Temario de la sesión

1. Estimación puntual
    1.1. Métodos de los Momentos (EMM)
    1.2. Estimación de Máxima Verosimilitud (EMV)
    1.3. Comparación entre EMM y EMV
2. Intervalos de confianza
    2.1. Intervalos para la media y la proporción
    2.2. Intervalos para varianza y desviación estándar
3. Pruebas de hipótesis
    3.1. Hipótesis simples y compuestas: notación formal
    3.2. Prueba z y prueba t (una y dos muestras)
    3.3. Pruebas para proporciones y varianzas
    3.4. Ejemplos con funciones de R (t.test(), prop.test(), var.test())
4. Distribuciones
    4.1. Distribución empírica
    4.2. Prueba de bondad de ajuste
    4.3. Q-Q Plot

1. Estimación puntual

La estimación puntual consiste en asignar un único valor numérico a un parámetro poblacional desconocido θ, usando la información proveniente de una muestra aleatoria $X_1, X_2, \ldots, X_n$.

Un estimador puntual $\hat{\theta}$ es una función de los datos que sirve como aproximación al verdadero valor del parámetro.

Propiedades deseables de los estimadores:

- Insesgadez: $\mathbb{E}[\hat{\theta}] = \theta$.
- Consistencia: $\hat{\theta} \xrightarrow{p} \theta$ cuando $n \to \infty$.
- Eficiencia: entre dos estimadores insesgados, preferimos el de menor varianza.
- Suficiencia: un estimador es suficiente si condensa toda la información de la muestra respecto a θ.

En esta sección exploraremos dos de los métodos más usados para obtener estimadores: Método de los Momentos (EMM) y Máxima Verosimilitud (EMV), comparando sus ventajas y limitaciones con ejemplos en R.

1.1. Métodos de los Momentos (EMM)

El Método de los Momentos se basa en igualar los momentos poblacionales teóricos con los momentos muestrales observados.

Definición

- El k-ésimo momento poblacional está definido como:

$$\mu'_k = \mathbb{E}[X^k]$$

- El k-ésimo momento muestral es:

$$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

  El EMM propone resolver el sistema: $m_k = \mu'_k(\theta), \quad k = 1, 2, \ldots, p$ para encontrar $\hat{\theta}$, donde p es el número de parámetros a estimar.

Ejemplo 1: Distribución Bernoulli

- Poblacional: $\mathbb{E}[X] = p$.
- Muestral: $m_1 = \bar{X}$.
- Igualando: $\tilde{p} = \bar{X}$.

```
set.seed(123)
x <- rbinom(50, size = 1, prob = 0.3)  # muestra
(p_EMM <- mean(x))   # estimador EMM de p
```

```
## [1] 0.3
```

Ejemplo 2: Distribución Exponencial

- Poblacional: $\mathbb{E}[X] = 1/\lambda$.
- Muestral: $m_1 = \bar{X}$.
- Igualando: $\tilde{\lambda} = \frac{1}{\bar{X}}$

```
set.seed(123)
x <- rexp(50, rate = 2)  # lambda = 2
(lambda_EMM <- 1/mean(x))
```

```
## [1] 1.769331
```

Ejemplo 3: Distribución Normal

- Momentos poblacionales:
$$\mu_1' = \mathbb{E}[X] = \mu, \qquad \mu_2' = \mathbb{E}[X^2] = \sigma^2 + \mu^2$$
- Momentos muestrales:
$$m_1 = \bar{X}, \qquad m_2 = \frac{1}{n}\sum X_i^2$$

  Resolviendo el sistema:
$$\tilde{\mu} = \bar{X}, \qquad \tilde{\sigma}^2 = m_2 - (\bar{X})^2$$

```
set.seed(123)
x <- rnorm(50, mean = 5, sd = 2)
mu_EMM <- mean(x)
sigma2_EMM <- mean(x^2) - mu_EMM^2
c(mu_EMM, sigma2_EMM)
```

```
## [1] 5.068807 3.360362
```

1.2. Estimadores de Máxima Verosimilitud (EMV)

La Máxima Verosimilitud (EMV) es uno de los métodos más usados en estadística para estimar parámetros, porque en la práctica suele producir estimadores con buenas propiedades asintóticas (consistencia, normalidad asintótica y eficiencia).

Definición general

Sea $X_1, X_2, \ldots, X_n$ una muestra i.i.d. con función de densidad (o masa) $f(x|\theta)$, donde θ es el vector de parámetros desconocidos.

La función de verosimilitud se define como:

$L(\theta; \vec{x}) = \prod_{i=1}^{n} f(x_i|\theta)$

En la práctica se trabaja con el logaritmo de la verosimilitud:

$\ell(\theta; \vec{x}) = \ln[L(\theta; \vec{x})] = \sum_{i=1}^{n} \ln[f(x_i|\theta)]$

El estimador de máxima verosimilitud es:

$\hat{\theta} = \arg\{\sup\{L(\theta; \vec{x}\}\}$

Nota: Si $\hat{\theta}(x)$ maximiza la función de verosimilitud L, por monotonía creciente de la función logaritmo $\hat{\theta}$ también maximiza la función log de verosimilitud ℓ.

