Estimación y contraste

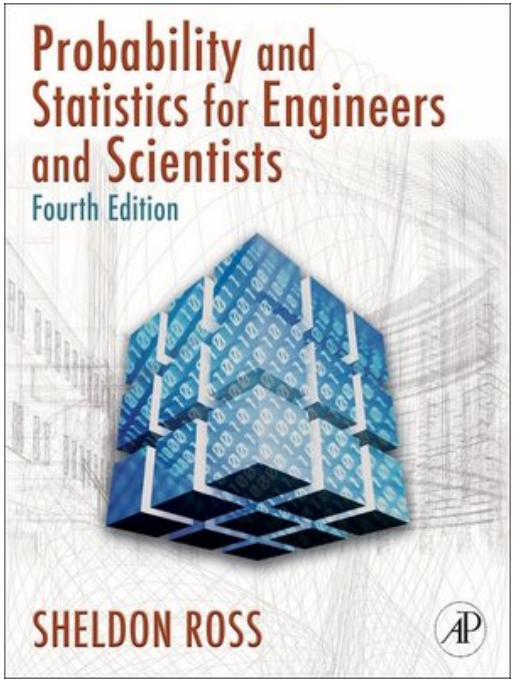
Estadística, Grado en Sistemas de Información

Constantino Antonio García Martínez

Universidad San Pablo Ceu

Contenidos

References



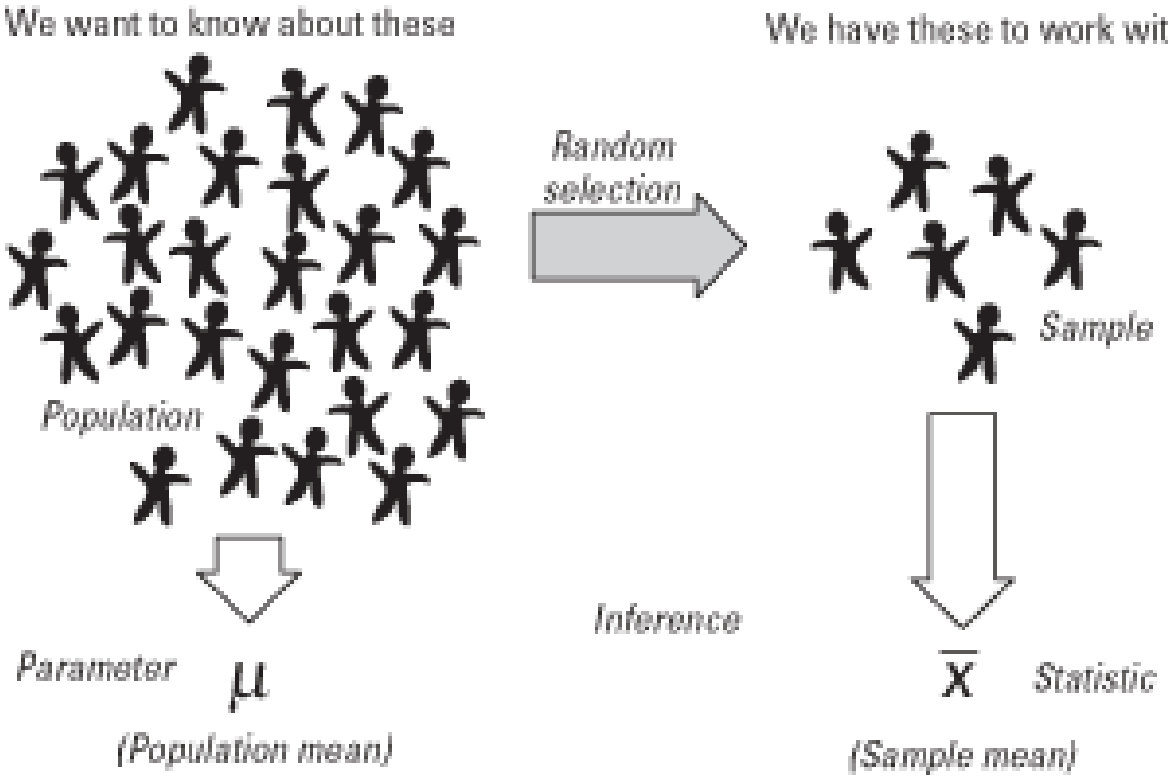
S. Ross. Introduction to Probability and Statistics for Engineers and Scientists. Chapter 7.

