MÉTODOS DE ANÁLISIS DE DATOS EN ECOLOGÍA CON R

Pilar Castro Díez, Asunción Saldaña López, Josabel Belliure, Tíscar Espigares Pinilla, Paloma Ruiz Benito, Ana Margarida Coelho dos Santos, Ignacio Morales Castilla.

Departamento de Ciencias de la Vida

Universidad de Alcalá

Este manual ha sido realizado dentro del Proyecto de Innovación Docente de la UAH "***Elaboración de materiales de apoyo a la docencia de prácticas de asignaturas de Ciencias: Manuales y vídeos tutoriales***".

Este manual está orientado a estudiantes de cualquier Grado de Ciencias que requiera el uso de estadística básica, aunque los ejemplos que se trabajan en él derivan del ámbito de la Ecología.

Julio de 2019

Universidad de Alcalá

Contenido

1. INTRODUCCIÓN

La Estadística proporciona a la Ecología (y a otras ciencias experimentales) las herramientas necesarias para el análisis de los datos. Dado que no podemos hacer estudios en toda la población (no es posible contar todos los ácaros que hay en un suelo, ni medir el área foliar de todas las hojas de un bosque, ni medir la longitud de todas las carpas que tiene un lago), la estadística nos permite cuantificar la probabilidad de cometer error al extrapolar los resultados obtenidos de una muestra al conjunto de la población.

La **estadística descriptiva** reúne un conjunto de técnicas que facilitan la organización, resumen y comunicación de datos; la **estadística inferencial** permite hacer pruebas de contraste de hipótesis.

1.1. Exploración de los datos

Cuando tenemos una colección de datos resultantes de un experimento o muestreo, conviene realizar una primera exploración de cómo son esos datos, antes de realizar ningún análisis complejo. La **estadística descriptiva** aporta parámetros que nos dan una idea inicial sobre cómo son esos datos. Concretamente, disponemos de "medidas de tendencia central" y de "medidas de dispersión" de los datos alrededor de ese valor central.

MEDIDAS DE TENDENCIA CENTRAL

Estas medidas indican alrededor de qué valor se agrupan los datos observados. Distinguimos:

1. Media aritmética (X): es el centro de gravedad de la serie de datos y se calcula como

$$X = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{(Ecuación 1)}$$

donde x_i representa cada uno de los valores de la variable y n el número de réplicas.

2. Mediana: es el punto medio de una serie ordenada de datos
3. Moda: es el valor más frecuente de la serie de datos.

MEDIDAS DE DISPERSIÓN

Estas medidas indican si los valores de la variable están muy dispersos o se concentran alrededor de la medida de centralización. Son:

1. Rango (R): Diferencia entre el valor máximo (x_{max}) y el mínimo (x_{min}) observado.

$$R = x_{max} - x_{min} \quad \text{(Ecuación 2)}$$

2. Varianza (s^2): Expresa la dispersión de valores entorno a la media (X)

$$s^2 = \frac{\sum(x_i - X)^2}{n-1} \quad \text{(Ecuación 3)}$$

3. Desviación estándar (s): Es la raíz cuadrada de la varianza.

Para tener una representación visual de estas medidas, es recomendable representar gráficamente la media junto con medidas de dispersión.

DISTRIBUCIÓN DE LOS DATOS

Tras calcular los parámetros de la estadística descriptiva, debemos explorar cómo se distribuyen los datos. Los histogramas de frecuencias (Fig. IV.1) son una herramienta de representación de datos que nos permiten observar cómo se distribuyen los mismos. Están formados por rectángulos adyacentes que tienen por base cada uno de los intervalos de la variable medida y por altura las frecuencias absolutas (nº de veces que aparecen datos dentro de ese intervalo). El número de intervalos a utilizar (k) se puede calcular según la regla de Sturges (1926): $k = 1 + 3.322 * \log(n)$, donde n es el tamaño de muestra.

De entre todas las distribuciones posibles que puedan seguir unos datos, la **distribución normal** es la más interesante desde el punto de vista estadístico, pues reúne unas propiedades que han hecho posible que a partir de ella se desarrollaran numerosos métodos de análisis de datos.

Propiedades de la distribución normal:

- Los valores cercanos a la media son los más abundantes, y a medida que nos alejamos de la media, los datos presentan una frecuencia cada vez menor.
- Es simétrica alrededor de la media. Por tanto, media, mediana y moda coinciden.
- Se caracteriza por dos medidas: media y desviación típica
- Tiene forma de campana, sin un pico excesivo.
- El 50% de las observaciones se encuentran por debajo de la media y el 50% por encima.
- El 68% de las observaciones se encuentran dentro del intervalo $x \pm s$
- El 95% de las observaciones se encuentran dentro del intervalo $x \pm 1{,}96 * s$
- El 99% de las observaciones se encuentra dentro del intervalo $x \pm 2{,}57 * s$.

Cómo dibujar un histograma de frecuencias

Para saber si una serie de datos sigue una distribución normal o no, podemos dibujar histrogramas de frecuencia o un gráficos *qq*. Por último, podemos utilizar test estadísticos para saber si la distribución de nuestros datos se ajusta a algún modelo de distribución. A continuación mostraremos cómo hacer cada una de estas pruebas con R studio.

Análisis en R de la distribución de datos *

** Los análisis en R mostrarán en verde los comentarios al texto (al estar precedidos de "#" R los interpreta como comentarios, no como comandos), en azul el código de R a incluir en el programa y en negro los resultados al mismo.*

Histograma de frecuencias para explorar la normalidad de una serie de datos

```
#Generamos datos un conjunto de datos imaginarios de longitud de tarso (en mm) en una población de aves. Nuestros datos deben seguir una distribución normal, una media 69.7 cm y una desviación estándar de 14.65. A continuación representamos un histograma de distribución de frecuencias de esos datos.

data<-rnorm(n = 1000, mean = 69.7, sd = 14.65) #genero una serie de datos aleatorios que cumplan los requisitos de media y desviación estándar indicados y que sigan una distribución normal.
hist(data, #histograma con los datos Figura IV.1
     xlab = "Longitud de tarso (mm)", ylab = "Probabilidad", #añado los nombres de los ejes.
     main = "", prob = TRUE,ylim = c(0,0.04))
lines(density(data), col="blue", lwd=2) #mostramos la probabilidad en el eje Y y aumentamos el límite para que se vea bien la curva
abline(v=69.7, col="red")#Dibujamos una línea roja en la media
```

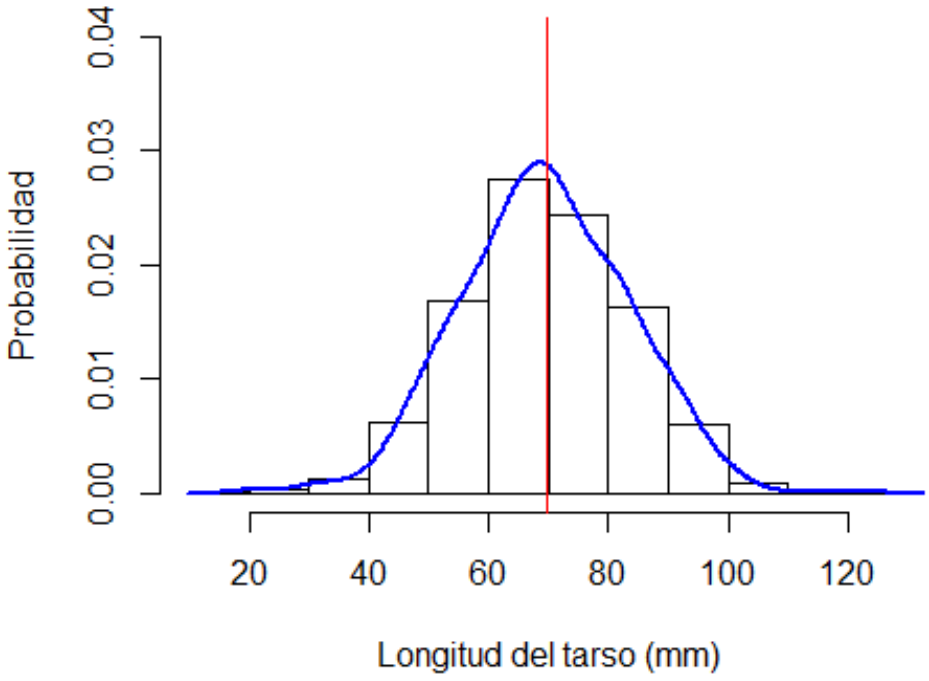

Figura IV.1. *Histograma de probabilidad de la variable "Longitud del tarso (mm)" (eje X) de una población de aves. El eje Y muestra la probabilidad con la que aparecen valores en cada intervalo de X. En este caso se trata de una distribución normal.*

Cómo dibujar un gráfico qq para explorar la normalidad

Un gráfico qq-normal confronta los cuantiles teóricos en caso de que la distribución sea normal, con los cuantiles reales de los datos (Fig.IV.2). Si la distribución se ajusta a la normalidad los puntos se distribuyen a lo largo de una línea recta diagonal.

Podemos crear un gráfico qq con los datos creados en el ejemplo anterior con el siguiente código:

```
Gráfico qq para explorar la normalidad de una serie de datos

data<-rnorm(n = 1000, mean = 69.7, sd = 14.65) # genero una serie de datos aleatorios
que cumplan los requisitos de media y desviación estándar indicados y que sigan una
distribución normal.

qqnorm(data, col="springgreen4") # dibujo el gráfico qq,indicando que los puntos tengan
color verde
qqline(data) # añado la línea diagonal, que indica la distribución teórica que seguirían
los puntos en caso de ser normales.
```

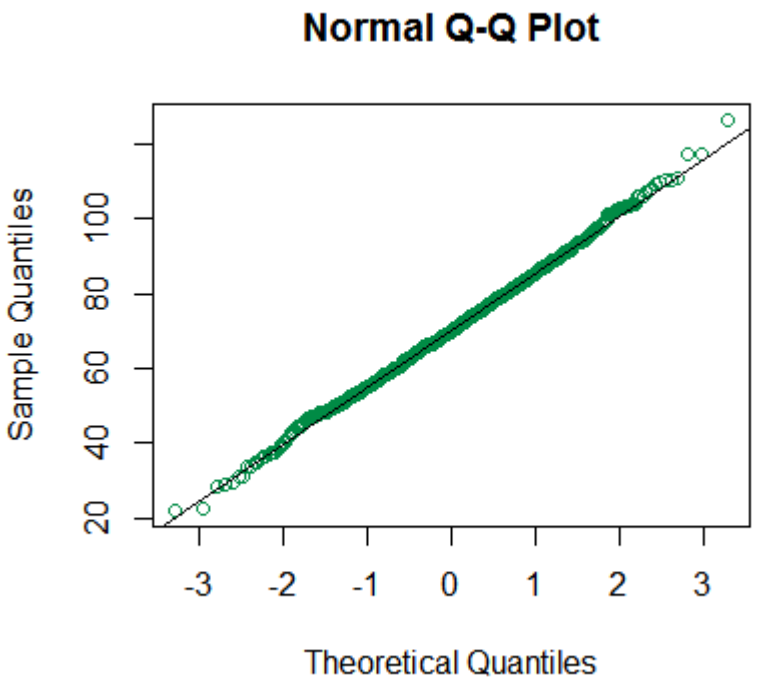

***Figura IV.2**. Gráfico qq para explorar la normalidad de la variable "Longitud del tarso (mm)". El eje X indica los cuantiles teóricos si la distribución es normal y el eje Y los cuantiles de la muestra. Como los datos (en verde) se ajustan a la línea diagonal teórica de una distribución normal, podemos considerar que la distribución de los datos es normal.*

Cómo hacer un test estadístico para comprobar la normalidad de los datos

Existen diversos test estadísticos que nos indican la probabilidad de que la distribución que sigue una serie de datos difiera de una distribución normal (hipótesis nula o H_o). El resultado del test nos devuelve un valor de probabilidad (*valor de p*). Si $p \leq 0.05$, rechazamos *Ho* y concluimos que nuestros datos no se ajustan a una distribución normal. Si por el contrario $p > 0.05$ aceptamos H_o) y concluimos que nuestros datos siguen una distribución normal (ver más detalles en el apartado 1.2).

