Bioestadística con R

Roberto Bustillos, MVZ, M.Sc - Basado en un material de D. Conesa

Universidad Central del Ecuador

Facultad de Medicina Veterinaria y Zootecnia

Maestría en Epidemiología y Salud Pública Veterinaria, 2016-2018

Contenido

Introducción a la modelización estadística.

- Los datos obtenidos cuando realizamos cualquier experimento presentan variabilidad y la Estadística nos permite analizar los datos que exhiben variabilidad.
- En ese proceso podemos distinguir entre Modelización estadística, Estadística descriptiva, e inferencia estadística.

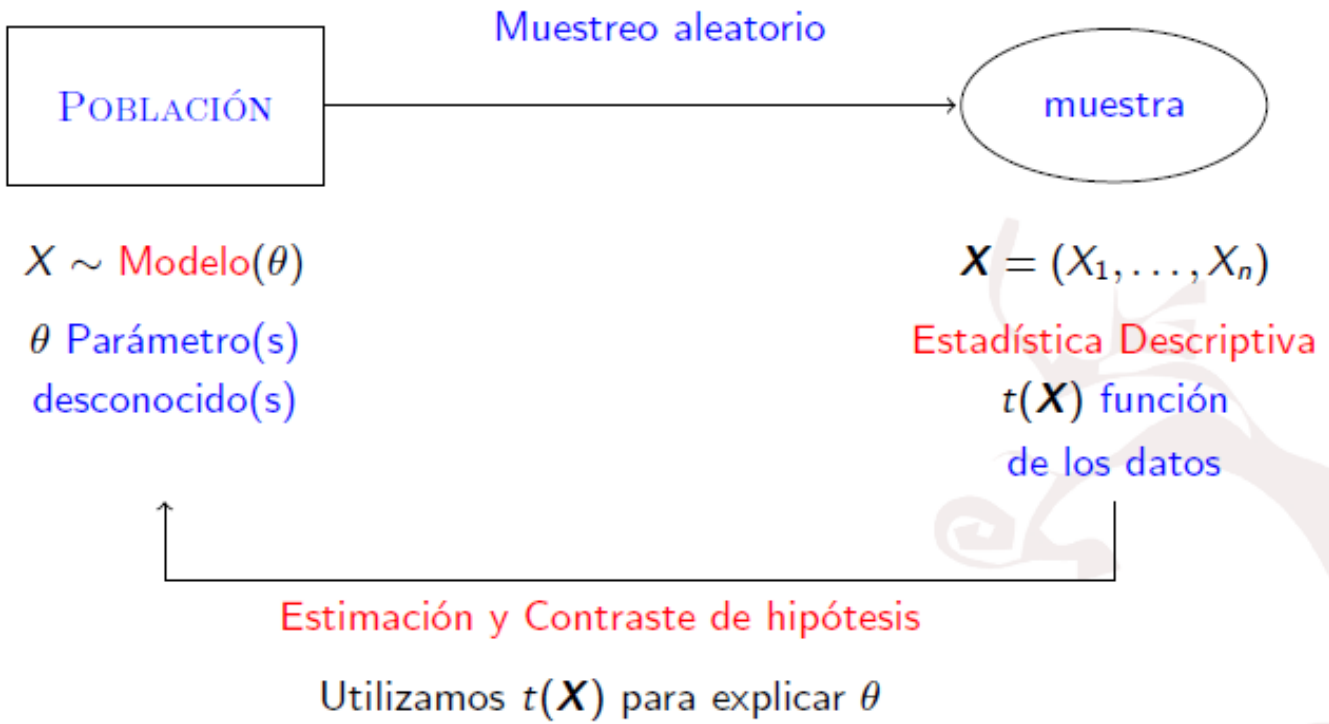


Figure: Conesa, 2016

Modelización en la Estadística.

- En general, un modelo es una representación en pequeña escala de la realidad.
- La Estadística nos permite incorporar la variabilidad presente en la vida real en nuestros modelos a través de la aleatoriedad.
- Los modelos estadísticos son la base en la que se sustentan la mayoría de las técnicas de análisis de datos habituales.
- "La formulación del problema es más esencial que su propia solución, que puede ser simplemente una habilidad matemática o experimental" (Albert Einstein).

Formulación de un problema.

