MUBD

Màster Universitari en Enginyeria de Dades Massives (Big Data)

Estadística

laSalle ENG
Universitat Ramon Llull

Índice

Inferencia estadística

Inferencia estadística

Introducción

- La metodología estadística permite la inducción: **inferir** las características de la población a partir de las observaciones de una muestra

- La Inferencia Estadística define y **cuantifica los riesgos** de este proceso

- Método científico y técnico (estadístico):
  - **Deductivo** → Diseño de la recogida de datos (Población → Muestra)
  - **Inductivo** → Inferir o estimar parámetros (Muestra → Población)

Inferencia estadística

Conceptos básicos

- **Población**: conjunto de elementos en estudio sobre los cuales deseamos extraer conclusiones
- **Parámetro**: indicador de la población se desea conocer.
- **Muestra**: pieza representativa de la población que sirve para obtener conocimiento de la misma
- **Estadístico**: función de los datos de la muestra
- **Estimador**: estadístico que sirve para estimar un parámetro de la población

Parámetros y estimadores

Ejemplos

Parámetro (θ) (**POBLACIÓN**)	Estimador ($\hat{\theta}$) (**MUESTRA**)
μ (Esperanza, media poblacional)	$\bar{x}$ (media muestral)
σ^2 (varianza poblacional) σ (desviación típica poblacional)	s^2 (varianza muestral) s (desviación típica muestral)
π (probabilidad)	p (proporción)

- Las palabras media y desviación se utilizan indistintamente para referirse al parámetro o al estimador llevando a confusión

Estimación puntual.

Propiedades de los estimadores

- Inevitablemente, las estimaciones puntuales **fallan** o, mejor dicho, como dependen de la muestra recogida, **fluctúan** entre las distintas muestras (aunque nosotros sólo observamos una de las posibles muestras)
- Las 2 obsesiones de la Estadística son:
  - **Cuantificar** los errores de estimación
  - **Minimizar** estos errores
- El **error típico** o estándar informa del **error esperado** al comparar el valor del estimador obtenido en el estudio con el valor del parámetro poblacional.
- No obstante, para una muestra concreta, el error exacto es desconocido pudiendo ser superior o inferior al error típico

Estimación puntual

Ejemplo

Los datos de consumo diario (en euros) de teléfono en una mediana empresa durante 9 días son: 587, 470, 676, 451, 436, 672, 584, 697 i 408

```
nterm <- c(587,470,676,451,436,
           672,584,697,408)
```

La **media muestral** ($\bar{x}$) estima la media poblacional (μ) de consumo diario de llamadas para esta empresa:

$\bar{x} = (\sum x_i)/n = 553.44$

```
mean(nterm)
```

La **desviación muestral** (s) estima la desviación poblacional (σ):

$s = \sqrt{(\sum(x_i - \bar{x})^2)/(n-1)} = 114.10$

```
sd(nterm)
```

El **error típico** de la media muestral (se) estima cuánto me equivoco en promedio en la estimación de la media poblacional:

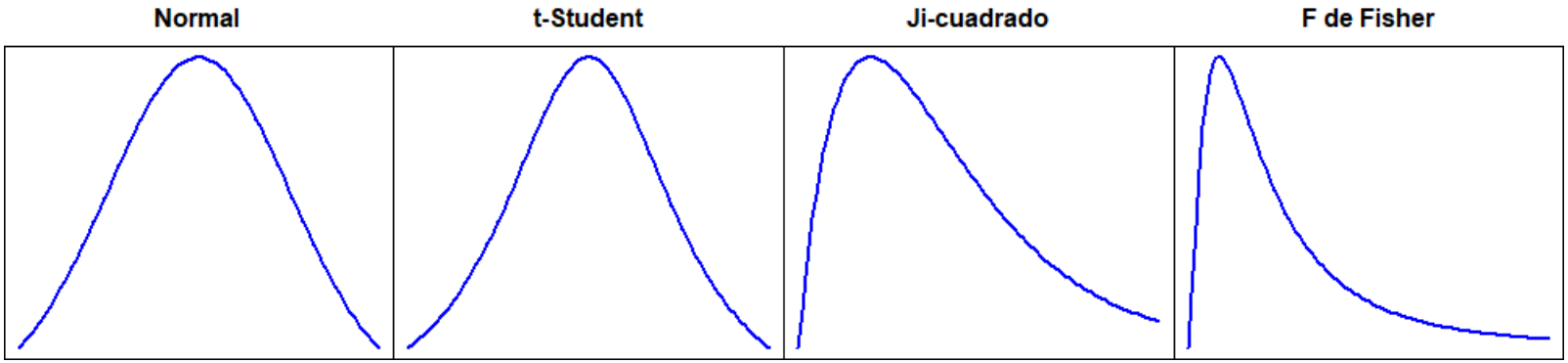
$\text{se} = \sqrt{(\sum(x_i - \bar{x})^2)/(n-1)} \cdot 1/\sqrt{n} = 38.03$

```
sd(nterm)/sqrt(length(nterm))
```

