Tema 5. Teoría Elemental del Muestreo e Inferencia Paramétrica con R

Estadística

Ángel Serrano Sánchez de León

Índice

- Distribuciones muestrales
  - Media
  - Proporción
- Dibujando la normal estándar
- Entendiendo el nivel de confianza
- Estimación de intervalos de confianza (IC)
  - Media de una población normal con σ^2 conocida
  - Media de una población no necesariamente normal con σ^2 desconocida a partir de una muestra grande
  - Proporción de éxito de una población binomial con una muestra grande
  - Diferencia de proporciones, muestra grande
- Tamaño de la muestra

Distribución muestral de una media

- Sea una población formada por N=5 individuos para los cuales una variable X toma los siguientes valores: 2, 3, 6, 8, 11.

```
> x <- c(2,3,6,8,11)
```

- La media μ y la varianza σ^2 (sesgada) de la población son las siguientes:

```
> mu <- mean(x)
> mu
[1] 6
> sigma2 <- var(x)*4/5  # var(x) = varianza insesgada de x
> sigma2
[1] 10.8
```

Distribución muestral de una media

- Hagamos muestras de n=2 individuos con reemplazo. En total habrá 5^2 casos (variaciones con repetición).

```
> muestras <- matrix(c(2,2,2,3,2,6,2,8,2,11,3,2,3,3,3,6,3,8,3,11,6,
  2,6,3,6,6,6,8,6,11,8,2,8,3,8,6,8,8,8,11,11,2,11,3,11,6,11,8,11,
  11),nrow=2)
> muestras
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14]
[1,]    2    2    2    2    2    3    3    3    3     3     6     6     6     6
[2,]    2    3    6    8   11    2    3    6    8    11     2     3     6     8
     [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23] [,24] [,25]
[1,]     6     8     8     8     8     8    11    11    11    11    11
[2,]    11     2     3     6     8    11     2     3     6     8    11
> n <- nrow(muestras)  # Tamaño de las muestras
> n
[1] 2
```

Distribución muestral de una media

- El valor de la media muestral $\overline{X}$ en cada una de las muestras es el siguiente:

```
> xbarra <- colMeans(muestras)
> xbarra
 [1] 2.0 2.5 4.0 5.0 6.5 2.5 3.0 4.5 5.5 7.0 4.0 4.5 6.0 7.0 8.5
[16] 5.0 5.5 7.0 8.0 9.5 6.5 7.0 8.5 9.5 11.0
```

- La media (=valor esperado) de $\overline{X}$ es:

```
> mean(xbarra)
[1] 6
```

- Coincide con el valor medio μ de la población:

```
> mu
[1] 6
```

$$\mathrm{E}(\overline{X}) \equiv \mu_{\overline{X}} = \mu$$

5

Distribución muestral de una media

- La varianza sesgada de la media muestral $\overline{X}$ es:

```
> var(xbarra)*24/25
[1] 5.4
```

- Este valor coincide con:

```
> sigma2 / n
[1] 5.4
```

$$\mathrm{Var}(\overline{X}) \equiv \sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$$

Distribución muestral de una proporción

- Sea una población formada por N=5 individuos a los que se pregunta si les gusta una bebida azucarada con gas.
- Si la respuesta es sí, la variable vale 1. En caso contrario, vale 0.

```
> r <- c(1,1,1,0,0) # A 3 les gusta, a 2 no
```

- La proporción p de personas de la población a las que les gusta la bebida es:

```
> p <- sum(r)/length(r)
> p
[1] 0.6
```