- Entender el background físico del problema.
- Comprender claramente el objetivo.
- !Poner el problema en términos estadísticos! Es un paso clave, una vez logrado, la solución suele ser rutinaria.

Comentarios sobre la toma de datos.

En el proceso de modelización es importante entender como se van a tomar o han tomado los datos.

- ¿Los datos son observacionales o experimentales?
- Es decir: ¿Provienen de una muestra o encuesta convencional (recogidos por mera observación sin control sobre las condiciones) o han sido obtenidos como resultado de un diseño experimental (con control sobre las condiciones)?
- Las conclusiones dependerán de que tipo son: en el primer caso podremos hablar de asociación entre los datos, mientras que en le segundo caso podremos hablar de una relación causa - efecto.
- ¿Hay información o datos que no hemos observado?
- ¿Hay valores faltantes?
- ¿Cómo se han codificado los datos?
- ¿Cuáles son las unidades de medida?
- ¿Hay errores en la entrada de datos? Conviene realizar análisis

Modelización: tipos de variables.

- Cuando modelizamos un problema real:
  - ¿qué queremos explicar?
  - y ¿en base a qué?
- Esto nos clasifica las variables en:
  - Variables respuesta: las que queremos explicar.
  - Variables explicativas (o independientes, o predictoras): las que nos sirven para explicar las variables respuesta.
- Las variables también se clasifican por su tipo de atributo:
  - Cualitativas.
  - Cuantitativas.

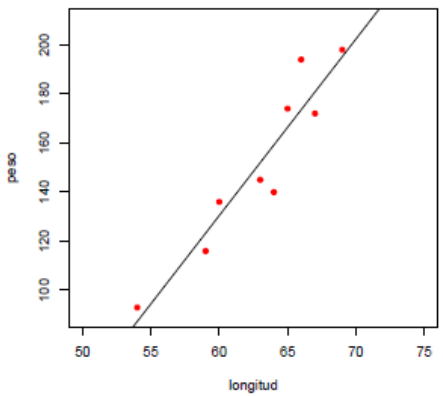
Modelos estadísticos.

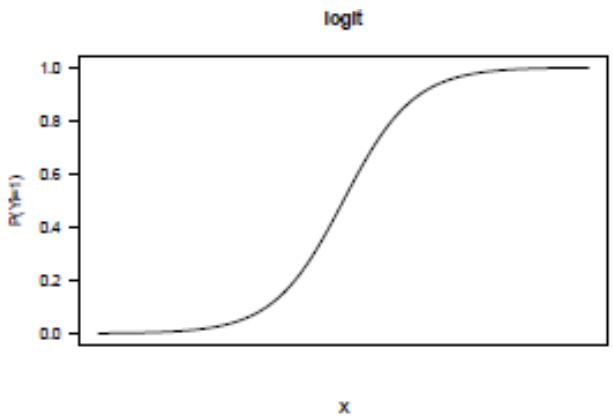
- La mayoría de los modelos estadísticos tienen una estructura del tipo:
  - Variable respuesta que se quiere explicar.
  - Una componente sistemática que contiene la información "general" del sistema bajo estudio, y que se expresa como una combinación de variables explicativas en forma de ecuación paramétrica. Indica como afectan las explicativas a la respuesta.
  - Una componente aleatoria que refleja la variabilidad intrínseca en cada situación (en cada dato) particular.
- Dependiendo del tipo de variable, las explicativas son:
  - Cualitativas $\longrightarrow$ Factores con sus correspondientes niveles.
  - Cuantitativas $\longrightarrow$ Covariables.

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad con \quad i = 1, ..., n \quad y \quad \epsilon_i \sim N(0, \sigma^2)$$

Modelos estadísticos paramétricos.

- La mayoría de las veces la componente sistemática viene expresada como una combinación lineal (pero puede ser no lineal).





- La forma de incluir la aleatoriedad en la componente aleatoria es asignando una distribución de probabilidad a la variable respuesta.
- Si la variable es normal (la mayoría de casos) y la relación lineal nos encontramos ante los modelos lineales.