Ejemplo 1: Bernoulli

Verosimilitud:

$$L(p) = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

Log-verosimilitud:

$$\ell(p) = \sum x_i \ln p + (n - \sum x_i) \ln(1-p)$$

Maximizando se obtiene:

$$\hat{p} = \bar{X}$$

```
set.seed(123)
x <- rbinom(50, size = 1, prob = 0.3)
(p_EMV <- mean(x))  # EMV de p
```

```
## [1] 0.3
```

Ejemplo 2: Exponencial

Densidad:

$$f(x|\lambda) = \lambda e^{-\lambda x},\ x > 0.$$

Log-verosimilitud:

$$\ell(\lambda) = n \ln \lambda - \lambda \sum x_i$$

Derivando y maximizando:

$$\hat{\lambda} = \frac{n}{\sum x_i} = \frac{1}{\bar{X}}$$

```
set.seed(123)
x <- rexp(50, rate = 2)
(lambda_EMV <- 1/mean(x))  # EMV
```

```
## [1] 1.769331
```

Ejemplo 3: Normal

Densidad:

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Log-verosimilitud:

$$\ell(\mu,\sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum(x_i-\mu)^2$$

Maximizando:

$$\hat{\mu} = \bar{X}, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum(x_i-\bar{X})^2$$

```
set.seed(123)
x <- rnorm(50, mean = 5, sd = 2)
mu_EMV <- mean(x)
sigma2_EMV <- mean((x - mu_EMV)^2)  # divisor n, no (n-1)
c(mu_EMV, sigma2_EMV)
```

```
## [1] 5.068807 3.360362
```

1.3 Comparación entre EMM y EMV:

El Método de los Momentos (EMM) y el Método de Máxima Verosimilitud (EMV) son dos enfoques distintos para estimar parámetros. Aunque en muchos casos producen el mismo estimador, en general presentan diferencias tanto en la forma como en las propiedades estadísticas de los estimadores obtenidos.

```
resultados <- data.frame(
  Distribucion = c("Bernoulli", "Exponencial", "Normal"),
  Parametro_Real = c("p= 0.3", "lambda=2", "mu=5"),
  EMM = c(p_EMM, lambda_EMM, mu_EMM),
  EMV = c(p_EMV, lambda_EMV, mu_EMV)
)

# Agregamos sigma^2 para Normal como fila adicional con rbind()
resultados <- rbind(
  resultados,
  data.frame(Distribucion = "Normal", Parametro_Real = "sigma^2 = 4",
             EMM = sigma2_EMM, EMV = sigma2_EMV)
)

resultados
```

```
##   Distribucion Parametro_Real      EMM      EMV
## 1    Bernoulli         p= 0.3 0.300000 0.300000
## 2  Exponencial       lambda=2 1.769331 1.769331
## 3       Normal           mu=5 5.068807 5.068807
## 4       Normal    sigma^2 = 4 3.360362 3.360362
```

Nótese como coinciden ambos estimadores, esto pues la respuesta de ambos estimadores es la misma teóricamente, pero ahora veamos un ejemplo de una distribución en la que los metodos no coinciden:

Ejemplo:

```r
set.seed(123)
theta <- 5
n  <- 20
x  <- runif(n, 0, theta)

theta_EMM   <- 2 * mean(x)        # Método de los Momentos
theta_EMV   <- max(x)             # Máxima Verosimilitud

c(EMM = theta_EMM, EMV = theta_EMV,  Verdadero = theta)
```

```
##       EMM       EMV Verdadero
##  5.508084  4.784167  5.000000
```

Tabla de resumen:

Distribución	**Parámetro(s)**	**EMM**	**EMV**
Bernoulli	p	$\bar{X}$	$\bar{X}$
Binomial(n, p)	p	$\bar{X}/n$	$\bar{X}/n$
Geométrica	p	$1/\bar{X}$	$1/\bar{X}$
Binomial Negativa(r, p)	p	$r/(\bar{X}+r)$	$r/(\bar{X}+r)$
Poisson	λ	$\bar{X}$	$\bar{X}$
Exponencial	λ	$1/\bar{X}$	$1/\bar{X}$
Normal	μ, σ^2	$\bar{X},\ m_2 - \bar{X}^2$	$\bar{X},\ \frac{1}{n}\sum(x_i - \bar{X})^2$
Uniforme$(0, \theta)$	θ	$2\bar{X}$	$\max(X_i)$
Gamma(α, β)	α, β	Ecuaciones con $\bar{X}, s^2$	Optimización numérica
Weibull(k, λ)	k, λ	Sistema con momentos	Optimización numérica
Lognormal(μ, σ^2)	μ, σ^2	A partir de $\ln X$: media y var	Optimización o transformación
Beta(α, β)	α, β	Sistema por momentos (no cerrado)	Optimización numérica

Cuadro 1: Comparación entre EMM y EMV

2. Intervalos de confianza

En la práctica, un estimador puntual no siempre es suficiente, ya que un solo número no refleja la incertidumbre asociada al muestreo. Por eso se construyen los intervalos de confianza (IC), que proveen un rango de valores plausibles para un parámetro desconocido θ, con un nivel de confianza $1-\alpha$.

