ENES
JURIQUILLA

CURSO INTER-SEMESTRAL
INTRODUCCIÓN A R

PUBLICO: PROFESORES, INVESTIGADORES Y ESTUDIANTES
REQUISITOS: NINGUNO
DURACIÓN DEL CURSO: 20 HRS
FECHAS: 17-21 DE ENERO 2022

Profesores:
Dr. Alberto Prado
Dr. Ulises Olivares Pinto

ESTADÍSTICA DESCRIPTIVA
4 HORAS

- Repaso del día 1
- Estadística descriptiva
  - Boxplot
  - T test
- Más sobre histogramas y distribuciones probabilísticas
- Cargar librerías
- Estandarizar datos
- Correlaciones
- Regresiones lineales
- Agregar colores y formas a las gráficas

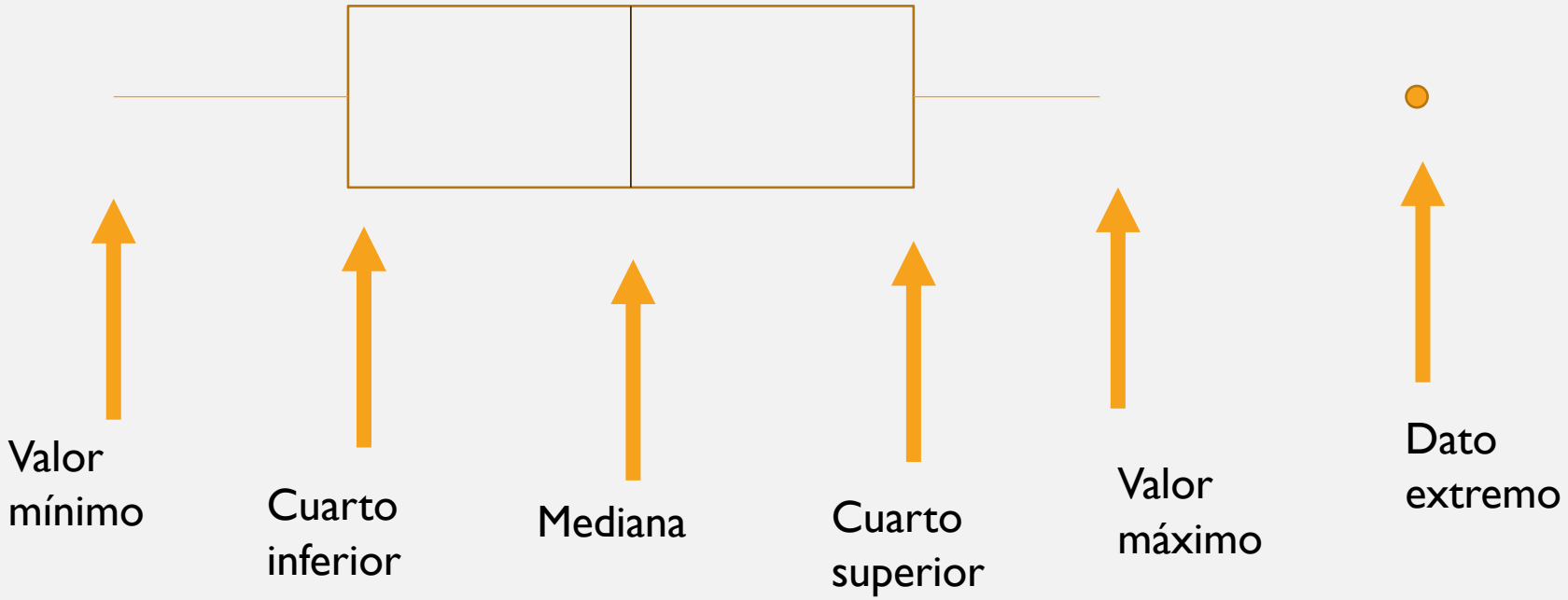
Ejercicio:

- Importar el archivo "mtcars.csv"
- Checar estructura
- Cambiar nombre de las filas por columna "X" y eliminar la columna.
- Cambiar el nombre de las columnas 1,4, 6 y 9 al español.
- Hacer un histograma de las millas por galón (mpg) con la función **hist()**
- Hacer un gráfico dispersión de los caballos de fuerza en función del peso.
- Colorear por tipo de transmisión.