Examples

Explicar el peso de un ternero por su altura y su edad.

$$Y_i = \beta_0 + \beta_1 X^{(1)} + \beta_2 X^{(2)} + \epsilon_i$$

- Esta explicación no es tan clara si la variable es discreta o cualitativa pero la observamos categorizada.
- La variable respuesta no tiene porque ser normal, podría ser binomial, Bernoulli, gamma, Poisson, etc.
- En cualquier caso, todos los modelos siempre vienen expresados en función de parámetros, siendo la inferencia sobre ellos (estimación y contraste de hipótesis) nuestro objetivo final.

Elementos básicos inferencia I.

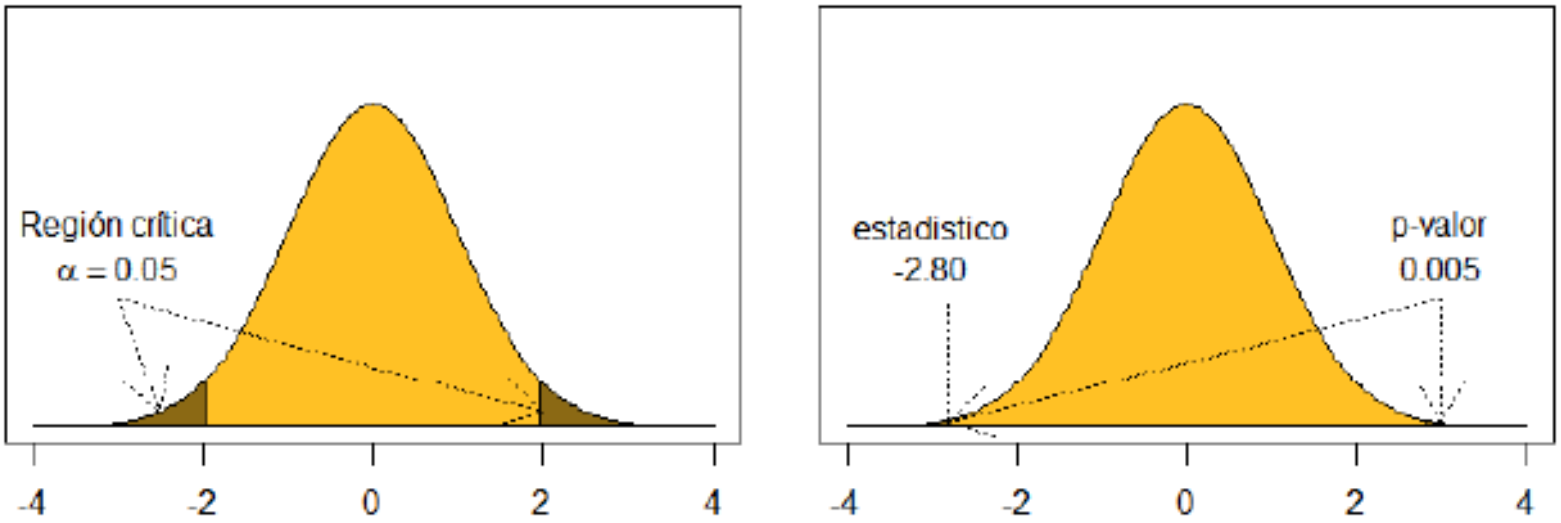
Estimación

- Utilizamos la estimación puntual de un parámetro para tener una primera aproximación sobre su valor.
- En la mayoría de los modelos se conoce el mejor estimador y un Intervalo de confianza para los parámetros de interés.

Elementos de un contraste

- Datos (obtenidos de forma muy diversa)
- Hipótesis nula (H_0)
- Hipótesis alternativa (H_A)
- Estadístico de contraste T (y su distribución bajo H_0)
- Valor observado del est. de contraste: t
- P-valor: Prob. si H_0 es cierta de que el valor de T sea más extremo que t en la dirección de la hip. alternativa

Elementos básicos inferencia I.



-Figura 1. Región crítica al nivel de significación $\alpha = 0.05$, a la izquierda, y p-valor, a la derecha, para los datos de la primera visita en el ejemplo e hipótesis nula $\mu = 200$.-

Elementos básicos inferencia II.

Contrastes Paramétricos

- Asumen que los datos tienen una determinada distribución
- El contraste es sobre alguno de los parámetros de una distribución
- Ejemplo: Test de la t de Student para una muestra

Contrastes No Paramétricos

- No asumen ninguna distribución para los datos
- En principio, son más flexibles

Entonces, ¿cuál usamos?