Formalmente, un intervalo de confianza al nivel $1-\alpha$ es un par de estadísticos $[L(X), U(X)]$ tales que:

$$\Pr\big(L(X) \leq \theta \leq U(X)\big) = 1-\alpha$$

El nivel de confianza más usado es el 95% ($\alpha = 0.05$), aunque también son comunes el 90% y el 99%.

Los intervalos dependen de la distribución muestral del estimador, algunos criterios son:

- Si la varianza poblacional es conocida y la muestra es grande, se usa la distribución Normal.
- Si la varianza es desconocida y el tamaño de muestra es pequeño, se usa la distribución t de Student.
- Para proporciones, se usa la aproximación normal a la Binomial.

2.1. Intervalos para la media y la proporción

2.1.1. Intervalo para la media (casos clásicos)

1. Varianza conocida (σ^2 conocida):
$$IC_{1-\alpha}: \quad \bar{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

2. Varianza desconocida (σ^2 desconocida):
$$IC_{1-\alpha}: \quad \bar{X} \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}$$

donde s es la desviación estándar muestral y $t_{\alpha/2,n-1}$ es el cuantil de la t de Student con $n-1$ grados de libertad.

Ejemplo en R (media):

```r
set.seed(123)
x <- rnorm(30, mean = 50, sd = 10)  # muestra pequeña

n <- length(x)
xbar <- mean(x)
s <- sd(x)

alpha <- 0.05
t_val <- qt(1 - alpha/2, df = n-1)

IC_media <- c(
  xbar - t_val * s/sqrt(n),
  xbar + t_val * s/sqrt(n)
)
IC_media
```

```
## [1] 45.86573 53.19219
```

2.1.2. Intervalo para una proporción

Si $X \sim Binomial(n, p)$, el estimador puntual es $\hat{p} = X/n$. Usando la aproximación normal:

$$IC_{1-\alpha}: \quad \hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Ejemplo en R (proporción):

```
set.seed(123)
n <- 200
x <- rbinom(1, size = n, prob = 0.4)  # número de éxitos
p_hat <- x/n

alpha <- 0.05
z_val <- qnorm(1 - alpha/2)

IC_prop <- c(
  p_hat - z_val * sqrt(p_hat*(1-p_hat)/n),
  p_hat + z_val * sqrt(p_hat*(1-p_hat)/n)
)
IC_prop
```

```
## [1] 0.3175626 0.4524374
```

2.2. Intervalos para la varianza

Sea $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$ i.i.d.

El estadístico:

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

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

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

es la varianza muestral.

A partir de las colas de la χ^2:

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

donde $\chi^2_{q,\, n-1}$ es el cuantil q de la distribución Chi–cuadrado con $n-1$ g.l.

```
set.seed(123)
x <- rnorm(25, mean = 100, sd = 15)  # muestra normal con sigma=15

n <- length(x)
s2 <- var(x)  # varianza muestral
alpha <- 0.05

# Cuantiles chi-cuadrado
chi_inf <- qchisq(1 - alpha/2, df = n-1)
```

```r
chi_sup <- qchisq(alpha/2, df = n-1)

# IC para varianza
IC_var <- c((n-1)*s2/chi_inf, (n-1)*s2/chi_sup)

# IC para desviación estándar
IC_sd <- sqrt(IC_var)

IC_var
```

```
## [1] 122.9556 390.2890
```

```r
IC_sd
```

```
## [1] 11.08853 19.75573
```

3. Pruebas de hipótesis

En estadística inferencial, muchas veces queremos tomar decisiones acerca de un parámetro poblacional a partir de una muestra. Para esto se plantean hipótesis estadísticas, que son afirmaciones verificables sobre el valor de un parámetro o sobre la forma de una distribución.

El procedimiento general de una prueba de hipótesis es:

1. Plantear hipótesis nula (H_0) y alternativa (H_1):
   - H_0: afirmación inicial o "status quo".
   - H_1: afirmación alternativanque se busca contrastar.
2. Elegir un estadístico de prueba cuya distribución bajo H_0 sea conocida (ejemplo: Z, t, χ^2, etc.).
3. Definir la región crítica o valor–p:
   - Si el valor observado del estadístico cae en la región crítica, se rechaza H_0.
   - Si no, se falla en rechazar H_0.
4. Nivel de significancia (α): probabilidad máxima tolerada de cometer un error tipo I (rechazar H_0 siendo verdadera). Valores comunes: $\alpha = 0.05$, $\alpha = 0.01$.
5. Interpretación: la prueba no "acepta" H_0, solo la mantiene o la rechaza.