Col	Nombre original	Descripción
[, 1]	mpg	Millas por galón
[, 2]	cyl	Número of cilindros
[, 3]	disp	Desplazamiento (cu.in.)
[, 4]	hp	Caballos de fuerza
[, 5]	drat	Vueltas del eje trasero con relación al de accionamiento
[, 6]	wt	Peso (1000 lbs)
[, 7]	qsec	Tiempo en segundos para recorrer 1/4 de milla
[, 8]	vs	Forma del motor (0 = Forma de V, 1 = recto)
[, 9]	am	Transmisión (0 = automatic, 1 = manual)
[,10]	gear	Número de velocidades
[,11]	carb	Número de carburadores

Gráfica de caja y bigotes

Cuartiles: Valores que dividen la muestra en 4 partes iguales



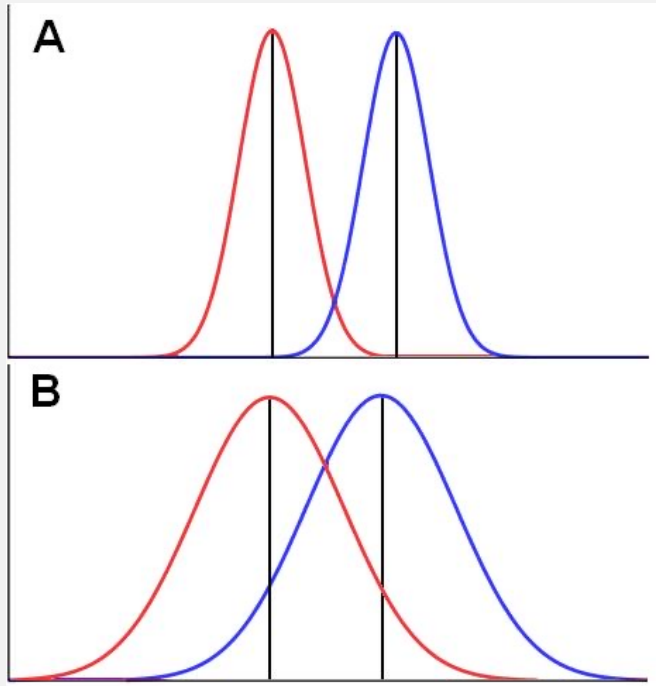
boxplot(y~x,data)

y variable numérica
x variable nominal (factor)

Ejercicio 1:

- Gráfico de caja y bigote del peso en función de la transmisión
- Cambiar los niveles de factor usando las función **factor()**
- **factor(mtcars$am, labels=c("automático", "manual"))**
- Volver a hacer el gráfico con la primer caja de color naranja y la segunda de color azul.

Prueba T Student



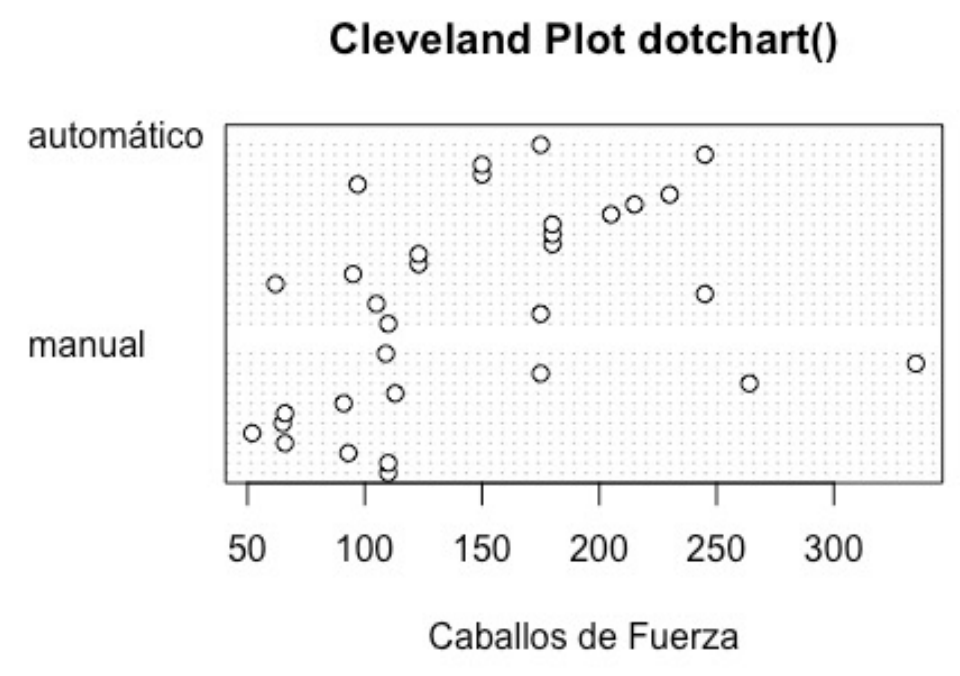
$$t = \frac{\text{Señal}}{\text{Ruido}}$$