Distribuciones

Enumeración

- Para poder cuantificar la incertidumbre de nuestras estimaciones, se deben conocer la distribución de los **estimadores/estadísticos** a lo largo de infinitas muestras.
- En el anexo de distribuciones se encuentran las distribuciones más usuales en inferencia estadística.



- Por ej., sabemos que la media muestral sigue una distribución Normal: si se cogiesen infinitas muestras, la distribución de sus medias tendría una forma de campana.
- Véase esta app sobre la distribución de la media muestral.

Estimación por intervalo de la media poblacional (μ)

Fórmula

- La expresión para calcular el Intervalo de Confianza (IC) con una confianza (1-α) es la siguiente:

$$IC(\mu, 1-\alpha) = \bar{x} \mp t_{n-1,1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

[**R**: función ***t.test***]

- La incertidumbre que tenemos sobre la estimación puntual viene dada por:
  - La confianza (*1-α*): La confianza que deseamos determina el valor $t_{n-1,1-\frac{\alpha}{2}}$. A mayor confianza, este valor será mayor y tendremos mayor precisión.
  - La variabilidad de la muestra (*s*): Si los datos de partida son más variables, también tendremos más incertidumbre.
  - Del tamaño muestral (*n*): A mayor tamaño, mayor precisión.
- Para que tenga sentido aplicar esta fórmula los datos de partida deben seguir una distribución Normal o tener una muestra no inferior a 100.

Estimación por intervalo de la media poblacional (μ)

R

- La instrucción *t.test* proporciona la estimación puntual y el intervalo de confianza a partir de unos datos. Supongamos que queremos estimar el consumo medio diario de los datos de consumo (en euros) de teléfono en una mediana empresa durante 9 días.

1-α	Sintaxis con R	IC(μ,1-α)
	`factura <- c(587,470,676,451,436,672,584,697,408)`	
95%	`t.test(factura,conf.level = 0.95)`	[465.7 ; 641.1]
99%	`t.test(factura,conf.level = 0.99)`	[425.8 ; 681.1]

- Se observa que a mayor confianza, menor precisión

Estimación por intervalo de la varianza (σ^2)

Cálculo

- En ocasiones, se puede estar interesado en conocer la variabilidad de un proceso. Podemos calcular el IC de la varianza poblacional (σ^2) con la siguiente expresión:

$$IC(\sigma^2, 1-\alpha) = \left[\frac{s^2(n-1)}{\chi^2_{n-1,1-\frac{\alpha}{2}}} \ , \frac{s^2(n-1)}{\chi^2_{n-1,\frac{\alpha}{2}}}\right]$$

- La amplitud depende de los mismos factores (e influyendo en el mismo sentido) que en el caso del IC para una media
- La premisa es que los datos provengan de una distribución Normal. Atención: es muy sensible al cumplimiento de esta premisa.
- A diferencia del IC para una media, este intervalo no es simétrico respecto a la estimación puntual

Estimación por intervalo de la varianza (σ^2)

Ejemplo

En un proceso industrial se desea estimar la variabilidad en las longitudes de una pieza determinada. Se recoge un lote de 25 piezas observando una variabilidad de s = 8 mm ¿Cuál es el IC del 95% para la variabilidad de la medida?