C.D. Barr, D.M. Diez, M. Çetinkaya-Rundel. OpenIntro Statistics. Chapters 4.

Introducción a la distribución de los estadísticos en el muestreo

La estadística y el muestreo

La estadística trata de obtener conclusiones a partir de la observación de datos. Generalmente estos datos se obtienen a partir del **muestreo** de una **población**.



A través del análisis de las muestras esperamos poder obtener conclusiones acerca de la población en su conjunto. Es fundamental saber **cómo se distribuyen las cantidades (estadísticos) que observamos tras muestrear**.

Parámetros de la población

La característica X de una población estará caracterizada cuando conozcamos su función de probabilidad $f(x)$. Muchas veces, podremos hacer suposiciones razonables acerca de la distribución en sí, pero no conoceremos los parámetros de la misma, p.ej.: de la media, varianza, etc.

Ejemplo: Parámetros de la población

Muchos parámetros físicos siguen una distribución Normal, por lo que podemos suponer que X, "La estatura de los españoles (mayores de edad) en 2017", sigue una distribución Normal. Sin embargo, desconocemos sus parámetros μ, y σ^2.

Parámetros de la población

Podemos tomar muestras aleatorias de la población para obtener valores que nos sirvan para estimar y testear hipótesis acerca de los parámetros poblacionales.

Ejemplo: Muestreo

Medimos la altura de 100 españoles. Como a priori no sabemos qué resultados vamos a obtener, las **100 medidas se caracterizan como 100 VAs**: $X_1, X_2, \cdots X_{100}$. A la hora de medir, cada una de las VAs da lugar a un único valor observado $x_1, x_2, \cdots, x_{100}$.

Cualquier cantidad obtenida a partir de las observaciones recibe el nombre de **estadístico muestral** o **estadístico**. ¡Fíjate que también es una VA!

$$T = g(X_1, X_2, \cdots, ..., X_n).$$

Ejemplo:

Generalmente eligimos estadísticos que se parecen a como se calculan los parámetros de interés. Por ejemplo, podemos usar $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ para estimar μ y $S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$ para estimar σ^2 (luego veremos que S^2 no es el mejor estadístico).

Población infinita/muestreo con reemplazamiento

Población infinita/muestreo con reemplazamiento

Distribución de la media

Distribución muestral de la media

Consideremos el estadístico de la media muestral

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

bajo la asunción de que las VAs X_i tienen media μ y varianza σ^2 finitas.

Teoremas
Si el muestreo es **con reemplazamiento**:

1. $\mathbb{E}[\bar{X}] = \mu$.
2. $\text{Var}(\bar{X}) = \mathbb{E}[(\bar{X} - \mu)^2] = \frac{\sigma^2}{n}$.
3. Si la población es Normal, dado que una suma de Normales es Normal se sigue que
$$\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n).$$
4. Si la población no es Normal, por el teorema central del límite, la VA
$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$
es asintóticamente una Normal estandarizada.

Distribución muestral de las proporciones

Veamos una aplicación: supongamos una población infinita con distribución de Bernouilli con probabilidades de éxito y fracaso p y q respectivamente.

Ejemplo:
Por ejemplo, una población podría ser todos los posibles lanzamientos de una moneda equilibrada.

Teorema
Sea el estadistico P: número de éxitos en n experimentos. Los momentos de P verifican

