

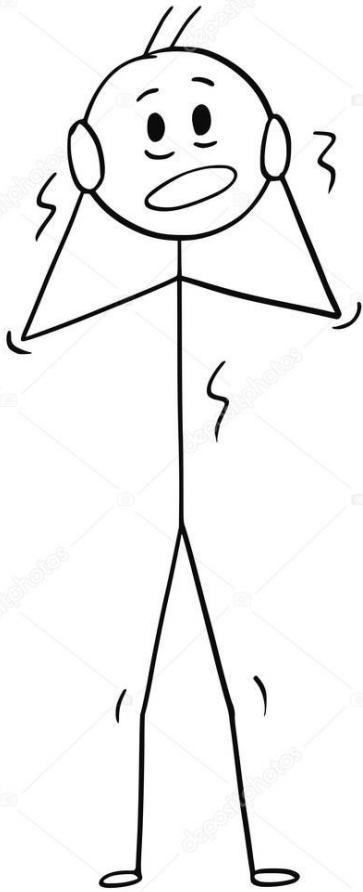
Estadística

Si quieras demostrar algo absurdo toma un montón de datos, tortúralos hasta que digan lo que quieras demostrar, y a la confesión así obtenida llámale "estadísticas".

(Darrel Huff, *How to lie with statistics*)

¿Porque usar R para análisis estadísticos?





Limpieza de datos

- Importar datos
- Explorar datos
- Necesidad de unir bases; filtrar o seleccionar datos

La limpieza de datos:

- Proceso muy tardado, pero sumamente importante
- Facilita las tareas de análisis de estos



Consideraciones.....

Hay que ser ordenados desde el primer momento



1. Nombres sin espacios, y jamás inicio con numero
2. Nombres cortos
3. No símbolos prohibidos
4. Valores faltantes con 0 o NA
5. Buena estructuración al llenar un dataframe



Acomodo de datos

nombre	tratamiento A	tratamiento B
Jose Lopez Alaman	0	2
Miguel Sanchez Perez	16	11
Lucia Carrillo Santana	3	1



	Jose Lopez Alaman	Miguel Sanchez Perez	Lucia Carrillo Santana
tratamiento A	0	16	3
tratamiento B	2	11	1

Acomodo de datos

nombre	tratamiento	resultado
Jose Lopez Alaman	A	0
Miguel Sanchez Perez	A	16
Lucia Carrillo Santana	A	3
Jose Lopez Alaman	B	2
Miguel Sanchez Perez	B	11
Lucia Carrillo Santana	B	1

Manipulación de dataframes

Instala la siguiente paquetería

```
>library(tidyverse)  
>install.packages("dslabs")  
>library(dslabs)
```

Revisemos el df murders

```
>murders
```

Seleccionamos solo cierta información

```
>murders %>% select(state, population, total) %>% head()
```

#necesitamos mas información de la tabla (por ejemplo tasa de homicidios por cada 100 mil habitantes)

```
>murders %>% mutate(ratio = total / population * 100000) %>%  
head()
```

#guardar esta tabla con la nueva columna

```
>murders <- murders %>% mutate(ratio = total / population * 100000)
```

#nos interesa identificar tasas de asesinato de menos de 1 por cada 10 mil personas en una región en específico

```
>data(murders)  
murders %>%  
mutate(ratio = total / population * 100000) %>%  
filter(ratio < 1 & region == "West")
```

Operadores
¿?

==	equal
!=	not equal
<	less than
<=	less than or equal
>	greater than
>=	greater than or equal
	or
!	not
%in%	in the set

Ejercicio

Agrega la columna ratio al data frame murders con el ratio de asesinatos por 50 mil habitantes. Luego, filtra los que tengan un ratio menor a 0.5 y sean de las regiones “South” y “West”. Reporta las columnas state, abb y ratio.