$$IC(\sigma^2, 0.95) = \left[\frac{s^2(n-1)}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{s^2(n-1)}{\chi^2_{n-1,\frac{\alpha}{2}}}\right] = \left[\frac{8^2(25-1)}{39.364}, \frac{8^2(25-1)}{12.401}\right] = [39.02, 123.86]$$

`[En R, no hay ninguna función base que lo calcule]`

Resultado:

$IC(\sigma^2, 0.95) = [39.02, 123.86]$

$IC(\sigma, 0.95) = \left[\sqrt{39.02}, \sqrt{123.86}\right] = [6.25, 11.13]$

La desviación típica de la longitud en este proceso está entre 6.25 y 11.13 mm. Es decir, las piezas se alejarán en promedio entre 6 y 11 mm del valor medio.

Estimación por intervalo de la probabilidad (π)

Cálculo

- Para estimar la probabilidad (π) de un evento se usa que, para muestras NO muy pequeñas, la proporción muestral se distribuye según una Normal. Por tanto el IC(1-α) para π es:

$$IC(\pi, 1-\alpha) = P \mp z_{1-\frac{\alpha}{2}}\sqrt{\frac{\pi(1-\pi)}{n}}$$

[**R**: función ***prop.test***]

- En este caso, la amplitud depende, sobre todo, de la confianza y del tamaño muestral.
- Ejemplo: En las 100 últimas valoraciones del servicio post-venta de nuestra empresa, en 56 el cliente ha quedado completamente satisfecho. La probabilidad de que un cliente quede completamente satisfecho es:

`prop.test(56,100,correct=FALSE)` → 0.56, IC95%=[0.46 , 0.65]

Pruebas de hipótesis (PH)

Razonamiento

- En las PH partimos de una hipótesis de partida (hipótesis nula) que se desea contrastar
- Es equivalente a calcular el IC y comprobar si el valor contrastado está dentro o fuera del intervalo. Si está fuera, se rechaza. En caso contrario, se dice que no existe evidencia para rechazar la hipótesis

Pruebas de hipótesis

Hipótesis nula y alternativa

- La hipótesis nula (H_0) se plantea formalmente con un parámetro al cuál le asignamos un valor a contrastar. En general, es una hipótesis conservadora.
- P.ej, se desea contrastar si la probabilidad de adquisición de un producto ofertado vía telefónica (con una muestra n = 100) es del 10%.

$$H_0: \pi = 0.10$$

- Además de H_0, se tiene una hipótesis alternativa H_1, que puede ser totalmente complementaria a la nula (enfoque bilateral), o parcialmente (unilateral):

$$H_1: \pi \neq 0.10 \qquad\qquad H_1: \pi < 0.10$$

- H_1 determina el sentido más opuesto a H_0, mientras que H_0 determina el valor del parámetro más conservador.

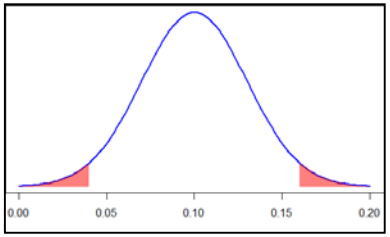
Pruebas de hipótesis

P-valor

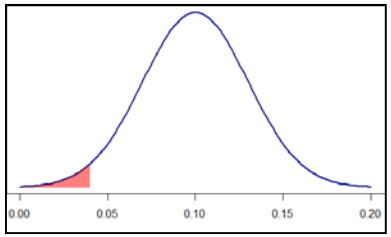
- El p-valor es la probabilidad de obtener unos resultados igual o más extremos que el observado siendo cierta $H_{0.}$
- Para un nivel de significación dado (generalmente $\alpha = 0.05$), los p-valores inferiores a este umbral implicarán un rechazo de la hipótesis nula
- El *P*-valor indica la frecuencia con la que puede pasar un evento en la muestra si la hipótesis H_0 es correcta:
  - *P-valor* pequeño ($p<0.05$) $\rightarrow$ Evidencia en contra de H_0
  - *P*-valor no pequeño ($p>0.05$) $\rightarrow$ **NO** demuestra la “certeza” de H_0