En nuestro caso vamos a aplicar el test de Shapiro-Wilk para evaluar la normalidad a la serie de datos de los ejemplos anteriores.

Test de Shapiro-Wilk para explorar la normalidad de una serie de datos

```
data<-rnorm(n = 1000, mean = 69.7, sd = 14.65) # genero una serie de datos aleatorios
que cumplan los requisitos de media y desviación estándar indicados y que sigan una
distribución normal.

shapiro.test(data) # Test de normalidad
Shapiro-Wilk normality test
data:  data
W = 0.99865, p-value = 0.6524
```

Observamos que el valor de p que devuelve el test es muy superior a 0.05, por lo que aceptamos la hipótesis nula de que los datos siguen una distribución normal. Este resultado es concordante con la exploración gráfica mostrada en las figuras IV.1 y IV.2.

HOMOGENEIDAD DE VARIANZAS (HOMOCEDASTICIDAD)

Cuando queremos comparar dos o más series de datos, correspondientes a la misma variable (por ej. la humedad del suelo entre zonas con suelo desnudo y zonas de tomillar), probablemente tendremos que saber si las varianzas difieren entre las series de datos. Esta información es necesaria para poder seleccionar el test estadístico adecuado (ver apartado 1.2). Para ello podemos aplicar manualmente el test de la F de Snedecor. Este test evalúa la hipótesis nula (H_0) de que las varianzas son iguales.

Cuadro IV.1. Prueba de comprobación de varianzas iguales: F de Snedecor

Se calculan las varianzas de cada una de las dos muestras: S^2_1 y S^2_2
Se calcula el estadístico F_{cal} a partir de la siguiente fórmula:

$$F_{cal} = \frac{S^2_{mayor}}{S^2_{menor}} \text{ (ecuación 4)}$$

Grados libertad: n_1-1, n_2-1 (n_1 tamaño de la muestra de varianza mayor)
H_o: varianzas iguales. Si $F_{cal} \geq F_{crítica}$ (La $F_{crítica}$ se busca en las tablas, ver sección dedicada al Anova), se rechaza la H_o, es decir, se concluye que las varianzas no son iguales.

Alternativamente, podemos realizar un test similar de homocedasticidad usando RStudio. La hipótesis nula que se evalúa, al igual que antes, es que las varianzas son iguales. El test nos devuelve un valor de *p*, que si es mayor de 0.05 nos llevará a aceptar la hipótesis nula. En caso contrario, rechazaremos H_0 y concluiremos que las varianzas no son iguales.

Test de Barlett para explorar la homogeneidad de varianzas

```
# genero una matriz (data frame llamado "datos") con una variable que es "longitud del tarso" de una especie de ave y otra que es "habitat" con dos categorías. En cada tipo de hábitat (bosque y matorral) hay 50 medidas de longitud de tarso.

longitud<-rnorm(n = 100, mean = 69.7, sd = 14.65)
habitat<-c("bosque", "matorral")
habitat<-rep(habitat, each=50)
datos<-data.frame(habitat, data)

# Aplico el test de Barlett para ver si la varianza de la longitud de tarso difiere entre ambos tipos de hábitat.
bartlett.test(longitud ~ habitat, data=datos)
Bartlett test of homogeneity of variances
data:  longitud by habitat
Bartlett's K-squared = 0.0056683, df = 1, p-value = 0.94
```

El resultado ofrece un valor de p muy superior a 0.05, por lo que no rechazo la hipótesis nula y concluyo que la varianza de longitud de tarso es igual en ambos tipos de hábitat.

1.2. Pruebas de contraste de hipótesis

Se han desarrollado numerosos tests estadísticos que permiten realizar pruebas de contraste de hipótesis a partir de la distribución normal y la existencia de homogeneidad de varianzas: son las pruebas **paramétricas**. Sin embargo, no siempre los datos que obtenemos en un trabajo científico se ajustan a estos requisitos; en esos casos recurrimos a la **estadística no paramétrica**.

Para contrastar una hipótesis ecológica (H_{ecol}, por ej. el factor A afecta a la respuesta B), hemos de plantear una hipótesis nula (H_o), que supone la negación de la hipótesis ecológica (el factor A no afecta a la respuesta B). Cuando realizamos cualquier test estadístico de contraste de hipótesis, lo que hacemos es calcular la probabilidad de equivocarnos al rechazar H_0 (y por tanto aceptar nuestra hipótesis). Esa probabilidad es el ***valor de p*** (o ***p-valor***) que acompaña a un resultado estadístico. Por tanto, para poder rechazar H_o el valor de *p* debe ser bajo. Para tomar una decisión respecto a cuál sea la hipótesis 'verdadera', el investigador fija el nivel máximo de error que se permite asumir al aceptar H_{ec} (que se suele denotar como α). Por convenio el umbral de significación se suele fijar en 0, es decir, nos permitimos un error máximo del 5% en nuestra afirmación de la hipótesis ecológica. **Por tanto, si $p \leq 0.05$, rechazamos *Ho* y aceptamos *Hec*.**

En función del número de variables implicadas en un análisis estadístico, distinguimos dos tipos de métodos de análisis de datos:

- **Métodos bivariantes**: Permiten evaluar la relación entre dos variables (normalmente un factor y una variable respuesta). Los tipos de pruebas bivariantes que se desarrollan en este manual dependen de la naturaleza de las variables implicadas y se resumen en la Tabla 1.

- **Métodos multivariantes**: El análisis implica manejar al mismo tiempo tres o más variables. Este tipo de pruebas no se incluyen en este manual.

Tabla IV.1. *Resumen de los métodos estadísticos bivariantes en función de la naturaleza de las variables y del tipo de distribución (normal o paramétrica o no)*

Variable 1 (dependiente)	**Variable 2 (independiente)**		**Los datos siguen distribución normal y/o tienen homogeneidad de varianzas**	
			SI	**NO**
Cualitativa	Cualitativa		-	Test de la χ^2 (tablas de contingencia)
Cuantitativa	Cualitativa	2 categorías	t-Student	U de Mann-Whitney/ Wilcox
		> 2 categorías	Análisis de la varianza (ANOVA)	Kruskal-Wallis
Cuantitativa	Cuantitativa	Se asume que una var. es causa de la otra	Regresión	-
		No se asume relación causa-efecto	Correlación de Pearson	Correlación de Spearman

2. ASOCIACIÓN ENTRE VARIABLES CUALITATIVAS: TEST DE LA χ^2

El test de la χ^2 se utiliza para analizar la asociación entre dos **variables cualitativas** (por ejemplo, la presencia/ausencia de una especie y el tipo de suelo, color de la flor y presencia/ausencia de polinizadores, etc.). Este test parte de una tabla de contingencia, donde las columnas indican las categorías de la variable A y las filas las categorías de la variable B. En cada celda se anota la frecuencia de observaciones correspondiente. El test compara las frecuencias observadas con las frecuencias esperadas en caso de que no existiera asociación (es decir, si las observaciones se distribuyen al azar entre las categorías de las variables).

2.1. Requisitos e hipótesis de trabajo

La aplicación de este test requiere que las muestras estén tomadas al azar y que las frecuencias esperadas sean superiores a 5. Como se trata de un test que relaciona variables cualitativas, no hay ningún requisito acerca de la distribución de las variables.

Las hipótesis de trabajo serán del tipo:

- H_{ecol}: Existe asociación entre las variables (por ej. esperamos mayor frecuencia de polinizadores en las flores amarillas que en las azules)
- H_0: Las dos variables son independientes (por ej. a los polinizadores no les importa el color de las flores y aparecerán con la misma frecuencia en las amarillas y en las azules)

2.2. Contraste de hipótesis

Se compara el valor obtenido de χ^2_{cal} con el valor χ^2_{crit} correspondiente al número de grados de libertad apropiados y al valor de α previamente seleccionado (normalmente, α=0.05 ó 0.01):

Si $\chi^2_{cal} \geq \chi^2_{crit}$, se rechaza la H_0 (hay asociación entre las variables)

Si $\chi^2_{cal} < \chi^2_{crit}$, se acepta la H_0 (no hay asociación entre las variables)

2.3 Procedimiento de cálculo de la $\chi 2$

Supongamos, por ejemplo, que queremos saber si existe asociación entre la presencia de la especie A (un invertebrado acuático) y el tramo del río (alto, medio y bajo) para el caso del río Henares. Nuestras hipótesis son:

- H_{ecol}: Existe relación entre la presencia de la especie A y el tramo del río
- H_0: La presencia de la especie A es independiente del tramo del río

Para comprobar cuál se cumple hemos hecho un muestreo a lo largo del río y en cada tramo hemos registrado la presencia (+) o ausencia (-) de la especie en 15 muestras de agua tomadas al azar. Los resultados se muestran en la Tabla IV.2. A partir de estos datos construiríamos una tabla de contingencia con los datos observados en campo (Tabla IV.3)

Tabla IV.2: *Presencia (+) o ausencia (-) en cada una de las 15 réplicas de agua tomadas en cada tramo del río Henares.*

Tramo Alto	Tramo Medio	Tramo Bajo
+	-	-
+	-	+
+	-	-
-	+	-
+	-	-
+	-	-
+	-	-
+	-	-
+	+	-
+	-	-
-	-	-
+	-	-
+	-	-
+	-	-
+	-	-

Tabla IV.3. *Tabla de contingencia que muestra los valores observados de frecuencia de la especie A en cada tramo del río, según la tabla 2.*

		Tramo del río		
		Alto	**Medio**	**Bajo**
Especie A	+	13	2	1
	-	2	13	14

A continuación se calcula el estadístico χ^2_{cal} siguiendo la siguiente fórmula:

$$\chi^2_{(\alpha,gl.)} = \sum \frac{(o-e)^2}{e}$$

- o = frecuencias observadas en el inventario
- e = frecuencia esperada de una celda, suponiendo que no hubiese asociación

$$e = \frac{c_t * f_t}{N}$$

c_t = total de la columna donde está la celda
f_t = total de la fila donde está la celda
N = nº total de casos

- *gl.* (grados de libertad) = (nº columnas-1)*(nº filas-1)

Para calcular el estadístico χ^2_{cal} conviene añadir a la tabla de contingencia las frecuencias esperadas en cada celda (entre paréntesis), como se indica en la Tabla 4.

Tabla IV.4. *Tabla de contingencia que muestra la frecuencia de la especie A observada en cada tramo del río y la frecuencia esperada en caso de independencia entre variables (entre paréntesis)*

		Tramo del río			
		Alto	**Medio**	**Bajo**	**Total**
Especie A	+	13 (*5.3*)	2 (*5.3*)	1 (*5.3*)	16
	-	2 (*9.7*)	13 (*9.7*)	14 (*9.7*)	29
Total		15	15	15	45

$$\chi^2_{cal} = \frac{(13-5{,}3)^2}{5.3} + \frac{(2-5{,}3)^2}{5.3} + \frac{(1-5{,}3)^2}{5.3} + \frac{(2-9{,}7)^2}{9.7} + \frac{(13-9{,}7)^2}{9.7} + \frac{(14-9{,}7)^2}{9.7} = 25.8$$

$\chi^2_{\text{crít (2 g.l., } \alpha=0.05)} = 5.99$

$\chi^2_{\text{cal}} > \chi^2_{\text{crít}}$ ($p<0.05$)

Análisis en R del test de la χ^2):

```
#Primero creamos la tabla de contingencias, que corresponde a nuestros datos:
rio<-matrix(c(13,2,1,2,13,14),byrow=TRUE,ncol=3)
colnames(rio)=c("Alto","Medio","Bajo")
rownames(rio)=c("Presente","Ausente")
rio #Para ver si la tabla está bien
          Alto Medio Bajo
Presente   13     2    1
Ausente     2    13   14

#Aplicar el test de la chi-cuadrada
chisq.test(rio)
Pearson's Chi-squared test
data:  rio
X-squared = 25.797, df = 2, p-value = 2.501e-06
```

En consecuencia, se rechaza H_0 con una probabilidad de equivocarnos = 2.5×10^{-6} y concluimos que la especie A aparece preferentemente en los tramos altos del río, ya que es en éstos donde su frecuencia observada es mayor que la esperada.