Distribución muestral de una proporción

- Hagamos muestras de n=2 individuos con reemplazo. En total habrá 5^2 casos (variaciones con repetición).

```
> muestras <- matrix(c(1,1,1,1,1,1,1,0,1,0,1,1,1,1,1,1,1,0,1,
  0,1,1,1,1,1,1,1,0,1,0,0,1,0,1,0,1,0,0,0,0,0,1,0,1,0,1,0,0,0,0
  ),nrow=2)
> muestras
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14]
[1,]    1    1    1    1    1    1    1    1    1     1     1     1     1     1
[2,]    1    1    1    0    0    1    1    1    0     0     1     1     1     0
     [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23] [,24] [,25]
[1,]     1     0     0     0     0     0     0     0     0     0     0
[2,]     0     1     1     1     0     0     1     1     1     0     0
> n <- nrow(muestras)  # Tamaño de las muestras
> n
[1] 2
```

Distribución muestral de una proporción

- La proporción de personas a las que les gusta la bebida en cada una de las 25 muestras es:

```
> pMuestral<-colMeans(muestras)
> pMuestral
[1] 1.0 1.0 1.0 0.5 0.5 1.0 1.0 1.0 0.5 0.5 1.0 1.0 1.0 0.5 0.5 0.5 0.5 0.5 0.0
[20] 0.0 0.5 0.5 0.5 0.0 0.0
```

- Esta proporción muestral es una variable aleatoria que tiene la siguiente media (=valor esperado) :

```
> mean(pMuestral)
[1] 0.6
```

- Coincide con la proporción de éxito de toda la población:

```
> p
[1] 0.6
```

$$\mathrm{E}(\hat{P}) \equiv \mu_{\hat{P}} = p$$

Distribución muestral de una proporción

- La varianza sesgada de la proporción muestral es:

```
> var(pMuestral)*24/25   # var = varianza insesgada
[1] 0.12
```

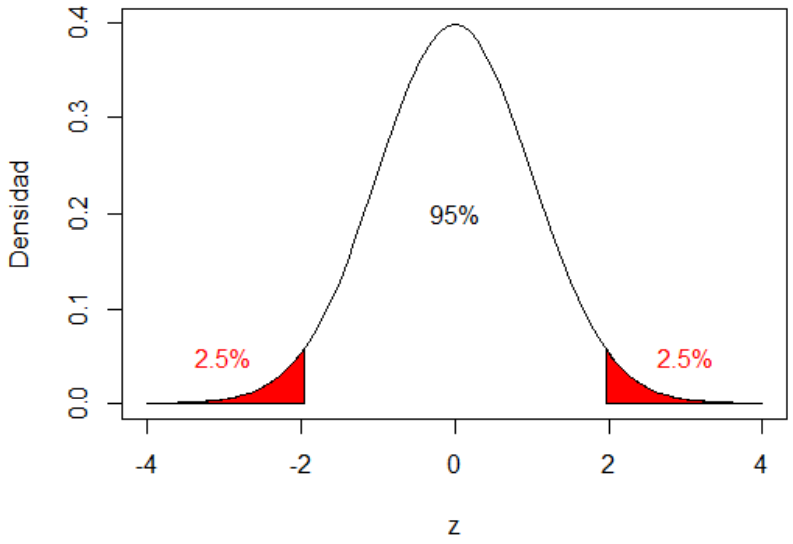
- Por otro lado, esta varianza coincide con el siguiente cálculo:

```
> p*(1-p)/n
[1] 0.12
```

$$\mathrm{Var}(\hat{P}) \equiv \sigma^2_{\hat{P}} = \frac{p(1-p)}{n}$$

Dibujando la normal estándar

```
> x <- seq(-4,4,0.1)
> y <- dnorm(x)  # Por defecto mu = 0, sigma = 1
> plot(x,y,type="l",xlab="z",ylab="Densidad")
> x1 <- qnorm(0.025) # -1.96
> x2 <- -x1  # 1.96
> xx1 <- seq(x1,-4,-0.1)
> yy1 <- dnorm(xx1)
> xx1 <- c(xx1,x1)
> yy1 <- c(yy1,0)
> polygon(xx1,yy1,col="red")
> xx2 <- seq(x2,4,0.1)
> yy2 <- dnorm(xx2)
> xx2 <- c(x2,xx2)
> yy2 <- c(0,yy2)
> polygon(xx2,yy2,col="red")
> text(0,0.2,"95%")
> text(-3,0.05,"2.5%",col="red")
> text(3,0.05,"2.5%",col="red")
```



