Trabajo estadística

yo

06/10/2017

Tarea 1. Cálculo de nuevas variables, recodificación y filtrado

Descripción del dataset

El fichero *JaenIndicadores.txt* contiene datos sobre indicadores importantes de los municipios de la provincia de Jaén en el año 2001, e incluye las siguientes variables:

- **Código INE del municipio.**
- **Nombre del municipio.**
- **Consumo de energía eléctrica en megavatios por hora.**
- **Consumo medio de agua en invierno, en metros cúbicos por día.**
- **Consumo medio de agua en verano, en metros cúbicos por día.**
- **Destino de los residuos sólidos urbanos: las posibilidades son vertedero controlado, vertedero incontrolado, compostaje.**
- **Cantidad de residuos sólidos urbanos, en toneladas.**

Ejercicios:

1. Importar el fichero *JaenIndicadores.txt* y denominar a la hoja de datos (data frame) *Datos.Jaen*

```
Datos.Jaen<-read.table("JaenIndicadores.txt",header = T,sep="\t",fileEncoding = "latin1" , na.strings =
head(Datos.Jaen)
```

```
##   CodigoINE           Municipio Consumo.de.energía.eléctrica
## 1     23001 Albanchez de Mágina                         2165
## 2     23002      Alcalá la Real                        93991
## 3     23003           Alcaudete                        34985
## 4     23004        Aldeaquemada                          853
## 5     23005             Andújar                       139971
## 6     23006              Arjona                        12576
##   Consumo.de.agua..Invierno Consumo.de.agua..Verano
## 1                       298                     400
## 2                      4882                    6342
## 3                      1537                    2633
## 4                       123                     500
## 5                      8896                   10326
## 6                      1134                    2542
##   Residuos.sólidos.urbanos..Destino Residuos.sólidos.urbanos..Cantidad
## 1              Vertedero controlado                             370,49
## 2                        Compostaje                            6774,11
## 3                        Compostaje                            3680,95
## 4              Vertedero controlado                             113,53
## 5              Vertedero controlado                            11775,5
## 6              Vertedero controlado                            1222,79
##   Población
## 1      1474
## 2     21523
## 3     11261
```

```
## 4       573
## 5     37903
## 6      5696
```

2. Recodificar la variable Poblacion en una variable cualitativa tipo factor llamada Tamaño con tres categorías:

- **Si la población es inferior a 2000 habitantes, Tamaño será "Pequeño".**
- **Si la población está entre 2000 y 4500 habitantes, Tamaño será "Mediano".**
- **Si la población es superior a 4500 habitantes, Tamaño será "Grande".**

```r
Datos.Jaen$Tamaño<-Datos.Jaen$Población

Datos.Jaen[which(Datos.Jaen$Tamaño < 2000),"Tamaño"]<-0
Datos.Jaen[which( 2000 <= Datos.Jaen$Tamaño
                  & Datos.Jaen$Tamaño <= 4500),"Tamaño"]<-1
Datos.Jaen[which(Datos.Jaen$Tamaño > 4500),"Tamaño"]<-2

Datos.Jaen$Tamaño<-factor(Datos.Jaen$Tamaño , levels = 0:2 , c("Pequeño","Mediano","Grande"))

head(Datos.Jaen$Tamaño)
```

```
## [1] Pequeño Grande  Grande  Pequeño Grande  Grande
## Levels: Pequeño Mediano Grande
```

3. Calcular los siguientes promedios que se especifican a continuación y añadirlos como nuevas variables al fichero Datos.Jaen obtenidas a partir de las variables existentes:

- **Variable elec.hab que contendrá el consumo de energía eléctrica por habitante, obtenida como Consumo.de.energia.electrica/Poblacion**

```r
Datos.Jaen$elec.hab<-Datos.Jaen$Consumo.de.energía.eléctrica / Datos.Jaen$Población

head(Datos.Jaen$elec.hab)
```

```
## [1] 1.468792 4.367003 3.106740 1.488656 3.692874 2.207865
```

- **Variable agua.hab que contendrá el consumo medio de agua por habitante y día, obtenida como (Consumo.de.agua..Invierno + Consumo.de.agua..V erano)/Poblacion**

```r
Datos.Jaen$agua.hab<- ( Datos.Jaen$Consumo.de.agua..Invierno
                        + Datos.Jaen$Consumo.de.agua..Verano ) / Datos.Jaen$Población
head(Datos.Jaen$agua.hab)
```

```
## [1] 0.4735414 0.5214886 0.3703046 1.0872600 0.5071366 0.6453652
```