Caso especial: En las tablas de contingencia de 2x2, como la de la Tabla IV.5, el estadístico χ^2_{cal} se puede calcular con las fórmulas que aparecen debajo.

Tabla IV.5. *Tabla de contingencia de 2 x.2*

		Variable 1		
		A	**B**	**Total filas**
Variable 2	+	(a)	(b)	(a+b)
	+	(c)	(d)	(c+d)
Total columnas		(a+c)	(b+d)	(a+b+c+d)

Si $N \geq 30$

$$\chi^2_{cal} = \frac{(a*d - b*c)^2 * N}{(a+b)*(c+d)*(a+c)*(b+d)}$$

Si $N < 30$ (Corrección de Yates)

$$\chi^2_{cal} = \frac{N*(|a*d - b*c| - N/2)^2}{(a+b)*(c+d)*(a+c)*(b+d)}$$

3. TESTS DE COMPARACIÓN DE DOS MEDIAS

Sirven para comparar la media o mediana de una variable respuesta cuantitativa entre dos grupos definidos por dos categorías de una variable independiente cualitativa. Por ejemplo, si queremos comparar el peso corporal de los conejos entre una población que vive en un retamar y otra que vive en una pradera sin retamas. En ese caso, la variable independiente cualitativa es la población (retamar o pradera) y la variable dependiente cuantitativa es el peso corporal.

3.1. Selección del test

Para seleccionar el test apropiado debemos saber si los valores de la variable respuesta cuantitativa siguen una distribución normal dentro de cada grupo. Esto se puede comprobar visualmente dibujando un histograma, un gráfico qq, o un test como Shapiro-Wilk (ver sección 2.1). Asimismo, han que comprobar si las varianzas de ambos grupos son similares, con la F de Snedecor (Cuadro IV.1) o con el test de Barlett (ver sección 1.2).

Si la variable cuantitativa sigue la distribución normal y las varianzas de ambos grupos son iguales, se utilizará el test paramétrico: **t de Student.** En cualquier otro caso se realizará el test no paramétrico: **U de Mann-Whitney**

3.2. Hipótesis de una o dos colas

Cuando la hipótesis ecológica establece que existen diferencias entre las medias (o medianas) de los dos grupos, sin presuponer cuál de las dos medias es mayor que la otra, se dice que la hipótesis es de "dos colas", ya que incluye dos posibilidades (que la media del grupo A sea mayor que la del B o viceversa).

H_{ecol}: $\mu_1 \neq \mu_2$

H_0: $\mu_1 = \mu_2$

Por el contrario, so la hipótesis ecológica establece que una de las dos medias es mayor que la otra, la hipótesis es de una cola, porque solo incluye una posibilidad. En este caso la hipótesis nula es la que abarca dos posibilidades (que la diferencia de las medias vaya en sentido contrario al esperado o que las medias sean iguales).

H_{ecol}: $\mu_1 > \mu_2$

H_0: $\mu_1 \leq \mu_2$

Es importante establecer esta diferencia, porque el resultado del test es distinto para una o para dos colas.

3.3. Cálculo de la t de Student para muestras independientes

Este es el cálculo que tenemos que aplicar cuando las muestras tomadas en las dos situaciones definidas por la variable categórica son independientes, es decir, no hay unas que a priori sean más similares a otras.

Si los datos cumplen los requisitos establecidos, se puede calcular el estadístico *tcal* a partir de la siguiente fórmula:

$$t_{cal} = \frac{\bar{x}_1 - \bar{x}_2}{Sc\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad \text{donde:} \qquad Sc = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}}$$

n_1 y n_2= tamaños de las muestras 1 y 2 respectivamente

$\bar{x}_1$ y $\bar{x}_2$ = medias de las muestras 1 y 2 respectivamente

s_1^2 y s_2^2 = varianzas de las muestras 1 y 2 respectivamente

A continuación se mide la significación del estadístico t_{cal}, comparando ese valor con el valor de un estadístico t_{crit} que se obtiene mirando las tablas correspondientes. Para identificar el t_{crit} que nos corresponde hemos de fijarnos en el número de colas que tiene nuestra hipótesis (una cola: *one-tailed*; dos colas: *two-tailed*), en el nivel de significación (α) con el que pretendemos rechazar la hipótesis nula (normalmente α = 0.05) y en los grados de libertad del test ($n_1 + n_2 - 2$).

- Si $|t_{cal}| \geq t_{crit}$ (α=0.05 o inferior) $\Rightarrow$ se rechaza H_0 y se acepta H_{ecol} (las medias son diferentes)
- Si $|t_{cal}| < t_{crit}$ (α=0.05) $\Rightarrow$ se acepta H_0 y se rechaza H_{ecol} (las medias son iguales)

Cuadro IV.2: Ejemplo de cálculo de la t de Student

Queremos saber si la humedad del suelo en un determinado lugar varía en función de la cubierta vegetal del mismo (tomillar o suelo desnudo), pues suponemos que la cubierta vegetal contribuye a aumentar la humedad del suelo por disminución de la evaporación. Para ello se ha realizado un muestreo en el que se ha medido la humedad de suelo (en % del volumen) en seis muestras distribuidas al azar bajo tomillares y en 8 muestras también distribuidas al azar en la misma zona, pero en condiciones de suelo desnudo.

Variables:

- Cobertura de suelo (cualitativa, independiente)
- Humedad del suelo (cuantitativa, dependiente)

Hipótesis

- H_{ecol}: la humedad de suelo es mayor bajo el tomillar: $\mu_{tomillar} > \mu_{suelo\ desnudo}$ (una cola).
- H_0: $\mu_{tomillar} \leq \mu_{suelo\ desnudo}$

Tabla de datos:

Cobertura	Humedad de suelo (%)								n	Media	s^2
tomillar	73.0	74.2	75.0	75.3	75.5	75.8			6	74.8	1.04
suelo desnudo	71.0	71.5	72.0	72.4	73.5	74.0	74.3	75.2	8	72.9	2.20

Cálculos:

$$t_{cal} = \frac{74.8 - 72.9}{1.42\sqrt{\frac{1}{6} + \frac{1}{8}}} = 2.36$$

$t_{cal} = 2.36 > t_{crít\ (\alpha=0.05,\ 12\ gl,\ una\ cola)} = 1.782$

Interpretación:
Se rechaza la H_0, y se acepta la H_{ecol}, es decir, se concluye que existen diferencias significativas en la humedad del suelo en función de la cobertura vegetal, siendo mayor en condiciones de cubierta vegetal de tomillar que en condiciones de suelo desnudo.

Análisis en R del test paramétrico de comparación de dos medias (t de student)

```
#Creamos un vector con datos para cada una de las dos situaciones que queremos comparar
tomillar<-c(73.0,74.2,75.0,75.3,75.5,75.8)
suelo<-c(71.0,71.5,72.0,72.4,73.5,74.0,74.3,75.2)

#Representamos gráficamente los datos para ver dónde parece que hay mayor humedad
boxplot(tomillar, suelo, ylab="humedad del suelo", col="tan", names=c("tomillar",
"suelo"))
#La figura resultante es la Fig. IV.3
#REQUISITOS PARA APLICAR LA T-STUDENT
#Comprobamos la normalidad de la variable en "tomillar" y en "suelo".
#Lo podemos hacer visualmente con histogramas y gráfico qq.
#También podemos aplicar el test de Shapiro –Wilk (ver scripts en sección 1.1)
#Analizamos si las varianzas de la variable en tomillar y en suelo son iguales.
#Aplicamos el test de Barlett (ver sección 1.1).
#Como se cumplen todos los requisitos, aplicamos la t de Student
t.test(tomillar, suelo)
Welch Two Sample t-test
data:  tomillar and suelo
t = 2.6896, df = 11.977, p-value = 0.01971
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval: 0.3438945 3.281105
sample estimates:  mean of x          mean of y
                         74.8000   72.9875
```

Los datos de humedad de suelo siguen una distribución normal con forma de campana de Gauss, tanto en el tomillar como en suelo desnudo (Figura IV.3a,b). Además, no muestran desviaciones notable de la normalidad en el gráfico qq. Igualmente, los test de Shapiro, que testan normalidad, sugieren que en ninguna de las dos situaciones (cobertura de tomillar y suelo desnudo) los datos presentan desviaciones de la normalidad (p>0.05). Cuando hemos realizado el test paramétrico t de Student, la variable dependiente "humedad del suelo" difiere significativamente entre las dos categorías de la variable independiente (tomillar y suelo desnudo) (P < 0.001).

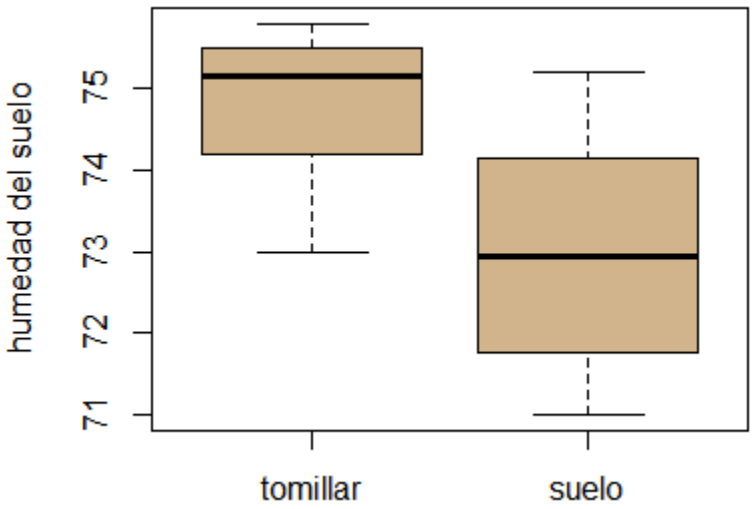

Figura IV.3. *Valores medios y desviaciones de la variable dependiente "humedad del suelo" en cada categoría de la variable independiente: tomillar y suelo desnudo.*

3.4. Cálculo de la t de Student para muestras pareadas

Los test de comparación de medias pareadas son necesarios cuando las muestras tomadas no son igualmente independientes entre sí. Por ejemplo, se quiere comparar la variable "riqueza de especies" entre dos situaciones: lugares invadidos por una especie invasora (+I) y lugares no invadidos (-I). El muestreo se realiza a lo largo de un universo de muestreo muy heterogéneo (por ej. con distintos tipos de sustrato, o distinta altitud sobre el nivel del mar). Por tanto, cabe esperar que dos muestras lejanas en zonas invadidas sean más diferentes entre sí que dos muestras cercanas, una en +I y otra en –I. Para evitar que el efecto de la heterogeneidad ambiental diluya el efecto que la invasión tiene sobre la riqueza de especies, diseñamos un muestreo pareado, de forma que cada muestra en +I se compara con un control próximo en –I (Fig. IV.3). A continuación se estimará la diferencia de riqueza entre cada muestra y su control y se realizará una t de Student para ver si la media de las diferencias difiere o no de cero.