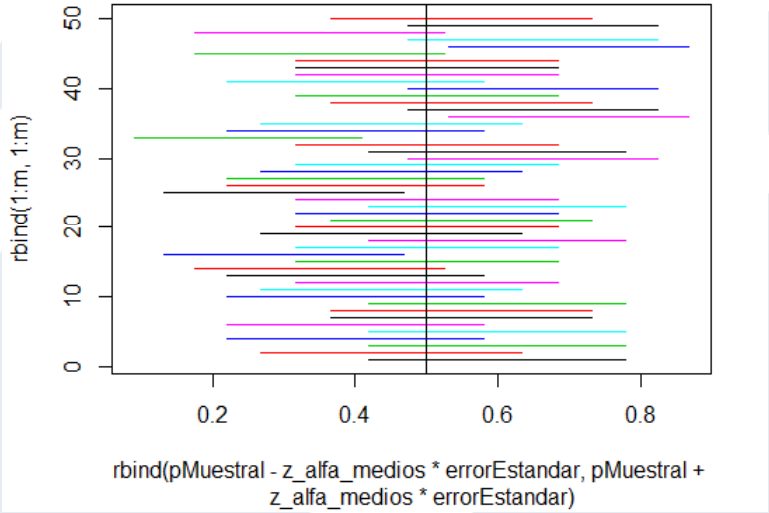
Entendiendo el nivel de confianza

- Supongamos que lanzamos una moneda $n = 20$ veces y contamos el número de caras. Repetimos el experimento 50 veces. Queremos comprobar si la moneda es perfectamente legal ($p = 0{,}5$).

```
> m <- 50 # Número de repeticiones
> n <- 20 # Tamaño de la muestra = número de tiradas en cada
  repetición
> p <- 0.5 # Probabilidad de cara
> pMuestral <- rbinom(m,n,p)/n  # Generamos las caras mediante
  números aleatorios binomiales. Al dividir por n, calculamos la
  proporción muestral de caras
> errorEstandar <- sqrt(pMuestral*(1-pMuestral)/n)
> nivelConfianza <- 0.90
> alfa <- 1-nivelConfianza  # 0.10
> z_alfa_medios <- qnorm(1-alfa/2)  # 1.64
> matplot(rbind(pMuestral - z_alfa_medios*errorEstandar,
  pMuestral + z_alfa_medios*errorEstandar), rbind(1:m,1:m),
  type="l", lty=1)
> abline(v=p) # Dibujamos la recta p = 0.5
```

Entendiendo el nivel de confianza

- Con el nivel de confianza del 90 %, el 10 % de los 50 de los intervalos de confianza generados (es decir, 5) **no** incluyen el valor verdadero de la proporción de éxitos de la población (0,5).



13

IC de la media de una población normal con σ^2 conocida

- Sea una población normal de μ desconocida y desviación típica $\sigma = 4$. Calcular el intervalo de confianza del 95 % del valor de la media dada la siguiente muestra:

  4, 13, 8, 12, 8, 15, 14, 7, 8

- Recordemos que en este caso el intervalo de confianza se calcula como:

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

- Donde α vale 0,05 y por tanto $z_{\alpha/2} = z_{0,025}$ corresponde al valor de z que deja un área del 2,5% bajo la cola derecha y de 97,5 % bajo la cola izquierda.

