Análisis de Datos con el Sistema Estadístico R

Lic. Patricia Vásquez Sotero

Inferencia en poblaciones normales
Análisis de la Varianza

Contenidos

- **Correlación en variables categóricas**
- **Contrastes de hipótesis paramétricos**
  - Inferencia sobre la media en poblaciones normales
    - Prueba para la media en una muestra
    - Prueba para la diferencia de medias en dos poblaciones independientes
    - Prueba para la diferencia de medias en dos poblaciones emparejadas
  - Inferencia sobre la varianza en poblaciones normales
    - Contraste sobre la varianza de una variable con distribución normal
    - Cociente de varianzas de poblaciones normales independientes
    - Cociente de varianzas de poblaciones normales en muestras emparejadas
- **Introducción al Análisis de la Varianza**

CORRELACIÓN EN VARIABLES CATEGÓRICAS

Introducción

- La correlación paramétrica se aplica para casos en donde la distribución de los datos sigue una curva Gausiana o normal.
- Las técnicas de correlación no-paramétrica, son libre de distribución, es decir, no existe la necesidad de que los datos tengan una distribución normal.
- Cuando tenemos variables que son consideradas como variables medidas en una escala ordinal no sería recomendable utilizar el coeficiente de correlación de Pearson.
- En este punto se tratan dos índices que proporciona R para estimar la relación que se establece entre variables de tipo ordinal:
  - ρ de Spearman
  - τ de Kendall

Coeficiente de correlación ρ de Spearman

El coeficiente ρ (rho) de Spearman

- También simbolizado como r_s o conocido como coeficiente de correlación por rangos.
- Es el coeficiente de correlación de Pearson aplicado sobre variables de tipo ordinal, bajo la hipótesis nula de asociación (independencia).

Fórmula

$$r_s = 1 - [6\ \Sigma d_i^2 / (n^3 - n)]$$

Donde, $\mathbf{d_i}$=diferencia de rangos entre los valores de las variables.

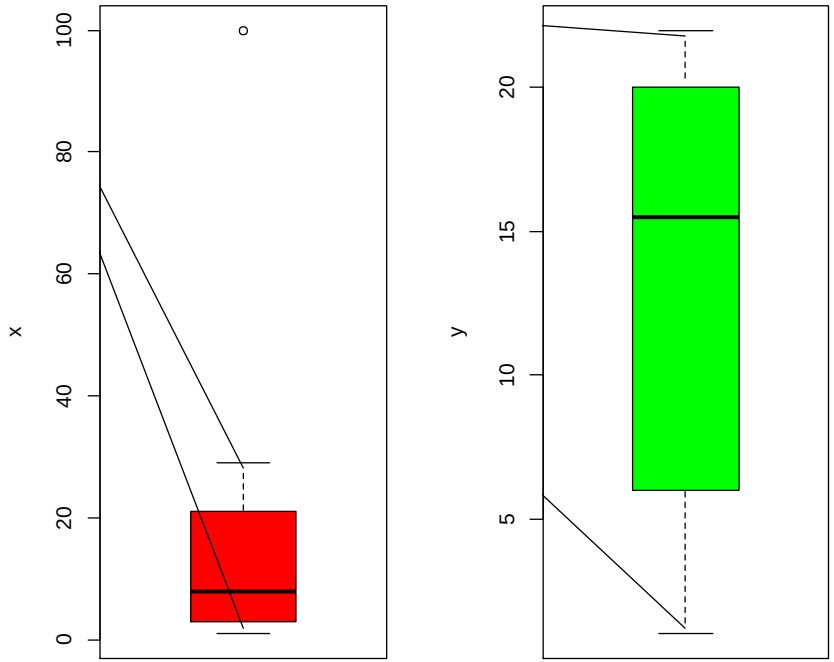
El signo de r_s indica la dirección de la relación y el valor absoluto del r_s indica la fuerza de la relación entre las variables.

Ejemplo en código R

```
x <- c(5,9,17,1,2,21,3,29,7,100)   # Horas de estudio que dedican 10 alumnos
y <- c(6,16,18,1,3,21,7,20,15,22) # Número de respuestas correctas en examen
# Analizar distribución de variables ¿normal?
par(mfrow=c(1,2));
boxplot(x, ylab="x", col="red"); boxplot(y, ylab="y", col="green")
```

Coeficiente de correlación ρ de Spearman

- Se aprecia que una de las dos variables es asimétrica (test de normalidad).
- En este caso utilizamos una medida de correlación no parmétrica.
- Para confirmar esta prueba, contrastamos las hipótesis H_0: “X e Y son independientes” frente a H_1: “Existe una asociación positiva entre ambas variables”, mediante el test de Spearman.



Ejemplo en código R

```
cor(x,y, method="spearman")        # Coeficiente de correlación
cor.test(x,y, method="spearman")   # Contraste de independencia
```

Coeficiente de correlación ρ de Spearman

El resultado es el siguiente,

```
> cor(x,y,method="spearman")
[1] 0.9757576
> cor.test(x,y,method="spearman")

        Spearman's rank correlation rho

data:  x and y
S = 4, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.9757576
```

Decisión: Obtenemos $r_S = 0.976$. Dado que se obtiene un p-valor < 0.05, rechazamos la hipótesis nula.

Conclusión: Por lo tanto, la asociación es significativa y deducimos que ambas variables están fuertemente relacionadas.

Coeficiente de correlación τ de Kendall

El coeficiente τ (tau) de Kendall

- Evalúa la relación que se establece entre variables de tipo ordinal de la concordancia y la discordancia entre ordenaciones de pares de observaciones.
- Indica la diferencia de la probabilidad de que las dos variables estén en el mismo orden menos la probabilidad de que estén en un orden diferente.

Fórmula

$$\tau = (S_a - S_b) / [n(n - 1) / 2]$$

Donde, S_a = Sumatoria de rangos más altos. S_b = Sumatoria de rangos más bajos. La diferencia es la puntuación efectiva de los rangos.

Ejemplo en código R

```
x <- c(5,9,17,1,2,21,3,29,7,100)     # Horas de estudio que dedican 10 alumnos
y <- c(6,16,18,1,3,21,7,20,15,22)    # Número de respuestas correctas en examen
cor(x,y, method="kendall")
cor.test(x,y, method="kendall")
```

Coeficiente de correlación τ de Kendall

El resultado es el siguiente,

```
> cor(x,y,method="kendall")
[1] 0.9111111
> cor.test(x,y,method="kendall")

        Kendall's rank correlation tau

data:  x and y
T = 43, p-value = 2.976e-05
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.9111111
```

Decisión: Obtenemos r_k = 0.911. Dado que se obtiene un p-valor < 0.05, rechazamos la hipótesis nula.

Conclusión: Por lo tanto, la asociación es significativa y deducimos que ambas variables están fuertemente relacionadas. Notar que esta misma conclusión se obtuvo mediante el coeficiente de correlación de Spearman.

INFERENCIA SOBRE LA MEDIA EN POBLACIONES NORMALES

Prueba para la media en una muestra

Análisis de una muestra: Test de la t

- Contrastar la media de una población: obtenemos una muestra
- La distribución de los elementos de la pob. es $X \sim N(\mu, \sigma^2)$
- $\boldsymbol{x} = (x_1, \ldots, x_n)$ es nuestra muestra (**obs. independientes**)
- σ^2 es desconocida

Test

$$\begin{cases} H_0: & \mu = \mu_0 \\ H_A: & \mu \neq \mu_0 \end{cases} \qquad \text{Estad. contraste: } T = \frac{\overline{x} - \mu_0}{\sqrt{var(\boldsymbol{x})/n}} \sim t(n-1)$$

Ejemplo en código R (ambiente.Rdata)

t.test (PH, mu=4)

t.test (SULFATO, mu=4)

Prueba para la media en una muestra

Resultados

R Console

```
> t.test(PH, mu=4)

        One Sample t-test

data:  PH
t = 61.62, df = 299, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 4
95 percent confidence interval:
 5.861713 5.984549
sample estimates:
mean of x
 5.923131
```

mu: Hipótesis nula (aquí se especifica el valor de la media a contrastar)

Regla de decisión: Rechazamos la hipótesis nula con un nivel de significancia del 5%.

```
> t.test(SULFATO, mu=4)

        One Sample t-test

data:  SULFATO
t = 1.2404, df = 299, p-value = 0.2158
alternative hypothesis: true mean is not equal to 4
95 percent confidence interval:
 3.870886 4.569396
sample estimates:
mean of x
 4.220141
```

mu: Hipótesis nula (aquí se especifica el valor de la media a contrastar)

Regla de decisión: No rezamos la hipótesis nula con un nivel de significancia del 5%.