3.1. Hipótesis simples y compuestas: notación formal

- Una hipótesis simple especifica completamente la distribución de los datos. Ejemplo:

$$H_0 : \mu = 100 \quad \text{en } X \sim N(\mu, \sigma^2 = 25)$$

  Aquí H_0 fija el valor de μ y se conoce la varianza, por lo tanto no queda parámetro libre.

- Una hipótesis compuesta deja parámetros sin fijar. Ejemplo:

$$H_0 : \mu = 100 \quad \text{en } X \sim N(\mu, \sigma^2 \text{ desconocida})$$

  Aquí la hipótesis no especifica toda la distribución, pues σ^2 es desconocida.

Tipos de alternativas

- Bilateral (dos colas):

$$H_0 : \theta = \theta_0, \quad H_1 : \theta \neq \theta_0$$

- Unilateral (cola superior):

$$H_0 : \theta \leq \theta_0, \quad H_1 : \theta > \theta_0$$

- Unilateral (cola inferior):

$$H_0 : \theta \geq \theta_0, \quad H_1 : \theta < \theta_0$$

Ejemplo en R: ilustración con media de una Normal

Supongamos que tenemos $X \sim N(\mu, \sigma^2 = 25)$ y queremos probar:

$$H_0 : \mu = 50 \quad \text{vs} \quad H_1 : \mu \neq 50$$

```r
set.seed(123)
x <- rnorm(30, mean = 52, sd = 5)

# Estadístico de prueba
n <- length(x)
xbar <- mean(x)
z <- (xbar - 50) / (5/sqrt(n))  # varianza conocida

alpha <- 0.05
z_crit <- qnorm(1 - alpha/2)

c(Estadistico = z, Region_critica = c(-z_crit, z_crit))
```

```
##     Estadistico Region_critica1 Region_critica2
##        1.932892       -1.959964        1.959964
```

Con un nivel de significancia del 5, el estadístico de prueba fue $z = 1.93$, y como cae dentro del intervalo $(-1.96, 1.96)$, no se rechaza H_0. Esto significa que la diferencia entre la media muestral y el valor hipotético ($\mu = 50$) no es estadísticamente significativa y puede deberse al azar. En la práctica, no es necesario calcular todo "a mano", pues R ya cuenta con funciones como `t.test()` que devuelven directamente el valor–p y el intervalo de confianza.

Ejemplo en R con t.test()

```r
set.seed(123)
x <- rnorm(30, mean = 52, sd = 5)

# Prueba bilateral de H0: mu = 50
t.test(x, mu = 50)
```

```
##
##  One Sample t-test
##
## data:  x
## t = 1.9703, df = 29, p-value = 0.05842
## alternative hypothesis: true mean is not equal to 50
## 95 percent confidence interval:
##  49.93287 53.59610
## sample estimates:
## mean of x
##  51.76448
```

Interpretación: como $p \approx 0.058 > 0.05$, tampoco se rechaza H_0, lo cual coincide con el análisis manual.

3.2. Prueba z y prueba t (una y dos muestras)

Las pruebas z y t se utilizan para contrastar hipótesis sobre medias poblacionales. La elección depende de si la varianza poblacional σ^2 es conocida y del tamaño de la muestra.

3.2.1. Prueba z para una muestra

Se usa cuando:

- Los datos provienen de una Normal o n es grande (TLC).
- La varianza poblacional σ^2 es conocida.

Estadístico:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

Ejemplo en R:

```r
set.seed(123)
x <- rnorm(40, mean = 51, sd = 5)
mu0 <- 50; sigma <- 5; n <- length(x)
z <- (mean(x) - mu0) / (sigma/sqrt(n))
p_value <- 2*(1 - pnorm(abs(z)))
c(Z = z, p_valor = p_value)
```

```
##         Z   p_valor
## 1.5506755 0.1209795
```

3.2.2. Prueba t para una muestra

Se usa cuando:

- La varianza poblacional σ^2 es desconocida.
- Se sustituye σ por la desviación estándar muestral s.

Estadístico:

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

Ejemplo en R:

```r
set.seed(123)
x <- rnorm(25, mean = 52, sd = 5)
t.test(x, mu = 50)  # prueba bilateral
```

```
##
##  One Sample t-test
##
## data:  x
## t = 1.9365, df = 24, p-value = 0.06467
## alternative hypothesis: true mean is not equal to 50
## 95 percent confidence interval:
##  49.87939 53.78731
## sample estimates:
## mean of x
##  51.83335
```

3.2.3. Prueba t para dos muestras (independientes)

Queremos contrastar:

$$H_0 : \mu_1 = \mu_2$$

Estadístico (varianzas iguales):

$$T = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

donde s_p^2 es la varianza combinada.