Medidas de dispersión

- `sd`: desviación estándar
- `quantile`: dividir en cuantiles
- `min`: mínimo valor (observaciones)
- `max`: máximo valor (observaciones)
- `range`: min y max

Tamaños de muestra

- `dim`: número de observaciones y variables
- `length`: (`murders$total`) número de observaciones
- `length`: (`murders`) número de variables

- Estadística descriptiva

Summary:

Arroja valores de mediana, promedio, mínimos y máximos y cuartiles

```
> murders
```

	state	abb	region	population	total
1	Alabama	AL	South	4779736	135
2	Alaska	AK	West	710231	19
3	Arizona	AZ	West	6392017	232
4	Arkansas	AR	South	2915918	93
5	California	CA	West	37253956	1257
6	Colorado	CO	West	5029196	65
7	Connecticut	CT	Northeast	3574097	97
8	Delaware	DE	South	897934	38

```
> summary(murders$total)
```

```
Min. 1st Qu. Median Mean 3rd Qu. Max.  
2.0 24.5 97.0 184.4 268.0 1257.0
```

```
> murders %>% group_by(region) %>%
```

```
summarise (m.total=mean(total))
```

```
>aggregate (total ~ state, data = murders, FUN= mean)
```

Ejercicio:

De los datos de la clase pasada (U3_2), obtén estadísticos descriptivos

**agregar summary de tabla

**agregar grafica de cuartiles

La inferencia estadística es un conjunto de métodos y técnicas que permiten deducir características de una población utilizando datos de una muestra aleatoria.

Principales usos de la estadística inferencial

Investigación Científica: Te ayuda a descubrir si tus resultados representan a toda una población.



Negocios: Toma decisiones clave con datos confiables sobre clientes, productos y empleados.



Política: Haz predicciones electorales y analiza datos de votación con estadísticas.