Prueba para la media en una muestra

Desde R Commander

- Importar los datos desde el archivo ambiente.sav para manejar la variable en cuestión.
- A continuación elegimos la opción del menú:

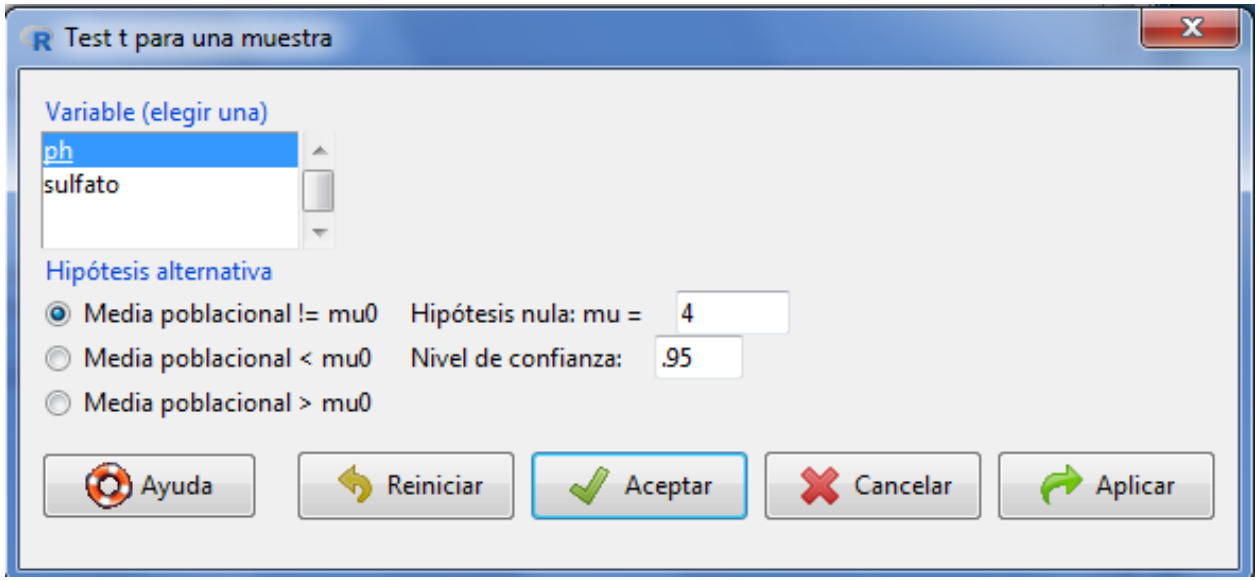
  Estadísticos → Medias → Test t para una muestra

- Esta opción abrirá la ventana que aparece en la Figura que se muestra en la siguiente página. Fijémonos con detalle en ella:
  - Nos pide en primer lugar que elijamos una (sólo una) variable, que debe ser aquella cuya media estemos analizando.
  - Nos pide que indiquemos cuál es la hipótesis alternativa. En nuestro caso hemos elegido la opción de un test bilateral.
  - Nos pide que especifiquemos el valor del valor hipotético con el que estamos comparando la media, en nuestro caso, 4.

Prueba para la media en una muestra

Desde R Commander

- Nos pide, por último, que especifiquemos un nivel de confianza. Este nivel de confianza es para el intervalo de confianza asociado al problema, el contraste se resolverá a través del p-valor. Si el enunciado del problema no dice nada, ponemos la opción habitual del 95%.



Prueba para la media en una muestra

Desde R Commander

El resultado es el siguiente:

```
        One Sample t-test

data:  ph
t = 61.62, df = 299, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 4
95 percent confidence interval:
 5.861713 5.984549
sample estimates:
mean of x
 5.923131
```

Análisis e interpretación:

- Nos informa del valor del estadístico de contraste (t=61.62), de los grados de libertad (df=299) y del p-valor (p-value<2.2e-16). Dado que el p-valor es inferior al 5%, rechazamos la hipótesis nula (μ=4) en favor de la alternativa (μ≠4). Concluimos que con los datos de la muestra tenemos suficientes evidencias de que el PH medio es distinto de 4.

Prueba para la media en una muestra

Desde R Commander

Análisis e interpretación (continuación…)

- Luego, proporciona un intervalo de confianza del 95%, para la media de la distribución normal que se le supone a los datos: 95 percent confidence interval: 5.861713 5.984549. Es decir,

$$P [\mu \in (5.861713, 5.984549)] = 0.95$$

La relación que guarda el intervalo de confianza con el contraste de hipótesis es la siguiente: fijémonos que el valor hipotético que hemos considerado para la media, 4, no está dentro de este intervalo, luego éste no es un valor de confianza para μ.

- Finalmente, proporciona los estadísticos muestrales utilizados, en este caso, la media muestral:

sample estimates: mean of x 5.923131

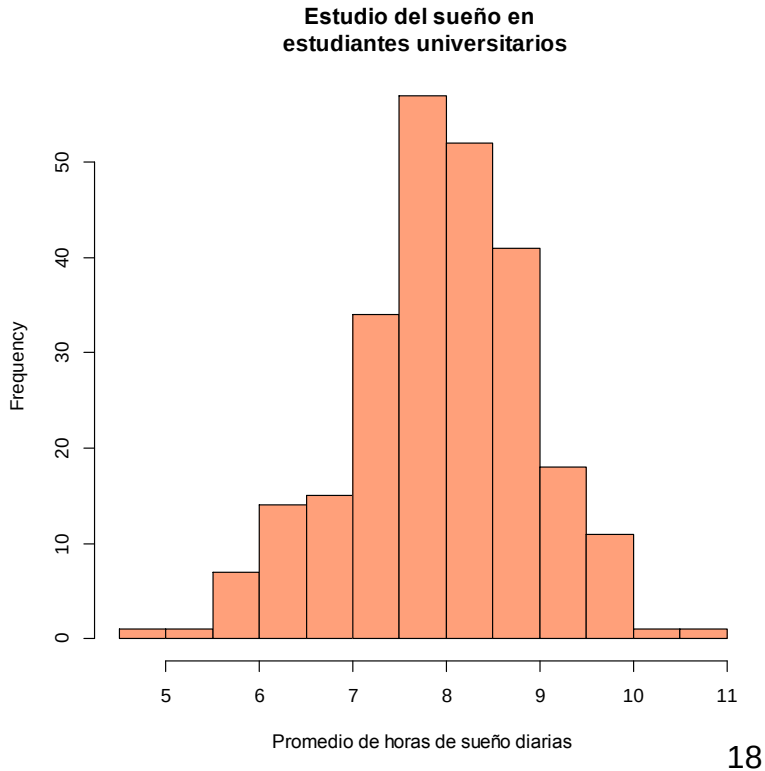
Prueba para la media en una muestra

Ejemplo. Utilizamos el conjunto de datos SleepStudy para contrastar la hipótesis de que los alumnos duermen por término medio 8 horas diarias.

En primer lugar leemos los datos y presentamos un histograma de esta variable

Análisis exploratorio

```
library(Lock5Data)  # Cargamos el
        paquete que contiene los datos
data(SleepStudy)    # Cargamos los
        datos del estudio de sueño
hist(SleepStudy$AverageSleep,
    col="lightSalmon", xlab="Promedio
    de horas de sueño diarias",
    main="Estudio del sueño en \n
    estudiantes universitarios")
```



Prueba para la media en una muestra

Ejemplo. Cont.

Prueba t - una muestra. Contraste bilateral

```
t.test (SleepStudy$AverageSleep, mu=8)
```

```
        One Sample t-test

data:  SleepStudy$AverageSleep
t = -0.56168, df = 252, p-value = 0.5748
alternative hypothesis: true mean is not equal to 8
95 percent confidence interval:
 7.846466 8.085392
sample estimates:
mean of x
 7.965929
```

Obsérvese que para llevar a cabo el contraste basta con especificar la media que se desea poner a prueba mediante mu=8. Como resultado del procedimiento se muestra el valor del estadístico t, sus grados de libertad (df) y el p-valor del contraste (0.57483), que indica que la hipótesis planteada es admisible. Además obtenemos también la estimación del número medio de horas de sueño en la muestra (7.96593) y un intervalo de confianza al 95%.

Prueba para la media en una muestra

Ejemplo. Cont.

Prueba t para una muestra. Contraste bilateral

Podemos solicitar un intervalo a otro nivel de confianza especificándolo en la llamada al t.test

```
t.test(SleepStudy$AverageSleep, mu=8, conf.level=0.9)
```

```
        One Sample t-test

data:  SleepStudy$AverageSleep
t = -0.56168, df = 252, p-value = 0.5748
alternative hypothesis: true mean is not equal to 8
90 percent confidence interval:
 7.865786 8.066072
sample estimates:
mean of x
 7.965929
```

Obsérvese que el intervalo de confianza al 90% contiene al valor de la media de prueba, que indica que la hipótesis planteada es admisible.

Prueba para la media en una muestra

Ejemplo. Cont.

Prueba t para una muestra. Contraste unilateral

Si nuestro planteamiento original hubiese sido determinar si existe evidencia suficiente de que estos estudiantes duermen en promedio más de 7 horas diarias, plantearíamos un test unilateral, especificando el sentido de la hipótesis alternativa a contrastar (en este caso $\mu > 7$)

```
t.test(SleepStudy$AverageSleep,  mu=7, alternative="greater")
```

```
        One Sample t-test

data:  SleepStudy$AverageSleep
t = 15.924, df = 252, p-value < 2.2e-16
alternative hypothesis: true mean is greater than 7
95 percent confidence interval:
 7.865786      Inf
sample estimates:
mean of x
 7.965929
```

Los datos muestran suficiente evidencia para admitir que los alumnos duermen en promedio más de 7 horas, con un nivel de significancia de 5%.