$$\mu_p = p \qquad \sigma_p = \sqrt{\frac{pq}{n}}.$$

Además, si $n \geq 30$, la distribución de la VA

$$Z = \frac{P - p}{\sqrt{\frac{pq}{n}}}$$

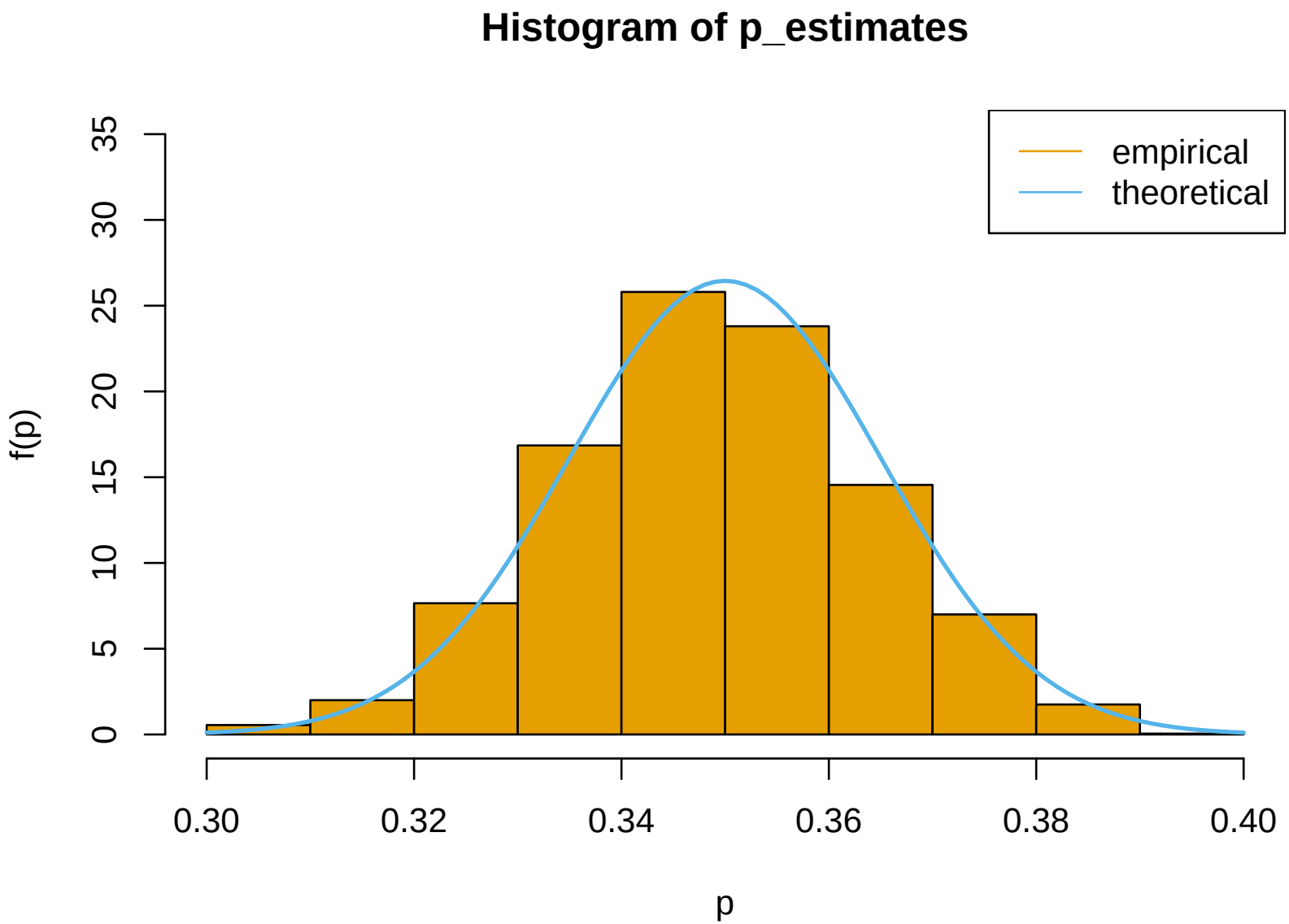
será aproximadamente una normal estándar.

Distribución de una proporción mediante simulaciones

```r
p = 0.35 # Real value of p
N = 1000  # Sample Size
nb_sim = 2000 # Number of simulations

p_estimates = replicate(nb_sim, {
  samples = rbinom(N, size = 1, prob = p)
  sum(samples) / N
})
hist(p_estimates, freq = FALSE, xlab = "p", ylab="f(p)", col = 2, ylim=c(0, 35))
x_axis = seq(0.3, 0.4, 0.001)
lines(x_axis, dnorm(x_axis, mean = p, sd = sqrt(p * (1 - p) / N)), col = 3, lwd = 2)
legend("topright", c('empirical', 'theoretical'), col = c(2, 3), lty = 1)
```

Distribución de una proporción mediante simulaciones



Población infinita/muestreo con reemplazamiento

Distribución de sumas y diferencias de medias

Distribución muestral de sumas y diferencias

Supongamos que tenemos dos poblaciones y que tomamos n_1 y n_2 muestras para calcular los estadísticos T_1 y T_2.