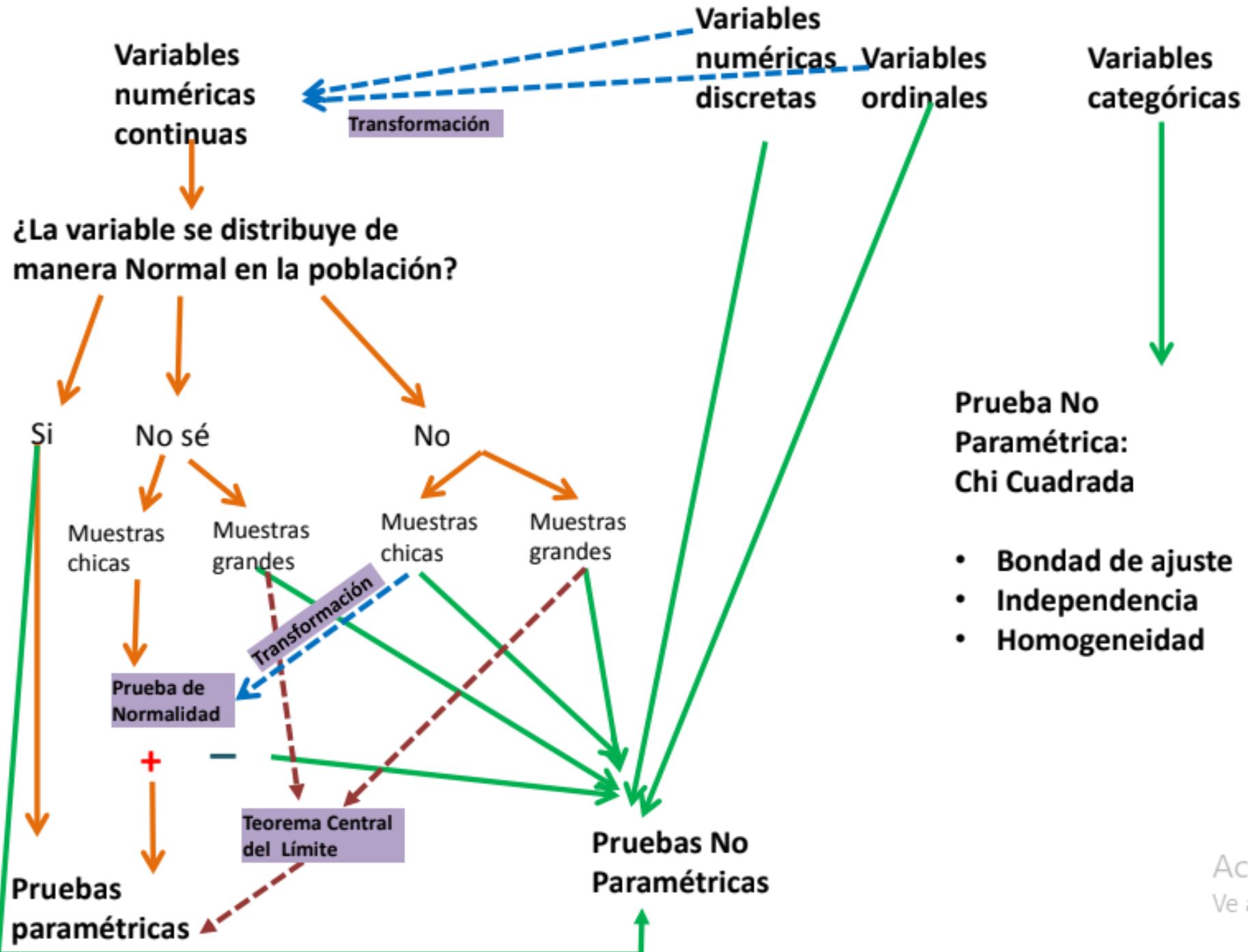


Salud: Evalúa la eficacia de tratamientos y analiza factores de riesgo para enfermedades.



Toma de Decisiones: En todos los campos, la estadística inferencial es una herramienta poderosa para la toma de decisiones basadas en datos.





Pruebas de hipótesis típicas

Antes de proseguir recordemos los conceptos de prueba de hipótesis:

1. Hipótesis nula (H_0): hipótesis que se desea contrastar, describe la conducta *default* del fenómeno de interés.
2. Hipótesis alternativa (H_1).
3. Estadística de prueba: es una estadística con base en la cuál tomamos la decisión de rechazar o no rechazar. Se calcula considerando la hipótesis nula como verdadera.
4. Valor-p: Nivel de significancia alcanzado, probabilidad de que la estadística de prueba sea al menos tan extrema como la observada con los datos si la hipótesis nula es verdadera.

Escenario	H_0 verdadera	H_0 Falsa
Rechazar H_0	Error Tipo 1 (α)	Decisión correcta
No rechazar H_0	Decisión correcta	Error tipo 2 (β)

Prueba T student

Comparación de dos medias (paramétrico)

**#Prueba de T

Hide

```
str(ChickWeight)
```

```
Classes 'nfnGroupedData', 'nfGroupedData', 'groupedData' and 'data.frame': 578 obs. of 4 variables:
 $ weight: num 42 51 59 64 76 93 106 125 149 171 ...
 $ Time   : num 0 2 4 6 8 10 12 14 16 18 ...
 $ Chick  : Ord.factor w/ 50 levels "18"<"16"<"15"<...: 15 15 15 15 15 15 15 15 15 15 ...
 $ Diet    : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 1 ...
 - attr(*, "formula")=Class 'formula' language weight ~ Time | Chick
 ... ...- attr(*, ".Environment")=<environment: R_EmptyEnv>
 - attr(*, "outer")=Class 'formula' language ~Diet
 ... ...- attr(*, ".Environment")=<environment: R_EmptyEnv>
 - attr(*, "labels")=List of 2
 ..$ x: chr "Time"
 ..$ y: chr "Body weight"
 - attr(*, "units")=List of 2
 ..$ x: chr "(days)"
 ..$ y: chr "(gm)"
```

```
ChickW<- ChickWeight %>% filter(Diet == c("1","2"))
t.test(weight~Diet, data = ChickW)
```

Welch Two Sample t-test

```
data: weight by Diet
t = -2.5871, df = 97.031, p-value = 0.01116
alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
95 percent confidence interval:
-49.849622 -6.568559
sample estimates:
mean in group 1 mean in group 2
101.3909      129.6000
```

Prueba de Mann-Whitney: es el equivalente para la prueba T para dos muestras independientes

No paramétrico

Ejemplo:

**#Prueba de Wilcoxon (Mann Whitney)

Hide

```
Hombres = c(19, 22, 16, 29, 24)  
Mujeres = c(20, 11, 17, 12)  
wilcox.test(Hombres, Mujeres)
```

Wilcoxon rank sum exact test

```
data: Hombres and Mujeres  
W = 17, p-value = 0.1111  
alternative hypothesis: true location shift is not equal to 0
```

**#El valor de pvalue es mayor a 0.05, por lo que no se rechaza la H0 nula

ANOVA

Prueba de ANOVA

[Hide](#)

```
aov(weight~Diet+Time,ChickW)
```

Call:

```
aov(formula = weight ~ Diet + Time, data = ChickW)
```

Terms:

	Diet	Time	Residuals
Sum of Squares	30893.9	420682.9	241867.7
Deg. of Freedom	1	1	167

Residual standard error: 38.05666

Estimated effects may be unbalanced

[Hide](#)

```
aov(weight~Diet*Time,ChickW)
```

Call:

```
aov(formula = weight ~ Diet * Time, data = ChickW)
```

Terms:

	Diet	Time	Diet:Time	Residuals
Sum of Squares	30893.9	420682.9	8091.8	233775.9
Deg. of Freedom	1	1	1	166

Residual standard error: 37.52717

Estimated effects may be unbalanced

```
aov.demo<-aov(weight~Diet*Time,ChickW)
summary(aov.demo)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Diet	1	30894	30894	21.937	5.83e-06	***
Time	1	420683	420683	298.719	< 2e-16	***
Diet:Time	1	8092	8092	5.746	0.0176	*
Residuals	166	233776	1408			

Signif. codes:	0	'****'	0.001	'**'	0.01	'*'
					0.05	'.'
					0.1	'.'
						1

##Prueba de Tukey

Hide

```
aov.demo<-aov(weight~Diet*as.factor(Time),ChickW)
TukeyHSD(aov.demo)
```

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

Fit: aov(formula = weight ~ Diet * as.factor(Time), data = ChickW)

\$Diet

diff	lwr	upr	p adj
------	-----	-----	-------

Test de Friedman

Comparación de 3 o más variables= equivalente a ANOVA

**Test de Friedman (valoracion de degustacion de un vino de acuerdo a la hora)

[Hide](#)

```
valoracion <- c( 9, 5, 2, 6, 3, 1, 5, 5, 5, 11, 5, 1, 8, 4, 3, 10, 4, 1, 7, 3, 4 )
hora <- factor( rep( c( "mañana", "tarde", "noche" ), 7 ) )
sujeto <- factor( rep( 1:7, each = 3 ) )
datos <- data.frame( valoracion, hora, sujeto )
head(datos)
```

	valoracion <code><dbl></code>	hora <code><fctr></code>	sujeto <code><fctr></code>
1	9	mañana	1
2	5	tarde	1
3	2	noche	1
4	6	mañana	2
5	3	tarde	2
6	1	noche	2

6 rows

[Hide](#)

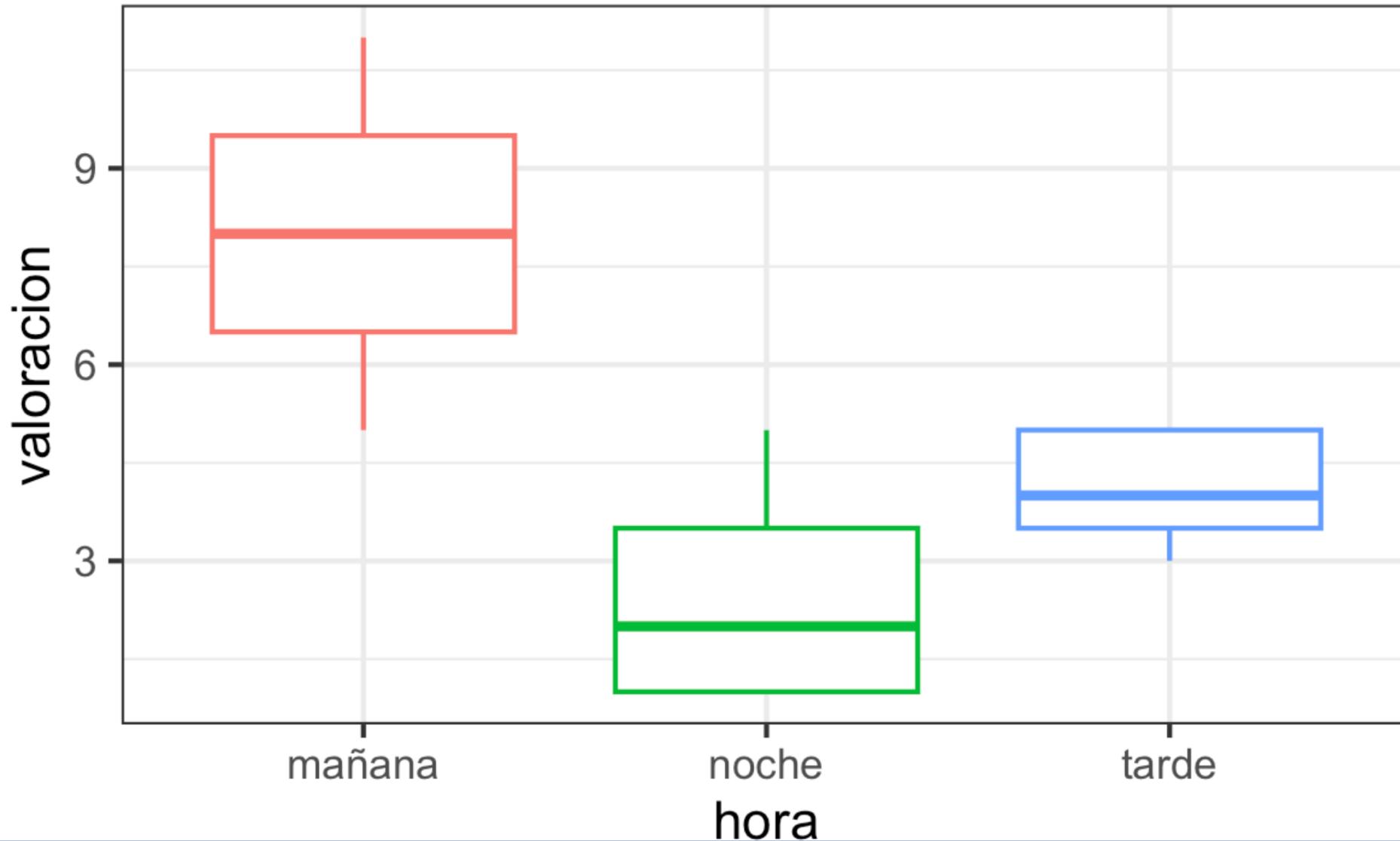
```
friedman.test(valoracion, hora, sujeto)
```