Prueba para la media en una muestra

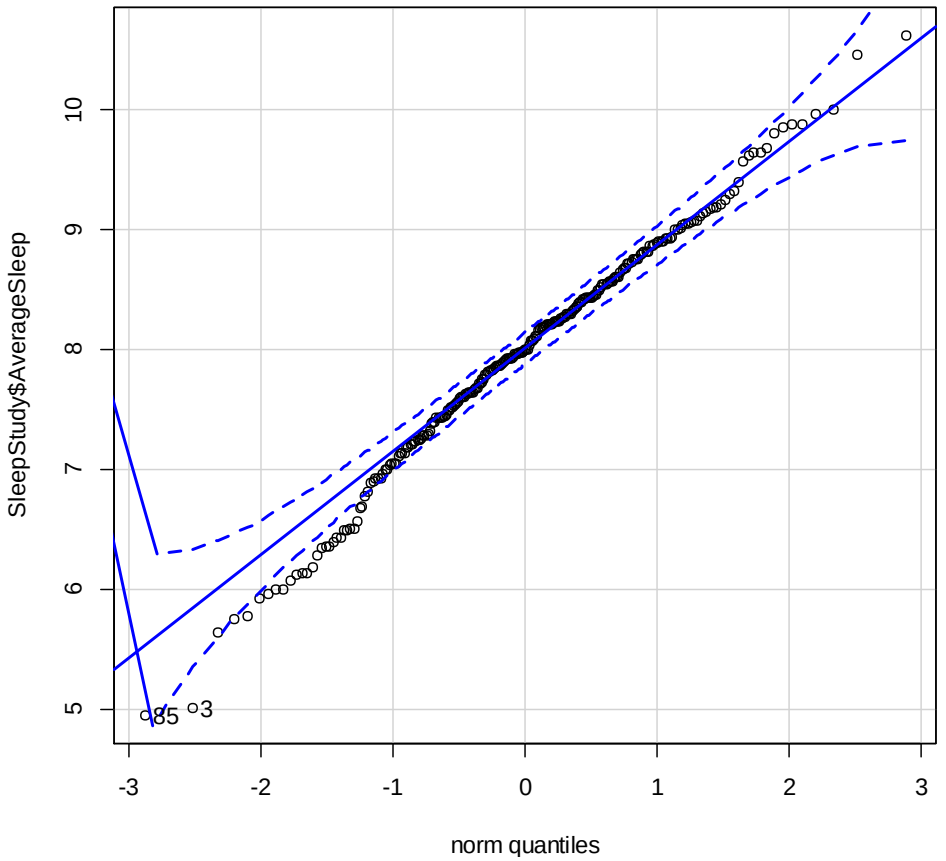
Ejemplo. Cont.

Evaluación gráfica de normalidad

El paquete **car** proporciona la función **qqPlot()** que permite evaluar gráficamente si se admite la hipótesis de normalidad de una variable

```
library(car)
qqPlot(SleepStudy$AverageSleep)
```

En este caso se aprecia una ligera asimetría en la cola inferior de la distribución. No obstante, el test de Shapiro-Wilk permite aceptar la normalidad de esta variable.



Diferencia de medias en dos poblaciones normales independientes

Descripción

- Comparar la media de dos poblaciones cualesquiera
- La distribución de los elementos de la pob. es $X \sim N(\mu_i, \sigma^2), i = 1, 2$
- $\boldsymbol{x_1} = (x_1^1, \ldots, x_{n_1}^1)$ y $\boldsymbol{x_2} = (x_1^2, \ldots, x_{n_2}^2)$, son nuestras muestras
- Se puede usar con datos no normales si la muestra es *grande* $(n > 30)$

Contraste

$$\begin{cases} H_0: & \mu_1 = \mu_2 \\ H_A: & \mu_1 \neq \mu_2 \end{cases}$$

$$T = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{var(\boldsymbol{x_1}, \boldsymbol{x_2})(\frac{1}{n_1} + \frac{1}{n_2})}} \sim t(n_1 + n_2 - 2)$$

$$var(\boldsymbol{x_1}, \boldsymbol{x_2}) = \frac{n_1 - 1}{n_1 + n_2 - 2} var(\boldsymbol{x_1}) + \frac{n_2 - 1}{n_1 + n_2 - 2} var(\boldsymbol{x_2})$$

Ejemplo en código R - Contraste unilateral

```
x <- c(0.80,0.83,1.89,1.04,1.45,1.38,1.91,1.64,0.73,1.46)
y <- c(1.15,0.88,0.90,0.74,1.21)
var.test(x,y); t.test(x, y, alternative="greater", var.equal=TRUE)
```

Diferencia de medias en dos poblaciones normales independientes

Verificamos que el test de igualdad de varianzas es significativo.

El resultado del test t es el siguiente,

```
        Two Sample t-test

data:  x and y
t = 1.6061, df = 13, p-value = 0.06613
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.03457769         Inf
sample estimates:
mean of x mean of y
    1.313     0.976
```

Decisión: La estadística de prueba es t = 1.6061. Dado que se obtiene el p-valor > 0.05, no rechazamos la hipótesis nula.

Conclusión: Los datos proporcionan evidencia estadística de que la diferencia de medias entre las variables de análisis es significativamente menor a cero, con un nivel de significación del 5%.

Diferencia de medias en dos poblaciones normales independientes

Nota: Por defecto, la función t.test() asume que la variable sobre la que se realiza el contraste tiene distinta varianza en los grupos que se comparan.

Ejemplo. Contrastar con los datos SleepStudy si existen diferencias en el promedio de horas de sueño diarias entre hombres y mujeres, asumiendo varianzas distintas.

Ejemplo en código R - Contraste bilateral

```
t.test(AverageSleep~Gender,data=SleepStudy)
```

```
        Welch Two Sample t-test

data:  AverageSleep by Gender
t = 0.58588, df = 227.08, p-value = 0.5585
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1690510  0.3121168
sample estimates:
mean in group 0 mean in group 1
       7.994768        7.923235
```

Como vemos, los datos proporcionan evidencias de que no existen diferencias significativas entre sexos (p-valor = 0.55854).

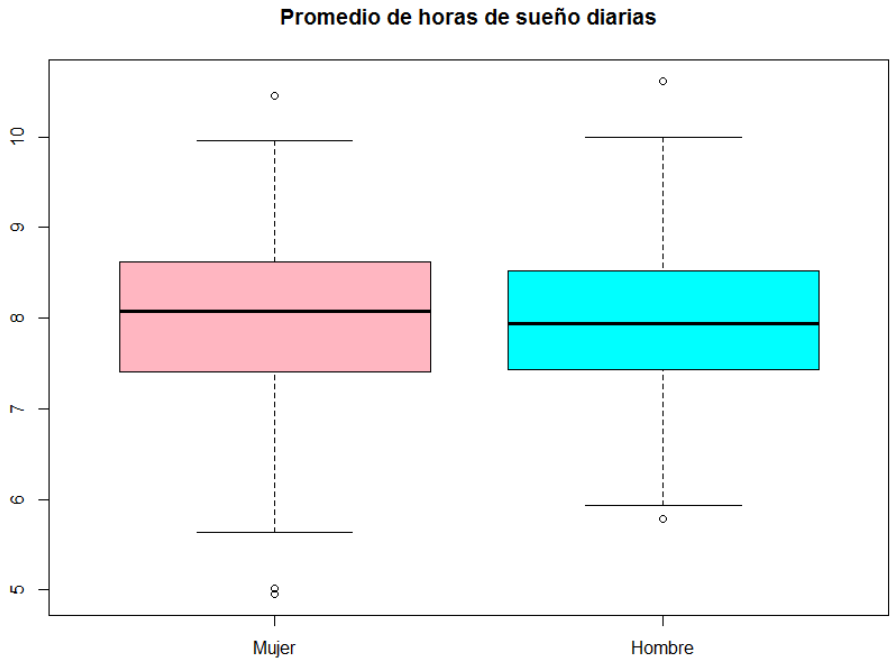
Diferencia de medias en dos poblaciones normales independientes

Ejemplo. Cont.

El boxplot que se presenta a continuación muestra que efectivamente ambos grupos son muy similares.

Ejemplo en código R - Boxplot

```
# Convierte Gender a factor
  SleepStudy$Gender=factor(Sleep
  Study$Gender, levels=0:1,
  labels=c("Mujer", "Hombre"))
boxplot(AverageSleep~Gender,
    data=SleepStudy,
    main="Promedio de horas de
    sueño diarias",
    col=c("lightpink","cyan"))
```



Diferencia de medias en dos poblaciones normales independientes

Ejemplo. Cont.

