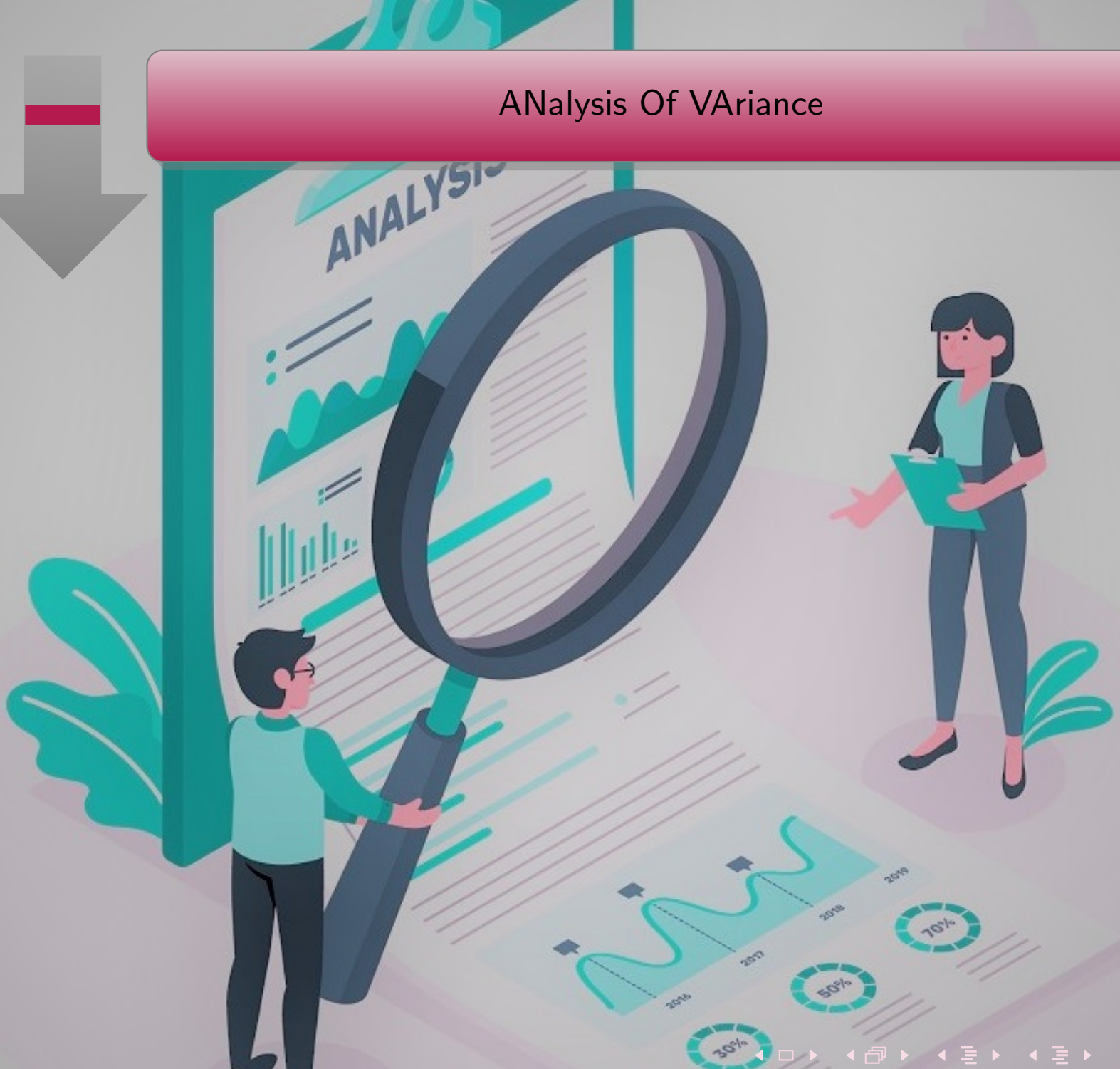
Fundamentos de Análisis de Datos
Análisis de la Varianza (ANOVA)

Dra. Andrea Alejandra Rey

Especialización en Ciencia de Datos - ITBA

ANalysis Of VAriance



TEST DE HIPÓTESIS

Terminología

Estadístico (muestral)

Es una función aplicada a los datos de una muestra y que arroja un valor cuantitativo.

Si $x_1, x_2, \ldots, x_n$ son observaciones recolectadas de una variable aleatoria X, algunos ejemplos de estadísticos son:

- $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i,$
- $s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2,$
- $\max(x_1, x_2, \ldots, x_n).$

Terminología

Parámetro

Es una medida poblacional que se usa en Estadística para crear modelos de la realidad.

Estimador

Es un estadístico usado para estimar un parámetro poblacional desconocido θ, el cual se indica por $\hat{\theta}$.

Prueba o test de hipótesis

- Es una tarea muy utilizada en Estadística, que consiste en evaluar la plausibilidad de una suposición midiendo y examinando una muestra aleatoria de la población que se analiza.
- La metodología empleada por el analista depende de la naturaleza de los datos utilizados y el motivo del análisis.
- El análisis involucra dos hipótesis: la nula versus la alternativa.

Prueba o test de hipótesis

Hipótesis

Afirmación que se desea testear.

Hipótesis nula H_0

Es el hecho aceptado o conjetura, que suele establecer la igualdad entre parámetros de una población.

Hipótesis alternativa H_1

Es la afirmación que establece todo lo contrario a la hipótesis nula.

H_0 y H_1 son mutuamente excluyentes por lo que sólo una de ellas puede ser verdadera.

Prueba o test de hipótesis

Conclusión o decisión

- Rechazar la hipótesis nula
- No rechazar la hipótesis nula

Errores de una prueba

Decisión		H_0 es verdadera	H_0 es falsa
	Rechazar H_0	Error de Tipo I	No hay error
	No rechazar H_0	No hay error	Error de Tipo II

Prueba o test de hipótesis

- Los errores de tipo I suelen identificarse como "falsos positivos".

$$P(\text{Error Tipo I}) = P(\text{Rechazar } H_0 \ / \ H_0 \text{ es Verdadera}) = \alpha$$

- Los errores de tipo II suelen identificarse como "falsos negativos".

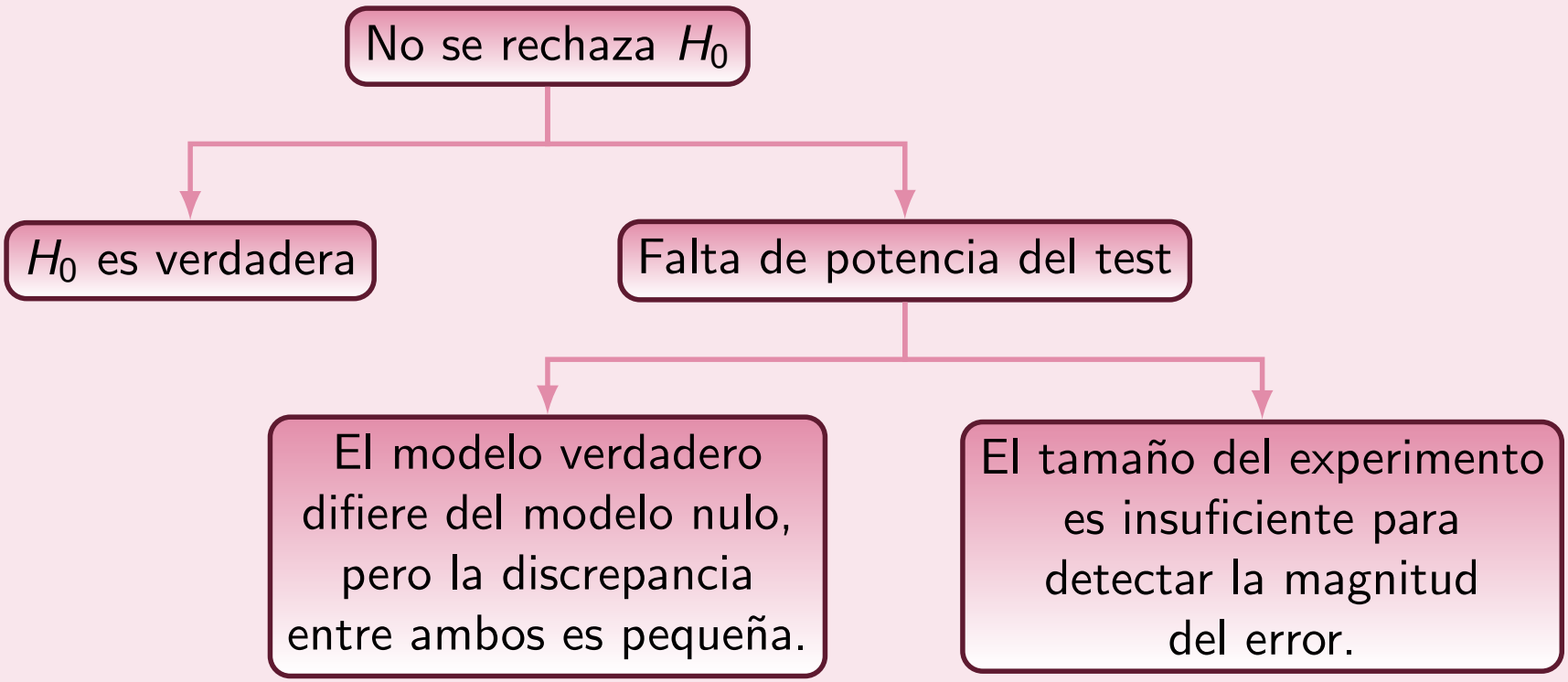
$$P(\text{Error Tipo II}) = P(\text{No Rechazar } H_0 \ / \ H_0 \text{ es Falsa}) = \beta$$

