

Tema 4: INFERENCIA ESTADISTICA - III

Intervalo de confianza para la varianza y la proporción

Biología sanitaria 2017/18. Universidad de Alcalá

M. Marvá. Actualizado: 2018-03-20

Más propiedades de la normal

Si $X_1 \sim N(\mu_1, \sigma_1)$ y $X_2 \sim N(\mu_2, \sigma_2)$ son independientes, entonces:

$$X = aX_1 + bX_2 \sim N(\mu_X = a\mu_1 + b\mu_2, \sigma_X = \sqrt{a^2\sigma_1^2 + b^2\sigma_2^2})$$

Esto se extiende a la combinación lineal de cualquier número de normales

Ejemplo: Dispones de dos medios de cultivo celular, A y B. Los fabricantes aseguran que la concentración de nutrientes es:

- En el medio A, en promedio, de 10gr/dm³, con una desviación típica de 0.5gr/dm³
- En el medio B, en promedio, de 12gr/dm³, con una desviación típica de 0.6gr/dm³

Si mezclas 4dm³ de A con 6dm³ de B:

- 1 Determina la concentración esperada y su desviación típica.
- 2 Si las concentraciones X_A y X_B siguen distribuciones normales, ¿qué distribución sigue la concentración del nuevo medio de cultivo (la mezcla)?

Inferencia sobre la varianza

Podemos estimar la media (centralización):

- Acompañarla de una medida de dispersión
- Conocer la dispersión de la población tiene interés de por sí

Dadas una v.a. X y sea $X_1, X_2, \dots, X_n$ una m.a.s. de X , entonces:

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

es un **estimador puntual** de la varianza.

El objetivo: encontrar un IC para σ .

Para ello: encontrar una distribución de probabilidad para una expresión que relacione s^2 y σ . intervalo de confianza para σ^2

Detalles en sección 6.5 del [libro](#)

Una función de densidad de probabilidad para la varianza

$$\begin{aligned}\frac{s^2}{\sigma^2} &= \frac{1}{\sigma^2} \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{(n-1)} = \frac{1}{(n-1)} \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \\&= \frac{1}{(n-1)} \sum_{i=1}^n \left(\frac{(X_i - \bar{X})^2}{\sigma^2} \right) = \frac{1}{(n-1)} \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 \\&= \frac{1}{(n-1)} \sum_{i=1}^n Z_i^2 = \frac{1}{n-1} (Z_1^2 + Z_2^2 + \cdots + Z_n^2)\end{aligned}$$

En resumen

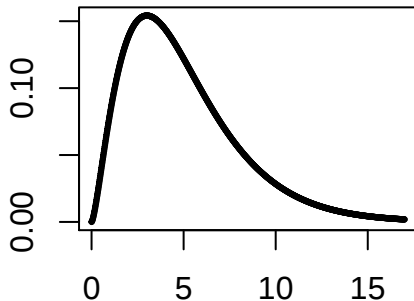
Si $X \sim N(\mu, \sigma)$, entonces para muestras de tamaño n

$$(n-1) \frac{s^2}{\sigma^2} \sim Z_1^2 + Z_2^2 + \cdots + Z_n^2$$

con $Z_i \sim N(0, 1)$

La v.a. $Y = Z_1^2 + Z_2^2 + \dots + Z_n^2$ con $Z_i \sim N(0, 1)$ independientes es de tipo **chi cuadrado** con **n-1** grados de libertad:

$$Y \sim \chi_{n-1}^2$$



Si $X \sim N(\mu, \sigma)$ y se utilizan m.a.s. de tamaño n , entonces:

$$(n-1) \frac{s^2}{\sigma^2} \sim \chi_{n-1}^2$$

La función de densidad está en la sección 6.5 del libro

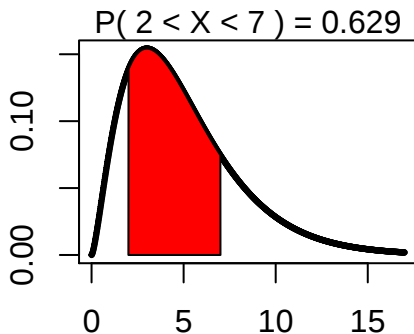
Con una χ^2_{n-1} se trabaja como con cualquier otra distribución continua.

Para calcular probabilidades con R, con muestras de tamaño 6 (5 grados de libertad):

- Probabilidades $P(2 < \chi^2_{19} < 7) = \text{pchisq}(5, \text{df} = 5) - \text{pchisq}(2, \text{df} = 5)$

```
pchisq(5, df = 5) - pchisq(2, df = 5)
```

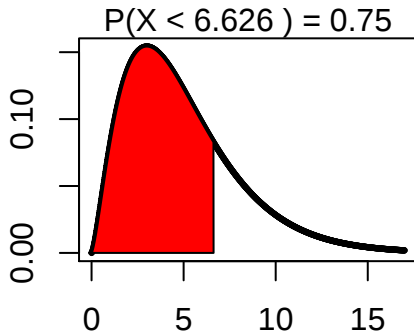
```
## [1] 0.4332648
```



- Percentil 75: x tal que $P(\chi_{19}^2 < x) = 0.75 = \text{qchisc}(0.75, \text{df} = 19)$

```
qchisq(0.75, df = 5)
```

```
## [1] 6.62568
```



Los **valores críticos** (probabilidades a la derecha) se definen de forma análoga ($n = 6$):