Para validar la aplicación del test, comprobamos la normalidad en cada grupo:

Ejemplo en código R - Normalidad

```
shapiro.test(SleepStudy$AverageSleep[SleepStudy$Gender=="Hombre"])
shapiro.test(SleepStudy$AverageSleep[SleepStudy$Gender=="Mujer"])
# o, de una manera más sintética
aggregate(AverageSleep~Gender,data=SleepStudy, function(x)
shapiro.test(x)$p.value)
```

```
> shapiro.test(SleepStudy$AverageSleep[SleepStudy$Gender=="Hombre"]

        Shapiro-Wilk normality test

data:  SleepStudy$AverageSleep[SleepStudy$Gender == "Hombre"]
W = 0.98566, p-value = 0.3402

> shapiro.test(SleepStudy$AverageSleep[SleepStudy$Gender=="Mujer"])

        Shapiro-Wilk normality test

data:  SleepStudy$AverageSleep[SleepStudy$Gender == "Mujer"]
W = 0.98607, p-value = 0.1338
```

Diferencia de medias en dos poblaciones normales relacionadas

Descripción

- Comparar la media de dos poblaciones (p.ej., medidas al mismo individuo en dos tiempos distintos)
- La distribución de los elementos de la población es normal
- $\boldsymbol{x_1} = (x_1^1, \ldots, x_n^1)$ y $\boldsymbol{x_2} = (x_1^2, \ldots, x_n^2)$, son nuestras muestras
- Trabajamos con $\boldsymbol{y} = \boldsymbol{x_1} - \boldsymbol{x_2}$ y aplicamos Test t para una muestra

Test

$$\begin{cases} H_0: & \mu_1 - \mu_2 = 0 \\ H_1: & \mu_1 - \mu_2 \neq 0 \end{cases}$$

Estad. contraste: $T = \frac{\bar{y}-0}{\sqrt{var(y)/n}} \sim t(n-1)$

Nota: La hipótesis nula (H_0) establece que no hay diferencia de medias, es decir, que la diferencia entre ellas es igual a 0.

Diferencia de medias en dos poblaciones normales relacionadas

Ejemplo en código R - Contraste unilateral

```
x <- c(1.83, 0.50,1.62,2.48,1.68,1.88,1.55,3.06,1.30)
y <- c(0.878,0.647,0.598,2.05,1.06,1.29,1.06,3.14,1.29)
shapiro.test(x); shapiro.test(y); var.test(x,y)
t.test(x, y, paired=TRUE, alternative="greater")
```

- Evaluación de la normalidad de los datos:

```
        Shapiro-Wilk normality test

data:  x
W = 0.95217, p-value = 0.7139


        Shapiro-Wilk normality test

data:  y
W = 0.81992, p-value = 0.03439
```

Conclusión: De acuerdo a los resultados, se confirma la normalidad de la variable x; pero no sucede lo mismo con la variable y.

Diferencia de medias en dos poblaciones normales relacionadas

- Sin embargo el resultado del test de igualdad de varianzas es el siguiente:

```
        F test to compare two variances

data:  x and y
F = 0.79547, num df = 8, denom df = 8, p-value = 0.754
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1794323 3.5265251
sample estimates:
ratio of variances
           0.79547
```

Conclusión: De acuerdo a los resultados, no rechazamos la hipótesis nula. Por tanto, los datos proporcionan evidencia estadística de que existe igualdad de varianzas entre las variables de análisis (F=0.7955, p-valor=0.754), con un nivel de significación del 5%.

Diferencia de medias en dos poblaciones normales relacionadas

- Los resultados de la prueba t con H_0: $\mu_x - \mu_y \leq 0$, son los siguientes:

```
        Paired t-test

data:  x and y
t = 3.0354, df = 8, p-value = 0.008088
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.1673028       Inf
sample estimates:
mean of the differences
              0.4318889
```

Decisión y Conclusión: De acuerdo a los resultados, rechazamos la hipótesis nula. Por tanto, los datos proporcionan evidencia estadística de que la diferencia de medias entre las variables de análisis es significativamente mayor a cero (t=3.0354, p=0.0081), con un nivel de significación del 5%.

INFERENCIA SOBRE LA VARIANZA EN POBLACIONES NORMALES

Contraste sobre la varianza de una variable con distribución normal

Descripción

- Evalúa una población con una v.a. normal cuya media y varianza son desconocidas.
- De la población se extrae una muestra aleatoria simple de tamaño n, cuyos registros forman el vector (x_1, x_2,..., x_n).

Contraste

$$\begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 \neq \sigma_0^2 \end{cases}$$

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

Ejemplo. Estamos interesados en determinar si en el estudio del sueño de patrones de estudiantes, citado más arriba, puede admitirse que la varianza del número de horas de sueño diario de los estudiantes es igual a 1.

Contraste sobre la varianza de una variable con distribución normal

Ejemplo. Cont.

Ejemplo en código R

```
library(TeachingDemos)
sigma.test(SleepStudy$AverageSleep, sigma=1)
```

```
        One sample Chi-squared test for variance

data:  SleepStudy$AverageSleep
X-squared = 234.59, df = 252, p-value = 0.4447
alternative hypothesis: true variance is not equal to 1
95 percent confidence interval:
 0.7875773 1.1175098
sample estimates:
var of SleepStudy$AverageSleep
                     0.9309155
```

Por tanto puede admitirse dicha hipótesis (p-valor=0.44474). La función sigma.test() nos proporciona además un intervalo de confianza al 95% para la varianza poblacional. Puede utilizarse la opción conf.level para especificar un nivel de confianza distinto.

Cociente de varianzas de poblaciones normales independientes

Descripción

- Contraste sobre el ratio de varianzas
- Poblaciones normales

Contraste

- $H_0 : \sigma_1^2/\sigma_2^2 = 1$
- $H_A : \sigma_1^2/\sigma_2^2 \neq 1$

$$T = \frac{var(x_1)}{var(x_2)} \sim F(n_1 - 1, n_2 - 1)$$

Ejemplo. El conjunto de datos de glucosa corresponde a los resultados de la prueba de glucose en sangre en diferentes periodos de un grupo de mujeres embarazadas.

Cociente de varianzas de poblaciones normales independientes

Ejemplo. Cont.

Ejemplo en código R

```
load (D:/glucosa.RData)
DIF2<-g2des-g2antes
attach(glucosa)
var.test(DIF2 ~ embarazo)
```

```
        F test to compare two variances

data:  DIF2 by embarazo
F = 1.0926, num df = 39, denom df = 39, p-value = 0.7836
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5778499 2.0657084
sample estimates:
ratio of variances
          1.092552
```

De acuerdo a los resultados, no rechazamos la hipótesis nula (pvalor=0.7839). Por tanto, los datos proporcionan evidencia estadística de que no existen diferencias significativa entre la variabilidad en ambos grupos de tiempos del examen de glucosa.

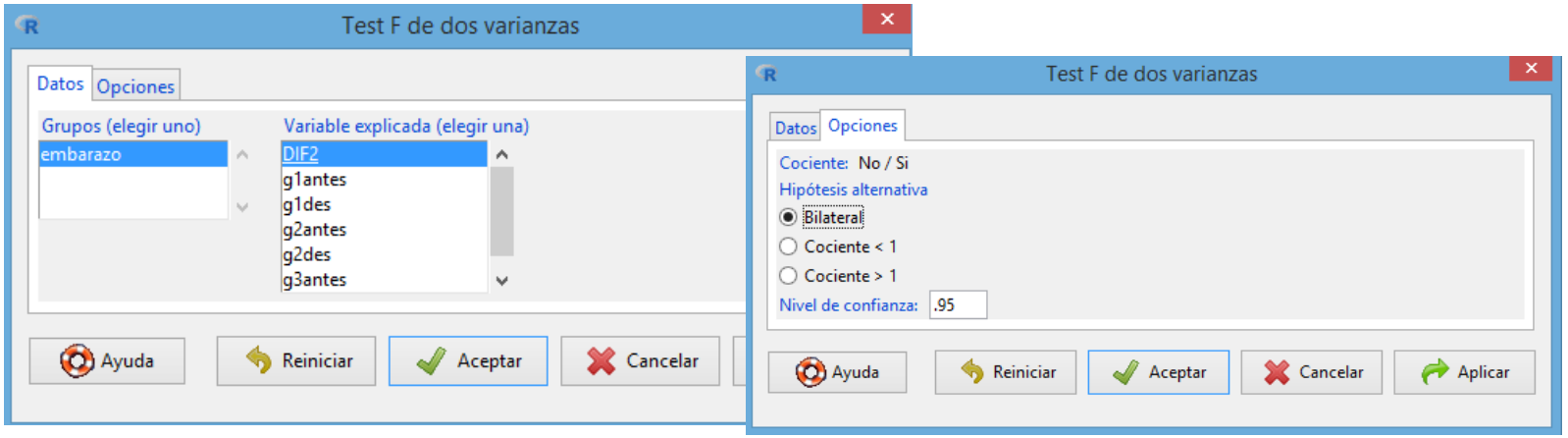
Cociente de varianzas de poblaciones normales independientes

Ejemplo. Cont.

Desde R Commander

Debemos importar los datos desde el archivo glucosa.sav para manejar las variables en cuestión, luego calculamos la variable DIF2. A continuación elegimos la opción del menú:

Estadísticos → Varianzas → Test F de dos varianzas



Cociente de varianzas de poblaciones normales emparejadas

Descripción

- Cuando se quiere comparar la varianza de muestras emparejadas puede utilizarse el **test de Pitman-Morgan**.
- En R este test se encuentra implementado en el paquete PairedData, en la función var.test().

Ejemplo. Contrastar si existen diferencias significativas entre las varianzas del número de pulsaciones por minuto de estudiantes según que estén haciendo un examen o atendiendo a una clase.