- Paramétricos, si se cumplen las hipótesis sobre los datos
- No paramétricos, en otro caso
- **OJO**: ¡¡Param./No param. no contrastan exactamente lo mismo!!

Examples

```
library(foreign)
ejemplo <- read.spss(file="ambiente.sav",to.data.frame=TRUE)
attach(ejemplo)
# Análisis descriptivo numérico
summary(ejemplo)
by(OZONO,OZONO,length) # N.- de lugares clasf. por ozono
by(SULFATO,OZONO,mean) # Media de sulfato por grupo de
ozono
by(PH,PROVIN,summary) # Est. resumen de PH por provincia
# Diagrama de cajas por factores
boxplot(SULFATO~PROVIN)
boxplot(PH~OZONO)
```

Examples

```
# Gráficos
hist(SULFATO,main="Histograma del SULFATO")
boxplot(PH,main="Diagrama de cajas del PH")
# Gráficos por grupos
par(mfrow=c(2,2))
hist(PH,main="Histograma del PH")
by(PH,PROVIN,function(X,xlim)hist(X,xlim=xlim),xlim=range(PH))
```

Modelos básicos con datos normales.

- De acuerdo con el principio de la navaja de Occam, la elección de un modelo estadístico debe ser siempre lo más simple posible.
- Vamos a describir los contrastes de hipótesis más básicos que permiten resolver algunos de los análisis de datos más habituales que involucran una y dos muestras cuando los datos son normales.
- En concreto:
  - Problemas de una muestra (interés sólo en la variable respuesta)
    - Varianza.
    - Media.
  - Problemas de dos muestras (interés en la variable respuesta pero explicada por un único factor de dos niveles)
    - Comparación de varianzas.
    - Comparación de medias.

Inferencia en problemas de una muestra.

Inferencia sobre la media de una población

Sea $X_1, \ldots, X_n$ una m.a. de una población $N(\mu, \sigma^2)$

- Si el interés es realizar inferencia sobre μ, podemos estimarla con $\bar{x}$.
- Además cuando σ^2 es desconocida, podemos hacer un contraste sobre μ del tipo $\begin{cases} H_0: & \mu \leq \mu_0 \\ H_A: & \mu > \mu_0 \end{cases}$ utilizando como regla de decisión:

  "Rechazar H_0 si $t_s = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} > t_{1-\alpha}(n-1)$"

  o bien obteniendo su correspondiente p-valor.
- Si σ^2 es conocida podemos contrastar dicho contraste rechazando si $\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} > Z_{1-\alpha}$, o utilizando su correspondiente p-valor.

Contraste unilateral de una media.

Inferencia sobre la media de una población

Sea $X_1, \ldots, X_n$ una m.a. de una población $N(\mu, \sigma^2)$

- Análogamente para el contraste $\begin{cases} H_0: & \mu \geq \mu_0 \\ H_A: & \mu < \mu_0 \end{cases}$ con σ^2 desconocida, podemos utilizar la regla:

  "Rechazar H_0 si $t_s = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} < t_\alpha(n-1)$",

  o el correspondiente p-valor.
- Si σ^2 es conocida se rechaza si $\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} < Z_\alpha$, o bien se utiliza el p-valor.

Contraste bilateral de una media.

Contrastes sobre la media de una población

Sea $X_1, \ldots, X_n$ una m.a. de una población $N(\mu, \sigma^2)$

- Para el contraste bilateral $\begin{cases} H_0: & \mu = \mu_0 \\ H_A: & \mu \neq \mu_0 \end{cases}$ con σ^2 desconocida, la regla de decisión es:

  "Rechazar H_0 si $t_s > t_{1-\alpha/2}(n-1)$ ó $t_s < t_{\alpha/2}(n-1)$",

  o bien utilizar el p-valor.
- Si σ^2 es conocida se rechaza si $\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} < Z_{\alpha/2}$ ó $\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} > Z_{1-\alpha/2}$, o bien se utiliza el p-valor.

Utilizando R para resolver contrastes de una media.