Ejemplo en R:

```r
set.seed(123)
x1 <- rnorm(20, mean = 50, sd = 5)
x2 <- rnorm(22, mean = 53, sd = 5)

t.test(x1, x2, var.equal = TRUE)   # asume varianzas iguales
```

```
##
##  Two Sample t-test
##
## data:  x1 and x2
## t = -1.3529, df = 40, p-value = 0.1837
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.6231400  0.9156359
## sample estimates:
## mean of x mean of y
##  50.70812  52.56187
```

```r
t.test(x1, x2, var.equal = FALSE)  # Welch (varianzas distintas)
```

```
##
##  Welch Two Sample t-test
##
## data:  x1 and x2
## t = -1.3403, df = 36.96, p-value = 0.1883
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.6562579  0.9487539
## sample estimates:
## mean of x mean of y
##  50.70812  52.56187
```

3.2.4. Prueba t para muestras apareadas

Se aplica cuando los datos provienen de pares dependientes (antes–después, gemelos, etc.). Se reduce a una prueba t para la diferencia $D = X_1 - X_2$.

Ejemplo en R:

```r
set.seed(123)
before <- rnorm(15, mean = 100, sd = 10)
after  <- before - rnorm(15, mean = 5, sd = 5)

t.test(before, after, paired = TRUE)
```

```
##
##  Paired t-test
##
```

```
## data:  before and after
## t = 2.6707, df = 14, p-value = 0.01827
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  0.7418115 6.7922699
## sample estimates:
## mean difference
##        3.767041
```

Comentarios:

- La prueba z casi no se usa en la práctica, porque rara vez conocemos la varianza.
- La prueba t es la más común: una muestra, dos muestras independientes, o datos apareados.
- En R basta con t.test(), ajustando los argumentos mu, var.equal, paired.

3.3. Pruebas para proporciones y varianzas

3.3.1. Pruebas para proporciones

Sea $X \sim Binomial(n, p)$ y queremos contrastar

$$H_0 : p = p_0 \quad \text{vs} \quad H_1 : p \neq p_0$$

El estimador puntual es $\hat{p} = X/n$. El estadístico (aproximación normal) es:

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

Ejemplo en R (una proporción)

```r
# 200 encuestados, 90 responden "sí"
x <- 90; n <- 200
prop.test(x, n, p = 0.5, alternative = "two.sided", correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  x out of n, null probability 0.5
## X-squared = 2, df = 1, p-value = 0.1573
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.3826407 0.5192438
## sample estimates:
##    p
## 0.45
```

Ejemplo en R (comparación de dos proporciones)

$$H_0 : p_1 = p_2$$

```r
# Grupo 1: 40 éxitos en 100 ensayos
# Grupo 2: 55 éxitos en 120 ensayos
x <- c(40, 55)
n <- c(100, 120)
prop.test(x, n, alternative = "two.sided", correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  x out of n
## X-squared = 0.75649, df = 1, p-value = 0.3844
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.18935609  0.07268943
## sample estimates:
##    prop 1    prop 2
## 0.4000000 0.4583333
```

3.3.2. Pruebas para varianzas

Si los datos provienen de normales, la varianza puede probarse mediante:

1. Una muestra (Chi–cuadrado):

$$H_0 : \sigma^2 = \sigma_0^2$$

El estadístico:

$$Q = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

En R se implementa manualmente con `pchisq()`.

```r
set.seed(123)
x <- rnorm(25, mean = 100, sd = 15)
s2 <- var(x); n <- length(x)
sigma0 <- 15^2

Q <- (n-1)*s2/sigma0
p_val <- 2*min(pchisq(Q, df=n-1), 1-pchisq(Q, df=n-1))
c(Estadistico = Q, p_valor = p_val)
```

```
## Estadistico     p_valor
##  21.5112540   0.7831743
```

2. Dos muestras (prueba F):

$$H_0 : \sigma_1^2 = \sigma_2^2$$

El estadístico:

$$F = \frac{S_1^2}{S_2^2} \sim F_{n_1-1, n_2-1}$$

Ejemplo en R (dos varianzas)

```r
set.seed(123)
x1 <- rnorm(20, mean = 50, sd = 5)
x2 <- rnorm(22, mean = 52, sd = 7)

var.test(x1, x2)
```

```
##
##  F test to compare two variances
##
## data:  x1 and x2
## F = 0.75104, num df = 19, denom df = 21, p-value = 0.5344
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3075013 1.8723129
## sample estimates:
## ratio of variances
##          0.7510425
```

4. Distribuciones y selección de modelos

En esta sección abordaremos temas que van más allá de las distribuciones univariadas clásicas, explorando cómo trabajar con distribuciones empíricas obtenidas directamente de los datos, distribuciones multivariadas (con énfasis en la Normal multivariada) y finalmente estrategias para seleccionar distribuciones adecuadas en función de los datos observados.

