ACTIVIDAD 3 - UD3

Apartado A

Pensemos que tenemos el CI de una muestra de 5 individuos: 110 100 115 105 104

Queremos (caso 1) calcular el intervalo de confianza sobre la media (al 95%) y (caso 2) indicar si tales sujetos han sido extraídos de una población con media 100.

Apartado B

Dado el dataset survey visto previamente, queremos contrastar si la diferencia del pulso entre hombres y mujeres es diferente o no ¿se puede considerar que el pulso de las mujeres es superior al de los hombres a un nivel de confianza del 90% ?

Apartado C

Ley de Benford e ha mandado una factura a la empresa Xdata que parece ser falsa. Esta factura tiene muchos números que no parecen generados de modo natural. Comprobamos si efectivamente la factura está generada artificialmente basándonos en la Ley de Benford https://es.wikipedia.org/wiki/Ley_de_Benford

Esta ley trata sobre la distribución de los primeros dígitos en: - facturas - artículos en revistas - direcciones de calles - precios de acciones - número de habitantes - tasas de mortalidad - longitud de los ríos - Física - constantes matemáticas - números primos

La ley Benford establece que la distribución natural de los primeros dígitos es

```
0.301,0.176,0.125,0.097,0.079,0.067,0.058,0.051,0.046
```

Las frecuencias de los primeros dígitos de las facturas de la empresa resultan ser

```
 7, 13, 12,  9,  9, 13, 11, 10, 16
```

¿Son facturas falsas?

Apartado D

Carga el dataset “PlantGrowth”

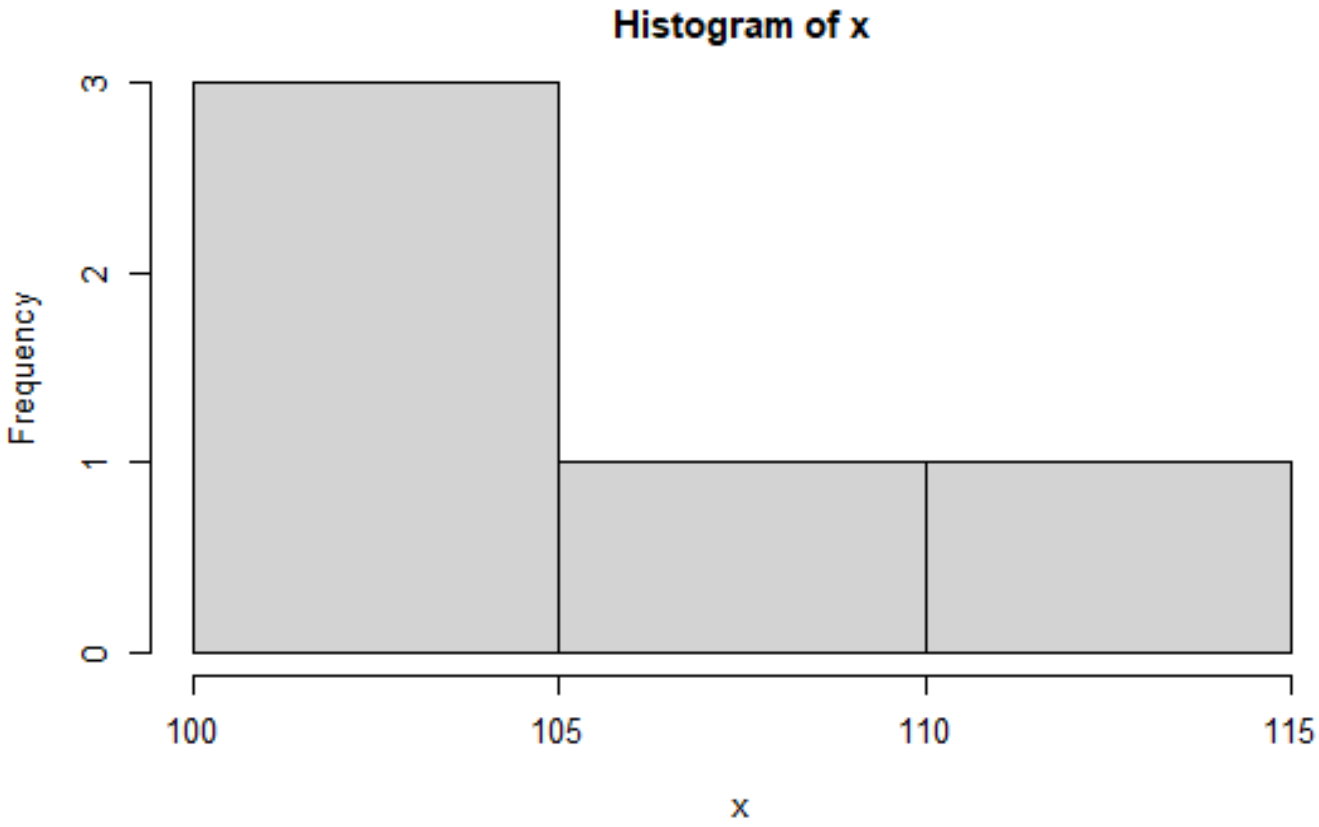
- ¿Se puede considerar que con los tres tratamientos las plantas tienen el mismo crecimiento?
- Haz un análisis exploratorio
- Comprueba las asunciones del modelo
- Realiza el one-way ANOVA
- ¿Qué conclusiones se pueden inferir de esta muestra?