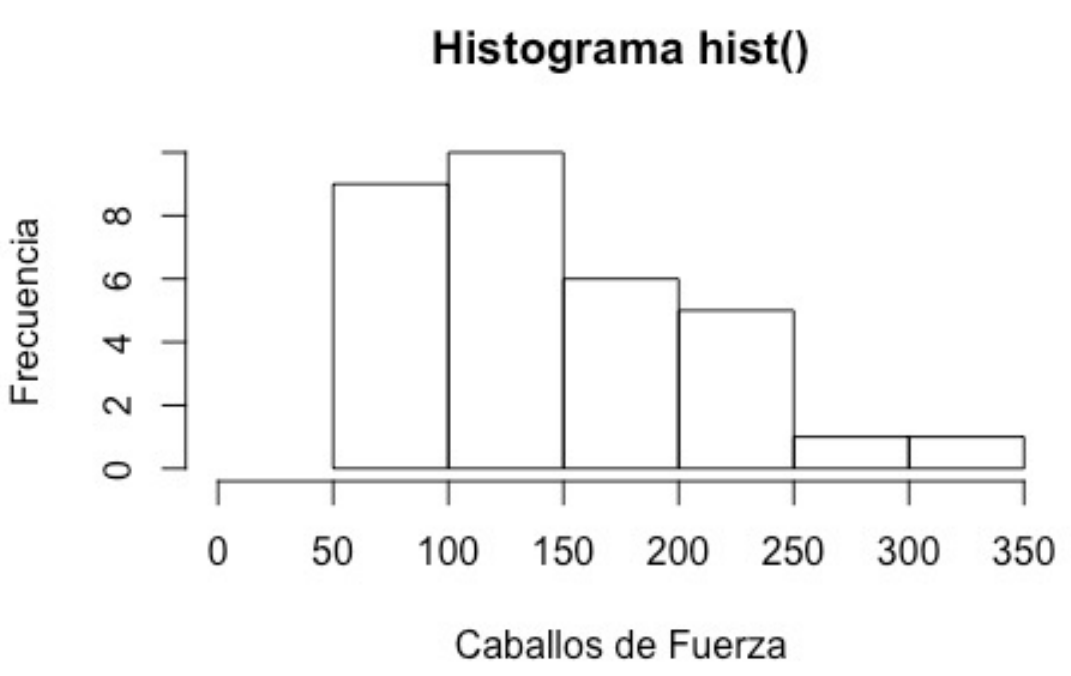
t.test(y~x,data)

Ejercicio 2:

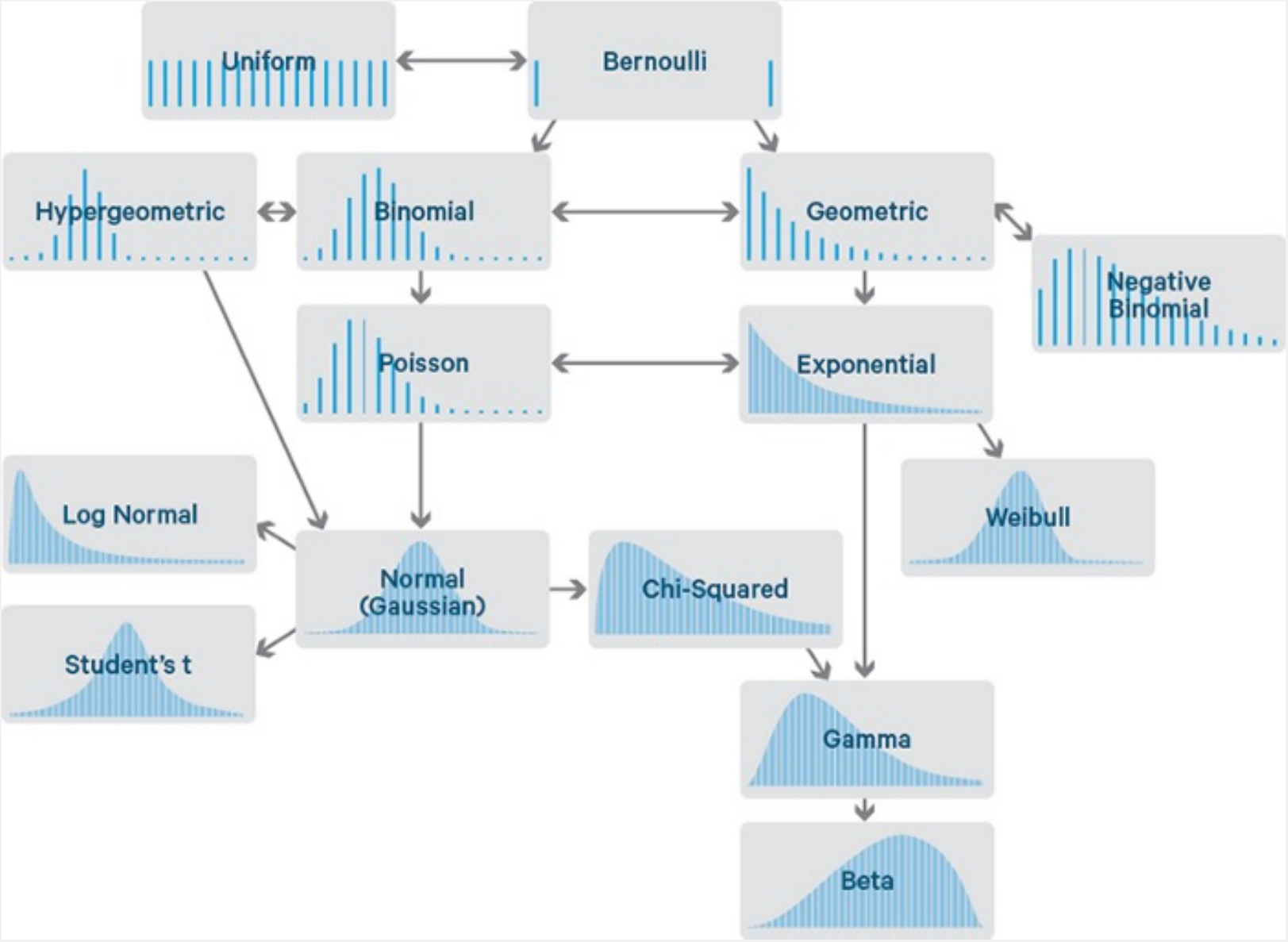
- Gráfico de caja y bigote de los caballos de fuerza en función de la forma del motor
- ¿Existe diferencia entre las medias?
- ¿Qué podemos decir de la distribución de los datos?