Pruebas de hipótesis

Tipos de planteamiento



- Bilateral
  - Nos interesa rechazar la hipótesis nula tanto si es un valor superior como inferior.
  - Ej: Medida de una pieza fabricada con unas especificaciones a cumplir, longitud=4.
  - Hipótesis: H_0: $\mu = 4$ vs H_1: $\mu \neq 4$
  - Se rechazará H_0 si se obtienen valores extremos en cualquier sentido



- Unilateral
  - Nos interesa rechazar la hipótesis nula sólo si es un valor superior o inferior.
  - Ej: número de ventas mensuales. Sólo nos interesa saber que llegamos a un umbral
  - Hipótesis: H_0: $\mu = 1000$ vs H_1: $\mu < 1000$
  - Se rechazará H_0 si se obtienen valores extremos sólo en el sentido de la hipótesis alternativa

Prueba de hipótesis sobre la μ y π

Sobre la media (μ)

- Para poner a prueba el valor de una media se utiliza el siguiente estadístico:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$ [**R**: función ***t.test***]

- Sigue una **t-Student con t_{n-1}** grados de libertad bajo la hipótesis nula si la variable de partida es Normal. La μ representa el valor que contrastamos, la *s* és la desviación típica y la *n* es el tamaño de la muestra.

Sobre una proporción (π)

- Para poner a prueba el valor de una probabilidad se utiliza el siguiente estadístico:

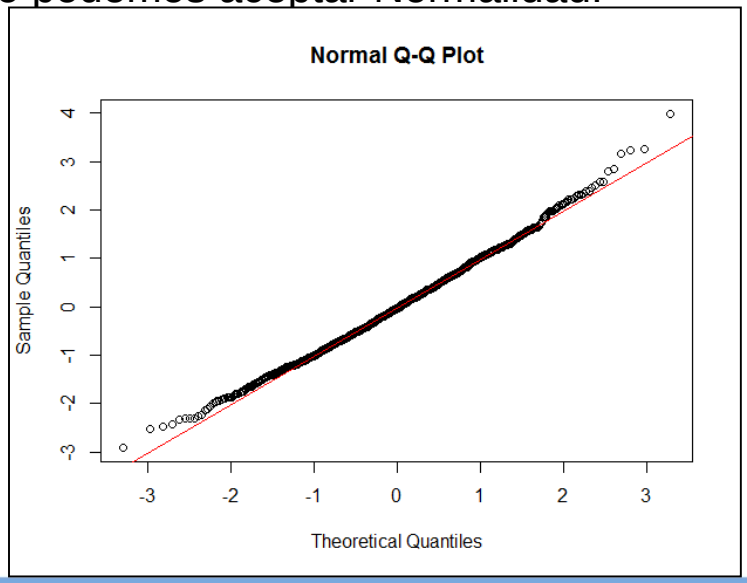
$$Z = \frac{P - \pi}{\sqrt{\pi \cdot (1-\pi)/n}} \sim N(0,1)$$ [**R**: función ***prop.test***]

- Sigue una **Normal estándar con** bajo la hipótesis nula si la variable de partida es Normal. Π es el valor de la probabilidad que se contrasta

Anexo: Premisa de Normalidad

¿Cómo comprobarla?

- Es necesario asumir que los datos de partida son normales para algunos análisis.
- Mediante estadísticos:
  - Kolmogorov-Smirnov (Estadístic D). Valores altos indican desajuste
  - Shapiro-Wilk (Estadístic W). Valores altos indican buen ajuste
- En ambos casos, p-valores < 0.05, indican que no podemos aceptar Normalidad. Ambos estadísticos deben interpretarse con cautela ya que son muy dependientes del tamaño muestral.
- Son más fiables los análisis visuales como el *qqplot (R: **qqnorm**)* que representa los cuantiles empíricos vs los cuantiles de una Normal teórica.