Apartado A

Pensemos que tenemos el CI de una muestra de 5 individuos: 110 100 115 105 104

Queremos (caso 1) calcular el intervalo de confianza sobre la media (al 95%) y (caso 2) indicar si tales sujetos han sido extraídos de una población con media 100.

```
x<- c(110,100,115,105,104)
hist(x)
```



Calculo primero la media para intuir por donde ir.

```
mean(x)
[1] 106.8
```

Con el test basado en t-Student, Ho= media es 0 y aquí ya vemos que la media es mucho mayor, así que el p-valor tiene que ser menor que 0.05 (le ponemos conf.level=0.95).

```
t.test (x, conf.level = 0.95)


    One Sample t-test


data:  x
t = 41.138, df = 4, p-value = 2.087e-06
```

```
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  99.59193 114.00807
sample estimates:
mean of x
    106.8
```

Confirmado con el test p= 2.087e-06 <0.05, nos quedamos con hipótesis alternativa, media distinta de 0. Hacemos el test poniendo la media que nos dio antes (106.8)

```
t.test(x, mu=106.8,conf.level = 0.95)


    One Sample t-test


data:  x
t = 0, df = 4, p-value = 1
alternative hypothesis: true mean is not equal to 106.8
95 percent confidence interval:
  99.59193 114.00807
sample estimates:
mean of x
    106.8
```

Tengo p-valor de 1, así que Ho es TRUE= la media es de 106,8 para 95% y el intervalo de confianza es de (99.59193- 114.00807)

indicar si tales sujetos han sido extraídos de una población con media 100.

Hemos calculado la media e la muestra que no es la media de la población. Para conocer la media de la población hay que hacer inferencia de datos, eso es lo que hace el test. Da 106.8 con intervalo de confianza de 99.59 y 114.008. Y el 100 está dentro de ese intervalo. Así que podría ser la media.