Ejemplo en código R

```
data(QuizPulse10)
library(PairedData) # Instalar paquetes si no están disponibles
p<-with(QuizPulse10,paired(Lecture,Quiz))
Var.test(p)
```

Cociente de varianzas de poblaciones normales emparejadas

Ejemplo. Cont.

```
        Paired Pitman-Morgan test

data:  Lecture and Quiz
t = 0.091513, df = 8, p-value = 0.9293
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3426116 3.2045321
sample estimates:
variance of x variance of y
     163.8778      156.4000
```

De acuerdo a los resultados, no rechazamos la hipótesis nula (pvalor=0.9293). Por tanto, concluimos que los datos proporcionan evidencia estadística de que no existen diferencias significativas entre las varianzas de los dos grupos de estudiantes, con un nivel de significancia de 5%.

INTRODUCCIÓN AL ANÁLISIS DE LA VARIANZA

Idea intuitiva del ANOVA

- La técnica de análisis de varianza (ANOVA) también conocida como análisis factorial y desarrollada por Fisher en 1930, constituye la herramienta básica para el estudio del efecto de uno o más factores (cada uno con dos o más niveles) sobre la media de una variable continua.

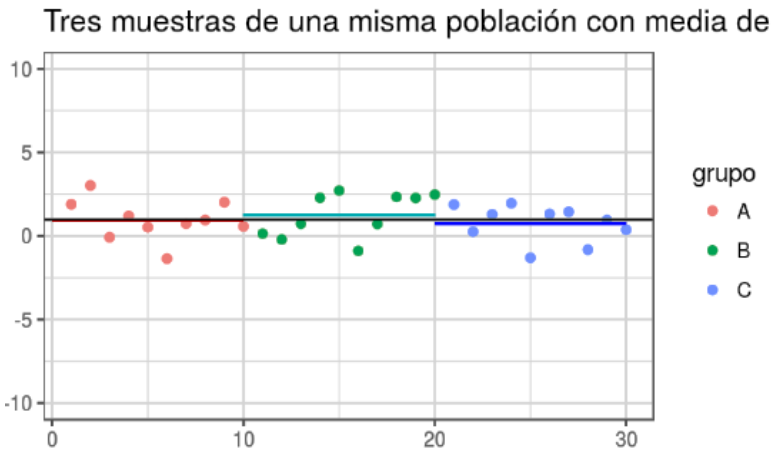
- Es por lo tanto el test estadístico a emplear cuando se desea comparar las medias de dos o más grupos. Esta técnica puede generalizarse también para estudiar los posibles efectos de los factores sobre la varianza de una variable.

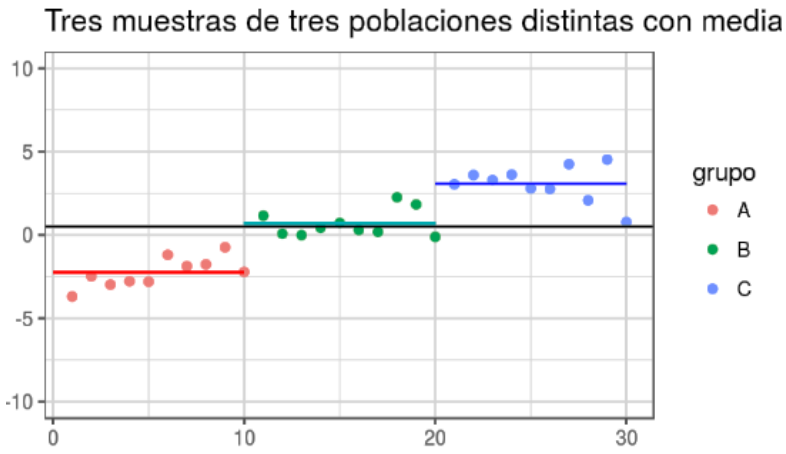
- La hipótesis nula de la que parten los diferentes tipos de ANOVA es que la media de la variable estudiada es la misma en los diferentes grupos, en contraposición a la hipótesis alternativa de que al menos dos medias difieren de forma significativa. ANOVA permite comparar múltiples medias, pero lo hace mediante el estudio de las varianzas.

Análisis de varianza de un factor

- El ANOVA de una vía es el tipo de análisis que se emplea cuando los datos no están pareados y se quiere estudiar si existen diferencias significativas entre las medias de una variable aleatoria continua en los diferentes niveles de otra variable cualitativa o factor.

- Conforme las medias de los grupos estén más alejadas las unas de las otras, la varianza entre medias se incrementará y dejará de ser igual a la varianza promedio dentro de los grupos.





La línea negra es la media para todas las observaciones.

Análisis de varianza de un factor

Objetivo

Explicar (controlar) las variaciones de una v.a. *Y* continua (numérica), mediante factores (variables cualitativas que definen categorías) que controlamos (no aleatorios).

- Este análisis permite poner en evidencia eventuales relaciones entre Y y estos factores. El factor puede ser la temperatura, la empresa que ha producido el bien, la región geográfica, etc.

- Hablaremos de análisis de la varianza de un factor (one-way), cuando sólo se contempla una sola variable explicativa y el arreglo de las observaciones en los niveles del factor es complemente al azar.

Análisis de varianza de un factor

Ejemplo. Queremos estudiar la influencia de la operadora sobre el importe de nuestra factura anual de teléfono (Y). Disponemos de datos que corresponden al gasto anual de teléfono en Soles (Y) de 15 clientes.

	Operadora 1	Operadora 2	Operadora 3
	750	800	950
	800	850	850
	810	880	820
	815	890	900
	815	900	820
Medias	798	864	868

Denotamos:

m_1 el valor medio de Y con la operadora 1.

m_2 el valor medio de Y con la operadora 2.

m_3 el valor medio de Y con la operadora 3.

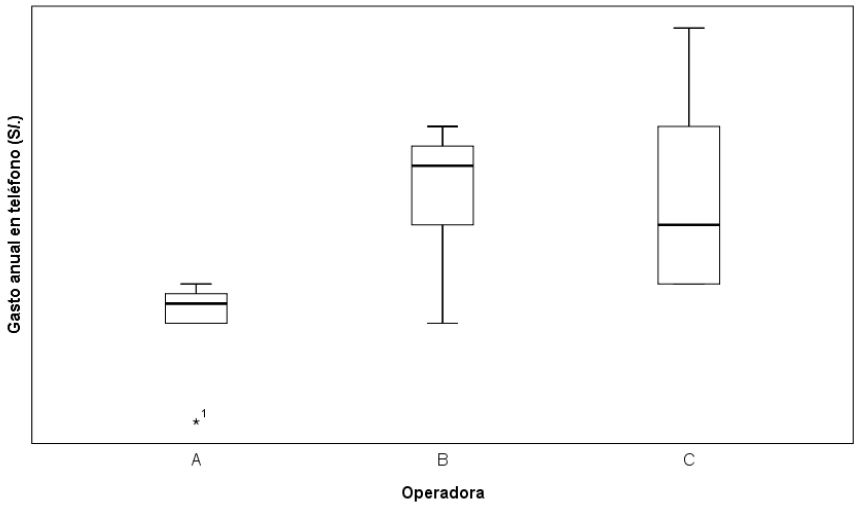
PREGUNTA: ¿$m_1 = m_2 = m_3$?

Análisis de varianza de un factor

Ejemplo. (Cont.)

Vocabulario

- *Y* = ”Gasto anual de teléfono”, es una variable cuantitativa (dependiente).
- La Operadora es una variable cualitativa (independiente) con la cual queremos explicar las variaciones de Y: **un factor**.
- El factor tiene un cierto número de **niveles**. El factor Operadora tiene aquí 3 niveles.



El modelo

- El modelo más simple de diseño consiste en suponer que se tiene

  - sólo un factor y
  - que las n unidades experimentales se asignan en forma completamente aleatoria a cada uno de los k niveles o tratamientos ($n = n_1 + n_2 + ... + n_j + ... + n_k$).
  - Los tamaños muestrales de cada tratamiento no tienen por qué ser iguales.

- El análisis de la varianza permite contrastar la hipótesis nula de que las medias de K poblaciones (K>2) son iguales, frente a la hipótesis alternativa de que por lo menos una de las poblaciones difiere de las demás en cuanto a su valor esperado.

Procedimiento

- Paso 1: Las hipótesis contrastadas en un ANOVA de un factor son:

H_0 : No hay diferencias entre las medias de los diferentes grupos :

$$\mu_1 = \mu_2 = \ldots = \mu_k$$

H_a : No todas las medias poblacionales son iguales.

El ANOVA se trata, por tanto, de una generalización de la *Prueba T para dos muestras independientes* al caso de diseños con más de dos muestras.

Procedimiento

- Paso 2: Supuestos del ANOVA.

El ANOVA requiere el cumplimiento los siguientes supuestos:

- Las poblaciones (distribuciones de probabilidad de la variable dependiente correspondiente a cada factor) son **normales**. Podemos utilizar el test de Kolmogorov Smirnov (ks.test()) o el de Shapiro Wilk (shapiro.test()).
- Las K muestras sobre las que se aplican los tratamientos son **independientes**.
- Las poblaciones tienen todas igual varianza (**homoscedasticidad**). Para comprobar si las varianzas de los grupos son homogéneas podemos utilizar el test de Bartlett (bartlett.test()) o el de Fligner-Killeen (fligner.test())

Procedimiento

- Paso 3: Estadística de prueba del ANOVA.

