estudiar la varianza de la distribución, se extrae una m.a.s de 6 bloques. Sabiendo que la varianza muestral es igual a 40, estimar la varianza poblacional mediante un intervalo de confianza al 90%.