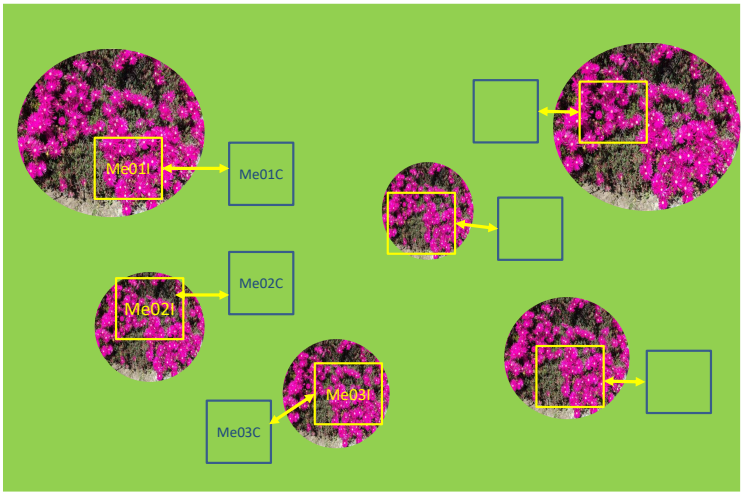

Figura IV.4. *Representación de un universo de muestreo en el que una comunidad ha sido invadida por una especie invasora (manchas de flores rosas). Para evitar la interferencia de factores ambientales no deseados, se ha diseñado un muestreo pareado, donde cada muestra en un lugar invadido (cuadros amarillos) será comparada con una muestra control en un lugar no invadido (cuadros azules).*

Cuadro IV.3: Ejemplo de cálculo de la t de Student con medias pareadas

Queremos saber si un tratamiento realizado durante un año en una población de pinares afectada por procesionaria ha tenido un impacto en la cantidad de cobertura arbórea. Por ello la cobertura arbórea se ha medido dos veces: antes y después del tratamiento en 10 sitios distintos. Esto nos proporciona 10 valores antes del tratamiento y 10 valores después del tratamiento, midiendo dos veces la cobertura arbórea del mismo sitio. Por ello, aquí debe usarse una t de Student de medias pareadas para comparar la media antes y después del tratamiento (las muestras correspondientes al mismo sitio están más relacionadas entre sí que muestras de sitios distintos).

Variables:

- Cobertura arbórea (cuantitativa, dependiente).
- Tratamiento (cualitativa, independiente con dos categorías: antes/después).

Hipótesis

- H_{ecol}: la cobertura arbórea es mayor después del tratamiento que antes del tratamiento: $cobertura_{después} > cobertura_{antes}$ (una cola).
- H_0: $cobertura_{después} \leq cobertura_{antes}$.

Tabla de datos:

Tratamiento	Cobertura arbórea (%)	n	Media	s^2
Antes	61.98777 63.72664 59.74309 62.70319 61.67165 61.40515 60.56058 61.16024 60.47696 61.69815	10	61.5	1.54
Después	96.94371 95.31954 99.16069 96.24361 96.30076 96.52271 97.83126 95.60667 94.57118 97.35336	10	95.9	1.20

Cálculos:

$$t = \frac{m}{s \times \sqrt{n}}$$

Donde, m es la media de las diferencias entre cada par de muestras, s es la desviación estándar de las diferencias y n es el tamaño de las diferencias.

$$t = \frac{-35.0720}{2.1099 \times \sqrt{10}} = -49.04632$$

$t_{cal} = -52.56 > t_{crít}$ (α=0.05, 9 gl, una cola)

Interpretación:
Se rechaza la H_0, y se acepta la H_{ecol}, es decir, se concluye que existen diferencias significativas en la cobertura arbórea antes y después de aplicar el tratamiento, siendo mayor después de aplicar el tratamiento.

Comparación de dos medias pareadas (t de Student) en R

```
#Creamos dos vectores con datos para cada una de las dos situaciones que queremos
#comparar
antes<-c(61.98777,63.72664,59.74309,62.70319,61.67165,
         61.40515,60.56058,61.16024,60.47696,61.69815)
despues<-c(96.94371,95.31954,99.16069,96.24361,96.30076,
           96.52271,97.83126,95.60667,94.57118,97.35336)

#Representamos gráficamente los valores medios y las desviaciones
boxplot(antes, despues, ylab="Cobertura de árboles", col= "tan",
        names=c("antes", "después")) #La figura resultante es la Fig. IV.5
```

```
#Comprobamos que se cumplen los requisitos de normalidad y homoscedasticidad (ver sección
1.1).
#Test de comparación de medias pareado
t.test(antes, despues, paired = TRUE)
data:  antes and despues
t = -52.564, df = 9, p-value = 1.64e-12
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval: -36.58138 -33.56263
sample estimates: mean of the differences -35.07201
```

La figura IV.5 sugiere que hay una gran diferencia de cobertura arbórea antes y después del tratamiento. Cuando hemos realizado el test paramétrico t de Student de medias pareadas se puede concluir que las medias muestran unas diferencias altamente significativas ($p << 0.001$).

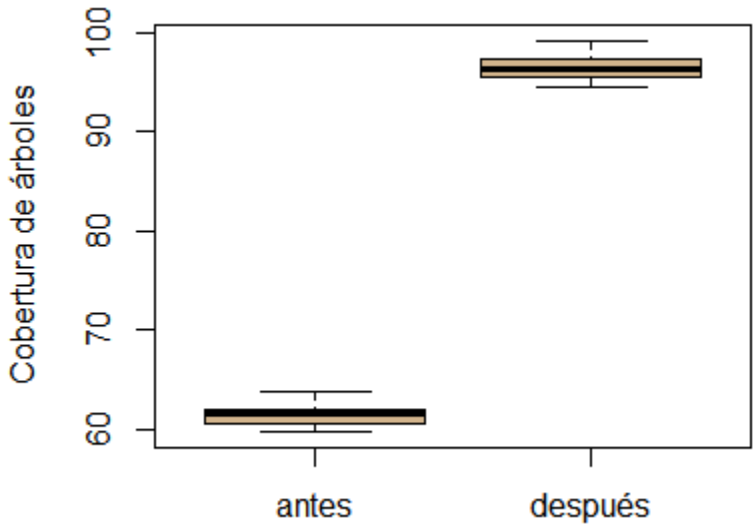

***Figura IV.5**. Valores medios y desviación de cobertura arbórea antes y después del tratamiento aplicado para controlar la procesionaria del pino.*

3.5. Comparación no paramétrica para muestras independientes

Si nuestros datos no cumplen los requisitos necesarios para aplicar una t-Student, podemos aplicar una U de Mann-Whitney, que compara las diferencias entre dos medianas (en lugar de las medias). Este test se basa en rangos (número de orden de los datos en función de su magnitud) en lugar de en los parámetros de la muestra (media, varianza), por ello se dice que es un test no paramétrico.

Los pasos a seguir para calcular la U de Mann-Whitney son:

1. Asignación de rangos a cada dato: se ordenan todos los valores (juntando los dos grupos) de menor a mayor. El rango de cada dato será el número de orden que le corresponde a cada dato. Cuando se repita el mismo valor numérico, el rango que se asigna a esos datos es la media aritmética de los rangos que les corresponderían en función del número de orden que ocupan.
2. Se suman los rangos de cada uno de los grupos que se comparan y se calcula la suma de los rangos de los datos de cada uno de los grupos (R_1 y R_2)

3. Se calculan los estadísticos U_1 y U_2 a partir de las siguientes fórmulas:

$$U_1 = n_1 \cdot n_2 + \frac{n_2(n_2+1)}{2} - R_2 \qquad U_2 = n_1 \cdot n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

   Se obtiene el estadístico U_{cal} escogiendo el valor más grande entre U_1 y U_2 .
4. Se comprueba la significación estadística del estadístico U_{cal} comparando este valor con el valor de un estadístico $U_{crít}$ obtenido a partir de las tablas correspondientes.
5. Si $U_{cal} \geq U_{crít}$ (α=0.05 o inferior) $\Rightarrow$ se rechaza H_0 y se acepta H_{ecol} (las medianas son diferentes)
6. Si $U_{cal} < U_{crít}$ (α=0.05) $\Rightarrow$ se acepta H_0 y se rechaza H_{ecol} (las medianas son iguales)

Cuadro IV.4: Ejemplo de cálculo de la U de Mann-Whitney

Se quiere estudiar si el número de especies de ácaros edáficos se ve influido por un incendio de baja intensidad. Para ello se simuló un incendio de baja intensidad en una parcela de un territorio homogéneo y se tomaron 6 muestras al azar de la zona incendiada y 7 muestras también al azar de la zona no incendiada, contándose el número de especies de ácaros edáficos en cada muestra.

Variables:
- Variable dependiente: número de especies de ácaros edáficos (cuantitativa)
- Variable independiente: ocurrencia de un incendio (cualitativa)

Hipótesis:
- H_{ecol}= La mediana del número de especies de ácaros edáficos varía dependiendo de que se haya producido un incendio: $M_{quemada} \neq M_{no\ quemada}$. (dos colas).
- H_0= La mediana del número de especies de ácaros edáficos es igual en la parcela quemada que en la no quemada: $M_{quemada} = M_{no\ quemada}$

Tabla de datos:

Parcela	Número de especies de ácaros edáficos	n
quemada	6 9 12 12 15 16	6
no quemada	10 13 16 16 17 19 20	7

Asignación de rangos a cada dato:

dato *	6	9	10	12	12	13	15	16	16	16	17	19	20
rango	1	2	3	4’5	4’5	6	7	9	9	9	11	12	13

* en negrita los valores correspondientes al inventario de la parcela quemada

Cálculo de U_{cal}
- - Se suman los rangos de cada grupo: R_1=28 R_2=63
- U_1=6x7+[(7x8)/2]-63=7
- U_2=6x7+[(6x7)/2]-28=35 ➔ U_{cal}
- U_{cal} = 35 < $U_{cít\ (\alpha=0.05)}$ = 36

Interpretación:
No se rechaza la H_0, concluimos que el número de especies de ácaros edáficos no se ve influido significativamente por la ocurrencia de un incendio de baja intensidad.

Si queremos realizar el test en RStudio, podemos aplicar el test de Wilcox.

Realización en R del test no paramétrico de comparación de dos muestras independientes

```
#Creamos dos vectores con datos para cada una de las dos situaciones que queremos
#comparar
quemada<-c(0,0,12,12,15,15)
noquemada<-c(10,13,16,16,17,19,20)

#Antes de nada, conviene representar gráficamente los datos para ver qué tendencias
# muestran
boxplot(quemada, noquemada, ylab="riqueza de ácaros", names=c("quemada", "no quemada"))

#Test de normalidad
#Podemos comprobar visualmente la normalidad con histogramas y gráficos qq (sección 1.2
#y figura IV.6)
#También podemos aplicar el test de normalidad de Shapiro
shapiro.test(quemada)
Shapiro-Wilk normality test
data:  quemada
W = 0.76178, p-value = 0.02591
shapiro.test(noquemada)
Shapiro-Wilk normality test
data:  noquemada
W = 0.94671, p-value = 0.6997
#El resultado sugiere que la variable no es normal en la categoría “quemada”

#Test de comparación de medias
wilcox.test(quemada, noquemada)
Wilcoxon rank sum test with continuity correction
data:  quemada and noquemada
W = 6, p-value = 0.03726
alternative hypothesis: true location shift is not equal to 0
```

El número de especies de ácaros parece ser mayor en la madera no quemada (Figura IV.5). Dado que la variable sigue una distribución no normal en la categoría “quemada” aplicamos el test de Wilcox. El resultado indica que las diferencias que muestra la figura sí son significativas (P > 0.05).

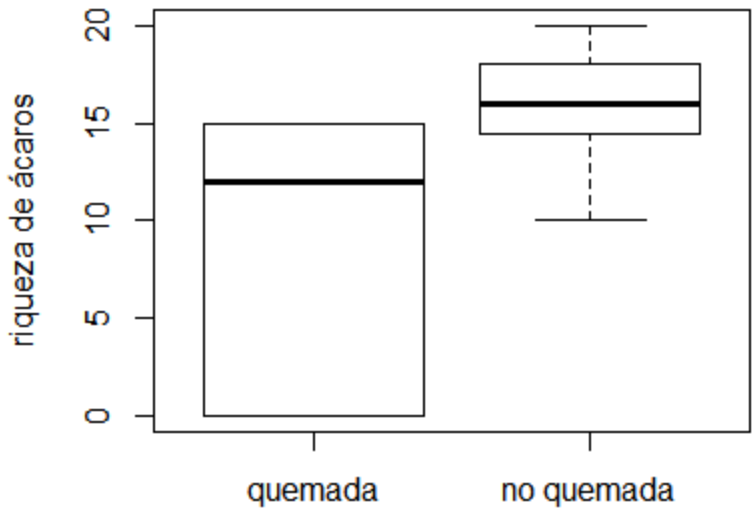

Figura IV.6*. Valores medios y desviación de la riqueza de ácaros encontrada en la zona quemada y no quemada.*

3.6. Comparación no paramétrica para muestras pareadas

Vamos a suponer ahora que en el ejemplo anterior queremos comparar la riqueza de ácaros en los mismos puntos de muestreo antes y después de un incendio. En este caso las muestras están pareadas, es decir, quiero comparar cada valor de riqueza en el mismo sitio antes y después del incendio.