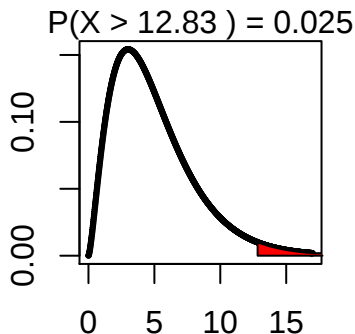
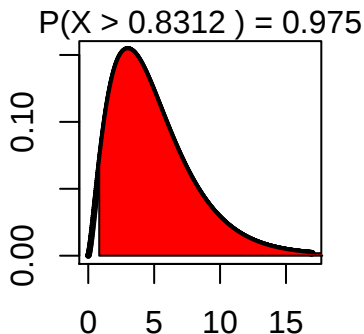
- $P(\chi_5^2 > \chi_{5,0.025}^2) = 0.025 = \text{qchisq}(0.025, \text{df} = 5, \text{lower.tail} = \text{F})$
- $P(\chi_5^2 > \chi_{5,0.975}^2) = 0.975 = \text{qchisq}(0.975, \text{df} = 5, \text{lower.tail} = \text{F})$

```
qchisq(0.025, df = 5, lower.tail = F)
```

```
## [1] 12.8325
```

```
qchisq(0.975, df = 5, lower.tail = F)
```

```
## [1] 0.8312116
```



Razonando de forma similar a como lo hicimos en el caso de la media se deduce lo siguiente:

Intervalo de confianza para σ^2 , poblaciones normales

Sea X una v.a. de tipo $N(\mu, \sigma)$. Si consideramos m.a.s. de tamaño n , el intervalo de confianza al nivel $nc = (1 - \alpha)$ para la media σ es:

$$\left(\frac{(n-1)s^2}{\chi_{n-1; \alpha/2}^2}, \frac{(n-1)s^2}{\chi_{n-1; 1-\alpha/2}^2} \right)$$

Ejemplo: Se ha medido las lipoproteínas de baja densidad (LDL) (en mg/dL) a 20 habitantes de Framingham. Determina el intervalo de confianza para la varianza al nivel de confianza del 90% suponiendo que la concentración de LDL se distribuye de forma normal

```
LDL = c(190, 157, 172, 130, 266, 193, 185, 170, 183, 152,  
        212, 97, 200, 162, 158, 132, 203, 111, 164, 244)
```

```
n = length(LDL) # IC varianzas
```

```
(n-1)*var(LDL)/qchisq(c(0.95, 0.05), df = n-1)
```

```
## [1] 1062.747 3166.443
```

```
# IC desviacion tipica
```

```
sqrt((n-1)*var(LDL)/qchisq(c(0.95, 0.05), df = n-1))
```

```
## [1] 32.59980 56.27116
```

```
library(asbio); ci.sigma(data = LDL, conf = 0.9) # IC varianza
```

```
##
```

```
## 90% Confidence interval for population variance
```

```
## Estimate      5%      95%
```

```
## 1686.050 1062.747 3166.443
```

Inferencia sobre la proporción:

Estimar la fracción p de individuos que presenta cierta característica:



Fracción de individuos **embridados**: polimorfismo en plumaje que consiste en un anillo ocular blanco. Se tomaron muestras y

Año	Embridados	No-embridados	% aves embridadas
2008	86	180	32.3
2009	138	270	33.8
2010	139	317	30.5

Para estimar p se toma una muestra y se determina si cada individuo presenta o no la característica.

La variable de interés es

$X =$ "el individuo presenta esa característica"

que es una variable Bernoullí $B(1, p)$, $p =$ prob de tenerla. Recuerda que

$$X = \begin{cases} 0 & \text{si NO presenta la característica} \\ 1 & \text{si presenta la característica} \end{cases}$$

Estimador puntual de p : **proporción muestral** para una muestra $X_1, \dots, X_n$ de tamaño n :

$$\hat{p} = \frac{X_1 + \dots + X_n}{n}$$

Es decir

$$\hat{p} = \frac{B(n, p)}{n}$$

Además,

$$E[\hat{p}] = \frac{1}{n} n\hat{p} \qquad \text{Var}[\hat{p}] = \frac{1}{n^2} n\hat{p}\hat{q}$$

Por el TCL (v1):

Intervalo de confianza (nivel $(1 - \alpha)$) para la proporción p , muestra grande

Si se cumplen, a la vez:

$$n > 30, \quad n \cdot \hat{p} > 5, \quad n \cdot \hat{q} > 5.$$

entonces el intervalo de confianza al nivel $(1 - \alpha)$ para la proporción p es:

$$\left(\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}, \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}} \right)$$

donde $\hat{p}$ es la proporción muestral y $\hat{q} = 1 - \hat{p}$

Se aproxima la binomial por la normal, en rigor hay que hacer la corrección por medio punto, pero como n es grande las diferencias apenas se notan. No suele hacerse.

Ejemplo: intervalo de confianza al 95% para la propoción de araos embridados en 2010

```
Embr = 139
noEmbr = 317
n = Embr + noEmbr
phat = Embr/n
z_alpha2 = qnorm(0.975)
# IC
phat + c(-1,1) * z_alpha2*sqrt(phat*(1-phat)/n)
```

```
## [1] 0.2625735 0.3470756
```

Alternativa con asbio

```
library(asbio)
ci.p(summarized = T,
     phat = Embr/(Embr + noEmbr),
     S.phat = sqrt(phat*(1-phat)/n),
     n = Embr + noEmbr)
```

```
##
## 95% Confidence interval for binomial parameter pi (method=Agresti-Coull)
## Estimate      2.5%      97.5%
## 0.3048246 0.2625735 0.3470756
```