Si los puntos están suficientemente alineados sobre la recta, se asume Normalidad

Anexo: ¿Qué hacer si no se cumplen las premisas?

Bootstrap

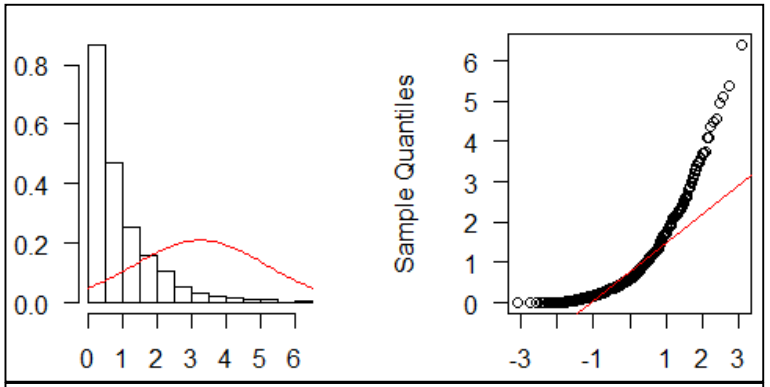
- Método de remuestreo sobre una misma muestra para obtener la distribución de un estadístico sin ninguna premisa adicional.
- Con la distribución de un estadístico obtenida por remuestreo se pueden calcular ICs
- Inconveniente: es costoso computacionalmente para muestras grandes
- Proceso:
  - Re-muestrear entre 1000 y 10000 muestras del mismo tamaño que la muestra original con reposición (un elemento puede estar repetido)
  - Calcular el indicador que deseemos para cada muestra (p. ej, la media)
  - Calcular el IC95% (o 90%) cogiendo los cuantiles 0.025 y 0.975 (o 0.05 y 0.95) de todos los indicadores calculados

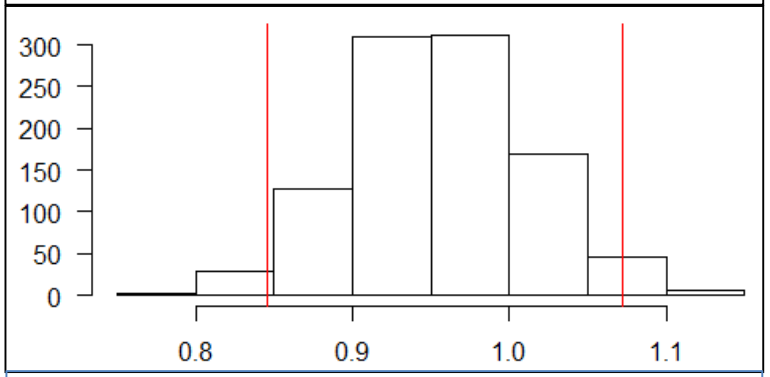
Anexo: ¿Qué hacer si no se cumplen las premisas?

Bootstrap - Ejemplo

- Se tiene una muestra (n=500) que no se ajusta a la Normalidad (ver primera figura) y se desea calcular un IC para la varianza (σ)
- Se generan 1,000 muestras que cogen elementos de la original con reposición
- Para cada muestra se calcula la desviación típica
- Se obtiene una distribución para s (estimador de σ). Ver segundo histograma.
- El IC95% se obtiene de los cuantiles 0.025 y 0.975 (líneas rojas) de esta distribución:

$$IC(\sigma\ ,\ 95\%) = [0.85\ ,\ 1.07]$$






```
n <- 500 ; nboot <- 1000
set.seed(12345)
x <- rexp(n)
par(mfrow=c(1,2),las=1)
hist(x,freq=FALSE)
curve(dnorm(x,mean(x),sd(x)),col=2,add=TRUE)
qqnorm(x);qqline(x,col=2)
m <- replicate(nboot,sd(sample(x,rep=TRUE)))
par(mfrow=c(1,1))
hist(m)
qic <- quantile(m,c(0.025,0.975))
abline(v=qic,col=2)
```

MUBD

Màster Universitari en Enginyeria de Dades Massives (Big Data)

Estadística

laSalle ENG
Universitat Ramon Llull