El ANOVA se basa en la **descomposición de la variación total de los datos Y_{ij}** con respecto a la media global μ (SCT), que bajo el supuesto de que H_0 es cierta es una estimación de σ^2 obtenida a partir de toda la información muestral, en dos partes:

- **Variación dentro de las muestras (SCR)** o Intra-grupos, cuantifica la dispersión de los valores de cada muestra con respecto a sus correspondientes medias.

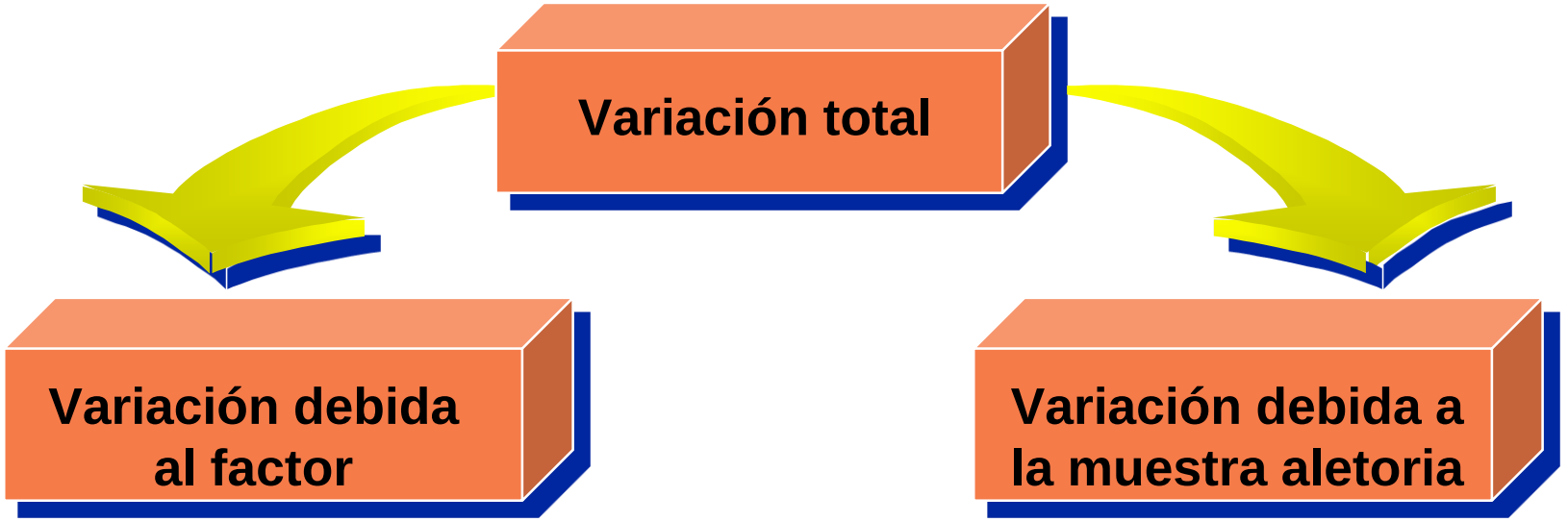
- **Variación entre muestras (SCE)** o Inter-grupos, cuantifica la dispersión de las medias de las muestras con respecto a la media global.

$$\underbrace{\sum (Y_{ij} - \mu)^2}_{SCT} = \underbrace{\sum (Y_{ij} - \mu_i)^2}_{SCR} + \underbrace{\sum (\mu_i - \mu)^2}_{SCE}$$

Procedimiento

- Paso 3: Prueba F ANOVA.

Partición de la variabilidad total



- Suma de cuadrados **entre**
- Suma de cuadrados del modelo
- Entre los grupos de variación

- Suma de cuadrados **dentro**
- Suma de cuadrados del error
- Dentro de los grupos de variación

Procedimiento

- Paso 3: Estadística de prueba ANOVA

ANOVA se define como análisis de varianza, pero en un sentido estricto, se trata de un análisis de la Suma de Cuadrados Medios.

- $\hat{S}_T^2 = \frac{TSS}{N-1}$ = Cuadrados Medios Totales = Cuasivarianza Total (varianza muestral total)
- $\hat{S}_t^2 = \frac{SST}{k-1}$ = Cuadrados Medios del Factor = Intervarianza (varianza entre las medias de los distintos niveles)
- $\hat{S}_E^2 = \frac{SSE}{N-k}$ = Cuadrados Medios del Error = Intravarianza (varianza dentro de los niveles, conocida como varianza residual o de error)

Procedimiento

Una vez descompuesta la estimación de la varianza, se obtiene el estadístico F_{ratio} dividiendo la intervarianza entre la intravarianza. Entonces, suponiendo H_0 cierta, el estadístico utilizado sigue una distribución ***F*** de Fisher-Snedecor con (*k-1*) y (*n-k*) grados de libertad, siendo *k* el número de muestras y *n* el número total de observaciones que participan en el estudio.

$$F_{ratio} = \frac{Cuadrados\ Medios\ del\ Factor}{Cuadrados\ Medios\ del\ Error} = \frac{\hat{S}_t^2}{\hat{S}_E^2} = \frac{intervarianza}{intravarianza} \sim F_{k-1,N-k}$$

- Paso 4: Calcular el p-valor y comparar con α.

  Por tanto, rechazamos H_0 si $F_{ratio} > F_{k-1,n-k}$, con un nivel de significancia α establecido.

- Paso 5: Escribir una conclusión.

Análisis de varianza de un factor

Estrategia ANOVA

- El primer paso es usar la prueba F ANOVA para determinar si hay diferencias significativas entre las medias.

- Si la prueba F ANOVA muestra que las medias no son todas iguales, entonces se pueden realizar pruebas de comparaciones múltiples para ver qué pares de medias difieren.

Análisis de varias muestras: ANOVA

Descripción

- Comparamos la media de p grupos
- Asumimos distribución Normal con varianzas iguales: $y_{ij} \sim N(\mu_i, \sigma^2)$
- Equivale al Test t para dos muestras independientes cuando $p = 2$

Contraste

- $H_0 : \mu_1 = \mu_2 = \ldots = \mu_p$
- H_A: No H_0

¿Cómo trabaja R?

- Toma como referencia un nivel del factor
- Utiliza variables *dummy* para el resto de niveles del factor
- Permite cambiar el nivel de referencia: `relevel()`

Análisis de varias muestras: ANOVA

Ejemplo en código R

```
load("ambiente.RData"); attach(ambiente)
bartlett.test(PH ~ PROVIN)
fligner.test(PH ~ PROVIN)
by(PH,PROVIN,shapiro.test)
```

- Resultados de las pruebas de homogeneidad da varianzas en los grupos

```
> bartlett.test(PH ~ PROVIN)

        Bartlett test of homogeneity of variances

data:  PH by PROVIN
Bartlett's K-squared = 0.77133, df = 2, p-value = 0.68

> fligner.test(PH ~ PROVIN)

        Fligner-Killeen test of homogeneity of variances

data:  PH by PROVIN
Fligner-Killeen:med chi-squared = 0.43641, df = 2, p-value = 0.804
```

Análisis de varias muestras: ANOVA

- Resultados de las pruebas de normalidad de PH por PROVIN

```
> by(PH,PROVIN,shapiro.test)
PROVIN: ALICANTE

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.98799, p-value = 0.5078

------------------------------------------------------------
PROVIN: CASTELLON

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.99214, p-value = 0.8313

------------------------------------------------------------
PROVIN: VALENCIA

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.98581, p-value = 0.3629
```

Análisis de varias muestras: ANOVA

Tabla ANOVA de 1 via

Variación	Suma Cuadrados	g.l.	Varianza	F
Entre grupos	$SE = \sum_i n_i(\overline{y}_{i.} - \overline{y}_{..})^2$	$p-1$	VE=SE/(p-1)	VE/VR
Residual	$SR = \sum_i \sum_j (y_{ij} - \overline{y}_{i.})^2$	$n-p$	VR=SR/(n-p)	
Total	$\sum_i \sum_j (y_{ij} - \overline{y}_{..})^2$	$n-1$		