Realización en R del test no paramétrico de comparación entre dos muestras pareadas

```
#Creamos dos vectores con datos para cada una de las dos situaciones que queremos
#comparar
antes<-c(10,10,16,19,20,20,24)
despues<-c(0,1,9,12,15,15,15)
#Represento la media y la dispersión de los datos para ver qué tendencia siguen
boxplot(antes, después, ylab=”riqueza de ácaros”, names=c(“antes”, “después”)
#La figura resultante es la IV.7
#Test de normalidad
#Podemos comprobar visualmente la normalidad con histogramas y gráficos qq (sección 1.2
#y figura IV.5)
#También podemos aplicar el test de normalidad de Shapiro
shapiro.test(antes)
Shapiro-Wilk normality test
data:  antes
W = 0.89258, p-value = 0.2884
shapiro.test(despues)
Shapiro-Wilk normality test
data:  despues
W = 0.80237, p-value = 0.04324
#El resultado sugiere que la variable no es normal en la categoría “despues”

#Test de comparación de medias
wilcox.test(antes, despues, paired= TRUE)
Wilcoxon signed rank test with continuity correction
data:  antes and despues
V = 28, p-value = 0.02178
alternative hypothesis: true location shift is not equal to 0
```

Los resultados indican que la riqueza de ácaros en los mismos puntos disminuyó tras el incendio (figura IV.6), y que la diferencia es significativa, ya que el resultado del test estadístico no da una p=0.02178, que se menor de 0.05.

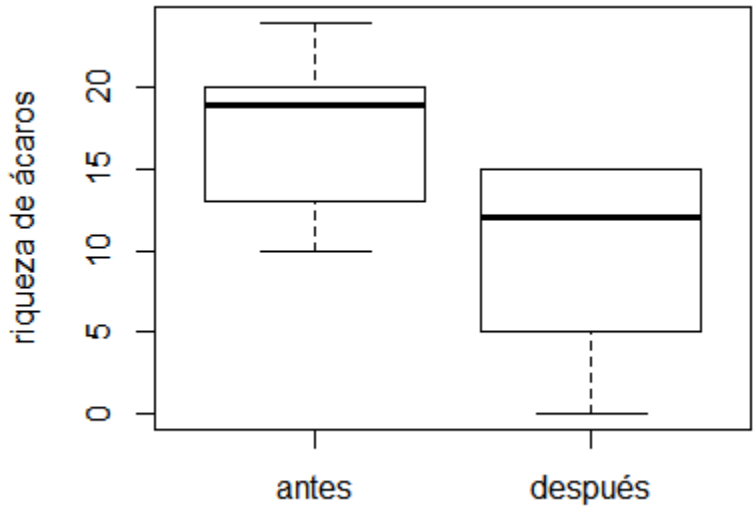

Figura IV. 7. *Mediana y desviación de la riqueza de ácaros en varios puntos de muestras antes y después de un incendio.*

4. TESTS DE COMPARACIÓN DE MÁS DE DOS MEDIAS

Sirven para comparar las medidas de tendencia central (media o mediana) entre más de dos grupos de datos para determinar si existen o no diferencias entre ellos. Por tanto relacionan una variable cualitativa de más de dos estados (variable independiente) con otra cuantitativa (variable dependiente). Por ejemplo, habría que aplicar este test para determinar si existen diferencias significativas en la densidad de escarabajos (variable dependiente, cuantitativa) que encontramos en cuatro tipos de suelo (variable independiente, cualitativa con cuatro estados).

4.1. Selección del test

Al igual que en la comparación de dos medias, para elegir el test hay que comprobar si los datos siguen una distribución normal dentro de cada grupo, y si las varianzas entre grupos son homogéneas (ver sección 1.1). Si ambos requisitos se cumplen se utilizará el test paramétrico: **ANOVA.** En cualquier otro caso se realizará el test no paramétrico: **Kruskal-Wallis**

4.2. Hipótesis

La hipótesis ecológica establece que existen diferencias entre las medias (o medianas, en el caso del test no paramétrico) de los grupos considerados, es decir, que **al menos** dos de las medias serán distintas. La hipótesis nula establece que no existen diferencias entre dichas medias.

H_{ecol}: No todas las medias/medianas son iguales

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

Si rechazamos la hipótesis nula, significa que al menos dos de los grupos difieren entre sí. Sin embargo, este test no indica qué grupos difieren o son iguales de qué grupos. Si quisiéramos conocer

las diferencias entre todos los pares de grupos posibles, tendríamos que aplicar algún test "post-hoc", que no vemos en este manual. En cualquier caso, es incorrecto aplicar varios test de Student con este fin.

4.3. Procedimiento de cálculo del ANOVA

La valoración de las diferencias entre las medias de los distintos grupos se basa en la descomposición de la variabilidad total del conjunto de datos en dos términos: variabilidad debida a las diferencias entre los grupos (variabilidad entre grupos), y variabilidad debida al azar (variabilidad dentro de grupos).

Variabilidad$_{total}$ = Variabilidad$_{entre\ grupos}$ + Variabilidad$_{dentro\ grupos}$

La variabilidad entre datos se puede estimar con la varianza (s^2), y con Suma de Cuadrados (SS), que es el cociente entre la varianza y los grados de libertad (g-l.). Por tanto:

SS$_{total}$ = SS$_{entre\ grupos}$ + SS$_{dentro\ grupos}$

Las sumas de cuadrados se obtienen a partir de las siguientes fórmulas:

$$SS_{total} = \sum x^2 - \frac{\left(\sum x\right)^2}{N}$$

$$SS_{entre\ grupos} = \left[\frac{\left(\sum x_1\right)^2}{n_1} + \frac{\left(\sum x_2\right)^2}{n_2} + \ldots + \frac{\left(\sum x_k\right)^2}{n_k}\right] - \frac{\left(\sum x\right)^2}{N}$$

Donde

k = número de grupos

N= número total de datos

n_1, n_2, ... n_k = número de datos en cada grupo.

x = cada uno de los datos de cada grupo

El cálculo de la suma de cuadrados se obtiene despejando de la ecuación:

SS$_{dentro\ grupos}$ = SS$_{total}$ – SS$_{entre\ grupos}$

- Cálculo de los grados de libertad de las sumas de cuadrados:

$$g.l.\ SS_{total} = N - 1 \qquad g.l.\ SS_{entregupos} = k - 1 \qquad g.l.\ SS_{dentrogrupos} = N - k$$

- Conversión de las sumas de cuadrados (SS) en varianzas:

$$s^2_{entregrupos} = \frac{SS_{entre\ grupos}}{g.l._{entre\ grupos}} = \frac{SS_{entre\ grupos}}{k-1} \qquad s^2_{dentrogrupos} = \frac{SS_{dentro\ grupos}}{g.l._{dentro\ grupos}} = \frac{SS_{dentro\ grupos}}{N-k}$$

- Cálculo del estadístico F:

$$F = \frac{s^2_{entregrupos}}{s^2_{dentrogrupos}}$$

Si en la población de la que proceden las muestras no hay diferencias reales entre los grupos definidos por la variable cualitativa, la varianza entre grupos será similar a la varianza dentro de grupos (por tanto, el cociente entre ambas estará cerca de 1). En el caso de que existan diferencias reales entre los grupos (lo que presupone la hipótesis ecológica) la varianza entre grupos será mayor que la varianza dentro de los grupos (el cociente entre ambas será mayor de 1). El estadístico que nos dice si las desviaciones respecto a ese valor de 1 son significativas es *F*.

El contraste de hipótesis se realiza comparando el valor de la F_{cal} con el valor $F_{crít}$ obtenido a partir de la tabla para el valor de α previamente establecido (normalmente α=0.05 o inferior). La búsqueda de dicha $F_{crít}$ requiere del número de grados de libertad del numerador y del denominador. La forma habitual de notación que se usa en las tablas lleva el valor de α entre paréntesis, y los grados de libertad del numerador y del denominador a continuación, en orden consecutivo y separados por comas. Por ejemplo, $F_{crít\ (0.05)\ 3,\ 22}$. significa el valor del estadístico F de las tablas para un α=0.05, con 3 grados de libertad en el numerador y 22 en el denominador.

- Si $F_{cal} \geq F_{crít} \Rightarrow$ se rechaza H_0 y se acepta H_{ecol} (alguna de las medias es diferente)
- Si $F_{cal} < F_{crít} \Rightarrow$ se acepta H_0 y se rechaza H_{ecol} (las medias son iguales)

Cuadro IV.5: Ejemplo de cálculo de ANOVA

Se quiere saber si el tipo de cobertura de suelo (suelo desnudo, piedras, hojarasca y pastizal) influye sobre la densidad de hormigueros. Para ello se ha realizado un muestreo en el que se ha medido el número de hormigueros en diez muestras distribuidas al azar dentro de cada una de las zonas con diferente cobertura.

Variables:
- cobertura de suelo (cualitativa, independiente)
- densidad de hormigueros (cuantitativa, dependiente)

Hipótesis:
- H_{ecol}: Alguna de las medias es diferente (la cobertura de suelo influye sobre la densidad de hormigueros)
- H_0: $\mu_{suelo\ desnudo} = \mu_{piedras} = \mu_{hojarasca} = \mu_{pastizal}$

Tabla de datos:

Cobertura	Densidad de hormigueros	n	Media	Σx	$(\Sigma x)^2$	Σx^2
suelo desnudo	78 88 87 88 83 82 81 80 80 89	10	83.6	836	698896	70036
piedras	78 78 83 81 78 81 81 82 76 76	10	79.4	794	630436	63100
hojarasca	79 73 79 75 77 78 80 78 83 84	10	78.6	786	617796	61878
pastizal	77 69 75 70 74 83 80 75 76 75	10	75.4	754	568516	57006
Total		40		3170		252020

Cálculo de la suma de cuadrados total:
$SS_T = 252020 - (3170)^2/40 = 797.5$

Cálculo de la variabilidad entre grupos ($SS_{entre\ grupos}$):
$SS_{entre} = 698896/10 + 630436/10 + 617796/10 + 568516/10 - 3170^2/40 = 341.9$

Cálculo de la variabilidad dentro de los grupos ($SS_{dentro\ grupos}$):
$SS_T = SS_{entre} + SS_{dentro}$ ⇒ $SS_{dentro} = SS_{total} - SS_{entre} = 797.5 – 341.9 = 455.6$

Determinar los grados de libertad de cada una de las sumas de cuadrados estimadas:
$SS_T = N - 1 = 40 - 1 = 39$ $SS_{entre\ grupos} = k - 1 = 4 - 1 = 3$ $SS_{dentro\ grupos} = N - k = 40 - 4 = 36$

Estimación de las varianzas dividiendo las SS por los grados de libertad:
$s^2_{entre\ grupos} = 341.9/3 = 113.97$ $s^2_{dentro\ grupos} = 455.6/36\ 12.66$

Cálculo del estadístico F_{cal} y comparación con el estadístico $F_{crít}$:
$F_{cal} = s^2_{entre\ grupos} / s^2_{dentro\ grupos} = 113.97/12.66 = 9.002$
$F_{crít\ (0.05)\ 3,\ 36} < 2.92$

Interpretación: $F_{cal} > F_{crít}$ ⇒ Rechazamos H_o La abundancia de hormigueros no es la misma en todas las zonas