- **Variable res.hab que contendrá los residuos sólidos urbanos por habitante, obtenida como Residuos.solidos.urbanos..Cantidad/Poblacion**

```r
Datos.Jaen$Residuos.sólidos.urbanos..Cantidad<-gsub(",", ".", Datos.Jaen$Residuos.sólidos.urbanos..Cant

Datos.Jaen$Residuos.sólidos.urbanos..Cantidad<-as.double(Datos.Jaen$Residuos.sólidos.urbanos..Cantidad)

Datos.Jaen$res.hab <- Datos.Jaen$Residuos.sólidos.urbanos..Cantidad / Datos.Jaen$Población

head(Datos.Jaen$res.hab)
```

```
## [1] 0.2513501 0.3147382 0.3268759 0.1981326 0.3106746 0.2146752
```

4. Crear una nueva hoja de datos con todas las variables que contiene actualmente el data frame Datos.Jaen, pero referida sólo a los municipios de tamaño mediano y denominarla

Datos.Jaen.Medianos

```
Datos.Jaen.Medianos<-Datos.Jaen[which(Datos.Jaen$Tamaño == "Mediano"),]

head(Datos.Jaen.Medianos)
```

```
##    CodigoINE              Municipio Consumo.de.energía.eléctrica
## 7      23007              Arjonilla                         8425
## 11     23011      Baños de la Encina                         5990
## 13     23014                 Begíjar                         6443
## 16     23017 Cabra del Santo Cristo                         4548
## 17     23018                  Cambil                         5606
## 18     23019      Campillo de Arenas                         4153
##    Consumo.de.agua..Invierno Consumo.de.agua..Verano
## 7                        799                    1260
## 11                       546                     920
## 13                       634                     755
## 16                       455                     994
## 17                       572                     988
## 18                       424                     720
##    Residuos.sólidos.urbanos..Destino Residuos.sólidos.urbanos..Cantidad
## 7               Vertedero controlado                             881.69
## 11              Vertedero controlado                             540.05
## 13              Vertedero controlado                             609.24
## 16              Vertedero controlado                             514.54
## 17                        Compostaje                             787.29
## 18                        Compostaje                             507.42
##    Población  Tamaño elec.hab  agua.hab   res.hab
## 7       3951 Mediano 2.132372 0.5211339 0.2231562
## 11      2700 Mediano 2.218519 0.5429630 0.2000185
## 13      3161 Mediano 2.038279 0.4394179 0.1927365
## 16      2229 Mediano 2.040377 0.6500673 0.2308389
## 17      3063 Mediano 1.830232 0.5093046 0.2570323
## 18      2119 Mediano 1.959887 0.5398773 0.2394620
```

5. Guardar la hoja de datos Datos.Jaen con las nuevas variables creadas en los apartados anteriores y la hoja que contiene los datos de las poblaciones medianas (Datos.Jaen.Medianos) en un archivo de datos de R y llamadlo JaenIndicadores.RData

```
save(Datos.Jaen , Datos.Jaen.Medianos , file = "JaenIndicadores.RData")
```

Tarea 2. Análisis Estadístico Descriptivo de Datos

Descripción del dataset

El fichero *Andalucia.txt* contiene datos sobre diversos indicadores de los municipios andaluces obtenidos del Instituto de Estadística y Cartografía de Andalucía. Los valores de estos indicadores están contenidos en las siguientes variables:

- **Código INE.**
- **Municipio.**
- **Tasa de actividad en 2001.**
- **Nº de líneas ADSL en funcionamiento en 2007.**
- **Edad media del municipio en 2007.**
- **Renta familiar disponible por habitante. Aparece agrupado en varias categorías de renta. Hay numerosos datos faltantes que aparecen señalados como "..".**

- **Crecimiento vegetativo en 2006. El crecimiento vegetativo es la diferencia entre el número de nacidos y el número de fallecidos.**
- **Número de parados en 2007.**

Ejercicios:

1. Importar el fichero *Andalucia.txt* y denominar a la hoja de datos (data frame) *Datos.Andalucia*. Comprobar si en el archivo *.txt* hay datos faltantes y cómo están codificados.

```
Datos.Andalucia<-read.table("Andalucia.txt",header = T,sep="\t",fileEncoding = "latin1" , na.strings = c
```

2. A partir de la variable código INE, construir una variable tipo factor que distinga la provincia de pertenencia de cada municipio, denominarla "Provincia" y añadirla al data frame.