Estos temas son especialmente relevantes cuando los supuestos tradicionales no se cumplen y se requiere un enfoque más flexible, ya sea para modelar fenómenos complejos, realizar simulaciones o ajustar modelos probabilísticos a datos reales.

4.1. Distribución empírica y pruebas de bondad de ajuste

Cuando no se conoce la distribución poblacional, una alternativa es trabajar con la distribución empírica derivada de la muestra observada.

Formalmente, la distribución empírica de una muestra $X_1, X_2, \ldots, X_n$ se define como:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i \leq x\}$$

donde $\mathbf{1}\{X_i \leq x\}$ es la función indicadora.

En R, la distribución empírica se implementa con ecdf():

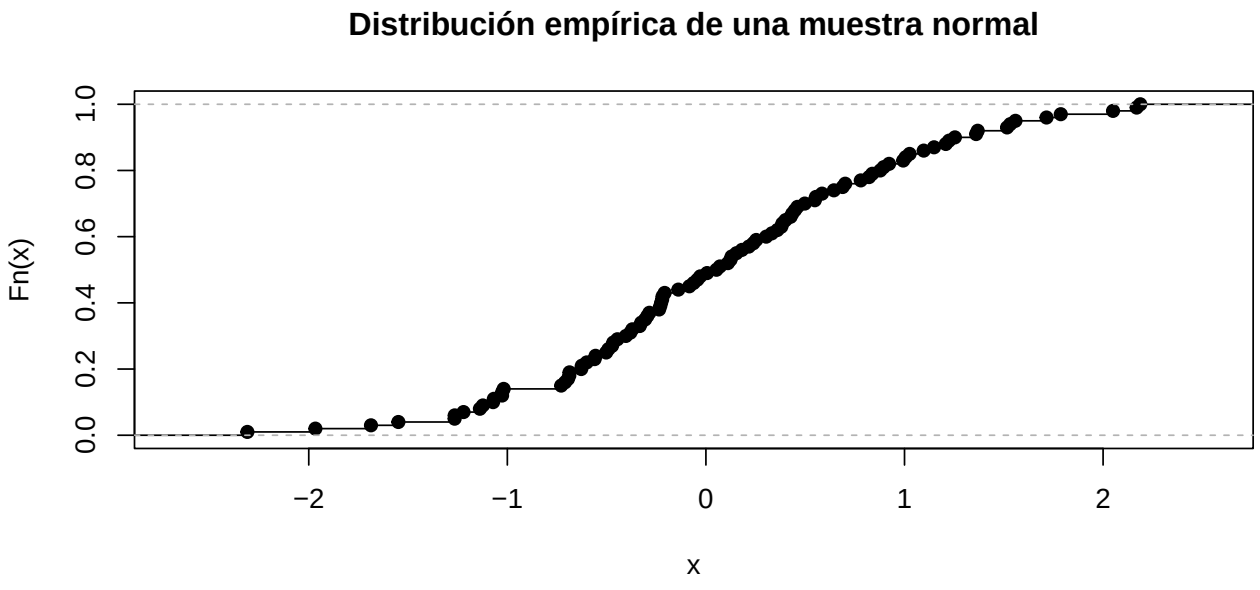
```
set.seed(123)
x <- rnorm(100, mean = 0, sd = 1)

Fn <- ecdf(x)
Fn(0)      # P(X <= 0)
```

```
## [1] 0.48
```

```
Fn(1.5)    # P(X <= 1.5)
```

```
## [1] 0.92
```

```
plot(Fn, main = "Distribución empírica de una muestra normal")
```



Comparación con una distribución teórica

La distribución empírica también sirve para comparar gráficamente contra una CDF teórica.