Realización en R del test ANOVA

```
#Creamos la matriz con los datos con dos columnas, en una tenemos la variable dependiente
“densidad” y en otra las categorías de la variable independiente “cobertura”
cobertura<-c("desnudo","piedras","hojarasca","pastizal")
cobertura<-rep(cobertura,each=10)
hormigueros<-data.frame(cobertura,
"densidad"=c(78,88,87,88,83,82,81,80,80,89,78,78,83,81,78,81,81,82,76,76,79,73,79,75,77
,78,80,78,83,84,77,69,75,70,74,83,80,75,76,75))

#Analizamos la estructura de los datos
str(hormigueros)
'data.frame':      40 obs. of  2 variables:
 $ cobertura: Factor w/ 4 levels "desnudo","hojarasca",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ densidad : num  78 88 87 88 83 82 81 80 80 89 ...

#Representamos gráficamente los valores medios de densidad de hormigueros en cada tipo
de cobertura de suelo (Fig. IV.8)
boxplot(densidad ~ cobertura,   data=hormigueros, col="tan", cex.axis=0.7, las = 2,
ylab="Densidad de hormigueros", cex.lab=0.75)
#Añadimos al boxplots los datos individuales para ver cómo se distribuyen
stripchart(densidad ~ cobertura,  data=hormigueros, col="red",
           vertical = TRUE, method = "jitter", cex=0.5,
           add=TRUE, pch=19)

#ANOVA
modelo <- lm(densidad ~ cobertura,  data=hormigueros)
anovaModelo <- anova(modelo)
anovaModelo
Analysis of Variance Table
Response: densidad
          Df Sum Sq Mean Sq F value   Pr(>F)
cobertura  3  341.9 113.967  9.0053 0.000139 ***
Residuals 36  455.6  12.656
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

#Comprobamos las asunciones del ANOVA:
```

```
# 1. Normalidad de la variable dependiente en cada categoría de la variable
    independiente. Puedo hacerlo visualmente con histogramas o gráficos qq.
    (ver sección 1.1).
    También puedo aplicar el test de normalidad de Shapiro-Wilk a los residuos del
    modelo
shapiro.test(modelo$residuals)
Shapiro-Wilk normality test
data:  modelo$residuals
W = 0.97657, p-value = 0.5641
# Aceptamos H0: los residuos siguen una distribución normal.

#2. Homocedasticidad (varianzas iguales entre las categorías de la variable indep.)
bartlett.test(densidad ~ cobertura,  data=hormigueros)
Bartlett test of homogeneity of variances
data:  densidad by cobertura
Bartlett's K-squared = 2.5279, df = 3, p-value = 0.4703
# Aceptamos H0: las varianzas son iguales entre categorías de cobertura.
```

La fig. IV.8 sugiere que la densidad de hormigueros difiere entre sitios con distinta cobertura de suelo, y que es mayor en suelo desnudo y mínima en pastizal. Los datos cumplen con los presupuestos de la ANOVA (normalidad de los residuos – test de Shapiro-Wilk; y homocedasticidad – Test de Bartlett), y como tal se puede aplicar un modelo de ANOVA. El modelo de ANOVA confirma que las diferencias que sugiere la figura son significativas ($p<0.001$).

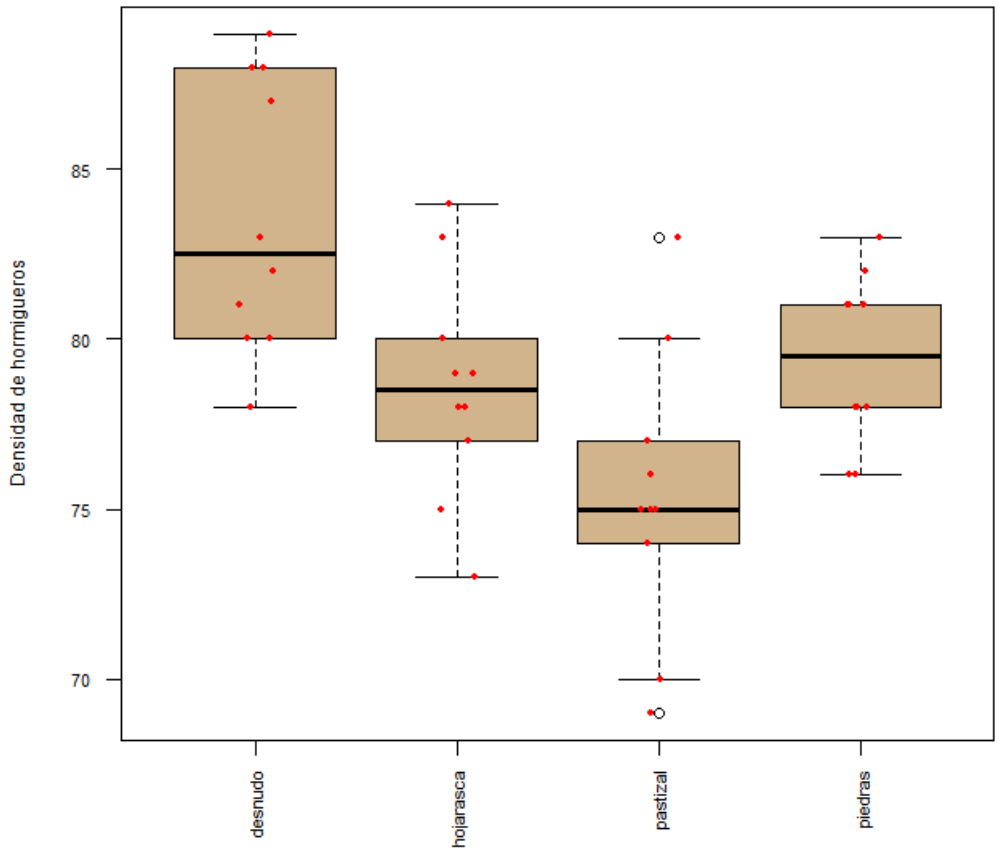

Figura IV.8. *Medias y desviaciones de la densidad de hormigueros en los distintos tipos de cobertura de suelo. Los puntos rojos representan los valores de los datos individuales.*

4.4. Procedimiento de cálculo del test no paramétrico: Kruskal-Wallis

Se emplea cuando los datos no siguen la distribución normal y/o tienen varianzas distintas, en sustitución del ANOVA paramétrico. Al igual que la U de Mann-Whitney se basa en rangos en lugar de los parámetros de la muestra (media, varianza) y compara medianas en lugar de medias.

Los pasos a seguir son los siguientes:

- Asignación de rangos: se realiza exactamente igual que para la U de Mann-Whitney.

- Cálculo del estadístico *H*:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

k = número de grupos
N = número total de datos
n_i = número de datos en el grupo i

Cuando existen rangos ligados (dos o más números con el mismo rango) se aplica un factor de corrección, siendo H_c el estadístico que se utiliza en lugar de H, calculado según la siguiente expresión:

$$H_c = \frac{H}{C} \qquad C = 1 - \frac{\sum_{i=1}^{m} (t_i^3 - t_i)}{N^3 - N}$$

t_i = número de rangos ligados en cada grupo
m = número de grupos de rangos ligados

El valor crítico del estadístico calculado (H o H_c) se consulta en la tabla de la χ^2 si N≥15, o si $k > 5$, para (k-1) grados de libertad. Si N<15 y k<5 se consulta en la tabla específica para H.

- Si $H_{cal} \geq H_{crít}$ ($\chi^2_{crít}$) ⇒ se rechaza H_0 y se acepta H_{ecol} (medianas diferentes)

Si $H_{cal} < H_{crít}$ ($\chi^2_{crít}$) ⇒ se acepta H_0 y se rechaza H_{ecol} (medianas son iguales)

Cuadro IV.6: ejemplo de cálculo de Kruskal-Wallis

Se quiere estudiar si el pH de cuatro charcas situadas sobre sustratos diferentes es distinto. Para ello se obtuvieron 8 muestras de agua procedentes de cada una de las charcas, midiéndose el pH en cada una de ellas. Los datos de pH se ordenaron de forma ascendente para cada charca. (Una muestra de agua de la charca nº 3 se perdió, de forma que n3=7; pero el test no requiere igualdad en el número de datos de cada grupo). Los rangos se muestran entre paréntesis.

Variables:

- Variable dependiente: pH (cuantitativa)
- Variable independiente: tipo de sustrato sobre el que cada charca (cualitativa)

Hipótesis:

- Hecol= el pH no es el mismo en las cuatro charcas
- H0= el pH es el mismo en las cuatro charcas

Tabla de datos:

Charca 1	Charca 2	Charca 3	Charca 4
7.68 (1)	7.71 (6*)	7.74 (13.5*)	7.71 (6*)
7.69 (2)	7.73 (10*)	7.75 (16)	7.71 (6*)
7.70 (3.5*)	7.74 (13.5*)	7.77 (18)	7.74 (13.5*)
7.70 (3.5*)	7.74 (13.5*)	7.78 (20*)	7.79 (22)
7.72 (8)	7.78 (20*)	7.80 (23.5*)	7.81 (26*)
7.73 (10*)	7.78 (20*)	7.81 (26*)	7.85 (29)
7.73 (10*)	7.80 (23.5*)	7.84 (28)	7.87 (30)
7.76 (17)	7.81 (26*)		7.91 (31)
n1=8	n2=8	n3=7	n4=8
R1=55	R2=132.5	R3=145	R4=163.5

* Rangos ligados

$$N = 8+8+7+8 = 31$$

$$H = \frac{12}{N(N+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i} - 3(N+1) = \frac{12}{31(32)}\left[\frac{55^2}{8}+\frac{132.5^2}{8}+\frac{145^2}{7}+\frac{163.5^2}{8}\right] - 3(32) = 11.876$$

Número de grupos de rangos ligados $= m = 7$

$$\sum_{i=1}^{m}(t_i^3 - t_i) = (2^3-2)+(3^3-3)+(3^3-3)+(4^3-4)+(3^3-3)+(2^3-2)+(3^3-3) = 168$$

$$C = 1 - \frac{\sum_{i=1}^{m}(t_i^3 - t_i)}{N^3 - N} = 1 - \frac{168}{31^3 - 31} = 1 - \frac{168}{29760} = 0.9944$$

$$H_c = \frac{H}{C} = \frac{11.876}{09944} = 11.943 \qquad \nu = k - 1 = 3 \qquad \chi^2_{0.05,\,3} = 7.815$$

$H_c cal > \chi^2_{crít}$ $\Rightarrow$ Se rechaza H_0

El pH no es el mismo en todas las charcas

Realización en R del test Kruskall-Wallis

```
## introducimos los datos
charca1 = c(7.68,7.69,7.70,7.70,7.72,7.73,7.73,7.76)
charca2 = c(7.71,7.73,7.74,7.74,7.78,7.78,7.80,7.81)
charca3 = c(7.74,7.75,7.77,7.78,7.80,7.81,7.84)
charca4 = c(7.71,7.71,7.74,7.79,7.81,7.85,7.87,7.91)
charcaID = c(rep("charca1",8),rep("charca2",8),
             rep("charca3",7),rep("charca4",8))
charcanum = c(rep(1,8),rep(2,8),
              rep(3,7),rep(4,8))
charcas = data.frame(charca = charcaID,
                     pH = c(charca1,charca2,charca3,charca4),
                     charca.num = charcanum)
charcas
   charca   pH charca.num
1  charca1 7.68          1
2  charca1 7.69          1
3  charca1 7.70          1
4  charca1 7.70          1
5  charca1 7.72          1
6  charca1 7.73          1
7  charca1 7.73          1
8  charca1 7.76          1
9  charca2 7.71          2
10 charca2 7.73          2
11 charca2 7.74          2
12 charca2 7.74          2
13 charca2 7.78          2
14 charca2 7.78          2
15 charca2 7.80          2
16 charca2 7.81          2
17 charca3 7.74          3...

#Examinamos la distribución de los datos, realizando un histograma para cada charca
par(mfrow=c(2,2)) # Esta instrucción es para que dibuje las cuatro figuras en
                  dos filas y dos columnas.
hist(charca1)
```

```
hist(charca1)
hist(charca1)
hist(charca1)
```