Apartado B

Dado el dataset survey visto previamente, queremos contrastar si la diferencia del pulso entre hombres y mujeres es diferente o no ¿se puede considerar que el pulso de las mujeres es superior al de los hombres a un nivel de confianza del 90% ?

```
library(MASS)
library(dplyr)
library(readxl)
data("survey")
survey
NA
```

```
summary(survey)
```

```
     Sex          Wr.Hnd          NW.Hnd        W.Hnd          Fold         Pu
lse             Clap       Exer       Smoke         Height            M.I
Age

 Female:118   Min.   :13.00   Min.   :12.50   Left : 18   L on R : 99   Min.
: 35.00   Left   : 39   Freq:115   Heavy: 11   Min.   :150.0   Imperial: 68
Min.   :16.75

 Male  :118   1st Qu.:17.50   1st Qu.:17.50   Right:218   Neither: 18   1st Qu
.: 66.00   Neither: 50   None: 24   Never:189   1st Qu.:165.0   Metric  :141
1st Qu.:17.67

 NA's  :  1   Median :18.50   Median :18.50   NA's :  1   R on L :120   Median
: 72.50   Right  :147   Some: 98   Occas: 19   Median :171.0   NA's    : 28
Median :18.58

              Mean   :18.67   Mean   :18.58                             Mean
: 74.15   NA's   :  1              Regul: 17   Mean   :172.4
Mean   :20.37

              3rd Qu.:19.80   3rd Qu.:19.73                             3rd Qu
.: 80.00                            NA's :  1   3rd Qu.:180.0
3rd Qu.:20.17

              Max.   :23.20   Max.   :23.50                             Max.
:104.00                                         Max.   :200.0
Max.   :73.00

              NA's   :1       NA's   :1                                 NA's
:45                                             NA's   :28
```

```
df<- subset(survey, select = c("Sex","Pulse"))


df <- na.omit(df)
```

Primero claculo con los datos del dataframe la media, para ver por dónde ir

```
desviacion_standar <- aggregate(Pulse~Sex,df,sd)

desviacion_standar

NA


media <- aggregate(Pulse~Sex,df,mean)

media

NA
```

Veo con esto las medias y las desviaciones típicas de cada distribución. Usaremos el test de Welch, suponemos distribución normal, y suponemos que las varianzas son diferentes.

La hipótesis de partida,Ho, es que las medias son iguales y H1, medias diferentes

```
Fx <- (filter(df,Sex=="Female"))


Fx <- Fx [,-1]

Fx

 [1]  92  64  74  80  66  89  64  76  72  72  80  70  60  50  70  72  70  64
64  64  68  40  88  68  76  68  98  76  70  75  92  70  60  68  72  80  80  85
76  75  60  70  70
```

```
[44] 100  92  68  74  90  86  80  68  84  65  68  92  64  80  92  74  80  60
81  70  65  50  48  68 104  84  70  87  79  79  72  76  80  80  80  61  76  86
83  76  74  83  68

[87]  70  88  96  80  70  80  85  88  85

 var(Fx)

[1] 130.1115

Fy <- (filter(df,Sex=="Male"))


Fy <- Fy [,-1]


Fy

 [1] 104  87  35  83  72  90  68  60  74  78  72  72  64  62  90  90  62  76
79  78  70  54  66  72  80  60  84  96  55  68  78  56  65  72  62  66  80  66
76  90  60  75  65

[44]  68  60  64  67  80  60  83 100  80  76  59  66  66  86  60  85  90  72
96  75  64  60  76  69  68  76  84  72  72  80  76  68  70  90  72  65  62  63
92  60  68  71  48

[87]  80  84  97  78  65  88  75  68  71  90

var(Fy)

[1] 143.992

mujeres<- c(Fx)

hombres<- c(Fy)

t.test(mujeres,hombres,alternative = "l")


    Welch Two Sample t-test


data:  mujeres and hombres

t = 1.1384, df = 188.7, p-value = 0.8718

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

     -Inf 4.728468

sample estimates:

mean of x mean of y

 75.12632  73.19792
```

Vemos que nos sale un p-value 0.87>0.05, la Ho verdadera. Las medias son iguales. Pero hemos visto con otras métricas que no lo son. Falla.

```
t.test(mujeres,hombres,alternative = "l",conf.level = 0.9)


    Welch Two Sample t-test


data:  mujeres and hombres
```

```
t = 1.1384, df = 188.7, p-value = 0.8718
alternative hypothesis: true difference in means is less than 0
90 percent confidence interval:
     -Inf 4.106928
sample estimates:
mean of x mean of y
 75.12632  73.19792
```

Con un 90% también falla, nos dice que son iguales y son diferentes.