Test paramétrico

El comando t.test nos permite realizar inferencia sobre los contrastes anteriores.

t.test(PH, mu=4)

t.test(SULFATO,mu=4, alternative="greater")

t.test(SULFATO,mu=4, alternative="less")

En general:

t.test(x, y=NULL, alternative=c("two.sided","less","greater"), mu=0,paired=FALSE, var.equal=FALSE,conf.level=0.95,...)

Análisis no param. de una muestra: Test de Wilcoxon.

Descripción

- Contraste sobre la centralidad de una población (mediana)
- Observaciones independientes: $X_1, \ldots, X_n$
- Distribución simétrica de la población

Contraste

- H_0: **Mediana** $= \mu_0$
- H_A: **Mediana** $\neq \mu_0$

Examples

```
vector <- c(9,10,8,4,8,3,0,10,15,9)
wilcox.test(vector,mu=5) # Es la mediana 5?
```

Examples

En un estudio de nutrición animal, un médico veterinario plantó 13 plantas de soya y les midió la altura al cabo de 16 días con la intención de comprobar si el crecimiento medio era superior a 20 cm.

Si los resultados fueron:

20.2, 22.9, 22.3, 20, 19.4, 22, 22.1,

22, 21.9, 21.5, 19.7, 21.5, 20.9

¿Qué podemos concluir al respecto?

Inferencia en problemas de dos muestras.

Contrastes sobre la varianza de dos poblaciones

Sea $X_{11}, \ldots, X_{1n_1}$ una m.a. de $N(\mu_1, \sigma_1^2)$ y $X_{21}, \ldots, X_{2n_2}$ una m.a. de $N(\mu_2, \sigma_2^2)$

- Para el contraste bilateral $\begin{cases} H_0: & \sigma_1^2 = \sigma_2^2 \\ H_A: & \sigma_1^2 \neq \sigma_2^2 \end{cases}$ la regla de decisión es:

  "Rechazar H_0 si $F_s = \frac{s_1^2}{s_2^2} > F_{1-\alpha/2}(n_1 - 1, n_2 - 1)$ ó $F_s < F_{\alpha/2}(n_1 - 1, n_2 - 1)$"

  o bien utilizar el p-valor.

- Si el contraste es unilateral $\begin{cases} H_0: & \sigma_1^2 \leq \sigma_2^2 \\ H_A: & \sigma_1^2 > \sigma_2^2 \end{cases}$ se rechaza si

  $F_s = \frac{s_1^2}{s_2^2} > F_{1-\alpha}(n_1 - 1, n_2 - 1)$ o utilizando el p-valor. Análogamente con el otro contraste unilateral.

Comando var.test de R

var.test(respuesta~factor)

var.test(x,y,ratio=1,alternative=c("two.sided","less","greater"), conf.level=0.95,...)

Examples

Un agropecuario desea evaluar un nuevo insecticida que, según la publicidad, reduce los daños causados por los insectos. Con esa finalidad, realiza el siguiente experimento con 57 de las plantas de soya de su parcela: trata 22 con el nuevo insecticida y las otras 35 con el antiguo. De los datos de la cosecha (en kg) de estas plantas se obtuvieron los siguientes estadísticos:

	Nuevo	**Antiguo**
Media	249	233
Desviación estándar	39	45

Como paso previo para conocer si realmente es mejor el nuevo insecticida, se necesita comprobar si las varianzas de los dos grupos son o no son iguales. ¿Qué podemos concluir?

Comparación de las medias de dos muestras independientes.

Contrastes sobre la media de dos poblaciones

Sea $X_{11}, \ldots, X_{1n_1}$ una m.a. de $N(\mu_1, \sigma_1^2)$ y $X_{21}, \ldots, X_{2n_2}$ una m.a. de $N(\mu_2, \sigma_2^2)$

- Si las varianzas son **iguales y desconocidas**, para el contraste bilateral

$$\begin{cases} H_0: & \mu_1 = \mu_2 \\ H_A: & \mu_1 \neq \mu_2 \end{cases}$$ la regla de decisión es:

"Rechazar H_0 si $t_s > t_{1-\alpha/2}(n_1+n_2-2)$ ó $t_s < t_{\alpha/2}(n_1+n_2-2)$",

donde $t_s = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ y $s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$ es la estimación de la Varianza común. Alternativamente se puede utilizar el p-valor.