Visualización de la distribución de los datos





Distribuciones



Ejercicio 3:

- Cargar el archivo "cars.csv" que contiene la velocidad máxima en millas por hora y la distancia de frenado en pies de 50 autos de 1920.
- Verificar la distribución de los datos usando **dotchart()** y **hist()**
- ¿Que tipo de distribución tienen?

Si quisiéramos estimar visualizar una curva de densidad de probabilidad
plot(density(cars$speed))

Cargar librerías

¿Qué es una librería o "package" ?

Una librería es un código compilado que define funciones con las cuales no contamos. Las librerías incluyen su manual de uso y son revisadas y actualizadas.

install.packages("MASS")

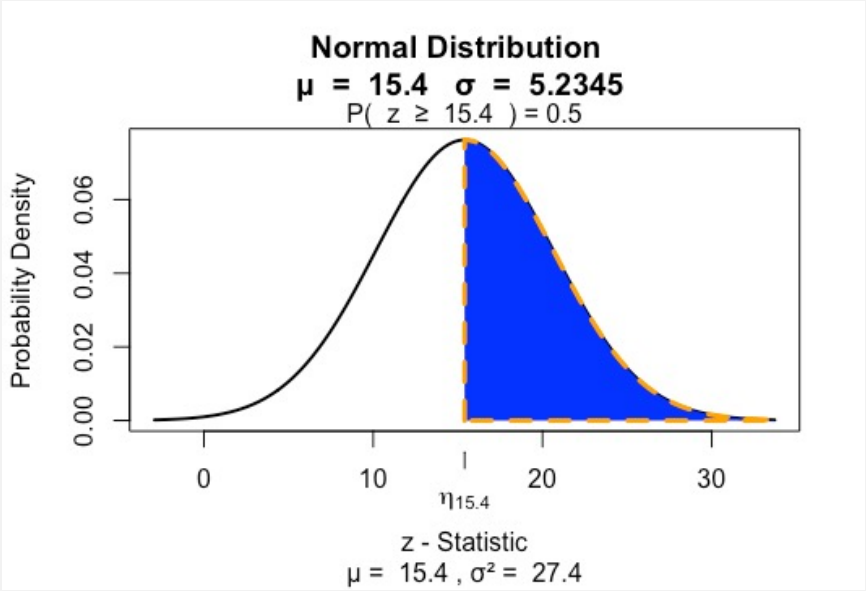
Una vez instalado hay que cargar el paquete
Library("MASS")

Podemos obtener información directamente en R con:
packageDescription ("MASS")
help(package = "MASS")

Ejercicio 4:

- Descargar la librería **MASS**
- Buscar en internet el manual de uso
- Buscar en el manual la información concerniente a la función **fitdistr()** que permite ajustar distribuciones a nuestros datos
- Obtener los parámetros que describen la distribución de nuestros datos.

Librería visualize



Ejercicio 5

- Descargar la librería **visualize**
- Usar la función **visualize.norm(mu= media, sd= desviación estandar)** para visualizar la función que se ajusto a los datos de velocidad
- Usar las argumentos **stat=25, section="upper"** para conocer la probabilidad de que un coche en 1920 alcanzará mas de 25 millas por hora.

Ejercicio 6

- Usar la funcion **visualize** para calcular la probabilidad de que un coche de 1920 frene en menos de 20 pies.
- Usa una distribución log-normal usando el argumento “lognormal

Estandarizar datos

Restar la media y dividir entre la desviación estándar

mean()

sd()

Ejercicio 7:

- Crear una nueva columna con los datos de distancia estandarizados
- Hacer un histograma
- Ajustar una distribución normal
- Usar la función **visualize.norm()** para calcular la probabilidad de que un coche de 1920 frene en menos de 20 pies.