$$\text{Potencia del test} = P(\text{Rechazar } H_0 \ / \ H_0 \text{ es Falsa}) = 1 - \beta$$

Prueba o test de hipótesis

Pregunta

¿Cómo se puede explicar la decisión de no rechazar H_0?



Prueba o test de hipótesis

Pregunta

¿En qué nos basamos para tomar una decisión?

Estadístico de prueba o de contraste

- Es una variable aleatoria de distribución conocida, que vincula a un parámetro de interés con un estimador de ese parámetro.
- Su valor observado cambia aleatoriamente dependiendo de la muestra.
- Contiene información acerca de los datos que es relevante para decidir si se puede rechazar la hipótesis nula.
- Cuando los datos muestran evidencia clara en contra de los supuestos de la hipótesis nula, la magnitud del estadístico toma valores muy extremos.
- Los tests de hipótesis utilizan diferentes estadísticos según el modelo de probabilidad asumido en la hipótesis nula.

Prueba o test de hipótesis

Región de rechazo o región crítica

Corresponde a la región de distribución muestral donde valores del estadístico de contraste se encuentran muy alejados de la afirmación establecida en H_0, de manera tal que es muy poco probable que ocurran si H_0 es verdadera. Su área se denomina **nivel de significación** y coincide con α.

Con el fin de que el error de tipo I sea "aceptable", los valores más utilizados en la práctica son $\alpha = 0.01, 0.05, 0.1$, que representan respectivamente una probabilidad del 1%, 5% y 10% de cometer tal error.

Región de no rechazo

Corresponde a la región de la distribución muestral donde los valores del estadístico de contraste están próximos a la afirmación establecida en H_0. Su área se denomina **nivel de confianza** y coincide con $1 - \alpha$.

Test de hipótesis

Prueba o test de hipótesis

Regla de decisión

Es el criterio utilizado para decidir si la hipótesis nula planteada debe o no ser rechazada.

Sea t_{obs} el valor del estadístico observado. Entonces:

$$t_{\text{obs}} \in \text{ región de rechazo} \Longrightarrow \text{ se rechaza } H_0$$
$$t_{\text{obs}} \in \text{ región de no rechazo} \Longrightarrow \text{ no se rechaza } H_0$$

Prueba o test de hipótesis

p-valor

- Un *p*-valor es una métrica que expresa la probabilidad de determinar si existe evidencia estadística para rechazar la hipótesis nula.
- Al considerar niveles de significación diferentes, la comparación de los resultados de dos pruebas diferentes puede dificultarse. En este sentido, los *p*-valores proporcionan una solución a este problema.
- Por lo general, se considera que un *p*-valor inferior a 0.05 es estadísticamente significativo, en cuyo caso se debe rechazar la hipótesis nula. Por el contrario, un *p*-valor superior a 0.05 significa que la desviación de la hipótesis nula no es estadísticamente significativa y la hipótesis nula no se rechaza.

TEST t

Test t

Se puede usar en las siguientes situaciones.

- Prueba t de una muestra: para comparar la media de un único grupo con un valor conocido.
- Prueba t de dos muestras independientes: para comparar las medias de dos grupos.
- Prueba t de dos muestras pareadas: para comparar las medias de un mismo grupo pero en diferentes momentos.

En el caso del test t para dos muestras independientes, el mismo puede aplicarse bajo ciertas condiciones.

Supuestos

- Independencia
- Normalidad
- Homocedasticidad

Independencia

- Las observaciones deben ser aleatorias.
- Los grupos (niveles del factor) deben de ser independientes.

Normalidad

- La variable cuantitativa debe distribuirse de forma normal en cada uno de los grupos, siendo menos estricta esta condición cuanto mayor sea el tamaño de cada grupo.
- Se recomienda estudiar los residuos de cada observación respecto a la media del grupo al que pertenecen.
- La presencia de outliers puede invalidar por completo las conclusiones de un ANOVA, por lo que se sugiere recalcular el ANOVA descartando aquellas observaciones con residuos extremos.

Pregunta

¿Cómo estudiamos la normalidad?

Gráfico Cuantil-Cuantil (Q-Q)

Se usa para observar la cercanía entre la distribución de un conjunto de datos y alguna distribución ideal o para comparar la distribución de dos conjuntos de datos.

En el caso particular donde se quiere contrastar con una distribución Normal, el gráfico se denomina **Q-Q plot Normal**.

El Q-Q plot tiene dos ejes:

- en el eje vertical se representan los cuantiles de la muestra,
- en el eje horizontal se representan los cuantiles teóricos.

Gráfico Cuantil-Cuantil (Q-Q)

Dada la muestra $X = \{x_1, x_2, \ldots, x_n\}$, se ordenan los datos $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$, para hallar los cuantiles muestrales.

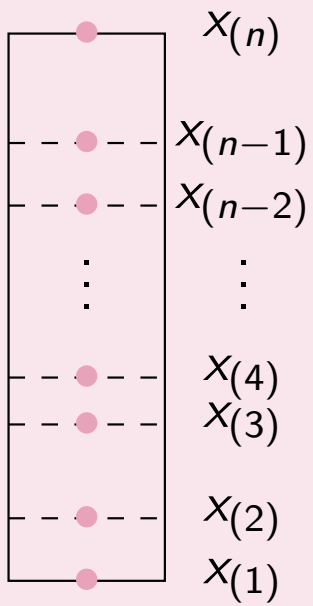
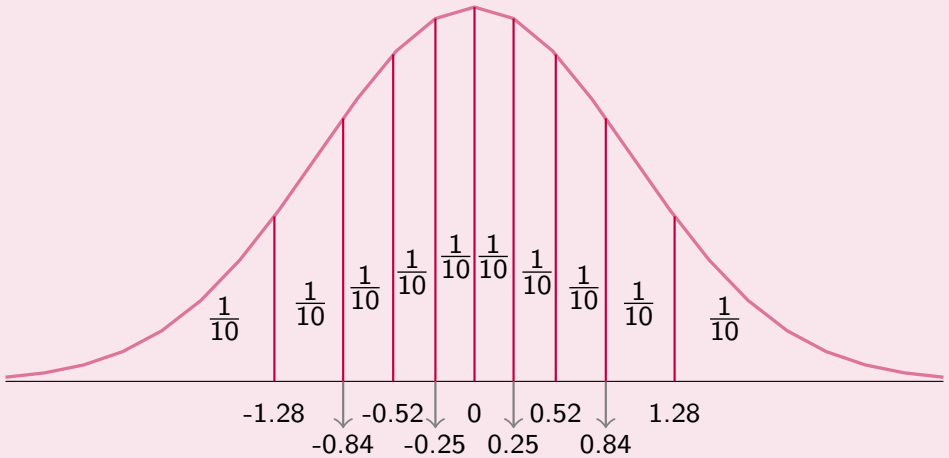


Gráfico Cuantil-Cuantil (Q-Q)

Para hallar los cuantiles teóricos dividimos, mediante rectas verticales, la gráfica de la distribución Normal estándar en $n+1$ regiones de igual área; es decir, cada una mide $1/(n+1)$.