14

IC de la media de una población normal con σ^2 conocida

```
> x <- c(4, 13, 8, 12, 8, 15, 14, 7, 8)
> n <- length(x)
> mediaMuestral <- mean(x)
> nivelConfianza <- 0.95
> alfa <- 1-nivelConfianza  # 0.05
> sigma <- 4
> errorEstandar <- sigma/sqrt(n)
> z_alfa_medios <- qnorm(1-alfa/2)  # 1.96
> semianchuraIC <- z_alfa_medios * errorEstandar
> intervaloC <- mediaMuestral + c(-semianchuraIC,
  semianchuraIC)
> intervaloC
[1] 7.275604 12.502174
```

15

IC de la media de una población normal con σ^2 conocida

- Versión alternativa:

```
> install.packages("TeachingDemos")
> require("TeachingDemos")
> z.test(x,sd=4) # Por defecto conf.level=0.95
One Sample z-test
data: x
z = 7.4167, n = 9.000, Std. Dev. = 4.000, Std. Dev.
  of the sample mean = 1.333, p-value = 1.201e-13
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval: 7.275604 12.502174
sample estimates: mean of x 9.888889
```

16

IC de la media de una población no necesariamente normal con σ^2 desconocida a partir de una muestra grande

- Cargar el dataset "morley", que incluye datos de 100 mediciones de la velocidad de la luz de Albert Michelson en 1879. Los datos en el dataframe están expresados en km/s menos 299000.
- Desconocemos la varianza poblacional y la estimamos con la varianza muestral (insesgada). La muestra es grande $n = 100 > 30$.

$$\mu = \overline{X} \pm z_{\alpha/2} \frac{S}{\sqrt{n}}$$

```
> x <- morley$Speed
> n <- length(x) # 100 datos (muestra grande)
> z.test(x,sd=sd(x)) # Por defecto conf.level=0.95
One Sample z-test data: x
z = 107.8843, n = 100.000, Std. Dev. = 79.011, Std. Dev. of the
   sample mean = 7.901, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval: 836.9142 867.8858
sample estimates: mean of x 852.4
```

- Según este experimento la velocidad de la luz está en el intervalo (299837, 299868) km/s con un 95 % de probabilidad (el valor aceptado hoy en día es: **299792** km/s).

17

Comprobando la normalidad

- Para estimar si unos datos siguen la distribución normal, se puede usar la **gráfica cuantil-cuantil**.
- Representa los cuantiles muestrales respecto de los cuantiles teóricos.
- Cuanto más parecido sea a una recta, más gaussiano es el comportamiento de la variable.

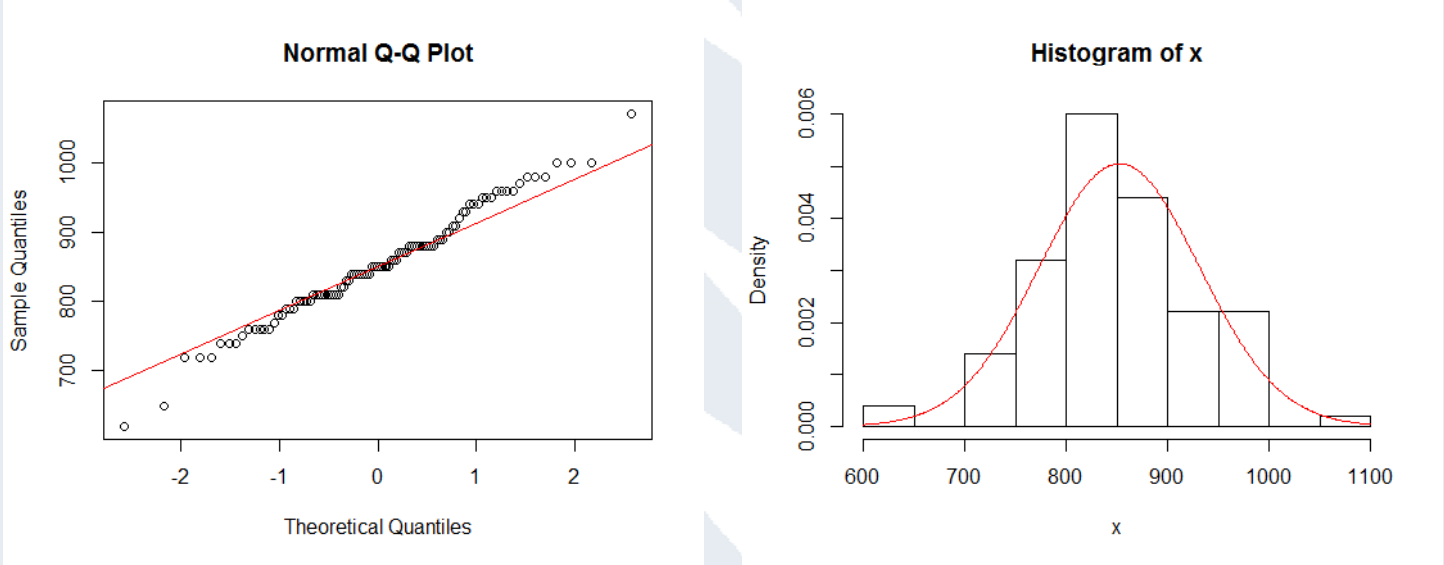
```
> qqnorm(x) # Gráfica cuantil-cuantil
> qqline(x,col="red") # Dibuja una recta
```