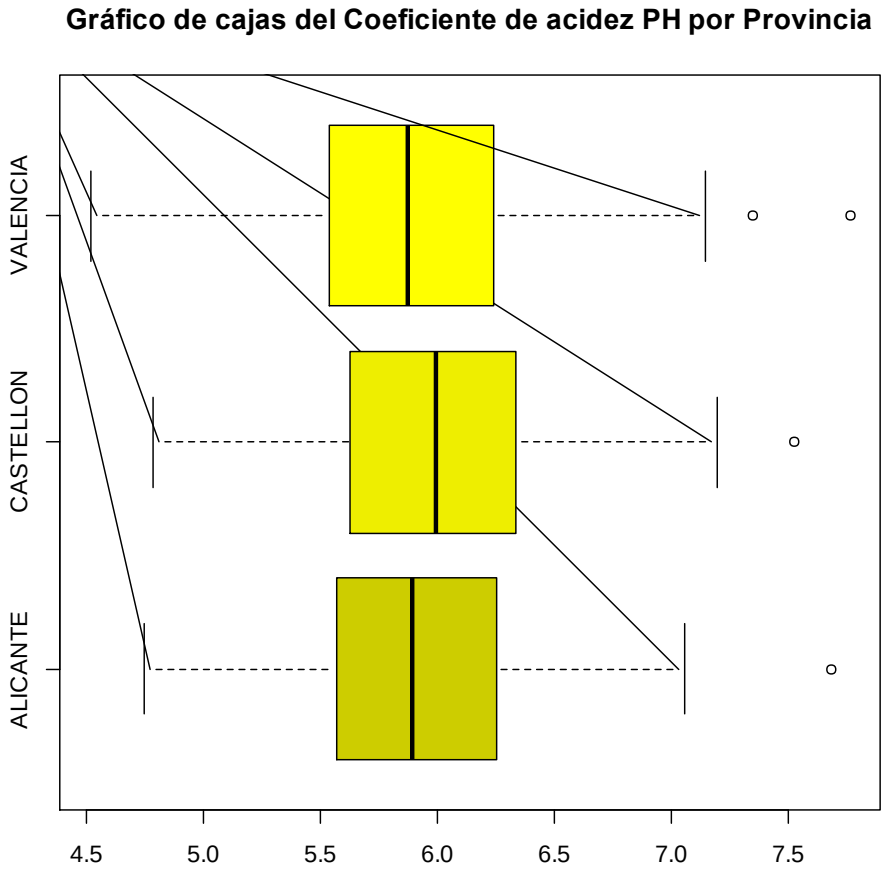
Ejemplo en código R

```
boxplot(PH ~ PROVIN, main="Gráfico de cajas del Coeficiente de acidez PH por Provincia", horizontal=TRUE, col=c("yellow3", "yellow2", "yellow"))
# ANOVA de una vía
anovaph <- aov(PH ~ PROVIN)
summary(anovaph)
```

Nota: Es muy importante indicar el orden en que se dan los argumentos. El primer argumento es siempre la variable dependiente (PH), que es seguido por el símbolo (~) y después la variable independiente (el factor o criterio de clasificación, PROVINCIA).

Análisis de varias muestras: ANOVA

- Resultados de la evaluación exploratoria de normalidad de PH por PROVIN



Análisis de varias muestras: ANOVA

- Resultados del ANOVA de una vía

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
PROVIN	2	0.32	0.1595	0.544	0.581
Residuals	297	87.05	0.2931		

Conclusión: En este caso, no hemos encontrado un efecto significativo de la variable PROVIN (p-valor > 0.05). Por tanto, no hay diferencias entre las provincias.

Análisis de varias muestras: ANOVA

Ejemplo. Se dispone de los datos de un experimento en el que los participantes recibían una entre tres posibles dosis de una droga experimental, y tras la cual se medía el grado de vigilancia en una tarea. ¿Existen diferencias significativas en administrar una u otra dosis?

Ejemplo en código R

```
load("doroga.RData") # Utilizar así, si se ha cambiado directorio de trabajo
droga
# Antes de hacer el ANOVA, es importante recalcar se puede ver el contenido de este u otros archivos almacenados (u objetos) en R en nuestra sesión. Se trata de emplear el comando objects
objects()
```

Análisis de varias muestras: ANOVA

Ejemplo. Cont.

```
> droga
   dosis vigilancia
1       a         30
2       a         38
3       a         35
4       a         41
5       a         27
6       a         24
7       b         32
8       b         26
9       b         31
10      b         29
11      b         27
12      b         35
13      b         21
14      b         25
15      c         17
16      c         21
17      c         20
18      c         19
> objects()
[1] "droga"
```

Análisis de varias muestras: ANOVA

Ejemplo. Cont.

Ejemplo en código R

```
# Para ejecutar el  ANOVA, R utiliza el comando aov
aov.droga <- aov(vigilancia~dosis, droga)
```

Es muy importante indicar el orden en que se dan los **argumentos.**

- El primer argumento es siempre la variable dependiente (vigilancia),
- que es seguido por el símbolo **(~)** y después la variable independiente (o, las variables independientes, en el caso de diseños factoriales).
- El argumento final para **aov** es el nombre del archivo R que está siendo analizado.
- **aov.dosis** es el nombre del archivo en el que se va a quedar el análisis (en otras palabras, se puede poner otro nombre, pero el dado es ya muy indicativo).

Análisis de varias muestras: ANOVA

Ejemplo. Cont.

Ejemplo en código R

Los resultados del ANOVA se pueden ver con el comando **summary**

```
summary(aov.droga)
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)
dosis        2  426.2  213.12   8.789 0.00298 **
Residuals   15  363.8   24.25
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Decisión: En este caso, hemos encontrado un efecto significativo de la variable Dosis (el valor de p ha sido menor de .05).

Conclusión: Los datos proporcionan evidencia suficiente para afirmar que existen diferencias significativas entre los tipos de dosis de una droga experimental.

Análisis de varias muestras: ANOVA

Ejemplo. Cont.

Ejemplo en código R

# Naturalmente es necesario indicar las medias por condición.
tapply(vigilancia, dosis, mean)

```
    a     b     c
32.50 28.25 19.25
```

Como vemos, es el grupo con dosis “c” el que tiene una media menor (19.25).

Los argumentos de **tapply** son los siguientes.

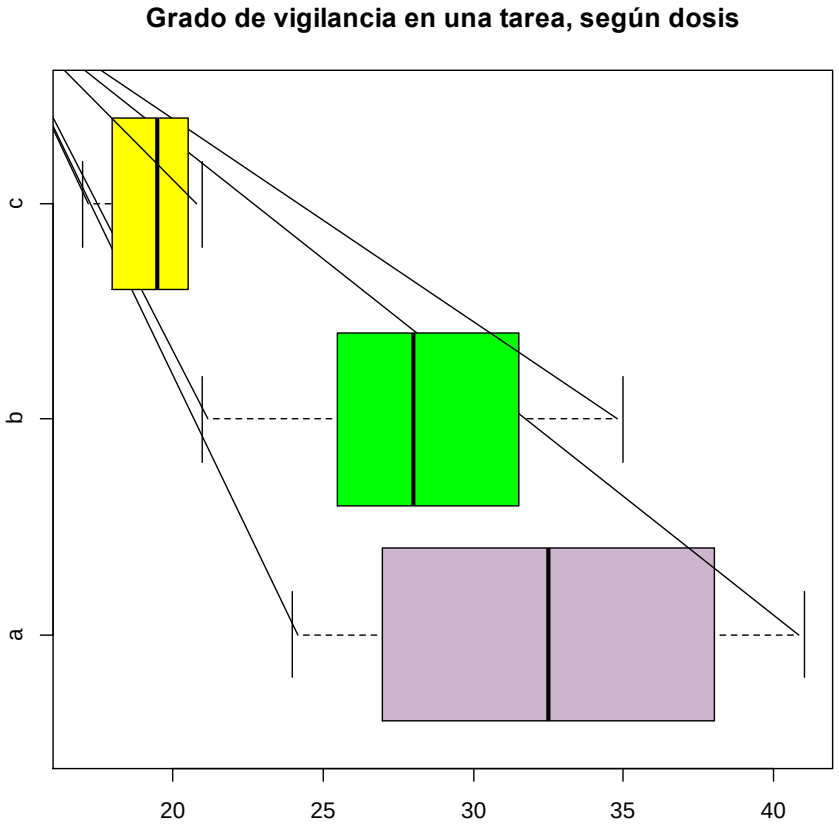
- El primer argumento es la columna de la variable dependiente; dado que dosis es un “objeto” de R con varias columnas, hemos de indicar cuál es la que queremos: vigilancia refleja el vector correspondiente a la columna “vigilancia”.
- El segundo argumento es la variable independiente, a la que se le aplica el mismo proceso (dosis refleja la columna de dosis).
- El tercer argumento es el estadístico que queremos calcular, en nuestro caso, la media aritmética (“mean” en inglés).

Análisis de varias muestras: ANOVA

Ejemplo. Cont.

Ejemplo en código R

```
# Un aspecto importante en
  cualquier experimento es
  observar los datos en cada
  grupo, para lo cual podemos
  tener el diagrama de caja para
  cada grupo
boxplot(vigilancia ~ dosis,
  main="Grado de vigilancia en
  una tarea, según dosis",
  horizontal=TRUE,
  col=c("thistle3", "green1",
  "yellow"))
```



Como vemos, es el grupo con dosis “c” el que tiene una media menor (19.25).

Análisis de varias muestras: ANOVA

Ejemplo. Cont.

En el gráfico anterior, se puede observar una diferencia clara entre el grupo de dosis “c” y los otros dos grupos. Naturalmente, viendo el gráfico y la existencia de un efecto significativo de Dosis en el ANOVA, lo que hemos de hacer ahora son **comparaciones múltiples.** (En caso de que el ANOVA no hubiera sido significativo, no se procede a efectuar estas pruebas.)

De esta manera podemos determinar entre qué condiciones experimentales hay diferencias significativas. Para ello, emplearemos el método de Tukey.

Ejemplo en código R

```
# Crearemos el objeto drogaTukey empleando la función TukeyHSD
drogaTukey <- TukeyHSD(aov.droga,'dosis')
# Para ver el resultado, simplemente hemos de teclear el nombre del objeto
drogaTukey
```

Análisis de varias muestras: ANOVA

Ejemplo. Cont.

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = vigilancia ~ dosis, data = droga)

$`dosis`
      diff       lwr       upr     p adj
b-a  -4.25 -11.15796  2.657961 0.2768132
c-a -13.25 -21.50659 -4.993408 0.0022342
c-b  -9.00 -16.83289 -1.167109 0.0237003
```

Las diferencias entre medias en las que el intervalo de confianza que engloba los límites inferior y superior no contienen el valor 0, son estadísticamente significativas con el método de Tukey. En nuestro caso, son las diferencias entre los grupos “b” y “c”, y entre los grupos “a” y “c”. Esto puede verse también si trazamos los intervalos de confianza gráficamente.