Ejercicio 8:

- Carga los datos de mm lluvia diaria llamado “lluvia.csv”
- Haz un histograma de la distribución de los mm de lluvia
- Ajusta una distribución exponencial
- Usar la función **visualize.exp()** para calcular la probabilidad de que en un día dado llueva más de 2 mm
- Nota: La distribución exponencial se describe con un solo parámetro, que es la tasa de cambio (rate). En la función **visualize.exp()** se especifica con el argumento **theta**

Correlaciones

- La correlación es una medida sin unidades de que tan fuerte dos variables se relacionan.
- **Nota: la correlación no implica causalidad**

cov (cars$dist,cars$speed) covarianza
cor(cars$dist,cars$speed) coeficiente de correlación

Los datos “advertising.csv” contienen información sobre las ventas en miles de dolares y el gasto que se hizo en tres diferentes medios de comunicación en los que se hizo la publicidad.

Ejercicio 9:

- Carga el archivo “advertising.csv”
- Revisa su estructura
- Haz una matríz de correlaciones con **cor()**
- ¿Qué variable se correlaciona más fuerte con las ventas?

Regresiones lineales

- El análisis de regresión es la parte de la estadístca que se encarga de estudiar la relación entre dos o mas variables asociadas de manera no determinística.
- Variable explicativa (independiente)
- Variable de respuesta (dependiente)

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

Dónde:

Y = variable dependiente

βo= coeficiente o intercepto en el eje Y

β1=pendiente de la regresión

X1= variable independiente

ε = margen del error

lm(y~x)

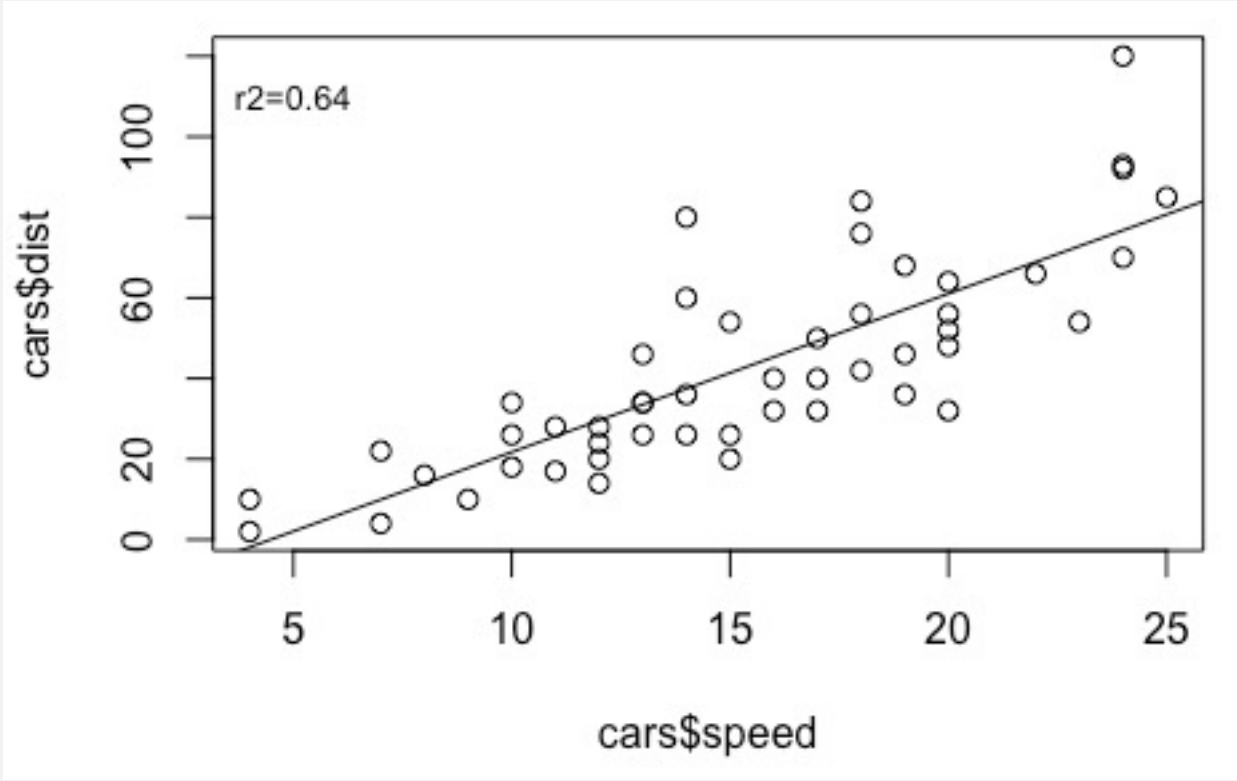
Ejemplo:

```
lm1 <- lm(cars$dist~cars$speed)
lm1
summary(lm1)
plot(cars$dist~cars$speed)
abline(a= intercepto, b= pendiente)
text(x=5,y=110,"r2=0.064", cex=0.8)
```

coordenadas

texto

tamaño



Ejercicio 10:

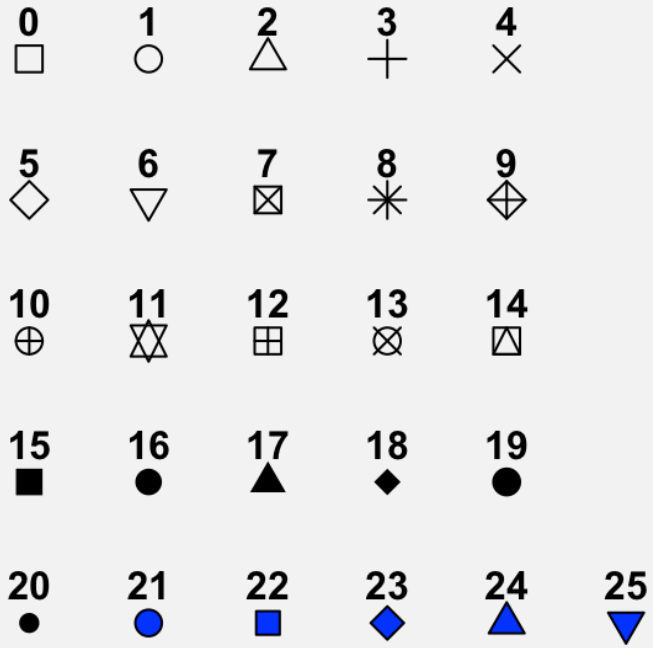
- Hacer tres regresión lineales para explicar las ventas
- Hacer tres gráficos de dispersión con sus regresiones lineales
- Agregar coeficiente de determinación (r cuadrada)
- Usar el coeficiente de determinación para escoger el medio de comunicación que explica mejor las ventas.
- Haz una regresión lineal de la longitud del pétalo en función de su ancho
- ¿Cuál modelo es el que explica mayor variación de los datos?

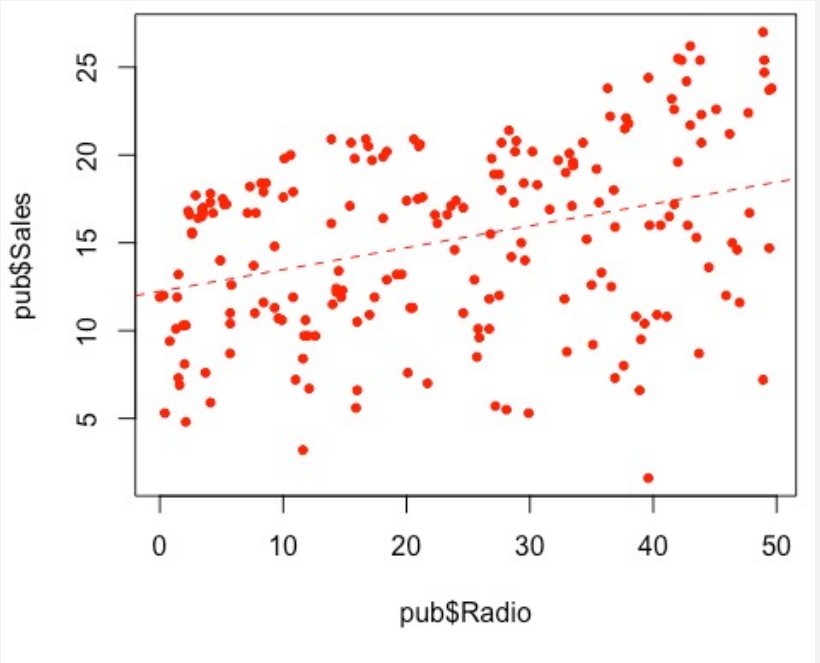
Colores


white aliceblue antiquewhite antiquewhite1 antiquewhite2
antiquewhite3 antiquewhite4 aquamarine aquamarine1 aquamarine2
aquamarine3 aquamarine4 azure azure1 azure2
azure3 azure4 beige bisque bisque1
bisque2 bisque3 bisque4 blanchedalmond
blue blue1 blue2 blue3 blue4
blueviolet brown brown1 brown2 brown3
brown4 burlywood burlywood1 burlywood2 burlywood3
burlywood4 cadetblue cadetblue1 cadetblue2 cadetblue3
cadetblue4 chartreuse chartreuse1 chartreuse2 chartreuse3
chartreuse4 chocolate chocolate1 chocolate2 chocolate3
chocolate4 coral coral1 coral2 coral3
coral4 cornflowerblue cornsilk cornsilk1 cornsilk2
cornsilk3 cornsilk4 cyan cyan1 cyan2
cyan3 cyan4 darkblue darkcyan darkgoldenrod
darkgoldenrod1 darkgoldenrod2 darkgoldenrod3 darkgoldenrod4 darkgray
darkgreen darkgrey darkkhaki darkmagenta darkolivegreen
darkolivegreen1 darkolivegreen2 darkolivegreen3 darkolivegreen4 darkorange
darkorange1 darkorange2 darkorange3 darkorange4 darkorchid
darkorchid1 darkorchid2 darkorchid3 darkorchid4 darkred
darksalmon darkseagreen darkseagreen1 darkseagreen2 darkseagreen3
darkseagreen4 darkslateblue darkslategray darkslategray1 darkslategray2
darkslategray3 darkslategray4 darkslategrey darkturquoise darkviolet
deeppink deeppink1 deeppink2 deeppink3 deeppink4
deepskyblue deepskyblue1 deepskyblue2 deepskyblue3 deepskyblue4