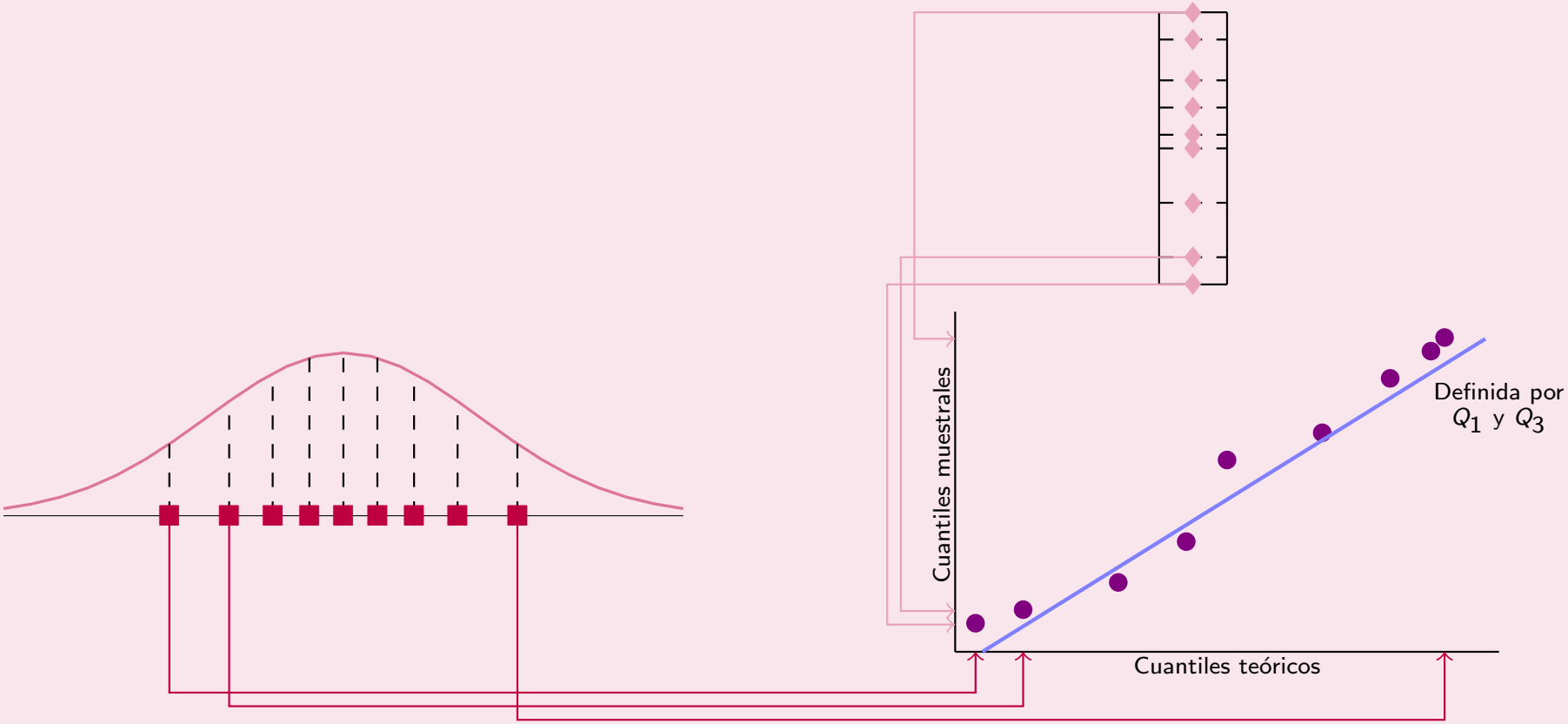
Ilustramos en el caso $n = 9$, donde tenemos 10 particiones de área 0.1.



Finalmente, ubicamos los puntos combinando los cuantiles muestrales y los teóricos.

Gráfico Cuantil-Cuantil (Q-Q)

Nuevamente ilustramos en el caso $n = 9$.



Si las observaciones siguen una distribución Normal aproximada, los puntos en la gráfica se ubicarán aproximados a una línea recta.

Test de Shapiro-Wilks

Se utiliza cuando el tamaño de la muestra es menor a 50 observaciones.

Hipótesis

H_0 : Los datos provienen de una distribución es Normal.

H_1 : Los datos no provienen de una distribución Normal.

Test de Shapiro-Wilks

Estadístico

Dada la muestra $X = \{x_1, x_2, \ldots, x_n\}$, se ordenan los datos $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. El estadístico es:

$$W = \frac{\left[\sum_{i=1}^{h} a_{n-i+1}(x_{(n-i+1)} - x_{(i)})\right]^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2},$$

donde a_{n-i+1} son los coeficientes de Shapiro-Wilks (que están tabulados) y h es la parte entera de $n/2$.

La distribución de este estadístico está tabulada.

Observación

Valores muy pequeños de W indican falta de normalidad. Por el contrario, valores de W cercanos a 1 indican un comportamiento Normal de la muestra.

Test de Lilliefors

Es una prueba de normalidad basada en la test de Kolmogorov-Smirnov, donde este último se usa para determinar la bondad de ajuste de dos distribuciones de probabilidad entre sí.

Se utiliza cuando el tamaño de la muestra es mayor a 50 observaciones.

Hipótesis

H_0 : Los datos provienen de una distribución es Normal.

H_1 : Los datos no provienen de una distribución Normal.

Test de Lilliefors

Estadístico

Dada una muestra $X = \{x_1, x_2, \ldots, x_n\}$, su distribución empírica acumulada para $i = 1, 2, \ldots, n$ es:

$$S(x_i) = \frac{\#\{x \in X : x \leq x_i\}}{n}.$$

Por otra parte, se consideran los valores estandarizados $z_i = (x_i - \bar{X})/s_X$ para $i = 1, 2, \ldots, n$. Usando la distribución Normal estándar, definimos:

$$F^*(x_i) = P(X \leq x_i) = P(Z \leq z_i) = \phi(z_i).$$

El estadístico es:

$$T_1 = \max_{1 \leq i \leq n} |S(x_i) - F^*(x_i)|.$$

La distribución de este estadístico está tabulada.

Test de Lilliefors

Observación

La hipótesis nula debe ser rechazada si el valor observado de T_1 es mayor que el valor crítico.

Homocedasticidad

- La varianza dentro de los grupos debe de ser aproximadamente igual en todos ellos, puesto que la hipótesis nula considera que todas las observaciones proceden de la misma población.
- Esta condición es más importante cuanto menor es el tamaño de los grupos.
- En diseños no equilibrados (el número de observaciones difiere mucho en cada grupo), la falta de homocedasticidad tiene mayor impacto. Si los grupos de menor tamaño son los que presentan mayor desviación estándar, aumentará el número de falsos positivos. Si por el contrario los grupos de mayor tamaño tienen mayor desviación estándar aumentarán los falsos negativos.

Pregunta

¿Cómo estudiamos la homocedasticidad?

Test de Levene

Permite contrastar la homocedasticidad independientemente del número de grupos involucrados en el estudio.

Sea σ_i^2 la varianza de la variable de estudio en el grupo i, para $i = 1, 2, \ldots, k$.

Hipótesis

$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$.

$H_1 :$ Existe al menos un par $i \neq j$ tal que $\sigma_i^2 \neq \sigma_j^2$ significativamente.