- También podemos superponer al histograma de área unidad de los valores una curva normal con la misma media y desviación que los datos.

```
> hist(x,freq=FALSE) # Histograma de área 1
> xx <- seq(600,1100)
> yy <- dnorm(xx,mean(x),sd(x))
> lines(xx,yy,col="red")  # Superponemos una curva
  normal con la misma media y desviación
```

18

Comprobando la normalidad



IC de la proporción de éxito de una población binomial con una muestra grande

- De una muestra de 100 personas, 64 afirman que le gusta un determinado producto. Calcular el intervalo de confianza del 95 % de la proporción de personas a las que les gusta el producto.
- La muestra es grande $n = 100 > 30$.
- No conocemos la proporción p de éxito de la población, pero podemos calcular la proporción muestral $\hat{P}$ a partir de los datos.

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

20

IC de la proporción de éxito de una población binomial con una muestra grande

- Sea un nivel de confianza del 95 %.

```
> n <- 100
> exitos <- 64
> pmuestral <- exitos/n
> pmuestral
[1] 0.64
> nivelConfianza <- 0.95
> alfa <- 1-nivelConfianza  # 0.05
> errorEstandar <- sqrt(pmuestral*(1-pmuestral)/n)
> z_alfa_medios <- qnorm(1-alfa/2) # 1.96
> semianchuraIC <- z_alfa_medios * errorEstandar
> semianchuraIC
[1] 0.09407827
> intervaloC <- pmuestral + c(-semianchuraIC, semianchuraIC)
> intervaloC
[1] 0.5459217 0.7340783
```

- O lo que es lo mismo: el intervalo de confianza del 95 % de la proporción de éxito la población es 0,64 ± 0,09, o bien: 64 ± 9 %.

IC de la proporción de éxito de una población binomial con una muestra grande

- Supongamos ahora un nivel de confianza del 99 %.

```
> nivelConfianza <- 0.99
> alfa <- 1-nivelConfianza  # 0.01
> z_alfa_medios <- qnorm(1-alfa/2)  # 2.57
> semianchuraIC <- z_alfa_medios * errorEstandar
> semianchuraIC
[1] 0.1236398
> intervaloC <- pmuestral + c(-semianchuraIC,
  semianchuraIC)
> intervaloC
[1] 0.5163602 0.7636398
```

- O lo que es lo mismo: 0,64 ± 0,12, o bien, 64 ± 12 %.
- A mayor nivel de confianza, mayor es el intervalo de confianza (y por tanto la estimación es más imprecisa).