Símbolos

Los símbolos se establecen con el argumento **pch**

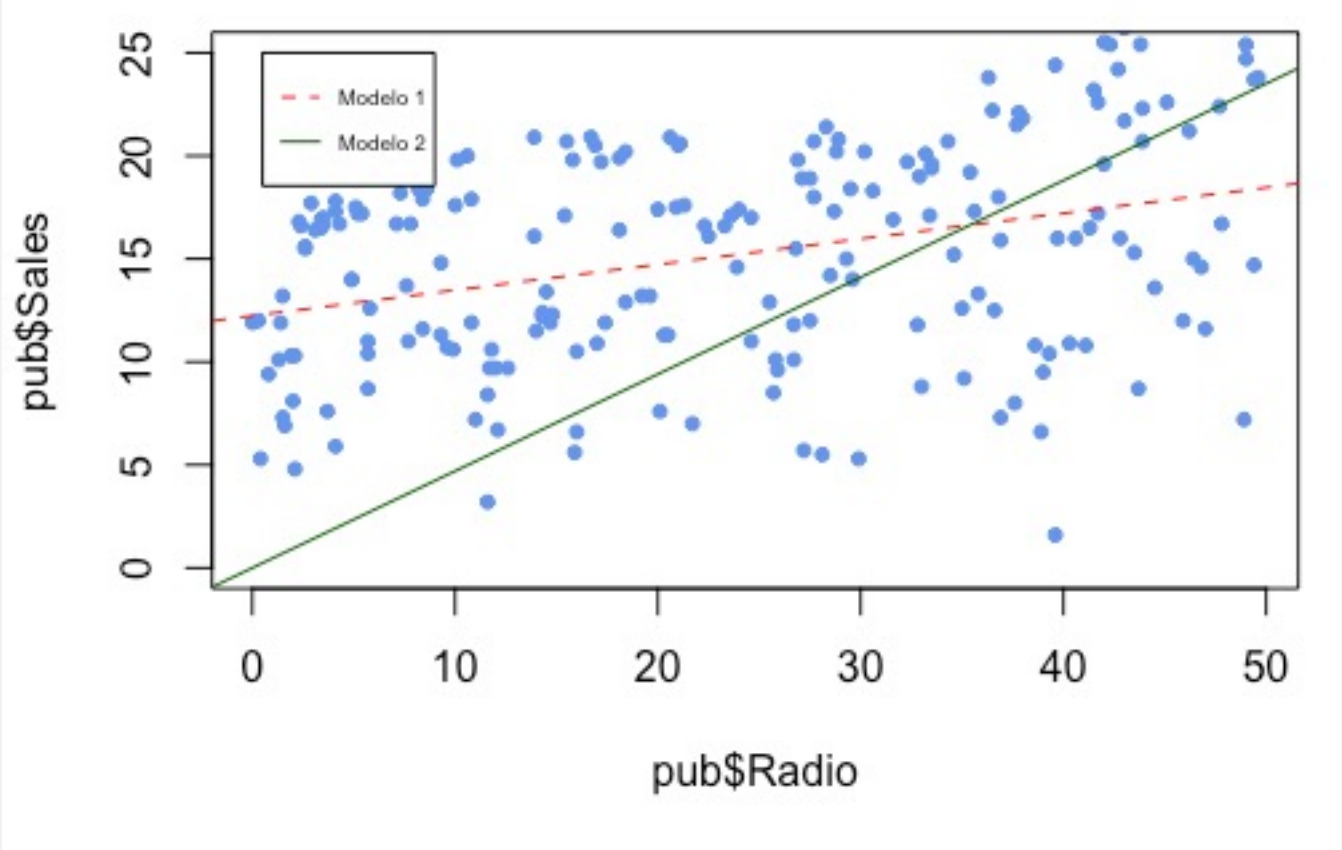





```
plot(pub$Sales~pub$Radio, col="red", pch=20)
```