Provincia	Almería	Cádiz	Córdoba	Granada	Huelva	Jaén	Málaga	Sevilla
Id cod INE	4	11	14	18	21	23	29	41

```
Datos.Andalucia$Provincia<-as.integer(Datos.Andalucia$Codigo.INE/1000)
Datos.Andalucia$Provincia<-factor(Datos.Andalucia$Provincia , levels = c(4,11,14,18,21,23,29,41) , label
```

Distribución de frecuencias absolutas:

```
Tabla<-table(Datos.Andalucia$Provincia)
Tabla
```

```
##
## Almería   Cádiz Córdoba Granada  Huelva    Jaén  Málaga Sevilla
##     102      44      75     168      79      97     100     105
```

Relativas:

```
prop.Tabla<-prop.table(Tabla)
prop.Tabla
```

```
##
##    Almería      Cádiz    Córdoba    Granada     Huelva       Jaén
## 0.13246753 0.05714286 0.09740260 0.21818182 0.10259740 0.12597403
##     Málaga    Sevilla
## 0.12987013 0.13636364
```

Diagrama de barras con las frecuencias absolutas:

```
barplot(Tabla,col="lightblue",xlab="Provincia" , ylab="Número de municipios")
```

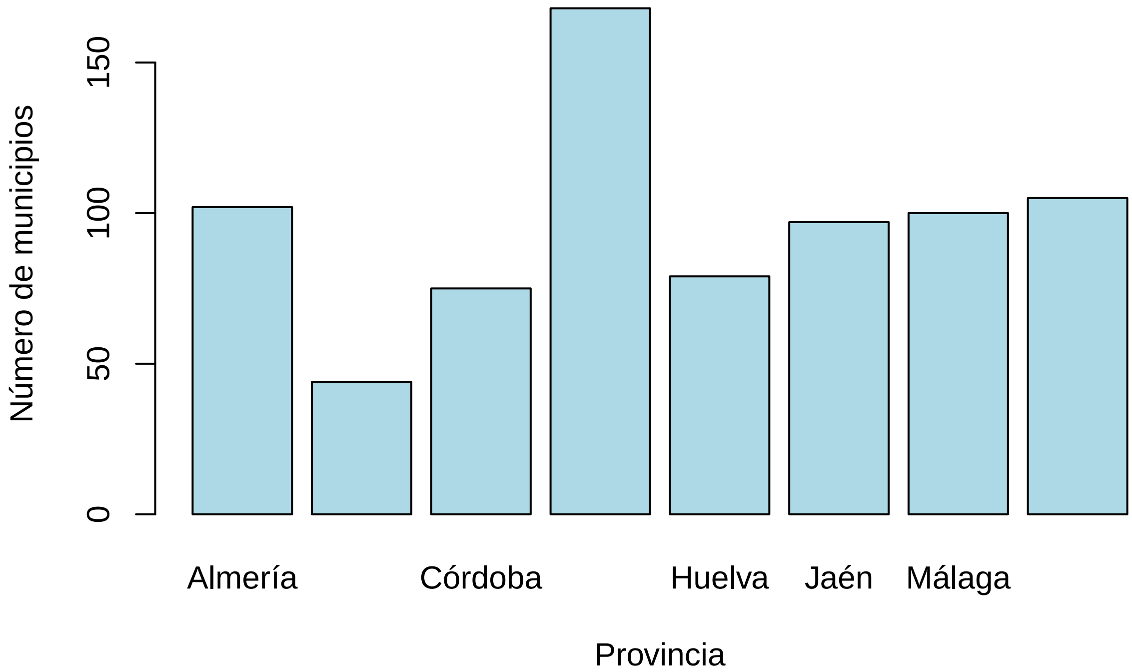
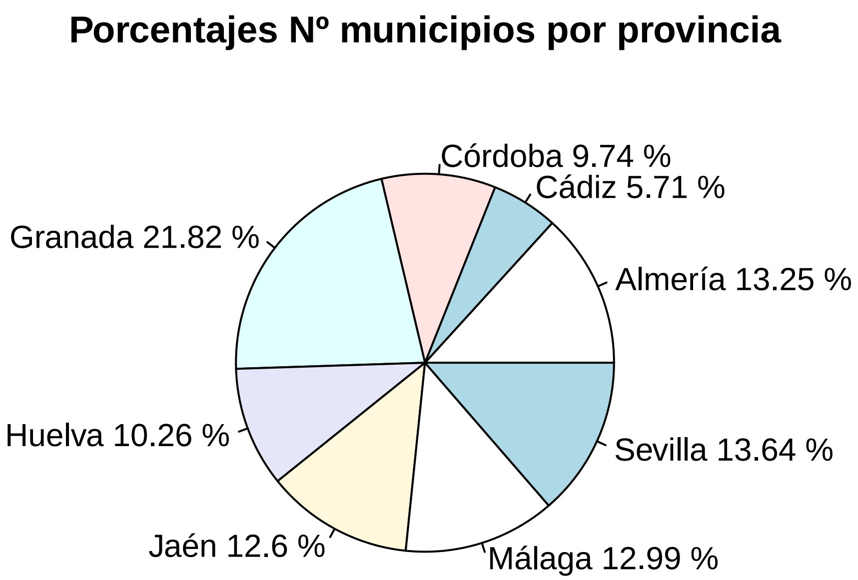