Apartado C

Ley de Benford

Se ha mandado una factura a la empresa Xdata que parece ser falsa. Esta factura tiene muchos números que no parecen generados de modo natural. Comprobamos si efectivamente la factura está generada artificalmente basándonos en la Ley de Benford https://es.wikipedia.org/wiki/Ley_de_Benford

Esta ley trata sobre la distribución de los primeros dígitos en: - facturas - artículos en revistas - direcciones de calles - precios de acciones - número de habitantes - tasas de mortalidad - longitud de los ríos - Física - constantes matemáticas - números primos

La ley Benford establece que la distribución natural de los primeros dígitos es

```
  0.301,0.176,0.125,0.097,0.079,0.067,0.058,0.051,0.046
```

Las frecuencias de los primeros dígitos de las facturas de la empresa resultan ser

```
  7, 13, 12,  9,  9, 13, 11, 10, 16
```

¿Son facturas falsas?

```
    Ley de Benford:
```

Primer dígito Proporción esperada

1 30.1%

2 17.6%

3 12.5%

4 9.7%

5 7.9%

6 6.7%

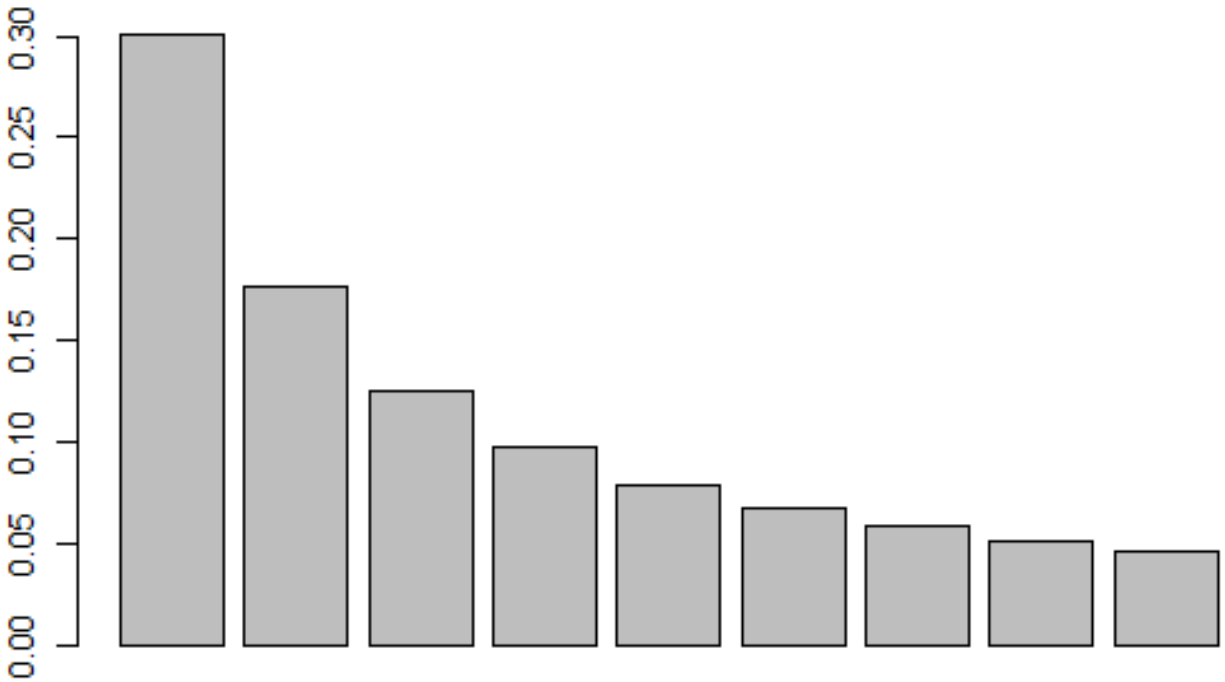
7 5.8%

8 5.1%

9 4.6%

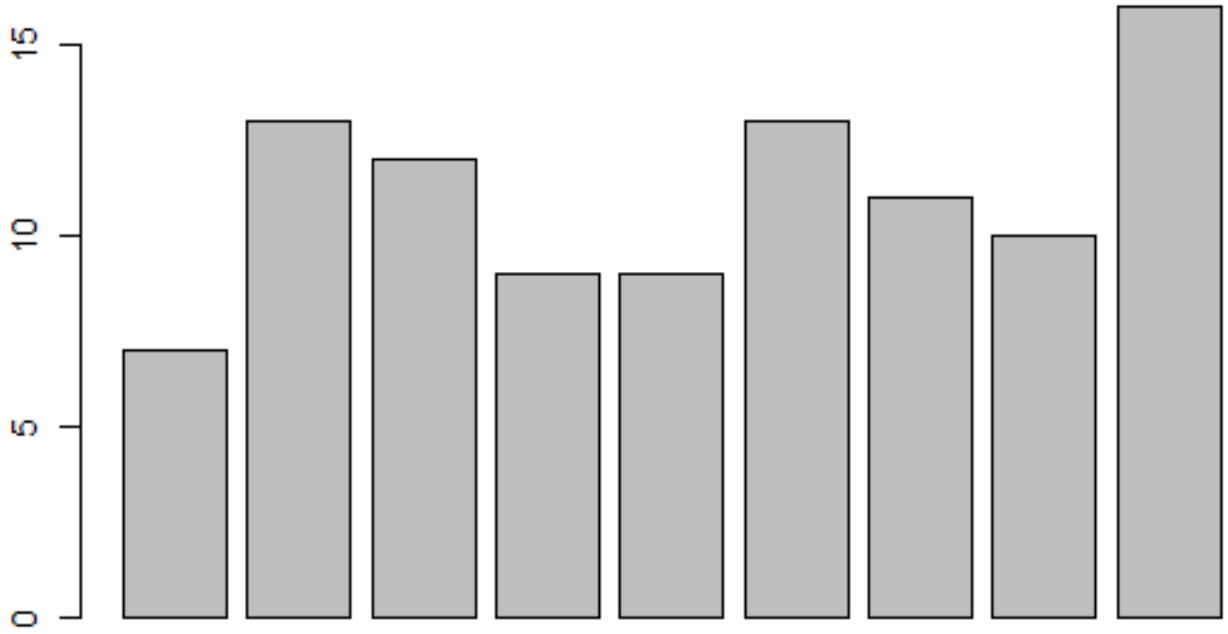
```
p<- c(0.301,0.176,0.125,0.097,0.079,0.067,0.058,0.051,0.046)
```

```
p
[1] 0.301 0.176 0.125 0.097 0.079 0.067 0.058 0.051 0.046

sum(p) # Comprobamos que suma 1
[1] 1
x<- c(7, 13, 12,  9,  9, 13, 11, 10, 16)

x
[1]  7 13 12  9  9 13 11 10 16
sum(p) # Comprobamos que suma 1
[1] 100
barplot(p)
```



Esta es la ley de Benford

```
barplot(x)
```



Esta es la gráfica de la muestra. No coinciden para nada.

```
p<- c(0.301,0.176,0.125,0.097,0.079,0.067,0.058,0.051,0.046)

x<- c(7, 13, 12,  9,  9, 13, 11, 10, 16)

N <- sum(p)
chisq.test(p,x = x)
Warning: Chi-squared approximation may be incorrect

    Pearson's Chi-squared test

data:  x and p
X-squared = 54, df = 48, p-value = 0.2559
```

La distribución teórica no coincide con la distribución que tenemos. Son facturas falsas.

Apartado D

Carga el dataset “PlantGrowth”