Test de Levene

Estadístico

Sea X una muestra de la variable independiente de tamaño N, dividida en k grupos. La cantidad de observaciones del grupo i se indica por n_i para $i = 1, 2, \ldots, k$. Además, X_{ij} denota la j-ésima observación del grupo i.

La media total se denota por $\bar{X}$ y la media del grupo i se indica con $\bar{X}_i$.

El estadístico es:

$$L = \frac{(N-k)\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_j}(X_{ij} - \bar{X}_i)^2} \sim \mathcal{F}_{k-1,N-k}.$$

Test de Levene

Distribución $\mathcal{F}$ de Snedecor

Si X es una variable aleatoria con distribución $\mathcal{F}$ de Snedecor con $\nu_1 > 0$ y $\nu_2 > 0$ grados de libertad, su función de densidad de probabilidad para $x \in [0, +\infty]$[†] es de la forma:

$$f_{\nu_1,\nu_2}(x) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right) \nu_1^{\nu_1/2} \nu_2^{\nu_2/2} x^{\nu_1/2-1}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)(\nu_2+\nu_1 x)^{(\nu_1+\nu_2)/2}},$$

donde Γ denota la función gama definida como $\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t}\, dt$.

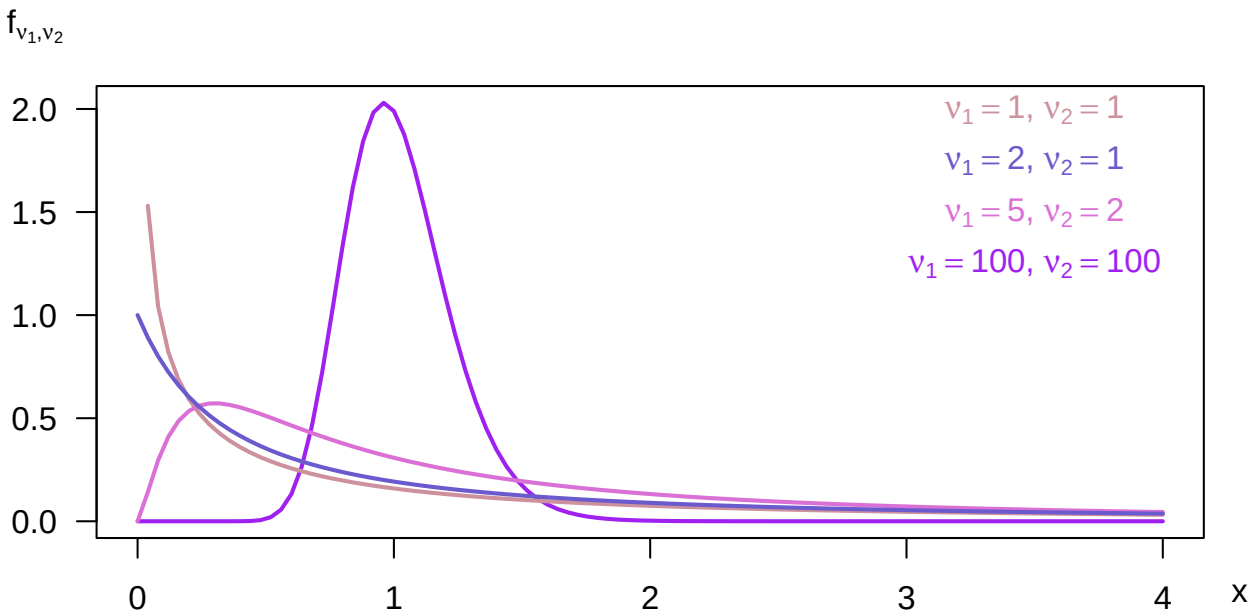
En este caso notamos $X \sim \mathcal{F}_{\nu_1,\nu_2}$.

[†]Si $\nu_1 = 1$ el soporte es $\mathbb{R}^+$.

Test de Levene

Distribución $\mathcal{F}$ de Snedecor

Gráfica de la función de densidad de probabilidad para distintos grados de libertad.



Test de Levene

Observación

La hipótesis nula debe ser rechazada si el valor observado de L es mayor que el valor crítico.

Test t

Sean μ_1 y μ_2 las medias de dos poblaciones diferentes.

Hipótesis

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Sean X_1 y X_2 muestras de cada población de tamaños n_1 y n_2, respectivamente.

Si las varianzas poblacionales son iguales, se define la **desviación estándar combinada** como:

$$s_p = \sqrt{\frac{(n_1 - 1)s_{X_1}^2 + (n_2 - 1)s_{X_2}^2}{n_1 + n_2 - 2}}.$$

Test t

Estadístico

Para varianzas poblacionales iguales:

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

Para varianzas poblacionales diferentes:

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s^2_{X_1}}{n_1} + \frac{s^2_{X_2}}{n_2}}}.$$

En ambos casos los estadísticos siguen una distribución t-Student.

Test t

Distribución t-Student

Si X es una variable aleatoria con distribución t-Student con $\nu > 0$ grados de libertad, su función de densidad de probabilidad para $x \in \mathbb{R}$ es de la forma:

$$f_\nu(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-(\nu+1)/2},$$

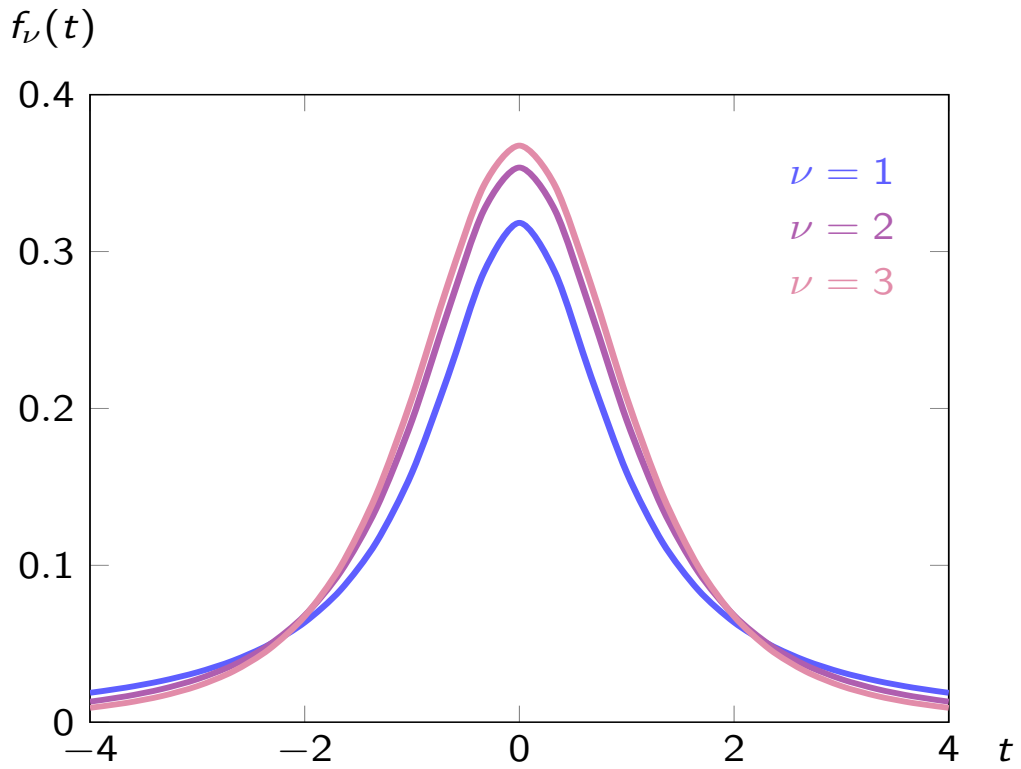
donde Γ denota la función gama definida como $\Gamma(x) = \int_0^{+\infty} t^{x-1}e^{-t}\,dt$.

En este caso notamos $X \sim T_\nu$.

Test t

Distribución t-Student

Gráfica de la función de densidad de probabilidad para distintos grados de libertad.