Teorema
Si T_1 y T_2 son independientes, entonces:

$$\mathbb{E}[T_1 \pm T_2] = \mu_{T_1} \pm \mu_{T_2} \qquad \text{Var}[T_1 \pm S_2] = \sigma^2_{T_1} + \sigma^2_{T_2}$$

En el caso concreto en el que $T_1 = \bar{X}_1$ y $S_2 = \bar{X}_2$:

$$\mathbb{E}[\bar{X}_1 \pm \bar{X}_2] = \mu_{\bar{X}_1} \pm \mu_{\bar{X}_2} = \mu_1 \pm \mu_2 \qquad \text{Var}[\bar{X}_1 \pm \bar{X}_2] = \sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Ejemplo:
Si $S_1 = \bar{X}_1$ y $S_2 = \bar{X}_2$ provienen de una población normal, o bien si $n_1, n_2 \geq 30$ entonces

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

Población infinita/muestreo con reemplazamiento

Distribución de la varianza muestral

La varianza muestral

Si $X_1, X_2, \cdots X_n$ son VAs obtenidas de un muestreo con reemplazamiento de la población, podemos estimar la varianza como

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

Sin embargo, este estimador es **sesgado** ya que

$$\mathbb{E}[S^2] = \frac{n-1}{n}\sigma^2.$$

Por ello definimos la **varianza muestral** como

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

Danger!

Algunos libros se refieren a S^2 como la varianza muestral y a $\hat{S}^2$ como la cuasi-varianza muestral. Nosotros NO usaremos esta nomenclatura, llamaremos a $\hat{S}^2$ **varianza muestral** y haremos como si S^2 nunca hubiese existido.

Ejercicio: var en R

La función *var* de `R`, ¿implementa $\hat{S}^2$ o S^2?

Ejercicio: Varianza muestral insesgada

Demuestra que $\hat{S}^2$ es insesgado.

Distribución muestral de la varianza

Teorema
Si se toman n muestras de una **población normal**, entonces la VA

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

tiene distribución Chi-cuadrado con $n-1$ grados de libertad.

Población infinita/muestreo con reemplazamiento

Distribución de la media cuando la varianza es desconocida

Distribución cuando la varianza es desconocida

Hemos visto que

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

se distribuye como una Normal si las muestras son normales, o que se puede aproximar como una normal si $n \geq 30$. Sin embargo, hemos asumido que σ^2 es conocida.

Teorema
Si n muestras se toman de una **población normal**, la VA

$$T = \frac{\bar{X} - \mu}{\hat{S}/\sqrt{n}}$$

se distribuye como una T de Student con $n - 1$ grados de libertad.

Población infinita/muestreo con reemplazamiento

Distribución del ratio de varianzas

Distribución en el muestro de ratios de varianzas

Para comparar varianzas podríamos estudiar $\hat{S}_1^2 - \hat{S}_2^2$ pero su distribución es complicada. En su lugar, empleamos $\hat{S}_1^2/\hat{S}_2^2$.

Teorema
Obtenemos dos muestras independientes de tamaño m y n de dos **poblaciones normales** con varianzas σ_1^2 y σ_2^2. Entonces, el estadístico

$$F = \frac{\hat{S}_1^2/\sigma_1^2}{\hat{S}_2^2/\sigma_2^2}$$

tiene una distribución F con $m-1, n-1$ grados de libertad.

Población Finita/Muestreo sin reemplazamiento

Muestreo sin reemplazamiento

Considera una población de N elementos, y supón que p es la proporción de de la población que tiene cierta característica de interés. Sea X la VA: número de individuos de la población que tienen dicha característica en un muestreo de tamaño n.

- Si n es con reemplazamiento/población infinita: X es Binomial y

$$\mu = np \qquad \sigma^2 = npq.$$

- Si n es sin reemplazamiento/población finita: X es hipergeométrica y

$$\mu = np \qquad \sigma^2 = npq\frac{N-n}{N-1}.$$

Al factor $\frac{N-n}{N-1}$ se le conoce como **factor de corrección para población finita**, y debemos emplearloe en los casos de **muestreo con reemplazamiento/población finita** para **corregir la varianza** del estadístico:

Muestreo sin reemplazamiento

Ejemplo: $\bar{X}$ Normal, σ desconocido y sin reemplazamiento

$$T_{n-1} = \frac{\bar{X} - \mu}{\frac{\hat{S}}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}$$

Ejemplo: $\hat{P}$ sin reemplazamiento

$$Z = \frac{\hat{P} - p}{\sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}}}$$