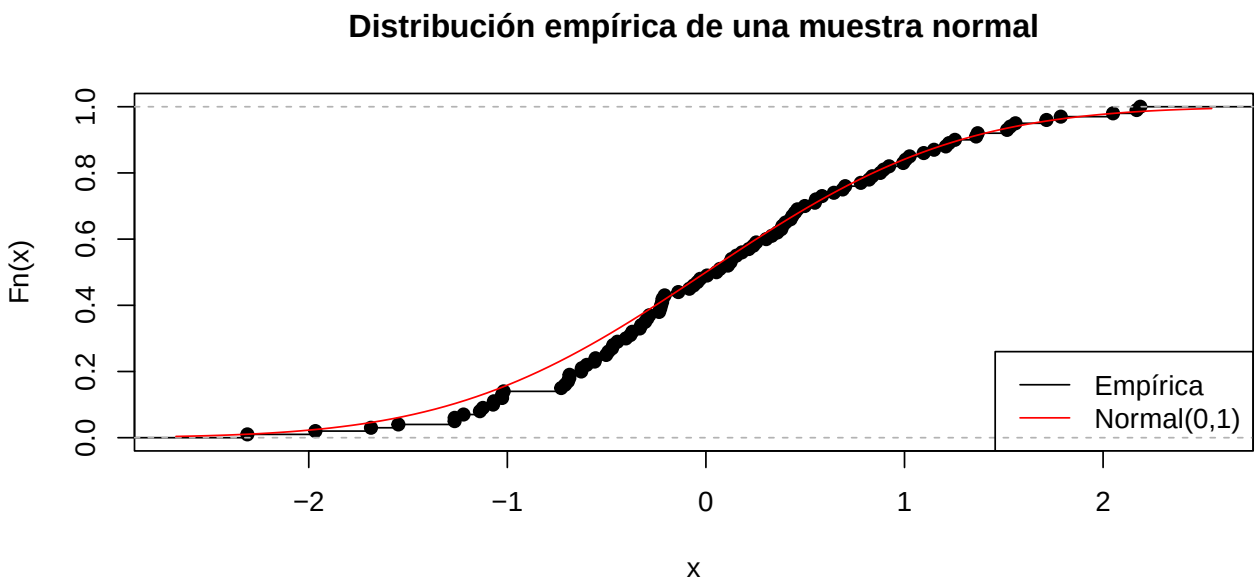
```
set.seed(123)
x <- rnorm(100, mean = 0, sd = 1)

Fn <- ecdf(x)
Fn(0)
```

```
## [1] 0.48
```

```
Fn(1.5)
```

```
## [1] 0.92
```

```
plot(Fn, main = "Distribución empírica de una muestra normal")
# Comparación contra la normal estándar
curve(pnorm(x), col = "red", add = TRUE)
legend("bottomright", legend = c("Empírica", "Normal(0,1)"),
       col = c("black", "red"), lty = 1)
```



4.2. Pruebas de bondad de ajuste

Además de la comparación visual, existen pruebas formales para contrastar si los datos siguen cierta distribución:

- Kolmogorov–Smirnov (KS test): compara la distribución empírica con una distribución teórica.

Hipótesis nula: los datos provienen de la distribución especificada.

```
set.seed(123)
x <- rnorm(50, mean = 0, sd = 1)

ks.test(x, "pnorm", mean = 0, sd = 1)
```

```
##
##  Exact one-sample Kolmogorov-Smirnov test
##
## data:  x
## D = 0.073034, p-value = 0.9347
## alternative hypothesis: two-sided
```

- Shapiro–Wilk test: específicamente para normalidad.

Hipótesis nula: los datos son normales.

```r
shapiro.test(x)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.98928, p-value = 0.9279
```

- Chi–cuadrado de bondad de ajuste: compara frecuencias observadas vs esperadas en clases.

```r
# Simulamos una muestra de Poisson
set.seed(123)
y <- rpois(100, lambda = 3)

# Tabla de frecuencias
obs <- table(factor(y, levels = 0:8))
esp <- dpois(0:8, lambda = 3) * length(y)

chisq.test(obs, p = esp/sum(esp))
```

```
## Warning in chisq.test(obs, p = esp/sum(esp)): Chi-squared approximation may be
## incorrect
```

```
##
##  Chi-squared test for given probabilities
##
## data:  obs
## X-squared = 3.0235, df = 8, p-value = 0.9329
```

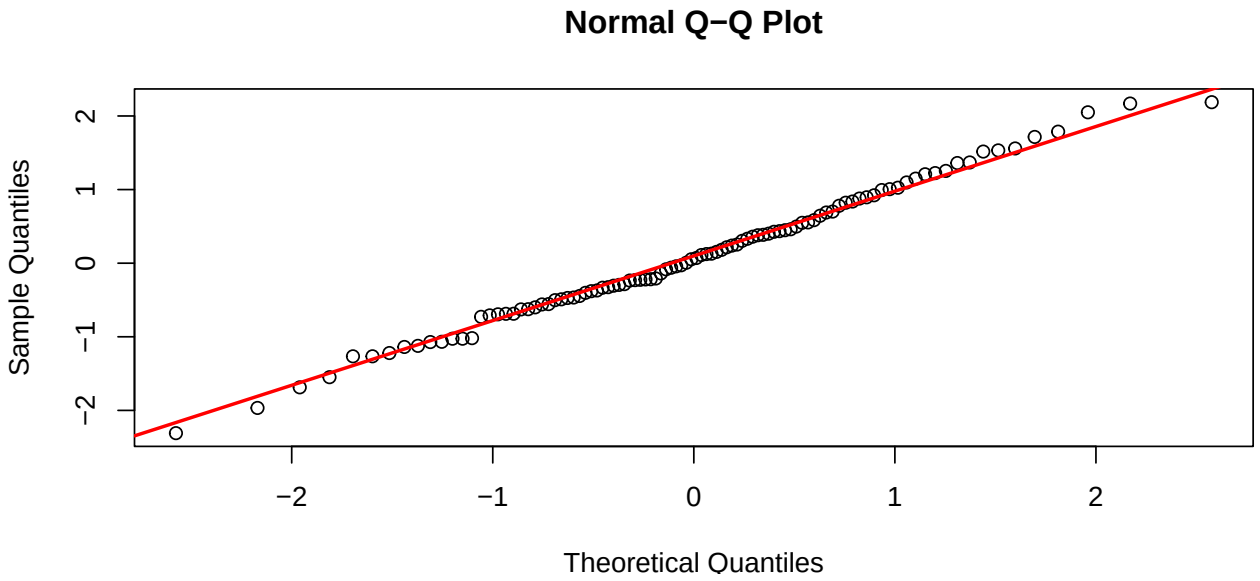
4.3. QQ-plot (Quantile-Quantile plot)