```
Friedman rank sum test

data: valoracion, hora and sujeto
Friedman chi-squared = 10.333, df = 2, p-value
= 0.005704
```

Al menos hay diferencia entre dos variables

```
#grafiquemos  
library(ggplot2)  
ggplot(data = datos, mapping = aes(x = hora, y = valoracion, colour = hora)) +  
  geom_boxplot() +  
  theme_bw() +  
  theme(legend.position = "none")
```



Prueba de Chi cuadrada (Pearson)

Para variables categóricas

PRUEBAS NO PARAMÉTRICAS

1. Para determinar si las frecuencias observadas dentro de cada categoría se ajustan a las frecuencias esperadas

Prueba de Chi cuadrada de bondad de ajuste

		genotipo	Frecuencia esperada	Frecuencia observada
		A A	25	20
		Aa	50	66
		aa	25	14
A	a			
AA	Aa			
a	aa			

2. Para determinar si en categorías combinadas, las frecuencias observadas dependen de la interacción entre los factores que definen las categorías

Prueba de Chi cuadrada de Independencia

	PRI	PAN	PRD
Hombre	20	33	47
Mujer	28	34	42

3. Para determinar si la proporción de frecuencias observadas en las categorías de una población, presentan la misma proporción en otras poblaciones

Prueba de Chi cuadrada de Homogeneidad

	Tipo sanguíneo			
	O	A	B	AB
Población 1	771	91	85	53
Población 2	803	82	89	26
Población 3	698	113	119	70
Población 4	790	61	88	61

**#Prueba de Chi cuadrada (ejemplo de encuesta de opinion sobre aborto)

Hide

```
M <- as.table(  
  rbind(c(762, 468),  
        c(484, 477))  
)
```

```
# Damos nombre a las columnas y las filas  
colnames(M) <- c("A favor", "En contra")  
rownames(M) <- c("Mujeres", "Hombres")  
M
```

rbind= une vectores en tabla

	A favor	En contra
Mujeres	762	468
Hombres	484	477

Hide

```
chisq.test(M)
```

Pearson's Chi-squared test with Yates'
continuity correction

```
data: M  
X-squared = 29.06, df = 1, p-value = 7.019e-08
```

→ Menor a 0.05