Test t

Grados de libertad

Para varianzas poblacionales iguales:

$$\text{gl} = n_1 + n_2 - 2.$$

Para varianzas poblacionales diferentes:

$$\text{gl} = \frac{\left(\frac{s_{X_1}^2}{n_1} + \frac{s_{X_2}^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_{X_1}^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_{X_2}^2}{n_2}\right)^2}.$$

Test t

Observación

Un valor absoluto de T grande indica una diferencia significativa entre los grupos. Por el contrario, un valor absoluto de T pequeño indica que los grupos son similares estadísticamente.

Pregunta

¿Qué podemos hacer si no se cumple el supuesto de normalidad?

Test de Kruskal-Wallis

Hipótesis

Son las mismas que las del test t.

Estadístico

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} n_i (\bar{r}_{i\cdot} - \bar{r})^2 \sim \chi^2_{k-1}$$

- k es la cantidad de grupos (por el momento 2),
- n es el número total de observaciones,
- n_i es el número de observaciones en el grupo i,
- r_{ij} es la posición de la observación j en el grupo i al ordenar los datos,
- $\bar{r}_{i\cdot} = \sum_{j=1}^{n_i} r_{ij}/n_i$,
- $\bar{r} = (n+1)/2$ es el promedio de los rankings.

Ejemplo*

*Base de datos `pirates` del paquete `yarrr` de `R`

Pregunta

¿Qué sucede si tenemos más de dos grupos?

ANÁLISIS DE LA VARIANZA (ANOVA)

ANOVA

Ronald Aylmer Fisher (1890 – 1962)

- Es una técnica desarrollada por Fisher en 1930.
- Es un test estadístico empleado cuando se desea comparar las medias de dos o más grupos.

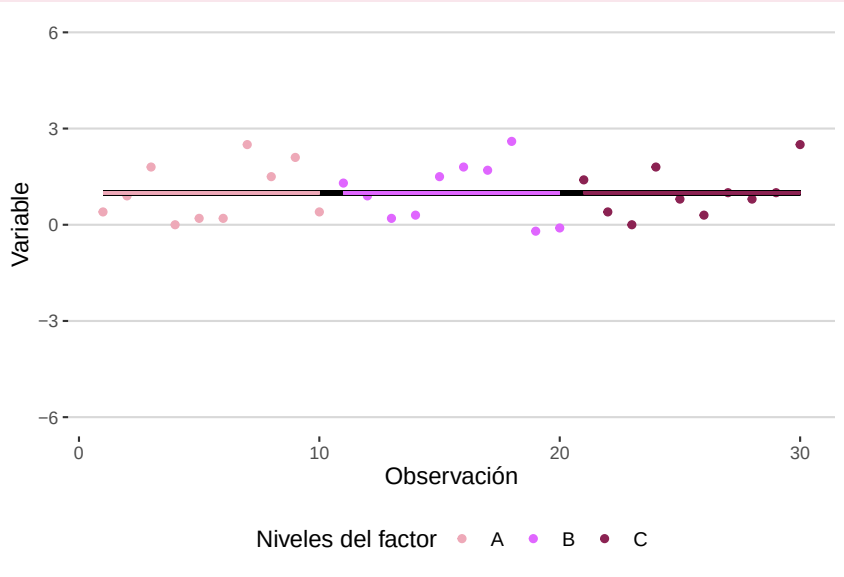
ANOVA

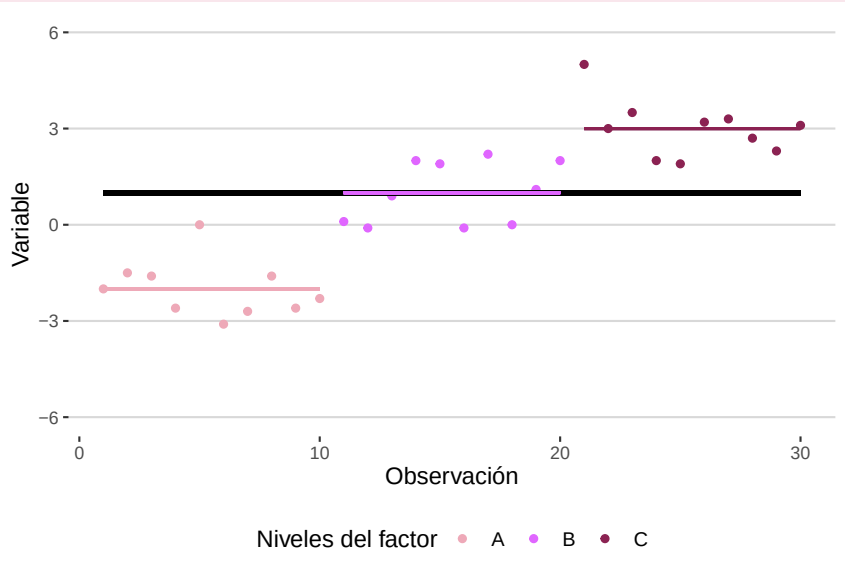
Factor o tratamiento

Variable independiente que afecta a la variable dependiente.

Niveles o grupos

Diferentes valores de la variable independiente que se utilizan en un experimento.





ANOVA

Utilidad del ANOVA

- Ayuda a detectar si la diferencia entre los valores medios es estadísticamente significativa, ya que esta diferencia podría ser consecuencia de un error de muestreo.
- Revela indirectamente si una variable independiente está influyendo en la variable dependiente.
- Permite seleccionar las mejores características para formar un modelo.
- Minimiza el número de variables de entrada para reducir la complejidad del modelo.
- Genera un error de tipo I menor que el obtenido al aplicar el test t múltiples veces cada dos muestras.

ANOVA

Ejemplos de aplicación

- Identificar características importantes para detectar correctamente qué correos electrónicos son *spam* y cuáles no.
- Comparar la producción de dos variedades diferentes de cierto cereal aplicando tres marcas diferentes de fertilizante.
- Comparar la efectividad de varios anuncios publicitarios en las ventas de un producto en particular.
- Comparar el rendimiento de diferentes lubricantes en distintos modelos de autos.
- Comparar la productividad de los trabajadores de una empresa en función del salario percibido y las aptitudes para la tarea desempeñada.

ANOVA SIMPLE
ANOVA DE UN FACTOR
ANOVA DE UNA VÍA

ANOVA de un factor

Se emplea para estudiar si existen diferencias significativas entre las medias de una variable aleatoria continua en los k diferentes niveles de otra variable cualitativa o factor.

El ANOVA puede aplicarse bajo ciertas condiciones.

Supuestos

- Independencia
- Normalidad
- Homocedasticidad

ANOVA de un factor

Sea μ_i la media de la variable de estudio en el grupo i, para $i = 1, 2, \ldots, k$.

Hipótesis

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$.

$H_1 :$ Existe al menos un par $i \neq j$ tal que $\mu_i \neq \mu_j$ significativamente.

El ANOVA requiere de una descomposición de la varianza basada en la siguiente idea.

ANOVA de un factor

Variabilidad total = Variabilidad debida a los diferentes niveles del factor + Variabilidad residual

= Variabilidad explicada por el factor + Variabilidad no explicada por el factor

= Varianza entre grupos + Varianza dentro de los grupos

ANOVA de un factor

Para comprender los cálculos necesitamos de algunas definiciones.

Sea X una muestra de la variable independiente de tamaño N, dividida en k grupos. La cantidad de observaciones del grupo i se indica por n_i para $i = 1, 2, \ldots, k$. Además, X_{ij} denota la j-ésima observación del grupo i.

La media total se denota por $\bar{X}$ y la media del grupo i se indica con $\bar{X}_i$.

Suma de cuadrados entre grupos

La SSB (Sum of Squares Between groups) se calcula como:

$$\text{SSB} = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2.$$

ANOVA de un factor

Suma de cuadrados dentro de grupos o de errores

La SSW (Sum of Squares Within groups) se calcula como:

$$\text{SSW} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2.$$

Suma de cuadrados total

La SST (Total Sum of Squares) se calcula como:

$$\text{SST} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2 = \text{SSB} + \text{SSW}.$$

ANOVA de un factor

Grados de libertad (gl)

- Grados de libertad entre grupos: $\nu_1 = k - 1$.
- Grados de libertad de los errores: $\nu_2 = N - k$.
- Grados de libertad total: $\nu_3 = N - 1$.