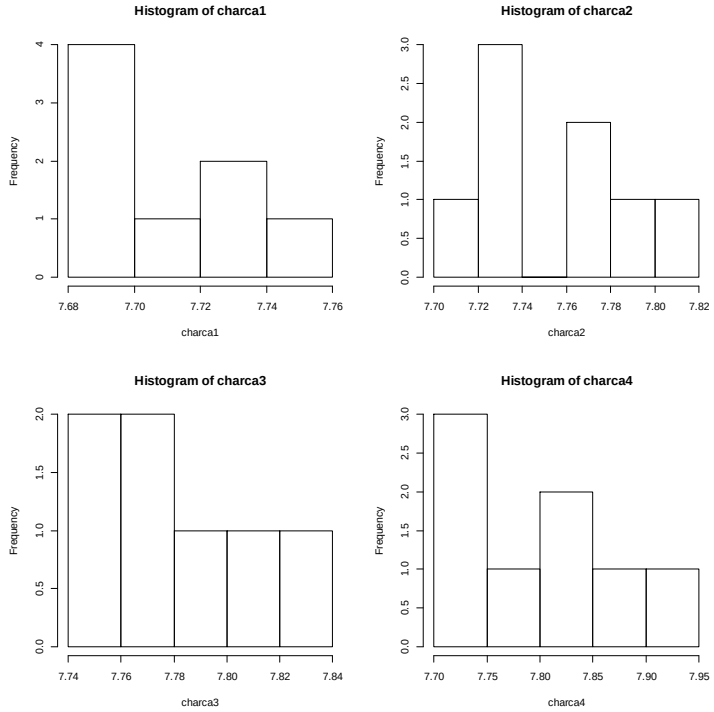

***Figura IV.9**. Histogramas mostrando la distribución de frecuencias del pH de cada una de las cuatro charcas del ejemplo.*

```
# examinamos la distribución del pH en cada charca con boxplot
dev.off()# Cierra la ventana del gráfico anterior y anula los parámetros gráficos
         previos
boxplot(pH ~ charca, data=charcas, col="tan", cex.axis=0.7, las = 2,
            ylab="pH de las charcas", cex.lab=0.75)
stripchart(pH ~ charca, data=charcas, col="red",
           vertical = TRUE, method = "jitter", cex=0.5,
           add=TRUE, pch=19)
```

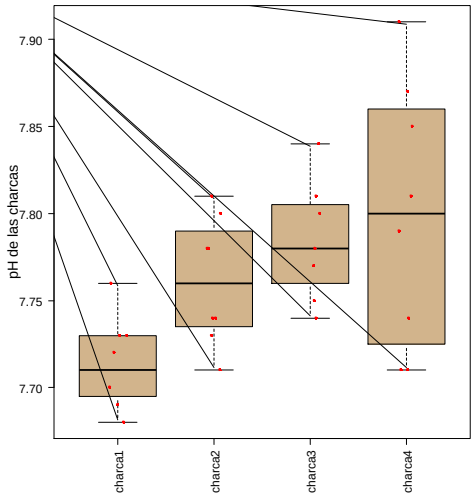

***Figura IV.10**. Diagrama de cajas y bigotes comparando las distribuciones de las medidas de pH en cada una de las cuatro charcas del ejemplo. Los puntos rojos muestran los valores correspondientes a cada observación.*

```
#Test de homocedasticidad (varianzas iguales)
# H0: las varianzas no difieren entre charcas. Test de Bartlett
bartlett.test(pH ~ charca, data=charcas)
Bartlett test of homogeneity of variances
data:  pH by charca
Bartlett's K-squared = 8.8272, df = 3, p-value = 0.03168
# las varianzas no son iguales, es decir, no se cumple la H0 de homocedasticidad ya que
p-value < 0.05, por tanto no podemos usar ANOVA, y procedemos a comparar las medias con
el test de Kruskal-Wallis

# Aplicamos el test Kruskal-Wallis
kruskal.charcas = kruskal.test(pH ~ charca.num, data = charcas)
kruskal.charcas
Kruskal-Wallis rank sum test
data:  pH by charca.num
Kruskal-Wallis chi-squared = 11.944, df = 3, p-value = 0.007579
# El valor de p-value < 0.05 nos informa de que no se cumple la H0 de que las medias de
pH de cada charca sean iguales.
```

La figura IV.10 sugiere que hay diferencias de pH entre las cuatro charcas comparadas, pero también indica que la dispersión de valores (medida con la varianza) es mayor en la charca 4. El test de Barlett confirma que las varianzas son heterogéneas. El resultado del Kruskal-Wallis confirma que las diferencias que muestra la figura son significativas.

5. ASOCIACIÓN ENTRE VARIABLES CUANTITATIVAS: COEFICIENTES DE CORRELACIÓN

El coeficiente de correlación cuantifica el grado de asociación entre dos variables cuantitativas. Se utiliza cuando no se asume que una variable es causa y la otra consecuencia. Por ejemplo, si queremos saber si el peso y con la longitud del pico covarían dentro de una población de aves (no se asume una relación de causalidad).

ρ es el coeficiente de correlación real que existe entre dos variables en el conjunto de la población.

r y r_s son los coeficientes medidos sobre la muestra.

Los coeficientes de correlación varían entre –1 y 1 del siguiente modo (Fig. IV.3):
a) $1 \geq \rho > 0$: correlación positiva.
b) $-1 \leq \rho < 0$: correlación negativa.
c) $\rho \approx 0$: no hay correlación, los valores de *x* e *y* varían de forma independiente.

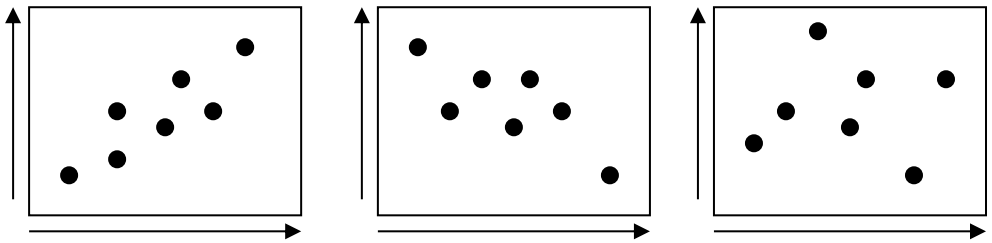

Figura IV.11. *Tres posibles tipos de asociación entre variables: positiva (izqda.) negativa (centro) y ausencia de asociación (derecha)*

Cuanto más cerca esté el coeficiente de 1 ó –1, más fuerte es la correlación

5.1. Hipótesis de una cola y de dos colas

Cuando la hipótesis ecológica indica que existe correlación (sin precisar el signo) se trata de una hipótesis de dos colas, ya que implica dos posibilidades (que la relación sea positiva o negativa). La hipótesis nula solo implica una posibilidad: que no exista correlación entre las variables.

H_{ec}: $\rho \neq 0$ $(\rho < 0$ ó $\rho > 0)$

H_0: $\rho = 0$

Por el contrario, si la hipótesis ecológica precisa el signo de la correlación, entonces se trata de una hipótesis de una cola. La hipótesis nula implica dos posibilidades: que no haya correlación o que ésta sea del signo contrario al esperado en la hipótesis ecológica.

H_{ecol}: $\rho > 0 \Rightarrow H_0$: $\rho \leq 0$

H_{ecol}: $\rho < 0 \Rightarrow H_0$: $\rho \geq 0$

Es importante saber de cuántas colas es la hipótesis a la hora de evaluar la significación del test.

5.2. Correlación paramétrica: r de Pearson

Para poder aplicar este test las dos variables (dependiente e independiente) deben seguir una distribución normal.

El cálculo del índice de correlación de Pearson se hace a partir de la siguiente fórmula:

$$r = \frac{n\sum_{i=1}^{i=n} x_i y_i - \sum_{i=1}^{i=n} x_i \times \sum_{i=1}^{i=n} y_i}{\sqrt{\left(n\sum_{n=1}^{n=i} x_i^2 - \left(\sum_{n=1}^{n=i} x_i\right)^2\right) \times \left(n\sum_{n=1}^{n=i} y_i^2 - \left(\sum_{n=1}^{n=i} y_i\right)^2\right)}}$$

n- nº de pares de muestras

x_i- valores de la variable *x*

y_i- valores de la variable *y*

A continuación, se comprueba la significación del índice de correlación calculado comparándolo con el valor de un estadístico r_{crit} obtenido a partir de la tabla correspondiente, para una $\alpha = 0.05$ o inferior y las colas que establezca la hipótesis.

Si $|r_{cal}| \geq r_{crit}$ (α=0.05 o inferior) → Se rechaza la hipótesis nula. → Existe correlación.

Cuadro IV.7: Ejemplo de cálculo de r de Pearson

Un ornitólogo está interesado en conocer la longitud del pico de una población de aves que estudia. Sin embargo, esa medida resulta más costosa de tomar que el peso corporal. Por ello quiere saber si ambas variables se correlacionan para estimar la primera a partir de la segunda.

Variables (Ambas son cuantitativas y normales):

- x: longitud del pico.
- y: peso corporal.

Hipótesis:

- H_{ecol}: $\rho \neq 0$ ($\rho < 0$ ó $\rho > 0$) (dos colas)
- H_0: $\rho = 0$

Tabla de datos:

Obs.	Longitud del pico (mm)	Peso corporal (g)	x^2	y^2	xy
1	33.5	51	1122	2601	1708
2	38.0	59	14444	3481	2242
3	32.0	49	1024	2401	1568
4	37.5	54	1406	2916	2025
5	31.5	50	992	2500	1575
6	33.0	55	1089	3025	1815
7	31.0	48	961	2304	1488
8	36.5	53	1332	2809	1935
9	34.0	52	1156	2704	1768
10	35.0	57	1225	2349	1995
SUMA	342	528	11752	27990	18119

Cálculos

n = 10; r = 0.779, $r_{cal} = 0.779 > r_{crit\ (0.01)\ n=10} = 0.765$. Se rechaza H_0 y se acepta H_{ecol}

Interpretación:

Se puede concluir que existe una correlación positiva entre el peso corporal y la longitud del pico de esa población de aves. Esto significa que los cambios en peso corporal de esas aves son un fiel reflejo de los cambios en la longitud del pico.

Realización en R de la correlación de Pearson

```
## Generamos una tabla de datos con los datos del ejemplo
datos.pico <- data.frame(
  Obs. = seq(1,10,1),
  longitud.pico = c(33.5,38.0,32.0,37.5,31.5,33.0,31.0,36.5,34.0,35.0),
  peso.corporal = c(51,59,49,54,50,55,48,53,52,57)
)
```

```
   Obs. longitud.pico peso.corporal
1     1          33.5            51
2     2          38.0            59
3     3          32.0            49
4     4          37.5            54
5     5          31.5            50
6     6          33.0            55
7     7          31.0            48
8     8          36.5            53
9     9          34.0            52
10   10          35.0            57
```

```
#Represento las variables gráficamente para ver si hay relación aparente entre ellas
plot(longitud.pico ~ peso.corporal, data=datos.pico, pch=16)
abline(lm(longitud.pico ~ peso.corporal,
          data=datos.pico), col="red")

#Compruebo que ambas variables cumplen el requisito de normalidad (sección 1.1).
#Test de correlación para comprobar la hipótesis nula (H0) de rho = 0
cor.test( ~ longitud.pico + peso.corporal, data=datos.pico,
          method = "pearson",  continuity = FALSE,
          conf.level = 0.95)
```

```
Pearson's product-moment correlation
data:  longitud.pico and peso.corporal
t = 3.5194, df = 8, p-value = 0.007853
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2942560 0.9452104
sample estimates:
      cor
0.7794691
```

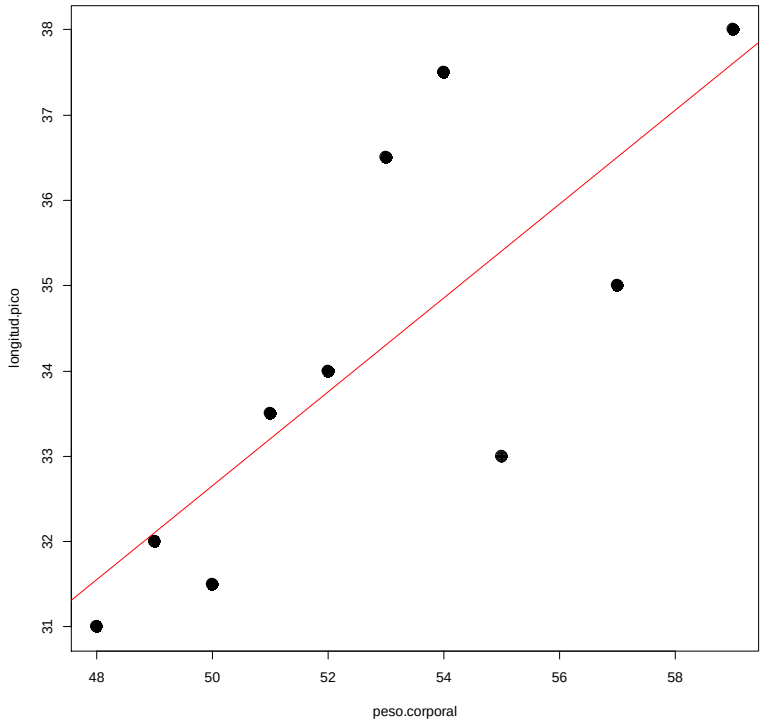

***Figura IV.12**. Gráfico de dispersión entre la longitud del pico y el peso corporal y ajuste del modelo lineal (ver línea roja).*

La figura IV.12 sugiere que existe una relación entre entre ambas variables y el resultado del test confirma que esa relación es significativa.