Otra herramienta gráfica muy útil para evaluar si los datos siguen una distribución teórica es el QQ-plot.

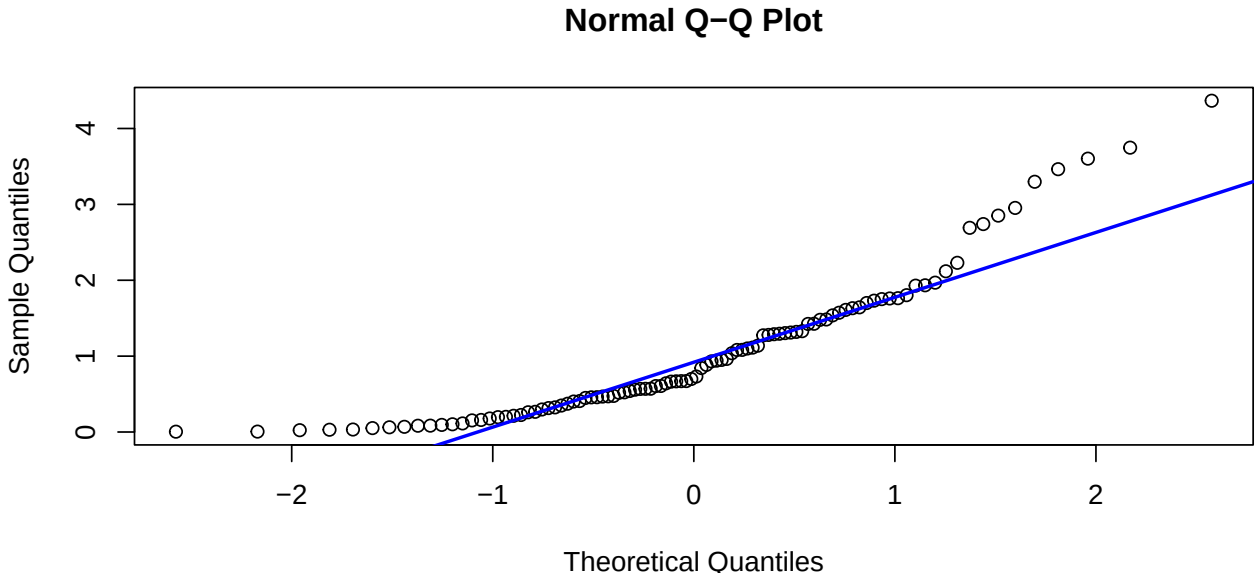
- En un QQ-plot se grafican los cuantiles empíricos de los datos contra los cuantiles teóricos de una distribución.
- Si los datos provienen de la distribución especificada, los puntos deberían alinearse aproximadamente sobre la diagonal de 45°.
- Desviaciones sistemáticas de la diagonal indican que la distribución teórica no describe bien los datos (colas más pesadas, sesgo, etc.).

Ejemplo en R: comparación con la normal estándar

```r
set.seed(123)
x <- rnorm(100, mean = 0, sd = 1)  # muestra normal

# QQ-plot básico
qqnorm(x)
qqline(x, col = "red", lwd = 2)
```




```r
# Comparación contra otra distribución
y <- rexp(100, rate = 1)
qqnorm(y)
qqline(y, col = "blue", lwd = 2)
```



En el primer gráfico, los puntos siguen bien la línea roja (normalidad razonable). En el segundo, los datos exponenciales muestran claras desviaciones en las colas.

Ejemplo con ggplot2 (más estilizado)

```r
library(ggplot2)

x <- rnorm(100, mean = 0, sd = 1)
```

```r
ggplot(data.frame(x), aes(sample = x)) +
  stat_qq(distribution = qnorm) +
  stat_qq_line(distribution = qnorm, color = "red") +
  labs(title = "QQ-plot contra la Normal(0,1)")
```