Observar que $\nu_1 + \nu_2 = \nu_3$.

ANOVA de un factor

Intervarianza o Mean Square Treatment

Es la varianza causada por el tratamiento y se calcula como:

$$\text{MST} = \frac{\text{SSB}}{\nu_1}.$$

Es la varianza explicada por la variable grupo.

Intravarianza o Mean Square Error

Es la varianza debida a la elección aleatoria y se calcula como:

$$\text{MSE} = \frac{\text{SSW}}{\nu_2}.$$

Es la varianza no explicada por la variable grupo.

ANOVA de un factor

Estadístico

$$F\text{-ratio} = \frac{\text{intervarianza}}{\text{intravarianza}} = \frac{\text{MST}}{\text{MSE}} \sim \mathcal{F}_{k-1,N-k}$$

Observación

Si no existe diferencia significativa entre los grupos testeados, el F-ratio será cercano a 1.

Tabla de ANOVA

Consideremos los siguientes datos:

Observación	Puntaje	Grupo
1	23	A
2	25	A
3	18	A
4	29	B
5	19	B
6	21	B
7	35	C
8	17	C

En este caso, $N = 8$, $k = 3$, $n_1 = n_2 = 3$ y $n_3 = 2$.

Tabla de ANOVA

Las medias son:

- $\bar{X} = (23 + 25 + 18 + 29 + 19 + 21 + 35 + 17)/8 = 23.375$
- $\bar{X}_1 = (23 + 25 + 18)/3 = 22$
- $\bar{X}_2 = (29 + 19 + 21)/3 = 23$
- $\bar{X}_3 = (35 + 17)/2 = 26$

Las sumas de los cuadrados son:

- $\text{SSB} = 3(22-23.375)^2+3(23-23.375)^2+2(26-23.375)^2 = \boxed{19.875}$
- $\text{SSW} = (23 - 22)^2 + (25 - 22)^2 + (18 - 22)^2 + (29 - 23)^2 + (19 - 23)^2 + (21 - 23)^2 + (35 - 26)^2 + (17 - 26)^2 = \boxed{244}$
- $\text{SST} = 19.875 + 244 = \boxed{263.875}$

Tabla de ANOVA

Fuente	Suma de cuadrados	gl	Media de Cuadrados	*F*-ratio
Entre grupos	19.875	2	9.9375	0.2036
Errores	244.000	5	48.8000	
Total	263.875	7		

Referencias:

$$\bullet = \bullet + \bullet$$

$$\bullet = \bullet / \bullet$$

$$\bullet = \bullet / \bullet$$

$$\bullet = \bullet / \bullet$$

Tamaño del efecto η^2

Permite medir cuánto afecta la variable independiente (factor) a la variable dependiente. En otras palabras, representa la cantidad de varianza explicada por la variable independiente.

Se define como:

$$\eta^2 = \frac{\text{SSB}}{\text{SST}}.$$

Los niveles de clasificación más empleados para el tamaño del efecto son:

- pequeño si $\eta^2 = 0.01$,
- mediano si $\eta^2 = 0.06$,
- grande si $\eta^2 = 0.14$.

Resultados de un ANOVA

En la comunicación de los resultados de un ANOVA se debe indicar:

- el valor obtenido para el estadístico F,
- los grados de libertad,
- el p-valor,
- el tamaño del efecto η^2.

Violación del supuesto de normalidad

Observación

A pesar de que el ANOVA es bastante robusto aún cuando se viola el supuesto de normalidad, en casos donde la simetría es muy pronunciada y el tamaño de cada grupo no es muy grande, se puede aplicar el test no paramétrico de Kruskal-Wallis. Sin embargo, es recomendable mantenerse con ANOVA a no ser que la falta de normalidad sea muy extrema.

Violación del supuesto de homocedasticidad

Observación

Si no se puede aceptar la homocedasticidad, se aplica la técnica conocida como ANOVA heterodástico, la cual emplea la corrección de Welch.

Sea X una muestra de la variable independiente de tamaño N, dividida en k grupos $X_1, X_2, \ldots, X_k$ de tamaños $n_1, n_2, \ldots, n_k$, respectivamente. Sean

$$w_i = \frac{n_i}{s_i^2},$$

$$w = \sum_{i=1}^{k} w_i,$$

$$\bar{X}' = \frac{1}{w} \sum_{i=1}^{k} w_i \bar{X}_i.$$

Violación del supuesto de homocedasticidad

Estadístico del F-test de Welch

$$W = \frac{\frac{1}{k-1}\sum_{i=1}^{k} w_i(\bar{X}_i - \bar{X}')^2}{1 + \frac{2(k-2)}{k^2-1}\sum_{i=1}^{k}\frac{1}{n_i-1}\left(1-\frac{w_j}{w}\right)^2} \sim \mathcal{F}_{\mathrm{gl}_n,\mathrm{gl}_d},$$

donde los grados de libertad son:

$$\mathrm{gl}_n = k-1,$$

$$\mathrm{gl}_d = \frac{k^2-1}{3\sum_{i=1}^{k}\frac{(w-w_j)^2}{w^2(n_i-1)}}.$$

Ejemplo*

*Datos simulados

COMPARACIÓN MÚLTIPLE DE MEDIAS

Contrastes POST-HOC

Si al aplicar un ANOVA rechazamos la hipótesis nula, inferimos que al menos dos de las medias comparadas son significativamente distintas entre sí, pero no sabemos cuáles son. Para identificarlas hay que comparar dos a dos las medias de todos los grupos introducidos en el análisis a través de un test t.

Hay que tener en cuenta que cuantas más comparaciones se realicen, mayor será la probabilidad de encontrar diferencias significativas. Por ejemplo, si $\alpha = 0.05$ de cada 100 comparaciones se esperan 5 significativas sólo por azar.

Contrastes POST-HOC

Los niveles de significancia pueden ser ajustados en función del número de comparaciones. Si no se hace ningún tipo de corrección, aumenta el error tipo I. Por el contrario, el hecho de ser muy estricto con las correcciones puede aumentar error tipo II. La necesidad de corrección o no, y de qué tipo, ha de estudiarse con detenimiento en cada caso.

Algunos de los principales métodos de comparación post-hoc son:

- intervalos LSD (Least Significant Difference) de Fisher,
- ajuste de Bonferroni,
- test de Tukey.

Intervalos LSD de Fisher

Consideramos $\bar{X}_i$ la media muestral del grupo i, para $i = 1, 2, \ldots, k$ y N el tamaño de la muestra completa.

Asumiendo la normalidad y la homocedasticidad de los grupos, para un nivel de significación α, se obtiene el intervalo LSD como:

$$\bar{X}_i \pm \frac{\sqrt{2}}{2} t^{\alpha}_{N-k} \sqrt{\frac{\text{SSW}}{N}}.$$

Observación

- Cuanto más se alejen los intervalos de dos grupos más diferentes son sus medias, siendo significativa dicha diferencia si los intervalos no se solapan.
- Se usa para identificar qué grupos tienen las medias más distantes y **no** para determinar significancia.

Ajuste de Bonferroni

- Es un método de comparaciones *a priori* o planeadas.
- Es un método que se utiliza para controlar el nivel de confianza simultáneo para un conjunto completo de intervalos de confianza.
- Ajusta el nivel de confianza para cada intervalo individualmente, de manera tal que el nivel de confianza simultáneo resultante sea igual al valor que se ha especificado.
- Suele ser bastante conservador y se utiliza especialmente cuando las comparaciones a realizar son pocas y los grupos son homogéneos en varianzas.

Ajuste de Bonferroni

La **tasa de error por familia** (TEF) es la máxima probabilidad de que un procedimiento que se componga de más de una comparación concluya de manera incorrecta que al menos una de las diferencias observadas es significativamente diferente de la hipótesis nula.