IC de la proporción de éxito de una población binomial con una muestra grande

```
> prop.test(64,100) # Por defecto conf.level=0.95
1-sample proportions test with continuity correction
data: 64 out of 100, null probability 0.5 X-squared = 7.29, df = 1, p-value
   = 0.006934
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval: 0.5372745 0.7318279
sample estimates: p 0.64

> prop.test(64,100,conf.level=0.99)
1-sample proportions test with continuity correction
data: 64 out of 100, null probability 0.5 X-squared = 7.29, df = 1, p-value
   = 0.006934
alternative hypothesis: true p is not equal to 0.5
99 percent confidence interval: 0.5062185 0.7556792
sample estimates: p 0.64
```

- Los resultados con prop.test no coinciden con los calculados anteriormente porque utiliza fórmulas más exactas que las que hemos usado nosotros para el intervalo de confianza.

IC de la diferencia de proporciones, muestra grande

- Encuesta electoral a n = 1111 personas, partido A = 375 votantes, partido B = 285 votantes. Calcular diferencia de proporción de votantes con un 95,45 % de confianza.

```
> votantes <- c(375,285)
> n <- 1111   # Tamaño de la muestra grande
> porcentajes <- votantes/n
> porcentajes
[1] 0.3375338 0.2565257
> diferencia <- porcentajes[1]-porcentajes[2]
> diferencia
[1] 0.0810081
> nivelConfianza
[1] 0.9545
> alfa <- 1-nivelConfianza
[1] 0.0455
> z_alfa_medios <- qnorm(1-alfa/2) # 2
> errorEstandar <- sqrt((porcentajes[1]*(1-porcentajes[1]) + porcentajes[2]*(1-
   porcentajes[2]))/n)
> semianchuraIC <- errorEstandar * z_alfa_medios
> semianchuraIC
[1] 0.03862283
> intervaloC <- diferencia + c(-semianchuraIC,semianchuraIC)
> intervaloC
[1] 0.04238527 0.11963093
```

$$p_1 - p_2 \approx \left(\hat{P}_1 - \hat{P}_2\right) \pm z_{\alpha/2}\sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}}$$

Tamaño de la muestra

- La semianchura del intervalo de confianza es una medida del **margen de error** ε del parámetro de la población.
- Debemos elegir el tamaño n de una muestra lo suficientemente grande para que este error sea menor que un valor umbral de interés ε_0.
- Caso de la media:

$$\varepsilon = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \varepsilon_0 \Rightarrow n \geq \frac{z_{\alpha/2}^2\sigma^2}{\varepsilon_0^2}$$

- Caso de una proporción:

$$\varepsilon = z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \leq \varepsilon_0 \Rightarrow n \geq \frac{z_{\alpha/2}^2\hat{P}(1-\hat{P})}{\varepsilon_0^2}$$

25

Tamaño de la muestra

- Ejemplo: calcular el tamaño mínimo de la muestra para que en una selección aleatoria de alumnos de la UFV la altura media de la población pueda ser determinada con un margen de error inferior a 2 cm con un nivel de confianza del 95 %, suponiendo que la desviación estándar poblacional es de 10 cm.

```
> nivelConfianza <- 0.95
> alfa <- 1-nivelConfianza  # 0.05
> z_alfa_medios <- qnorm(1-alfa/2)  # 1.96
> e0 <- 2   # Umbral
> sigma <- 10
> z_alfa_medios^2 * sigma^2 / e0^2
[1] 96.03647
```

Tamaño de la muestra

- Ejemplo: calcular el tamaño mínimo de la muestra para que en una encuesta el error sea como máximo del 3% en la zona más desfavorable de máxima indeterminación(p=q=0,5) con un nivel de confianza del 95,45 % (el habitual en sondeos electorales).

```
> nivelConfianza <- 0.9545
> alfa <- 1-nivelConfianza
> alfa
[1] 0.0455
> z_alfa_medios <- qnorm(1-alfa/2)   # 2
> pMuestral <- 0.5
> e0 <- 0.03  # Umbral
> z_alfa_medios^2 * pMuestral * (1- pMuestral) / e0^2
[1] 1111.114
```