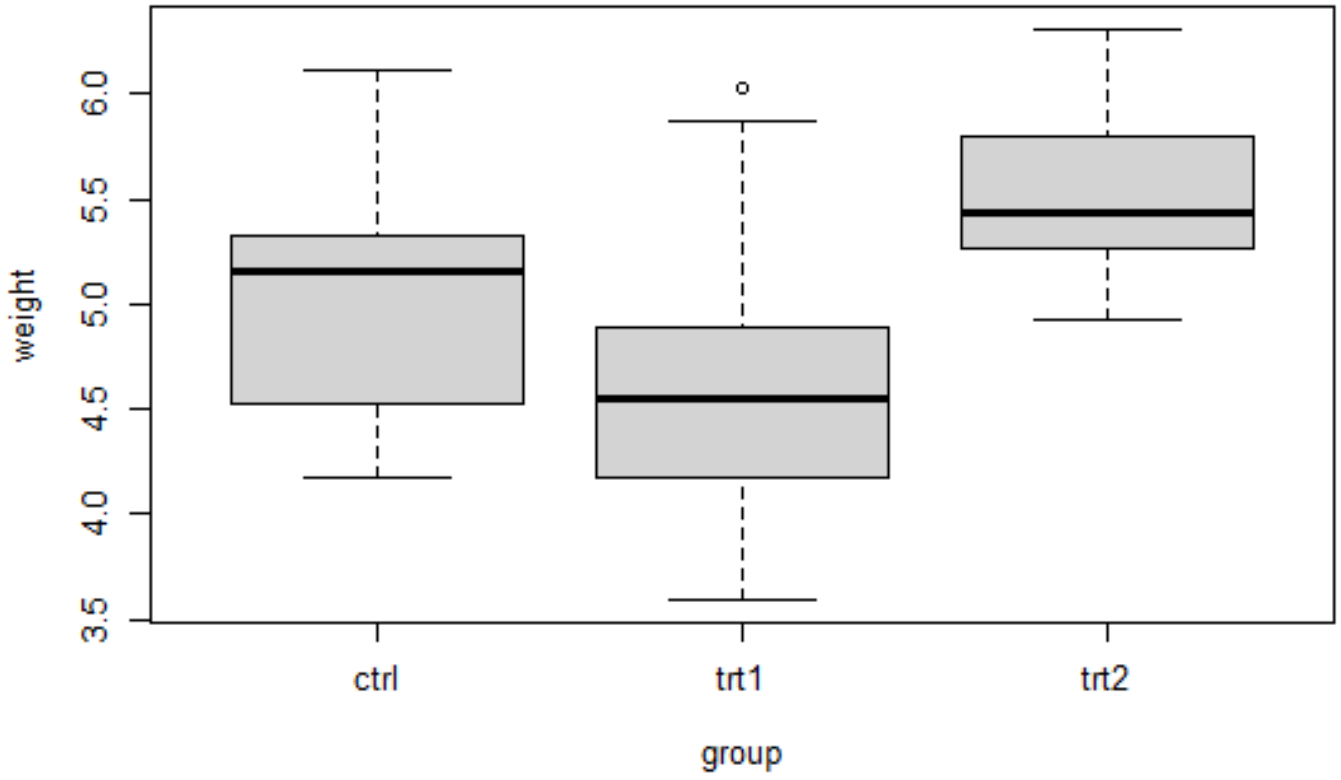
```
data("PlantGrowth")
PlantGrowth
```

- ¿Se puede considerar que con los tres tratamientos las plantas tienen el mismo crecimiento?
- Haz un análisis exploratorio
- Comprueba las asunciones del modelo
- Realiza el one-way ANOVA
- ¿Qué conclusiones se pueden inferir de esta muestra?

```
unique(PlantGrowth$group)
[1] ctrl trt1 trt2
Levels: ctrl trt1 trt2
```