Si el método realiza h comparaciones, para un nivel de significación α, se verifica que:

$$\text{TEF} = 1 - (1 - \alpha)^h.$$

Ajuste de Bonferroni

Supongamos que realizamos un ANOVA con nivel de significación $\alpha = 0.05$ y que involucra los grupos A, B, C y D.
El método realiza 6 comparaciones:

- A versus B
- A versus C
- A versus D
- B versus C
- B versus D
- C versus D

Luego,

$$\text{TEF} = 1 - (1 - \alpha)^h = 1 - (1 - 0.05)^6 = \boxed{0.265}.$$

Esto implica que al realizar 6 tests, existe un 26.5% de posibilidades de descubrir al menos un falso positivo.

Ajuste de Bonferroni

El ajuste de Bonferroni propone considerar un valor de significación corregido:

$$\alpha_{\text{corregido}} = \frac{\alpha}{h}.$$

Siguiendo con nuestro ejemplo,

$$\alpha_{\text{corregido}} = \frac{0.05}{6} = \boxed{0.008}.$$

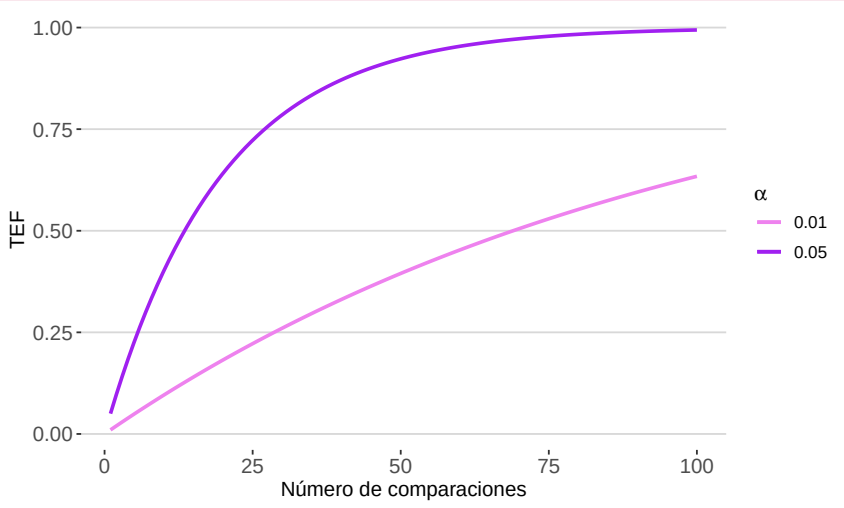
Observemos que en este caso:

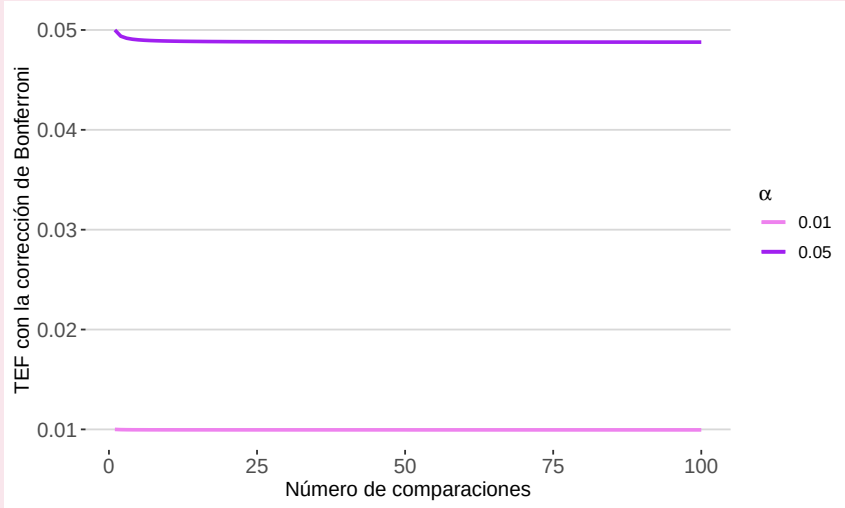
$$\text{TEF} = 1 - (1 - \alpha_{\text{corregido}})^h = 1 - (1 - 0.008)^6 = \boxed{0.047}.$$

Esto implica que hemos reducido las posibilidades de descubrir al menos un falso positivo al 4.7%. Este valor es aún inferior al establecido previamente.

Ajuste de Bonferroni

Variación de la TEF en función de la cantidad de tests con el valor de α original (izquierda) y con la corrección de Bonferroni (derecha).





Test de Tukey

- Es un método de comparaciones *a posteriori* o no planeadas.
- Se usa en experimentos que involucran un número elevado de comparaciones.
- Los cálculos requeridos son sencillos.
- Provee un nivel de significancia global de α cuando los tamaños de las muestras son iguales y de a lo sumo α en caso contrario.
- Se basa en la construcción de intervalos de confianza para todas las diferencias en parejas.

Test de Tukey

Sea n el número de muestras en cada uno de los k grupos.

Se define la Honestly-Significant-Difference como:

$$\text{HSD} = q\sqrt{\frac{\text{MSE}}{n}},$$

donde q se busca en la tabla Studentized Range en función de k y de los grados de libertad de los errores, $kn - k$.
Sólo están tabulados los valores para $\alpha = 0.01$ y $\alpha = 0.05$.

Sean $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k$ las medias de cada grupo.
La diferencia entre las medias de los grupos i y j es significativa si:

$$|\bar{X}_i - \bar{X}_j| > \text{HSD}.$$

Ejemplo*

*Datos simulados

ANOVA FACTORIAL COMPLETO
ANOVA DE DOS FACTORES
ANOVA DE DOS VÍAS

ANOVA factorial completo

Se utiliza cuando existe una variable continua dependiente de dos o más variables independientes, donde cada uno de estos factores puede tener varios niveles.

En el ANOVA factorial completo se utilizan todas las posibles combinaciones de los factores y sus niveles, y no sólo se mide la variable independiente frente a la independiente, sino también si los dos factores se afectan entre sí.

ANOVA de dos factores

Hipótesis

Se realizan tres tests en simultáneo.

Las hipótesis nulas para cada uno de los tests son:

$H_0^{(1)}$: Las medias poblacionales del primer factor son iguales.

$H_0^{(2)}$: Las medias poblacionales del segundo factor son iguales.

$H_0^{(3)}$: No existe interacción entre los dos factores.

$H_0^{(3)}$ es equivalente a decir que el efecto de una variable independiente no depende del efecto de la otra variable independiente.

ANOVA de dos factores

Supuestos

- Independencia
- Normalidad
- Homocedasticidad
- Las variables independientes deben estar en categorías o grupos separados.
- Los grupos de cada tratamiento deben tener la misma cantidad de elementos.

ANOVA de dos factores

Efecto principal

Involucra el tratamiento individual de cada una de las variables independientes, lo que equivale a aplicar el ANOVA de un factor a cada una de ellas.

Efecto de interacción

Es el efecto que una variable independiente tiene sobre la otra. En este caso, los grados de libertad coinciden con el producto de los grados de libertad de cada uno de los factores.

F-test

Por cada uno de los conjuntos de hipótesis existe un F-test.

ANOVA de dos factores

Sea X la variable bajo estudio.
Supongamos que tenemos los siguientes factores divididos por niveles:

$$\text{Factor 1} : A = \{A_1, A_2, \ldots, A_{k_1}\},$$
$$\text{Factor 2} : B = \{B_1, B_2, \ldots, B_{k_2}\},$$

y que n_i es el tamaño muestral de cada nivel del Factor i, para $i = 1, 2$.

Factor 2 / Factor 1	B_1	B_2	$\cdots$	B_{k_2}
A_1	≈	≈		≈
A_2	≈	≈		≈
A_3	≈	≈		≈
$\vdots$				
A_{k_1}	≈	≈		≈

- Si dividimos la muestra por filas, tenemos k_1 grupos.
- Si dividimos la muestra por columnas, tenemos k_2 grupos.
- Si dividimos la muestra por celdas, tenemos $k_1 k_2$ grupos.