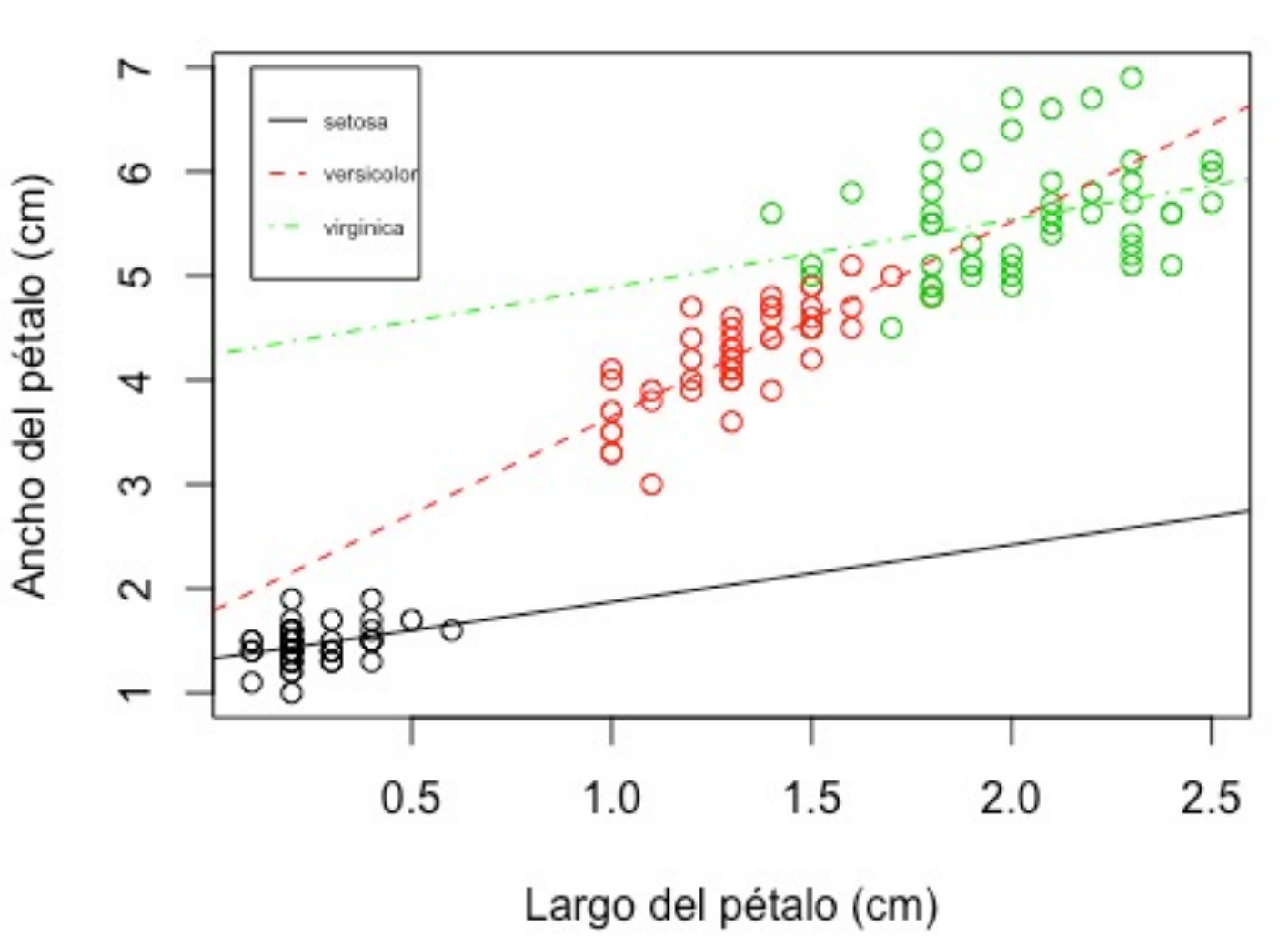
Añadir una leyenda

```
legend(x, y, legend=c("texto 1", "texto 2"), col=c("red", "darkgreen"), lty=c(2,1), cex=0.5)
```



Ejercicio 11:

- Usar los datos iris para hacer un gráfico de caja y bigote por especie del largo del pétalo donde cada caja tenga un color diferente. **boxplot(y~Species)**
- Encontrar las funciones de distribución de probabilidad para cada especie y graficarlas. **fitdistr()**
- Calcular la probabilidad de que la longitud del pétalo sea mayor a 5 cm. ***visualize***
- Haz una prueba T Student para saber si existe una diferencia entre la longitud del pétalo entre versicolor y virginica. **iris[iris$Species==“versicolor” | iris$Species==“virginica,]**
- Crea una matriz de correlación entre las cuatro variables numéricas, ¿Cuáles son las dos variables que se correlacionan de manera mas fuerte? ¿Cuáles son las dos variables que se correlacionan de manera mas débil?
- **cor()**
- Haz una gráfica de dispersión de los datos de longitud del pétalo en función de su ancho donde cada especie tenga un color y forma.
- Añade los tres modelos lineales con sus respectivos colores y con tipos de línea diferentes. **abline()**


setosa
versicolor
virginica
Ancho del pétalo (cm)
Largo del pétalo (cm)
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5