- Si el contraste es unilateral, $\begin{cases} H_0: & \mu_1 \leq \mu_2 \\ H_A: & \mu_1 > \mu_2 \end{cases}$ se rechaza H_0 si el estadístico del contraste $t_s > t_{1-\alpha}(n_1+n_2-2)$ (análogamente con el otro contraste unilateral). Alternativamente se puede utilizar el p-valor.

Utilizando R para realizar test t.

Comando t.test de R

t.test(respuesta~factor)

t.test(x,y=NULL,alternative=c("two.sided","less","greater"),mu=0,
paired=FALSE, var.equal=FALSE,
conf.level=0.95,...)

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")

Comparación de 2 medias no param.: Test Mann-Whitney

Descripción

- Comparamos si dos poblaciones tienen la misma mediana
- Muestras: $\boldsymbol{x_1} = (x_1^1, \ldots, x_{n_1}^1)$ y $\boldsymbol{x_2} = (x_1^2, \ldots, x_{n_2}^2)$

Contraste

- H_0: $\textbf{Mediana}_1 = \textbf{Mediana}_2$
- H_A: $\textbf{Mediana}_1 \neq \textbf{Mediana}_2$

Examples

```
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)
wilcox.test(x,y) # opcional, alternative="greater"
```

Examples

Un agropecuario desea evaluar un nuevo insecticida que, según la publicidad, reduce los daños causados por los insectos. Con esa finalidad, realiza el siguiente experimento con 57 de las plantas de soya de su parcela: trata 22 con el nuevo insecticida y las otras 35 con el antiguo. De los datos de la cosecha (en kg) de estas plantas se obtuvieron los siguientes estadísticos:

	Nuevo	**Antiguo**
Media	249	233
Desviación estándar	39	45

¿Es realmente mejor el nuevo insecticida?

Comparación no paramétrica de dos muestras emparejadas.

Contrastes sobre la media de dos poblaciones

Sea $X_1, \ldots, X_n$ una m.a. de $N(\mu_1, \sigma_1^2)$ e $Y_1, \ldots, Y_n$ una m.a. de $N(\mu_2, \sigma_2^2)$

- Los datos vienen dados como pares $(X_1, Y_1), \ldots, (X_n, Y_n)$ porque entre ellos existe una relación de dependencia. Si definimos $D = X - Y$, tendremos una nueva muestra formada por las diferencias que tendrá su media $(\bar{d})$ y desviación estándar (s_d).
- Así, para resolver el contraste $\begin{cases} H_0: & \mu_1 = \mu_2 \\ H_A: & \mu_1 \neq \mu_2 \end{cases}$ utilizaremos el contraste $\begin{cases} H_0: & \mu_d = 0 \\ H_A: & \mu_d \neq 0 \end{cases}$ cuya resolución es similar a la de los contrastes de una muestra (análogamente también los unilaterales).

t.test tiene una opción para muestras emparejadas

```
x <- c(1.83,0.50,1.62,2.48,1.68,1.88,1.55,3.06,1.30)
y <- c(0.87,0.64,0.59,2.05,1.06,1.29,1.06,3.14,1.29)
t.test(x,y,paired=TRUE, alternative="greater")
```

Comparación de las medias de dos muestras emparejadas.

Versión no paramétrica : Test de los signos

- Comparamos si dos poblaciones tienen la misma distribución
- Utilizamos el Test de Wilcoxon para una muestra con $\boldsymbol{y} = \boldsymbol{x_1} - \boldsymbol{x_2}$ y $\mu_0 = 0$

Contraste

- H_0: $\textbf{Mediana}_y = 0$
- H_A: $\textbf{Mediana}_y \neq 0$

Examples

wilcox.test(x,y,paired=TRUE, alternative="greater")

Examples

Un proceso habitual en las industrias lecheras consiste en suministrar balanceado en la alimentación de las vacas. Se piensa que un nuevo método vitamínico incrementará la producción. Para comparar los dos métodos, se analizaron 20 vacas del hato. La mitad de animales se alimentaron con balanceado y la otra mitad con el vitamínico. Se midió la cantidad de leche de cada grupo y se obtuvieron los siguientes resultados:

Grupo	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Vitamínico	35	48	65	33	61	54	49	37	58	65
Balanceado	33	40	55	41	62	54	40	35	59	56

¿Se puede decir que realmente el método vitamínico da mejores resultados que el balanceado?