ANOVA de dos factores

Podemos calcular las siguientes sumas de cuadrados:

- del Factor 1, notada por SS Factor 1, como la SSB de la división de la muestra por filas.
- del Factor 2, notada por SS Factor 2, como la SSB de la división de la muestra por columnas.
- del error, notada por SSE, como la SSW de la división de la muestra por celdas.
- de interacción, calculada como

$$\text{SS Interacción} = \text{SST} - \text{SS Factor 1} - \text{SS Factor 2} - \text{SSE}.$$

Tabla de ANOVA de dos factores

Fuente	SS	gl	Media de Cuadrados (MS)	*F*-ratio
Efecto principal 1	SS Factor 1	$k_1 - 1$	SS/gl	MS/MSE
Efecto principal 2	SS Factor 2	$k_2 - 1$	SS/gl	MS/MSE
Efecto de interacción	SS Interacción	$(k_1 - 1)(k_2 - 1)$	SS/gl	MS/MSE
Error	SSE	$n_1 k_1 + n_2 k_2 - k_1 k_2$	MSE = SS/gl	
Totales	Suma de la columna	$n_1 k_1 + n_2 k_2 - 1$		

Tamaño del efecto

En el caso del ANOVA de dos factores, se puede calcular el tamaño del efecto η^2 para cada uno de los dos factores, así como para la interacción entre ambos.

Ejemplo*

*Datos simulados

ANOVA CON VARIABLES DEPENDIENTES

ANOVA con variables dependientes

Se aplica cuando las variables a comparar son mediciones distintas pero sobre los mismos sujetos; es decir, no se cumple la condición de independencia.

Se conoce también con los nombres de ANOVA para medidas repetidas o ANOVA para datos pareados.

Las hipótesis H_0 y H_1 son las mismas que para el ANOVA de un factor.

ANOVA con variables dependientes

Supongamos que tenemos k mediciones de n sujetos. En total tenemos $N = kn$ observaciones.

Se calcula la misma tabla que para el ANOVA de un factor sólo que se debe separar:

$$\text{SSW} = \text{SSS} + \text{SSE},$$

donde SSW es la suma de cuadrados dentro de grupos que calculamos para el ANOVA de un factor y SSS es la suma de cuadrados de los sujetos que se calcula como:

$$\text{SSS} = \sum_{s=1}^{n} (\bar{X}_s - \bar{X})^2,$$

donde $\bar{X}_s$ indica la media de la variable observada en el sujeto s.

ANOVA con variables dependientes

Grados de libertad (gl)

- Grados de libertad entre grupos: $\text{gl}_B = k - 1$
- Grados de libertad dentro de los grupos: $\text{gl}_W = N - k$
- Grados de libertad entre sujetos: $\text{gl}_S = n - 1$
- Grados de libertad de los errores: $\text{gl}_E = \text{gl}_W - \text{gl}_S$
- Grados de libertad total: $\text{gl}_T = N - 1$

Estadístico observado

$$F\text{-ratio} = \frac{\text{SSB}/\text{gl}_B}{\text{SSE}/\text{gl}_E}$$

ANOVA con variables dependientes

Supuesto

- **Esfericidad:** la varianza de las diferencias entre todos los pares de variables a comparar es igual.

Pregunta

¿Cómo estudiamos la esfericidad?

Test de Mauchly

Sean $X_1, X_2, \ldots, X_k$ las variables dependientes.

Consideremos la muestra $\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$.

Definición

La **matriz suma de cuadrados y productos cruzados** es la matriz $S = \boldsymbol{X}'\boldsymbol{X} \in \mathbb{R}^{k \times k}$ tal que:

$$S_{ij} = X_i' X_j = \sum_{h=1}^{k} x_{hi} x_{hj}.$$

Observemos que si $i = j$, $S_{ii} = \sum_{h=1}^{k} x_{hi}^2$, con lo cual S tiene la suma de los cuadrados en su diagonal.

Test de Mauchly

Para cada $i = 1, 2, \ldots, k-1$ y cada $j = i+1, i+2, \ldots, k$, la varianza de la diferencia $X_i - X_j$ se denota por σ^2_{ij}.

Hipótesis

$H_0 : \sigma^2_{ij} = \sigma^2_{i'j'} \; \forall (i,j),(i',j') : i, i' = 1, 2, \ldots, k-1;$

$j = i+1, i+2, \ldots, k \wedge j' = i'+1, i'+2, \ldots, k.$

$H_1 :$ Existe al menos dos pares $(i,j) \neq (i',j')$ tales que $\sigma^2_{ij} \neq \sigma^2_{i'j'}$ significativamente.

Test de Mauchly

Sean n la cantidad de observaciones y r el rango de la matriz de diseño (la cantidad de observaciones linealmente independientes).
Sea $A = M'SM$, donde M es una matriz de contraste ortogonal de tamaño $k \times k$.

Estadístico

El estadístico es:

$$W = \frac{|A|k^k}{[\text{tr}(A)]^k},$$

donde $|A|$ indica el determinante de A y $\text{tr}(A)$ su traza (suma de los elementos de la diagonal).

Para $d = k - 1$, consideramos:

$$C = \left(\frac{2d^2 + d + 2}{6d} - n - r\right) \log(W) \sim \chi^2_{\frac{d(d+1)}{2}-1}.$$

Test de Mauchly

Observación

A pesar de ser una técnica muy popular, el test de Mauchly ha sido muy criticado por su inexactitud cuando la normalidad multivariada no puede ser asegurada.

Pregunta

¿Cómo se puede aplicar ANOVA si no se cumple el supuesto la esfericidad?

Veremos cómo modificar los grados de libertad del estadístico de prueba con el cual se contrasta el F-ratio para tomar la decisión del ANOVA.

Índice de Box (1954)

Box sugirió medir la esfericidad mediante un índice $\varepsilon \in [0, 1]$, donde un valor igual a 1 indica que los datos cumplen perfectamente la propiedad de esfericidad.
Supongamos que tenemos k tratamientos.
Sea $S \in \mathbb{R}^{k \times k}$ la matriz de covarianzas poblacional.

Índice de esfericidad

$$\varepsilon = \frac{\left(\sum_{i=1}^{k} S_{ii}\right)^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{k} S_{ij}^2}.$$

En caso de fallar la esfericidad, al aplicar el ANOVA a n registros, tenemos que:

$$F\text{-ratio} \sim \mathcal{F}_{(k-1)\varepsilon,(k-1)(n-1)\varepsilon}.$$

Observar que los grados de libertad deben ser redondeados al entero más próximo.

Corrección de Greenhouse-Geisser (1959)

El cálculo del índice de esfericidad involucra la matriz de covarianzas poblacional, la cual es desconocida en la mayoría de los casos.
Sea $\hat{S}$ la matriz de covarianzas muestral. Definimos lo siguiente:

$$\bar{s}_{i\cdot} = \frac{1}{k}\sum_{j=1}^{k} \hat{S}_{ij},$$

$$\bar{s}_{\cdot\cdot} = \frac{1}{k}\sum_{i=1}^{k} \bar{s}_{i\cdot},$$

$$s_{ij} = \hat{S}_{ij} - \bar{s}_{i\cdot} - \bar{s}_{j\cdot} + \bar{s}_{\cdot\cdot} \text{ para } i, j = 1, 2, \ldots, k.$$

Índice de esfericidad corregido

$$\hat{\varepsilon}_{GG} = \frac{\left(\sum_{i=1}^{k} s_{ii}\right)^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{k} s_{ij}^2}$$

Corrección de Huynh-Feldt (1976)

Si $\hat{\varepsilon}_{GG} \geq 0.75$, se recomienda usar la siguiente corrección menos conservativa; es decir, produce más rechazos.

Índice de esfericidad corregido

$$\hat{\varepsilon}_{HF} = \frac{n(k-1)\hat{\varepsilon}_{GG} - 2}{(k-1)[n-1-(k-1)\hat{\varepsilon}_{GG}]}$$

Ejemplo*

*Datos simulados