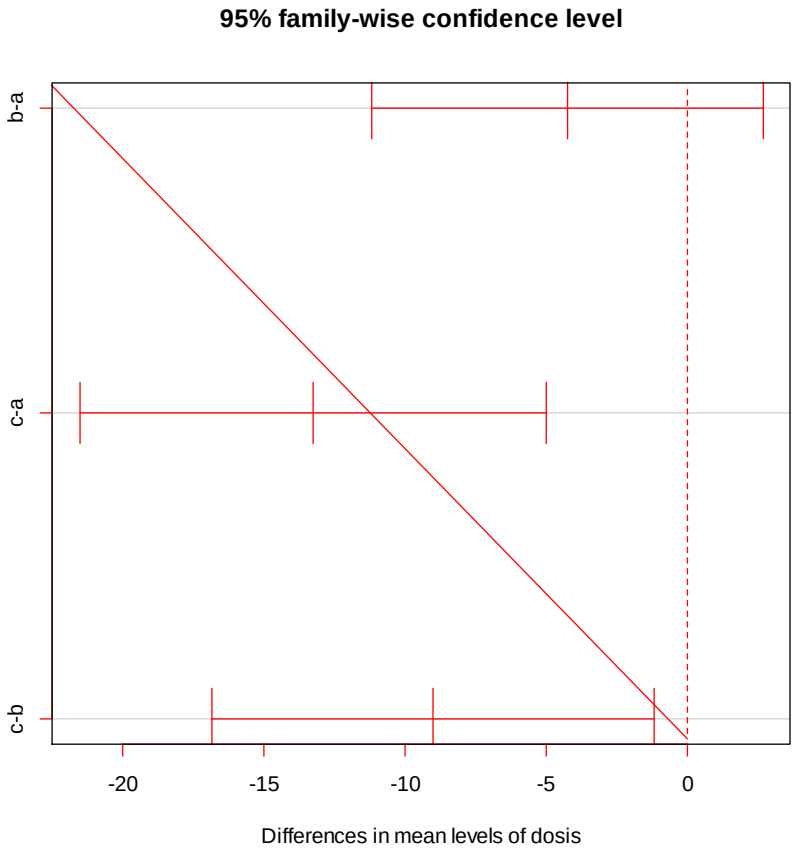
Análisis de varias muestras: ANOVA

Ejemplo. Cont.

Ejemplo en código R

# Utilizamos la **función plot** para trazar los intervalos de confianza
plot(drogaTukey)

Se puede apreciar que sólo el intervalo de confiaza de la diferencia entre los grupos “a” y “b” toca 0, es decir, que la diferencia entre las dos medias no es estadísticamente significativa.



Análisis de varias muestras: Test Kruskal-Wallis

- Este modelo estadístico corresponde a decidir si puede aceptarse la hipótesis de que k muestras independientes proceden de la misma población o de poblaciones idénticas con la misma mediana.

- Características:

1. Se emplea cuando se quieren comparar tres o más poblaciones.
2. Es el equivalente no paramétrico a un análisis de varianza de una sola vía (Análisis de Varianza Unifactorial por Rangos).
3. No requiere supuesto de normalidad.
4. No requiere supuesto de varianzas iguales (homogeneidad de varianzas).
5. Compara esencialmente los rangos promedios observados para las k muestras, con los esperados bajo H_0.
6. Nivel ordinal de la variable dependiente.

Análisis de varias muestras: Test Kruskal-Wallis

Descripción

- Extensión del Test de Mann-Whitney para 3 ó más grupos
- Comparamos las medianas de p grupos

Contraste

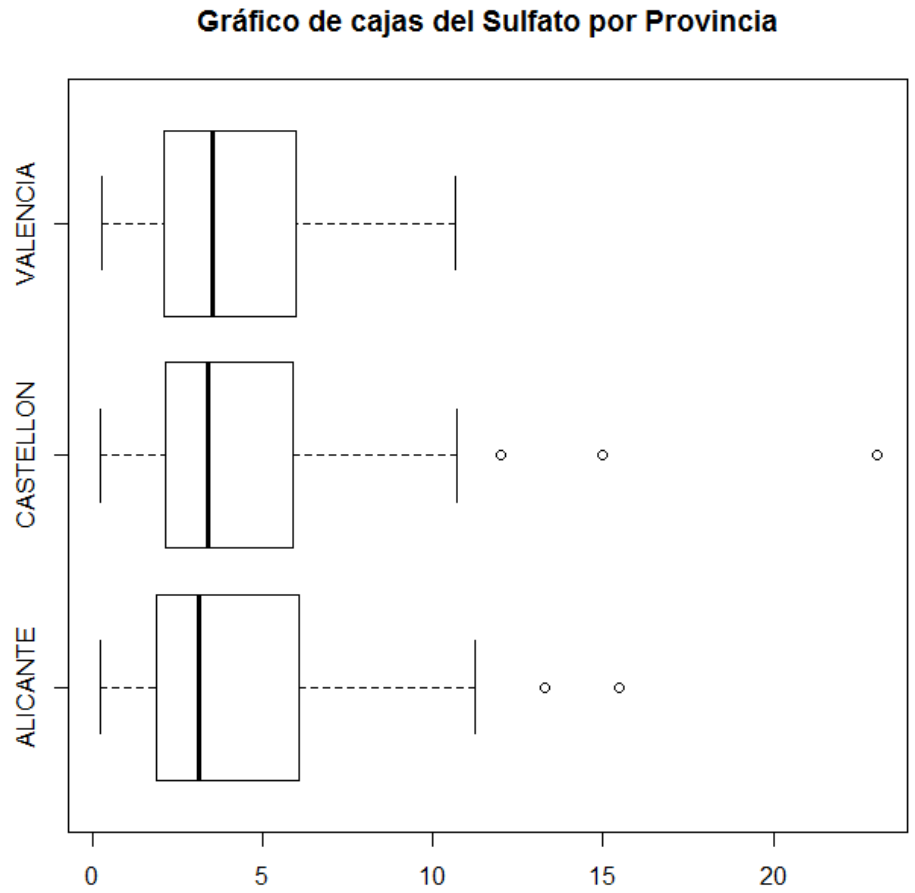
- H_0: $\text{Mediana}_1 = \ldots = \text{Mediana}_p$
- H_A: No H_0

Ejemplo en código R

```
boxplot(SULFATO ~ PROVIN, main="Gráfico de cajas del Sulfato por Provincia", horizontal=TRUE)
kruskal.test(SULFATO ~ PROVIN)
```

Análisis de varias muestras: Test Kruskal-Wallis

- Resultados



Análisis de varias muestras: Test Kruskal-Wallis

- Resultados

```
> kruskal.test(SULFATO ~ PROVIN)

        Kruskal-Wallis rank sum test

data:  SULFATO by PROVIN
Kruskal-Wallis chi-squared = 0.19376, df = 2, p-value = 0.9077
```

Conclusión: En este caso, aceptamos la hipótesis nula de igualdad de medianas del SULFATO por los niveles de PROVINCIA (el valor de p ha sido mayor a 0.05). Equivalentemente, existe evidencia estadística para afirmar que las poblaciones de las que proceden las k muestras son idénticas.

Análisis de varias muestras: Test Kruskal-Wallis

Ejercicio. Una EMPRESA MANUFACTURERA solicita y contrata personal para su equipo gerencial en tres escuelas diferentes. Se dispone de calificaciones de desempeño en muestras independientes de cada una de las escuelas.

Se obtienen las calificaciones de 7 empleados de la escuela A, 6 empleados de la escuela B y 7 empleados de la escuela C. La calificación de cada gerente está en escala de 0 a 100.

A	B	C
25	30	40
60	60	90
50	85	90
70	15	35
20	80	70
70	95	80
60		75

¿Existe evidencia para concluir que las escuelas son idénticas ?

Análisis de varias muestras: Test Kruskal-Wallis

Ejercicio. Cont.

Planteamiento

- El problema es de comparación de tres grupos independientes.
- Modelo:

  Variable independiente: Tipo de escuela

  Variable dependiente: Calificaciones de los gerentes (escala ordinal)

- Hipótesis:

  H_0: Las Escuelas son idénticas en términos de las evaluaciones de desempeño (no hay diferencias significativas).
  H_1: Por lo menos una de las Escuelas no es idéntica en términos de las evaluaciones de desempeño.

- Grado de significancia del 5%.

Continuar…

Comunicación constante con la Escuela del INEI

Correo de la Dirección Técnica de la ENEI
Sr. Eduardo Villa Morocho (Eduardo.villa@inei.gob.pe)

Coordinación Académica
Sra. María Elena Quirós Cubillas (Maria.Quiros@inei.gob.pe)

Correo de la Escuela del INEI
enei@inei.gob.pe

Área de Educación Virtual
Sr. Gonzalo Anchante (gonzalo.anchante@inei.gob.pe)