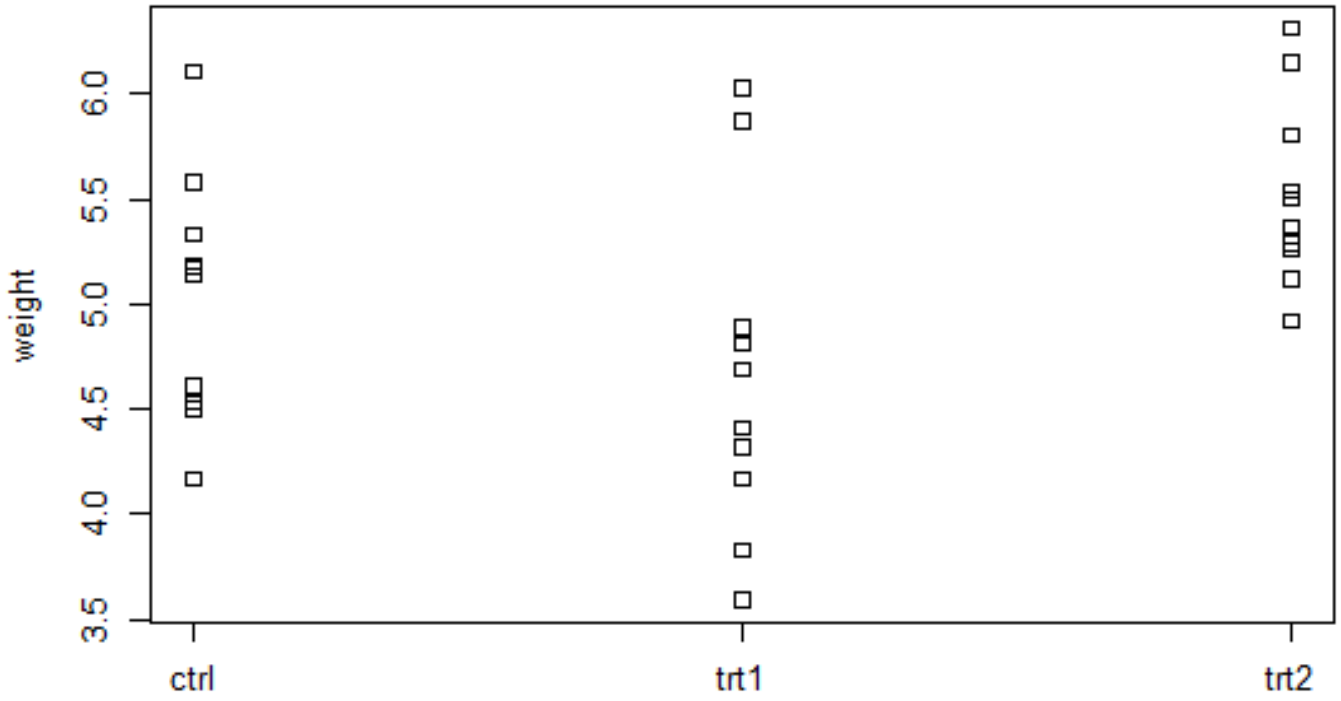
Tengo ya los tres grupos.

```
summary(PlantGrowth)
     weight       group
 Min.   :3.590   ctrl:10
 1st Qu.:4.550   trt1:10
 Median :5.155   trt2:10
 Mean   :5.073
 3rd Qu.:5.530
 Max.   :6.310

attach(PlantGrowth)
The following objects are masked from PlantGrowth (pos = 3):

    group, weight
boxplot(weight ~ group, ylab = "weight")
```



Ya intuimos con la representación que las medias no son iguales,aunque la escala en peso es pequeña. No tienen el mismo crecimiento a priori

```
stripchart(weight ~ group, vertical=TRUE)
```



Ahora se calculan las medias y las desviaciones típicas de cada grupo de tipo de crecimiento

```
meansd <- function(x) c(mean=mean(x), sd=sd(x))
by(weight, group, FUN=meansd)
group: ctrl
     mean        sd
5.0320000 0.5830914
------------------------------------------------------------------------------
------------------------------------------------------
group: trt1
     mean        sd
4.6610000 0.7936757
------------------------------------------------------------------------------
------------------------------------------------------
group: trt2
     mean        sd
5.5260000 0.4425733
```

El ANOVA se basa en igualdad o diferencia de varianzas en las muestras.

1. ANOVA si asumimos ue las varianzas son iguales:

```
oneway.test(weight ~ group, var.equal=TRUE)

    One-way analysis of means

data:  weight and group
F = 4.8461, num df = 2, denom df = 27, p-value = 0.01591
```

Nos dice que p=0.01 < 0.05, entonces la Ho= las medias son iguales es falsa. LAS MEDIAS SON DIFERENTES (no especifica si son las tres diferentes o si una de ellas sólo…)

2.ANOVA si asumimos varianzas diferentes

```
oneway.test(weight ~ group)

    One-way analysis of means (not assuming equal variances)

data:  weight and group
F = 5.181, num df = 2.000, denom df = 17.128, p-value = 0.01739
```

Aquí nos da un valor de p= 0.01 < 0.05, nos dice lo mismo que la prueba anterior, Ho no se cumple, mejor asumir H1= las medias son diferentes