5.3. Correlación no paramétrica: r *de Spearman*

Se aplica este test cuando una o ninguna de las dos variables implicadas sigue una distribución normal. Para calcular la *r* de Spearman hay que realizar los siguientes pasos:

- Ordenar los pares de datos en función del valor de *x* y asignar rangos a *x*.
- Repetir la ordenación en función de *y* y asignar rangos a *y*.
- Calcular el coeficiente:

$$r_s = 1 - \frac{6\sum_{i=1}^{i=n} d_i^2}{n^3 - n}$$

n = nº de pares de datos

d_i = diferencia de rangos en las variables del par i

Para comprobar la significación estadística del índice de correlación se consulta en la tabla correspondiente el valor crítico de r_s para n pares de datos, para p=0.05 o inferior y para el número de colas acorde con la hipótesis. Si $r_{s\ cal} \geq r_{s\ crít}$, se rechaza H_0.

Cuadro IV.8: Ejemplo de cálculo de r de Spearman

Se sospecha que la abundancia de la especie de gramínea *Poa bulbosa* en los pastizales mediterráneos depende en gran medida de la humedad que hay en el suelo. Para comprobar la hipótesis se realiza un muestreo con una cuadrícula de 20 cm de lado, que se dispone 12 veces al azar sobre la comunidad de pasto. En cada cuadrícula se mide la cobertura de la especie y la humedad del suelo mediante un TDR.

Variables: Ambas son cuantitativas y no siguen una distribución normal
- Cobertura de la especie
- Humedad del suelo.

Hipótesis
- H_{ec}: existe una correlación positiva entre la cobertura de *Poa* y la humedad $\rho > 0$ (de una cola)
- H_0: $\rho \leq 0$

Tabla de datos:

Obs.	Cobertura	Humedad	Rango cob.	Rango hum.	d	d^2
1	82	42	2	3	-1	1
2	98	46	6	4	2	4
3	87	39	5	2	3	9
4	40	37	1	1	0	0
5	116	65	10	8	2	4
6	113	88	9	11	-2	4
7	111	86	8	10	-2	4
8	83	56	3	6	-3	9
9	85	62	4	7	-3	9
10	126	92	12	12	0	0
11	106	54	7	5	2	4
12	117	81	11	9	2	4
					Suma	52

Cálculos

$$r_s = 1 - \frac{6 \times 52}{12^3 - 12} = 0.82 > r_s \text{ crit } (0.05) = 0.503$$

Interpretación:
Se rechaza H_0, hay correlación positiva entre la cobertura de Poa bulbosa y la humedad del suelo. Es importante destacar que este muestreo no es una demostración de una relación causa-efecto entre las variables, es decir, que con este muestreo no podemos concluir que la mayor humedad de suelo es la causa de la mayor abundancia de Poa bulbosa. Para determinar relaciones de causa-efecto se necesita realizar experimentos controlados y otros tests estadísticos que verifiquen ese tipo de relación.

Cálculo en R de la correlación de Spearman

```
#Genero una matriz con los datos que quiero analizar)
datos <- data.frame("cobertura" = c(82,98,87,40,116,113,111,83,85,126,106,117),
                    "humedad" = c(42,46,39,37,65,88,86,56,62,92,54,81))
str(datos)

#Dibujo el gráfico de dispersión para ver la relación entre las variables (Fig. IV.13)
plot(cobertura ~ humedad, data=datos, pch=16)
```

```
#Correlación entre los datos

#Compruebo si ambas variables cumplen el requisito de normalidad (sección 1.1).
#Como no es así, aplico el test de correlación de Spearman. Puedo hacerlo de dos formas

cor.test( ~ cobertura + humedad,
          data=datos,
          method = "spearman",
          continuity = FALSE,
          conf.level = 0.95)
Spearman's rank correlation rho
data:  cobertura and humedad
S = 52, p-value = 0.002027
alternative hypothesis: true rho is not equal to 0
sample estimates:   rho = 0.8181818
```

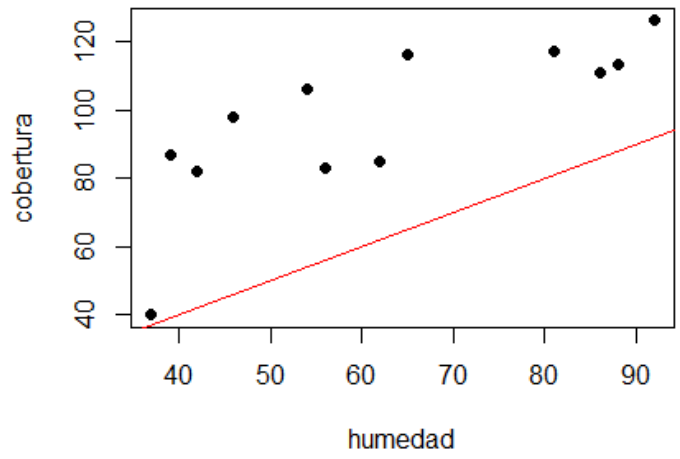

***Figura IV.13**. Gráfico de dispersión entre la cobertura y la humedad mostrando una línea 1:1 en rojo.*

La Fig. IV.13 sugiere que hay relación entra las variables, ya que a medida que aumenta la humedad también lo hace la cobertura. El resultado del test da un valor de $p<0.05$, que confirma que efectivamente ambas variables están correlacionadas significativamente.

6. REGRESIÓN

La regresión se aplica cuando tenemos dos variables y asumimos (en nuestra hipótesis) que una depende de la otra. Por ejemplo, la dosis de fertilizante que se aplica a una serie de plantas cultivadas en macetas esperamos que cause diferencias en la altura de las plantas. La variable independiente es la dosis de fertilizante y la dependiente la altura.

Para poder aplicar este test, los residuos del modelo deben seguir una distribución normal.

El parámetro que se calcula en regresión es el coeficiente de regresión o R^2, que equivale al cuadrado del coeficiente de correlación de Pearson y se interpreta como el porcentaje de la varianza de la variable dependiente que explica la variable independiente. La significación de este coeficiente se calcula igual que el de la *r* de Pearson.

Cuadro IV.9: Ejemplo de cálculo de regresión

Esperamos que la salinidad del suelo afecte negativamente al crecimiento de plantas de una especie herbácea (*Onobrychis sativa*). Para comprobarlo hemos preparado una serie de disoluciones con potencial osmótico decreciente (desde agua destilada donde el potencial osmótico es igual a 0 MPa) hasta la solución más concentrada de potencial osmótico = -0.328 MPa). Tras una semana, hemos medido la longitud media de las plántulas emergidas en cada placa.

Variables (Ambas son cuantitativas y normales):
- x: potencial osmótico (varía entre – y 0, MPa).
- y: Longitud (cm).

Hipótesis
- H_{ecol}: A mayor potencial osmótico (menos negativo) esperamos una mayor longitud de plántula.

Tabla de datos:

Longitud (cm)	Potencial osmótico (MPa)
0.2	0
0.17	-0.043
0.15	-0.127
0.1	-0.193
0.05	-0.263
0.02	-0.32
0.2	0
0.17	-0.043
0.15	-0.127
0.1	-0.193
0.05	-0.263
0.02	-0.328
0.2	0
0.17	-0.043
0.15	-0.127
0.1	-0.193
0.05	-0.263
0.02	-0.328

Hacemos una recta de regresión del tipo:

Longitud = a + b × Potencial osmótico

Donde a y b son parámetros de la recta de regresión y obtenemos:

Longitud = 0.2028 + 0.5522 × Potencial osmótico

Interpretación:

El parámetro *b* relacionado con la pendiente de la curva tiene un valor positivo que no cruza el cero, por lo que podemos aceptar nuestra hipótesis de un efecto positivo del potencial osmótico en la longitud de la plántula.

Cálculo en R de una regresión

```
# Introducimos los datos y creamos las dos variables a correlacionar.

Longitud=c(0.2,0.17,0.15,0.1,0.05,0.02,0.2,0.17,0.15,0.1,
      0.05,0.02,0.2,0.17,0.15,0.1,0.05,0.02)
PotOsm=c(0,-0.043,-0.127,-0.193,-0.263,-0.328,
      0,-0.043,-0.127,-0.193,-0.263,-0.328,0,
      -0.043,-0.127,-0.193,-0.263,-0.328)

# Creamos el modelo de regresión lineal
modeloL = lm(Longitud ~ PotOsm)

# Coeficientes de la recta
modeloL$coefficients
(Intercept)      PotOsm
  0.2028031   0.5522206

#Representamos gráficamente la nube de puntos: Figura IV.14
plot(PotOsm,Longitud, col = "darkgray",
     xlab="potencial osmotico (MPa)", ylab="Longitud (mm)",
     pch = 21, cex = .8)

#Añadimos la recta de regresión
abline(modeloL, col = "red", lwd = 3)

# COMPRUEBA SI SE CUMPLEN LOS REQUISITOS DE LA REGRESIÓN
# H0: los residuos son normales.
shapiro.test(modeloL$residuals)
Shapiro-Wilk normality test
data:  modeloL$residuals
W = 0.82134, p-value = 0.00309
```

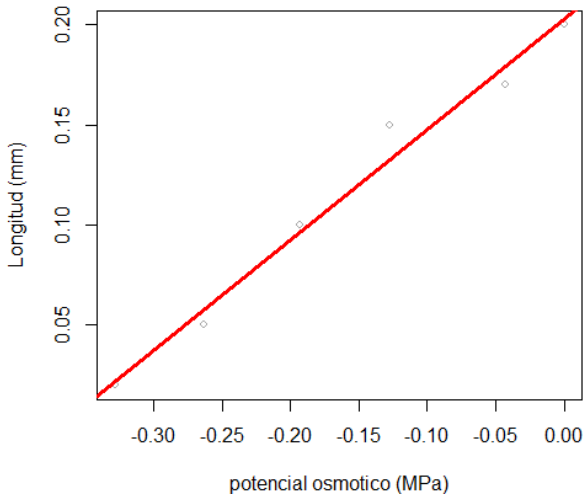

***Figura IV.14**. Gráfico de dispersión entre la longitud de la plántula y el potencial osmótico y ajuste del modelo lineal (línea roja)*

Se puede observar como el potencial osmótico tiene un efecto positivo en la longitud. El intercepto de la regresión tiene un valor de 0.20 y la pendiente de 0.55 (ver Figura IV.14). Sin embargo, al comprobar la asunción del modelo de normalidad de los residuos podemos afirmar que nuestros residuos no se distribuyen normalmente ya que el p-valor es menor de 0.05 ($p = 0.003$). Además, se observan gráficamente desviaciones de la normalidad (ver Figura IV.15). Por tanto, estrictamente no podríamos aplicar una regresión lineal y deberíamos adoptar alguna otra solución, por ej. aplicar una correlación de Spearman.