Diagrama de sectores con las frecuencias relativas en porcentajes de esta variable tipo factor:

```
tabla.porcentaje<-round(100*prop.Tabla,2)

sectores<-pie(tabla.porcentaje,labels=paste(names(tabla.porcentaje),tabla.porcentaje,"%"),main="Porcenta
```



¿Qué provincia tiene más municipios?

Granada.

¿Cuál tiene menos?

Cádiz.

¿Qué porcentaje representa en cada caso?

21.82 % y 5.71 %.

3. Obtener un resumen descriptivo de la variable tasa de actividad de 2001 que incluya parámetros de posición, dispersión, asimetría y curtosis, histograma y diagrama de caja. En función de este resumen, contestar a las siguientes preguntas:

```
Datos.Andalucia$Tasa.actividad.2001<-gsub(",", ".", Datos.Andalucia$Tasa.actividad.2001)

Datos.Andalucia$Tasa.actividad.2001<-as.numeric(Datos.Andalucia$Tasa.actividad.2001)
```

Parámetros de posición:

```
min.Tasa<-min(Datos.Andalucia$Tasa.actividad.2001, na.rm=T)

max.Tasa<-max(Datos.Andalucia$Tasa.actividad.2001, na.rm=T)

q1.Tasa<-quantile(Datos.Andalucia$Tasa.actividad.2001, probs=0.25,
na.rm=T)

q3.Tasa<-quantile(Datos.Andalucia$Tasa.actividad.2001, probs=0.75,
na.rm=T)

mean.Tasa<-mean(Datos.Andalucia$Tasa.actividad.2001,na.rm = T)

median.Tasa<-median(Datos.Andalucia$Tasa.actividad.2001,na.rm = T)

table.Tasa<-table(Datos.Andalucia$Tasa.actividad.2001)
moda.Tasa<-names(table.Tasa[which(table.Tasa==max(table.Tasa))[1]])

par.posicion<-round(as.numeric(c(min.Tasa,max.Tasa,q1.Tasa,q3.Tasa,mean.Tasa,median.Tasa,moda.Tasa)),2)
names(par.posicion)<-c("Mínimo","Máximo","Primer cuartil","Tercer cuartil","Media","Mediana","Moda")
par.posicion
```

```
##         Mínimo         Máximo Primer cuartil Tercer cuartil          Media
##          26.92          74.21          46.73          56.22          51.44
##        Mediana           Moda
##          51.88          55.09
```

Parámetros de asimetría:

```
sd.Tasa<-sd(Datos.Andalucia$Tasa.actividad.2001,na.rm=TRUE)

var.Tasa<-var(Datos.Andalucia$Tasa.actividad.2001,na.rm=TRUE)

cv.Tasa<-sd.Tasa/mean.Tasa

ri.Tasa<-q3.Tasa-q1.Tasa

rango.Tasa<-max.Tasa-min.Tasa

par.asimetria<-round(c(sd.Tasa,var.Tasa,cv.Tasa,ri.Tasa,rango.Tasa),2)
names(par.asimetria)<-c("Desviación típica","Varianza","Coeficiente de variación","Recorrido intercuart
par.asimetria
```

```
##         Desviación típica                  Varianza
##                      6.99                     48.81
##  Coeficiente de variación Recorrido intercuartílico
##                      0.14                      9.49
##                     Rango
```

```
##                        47.29
```

Parámetro de simetría

```r
#install.packages("e1071")
library(e1071)

akew.Tasa<-skewness(Datos.Andalucia$Tasa.actividad.2001 , na.rm=TRUE)
names(akew.Tasa)<-("Coeficiente de asimetría")

akew.Tasa
```

```
## Coeficiente de asimetría
##              -0.09917887
```

Parámetro de curtosis

```r
kurt.Tasa<-kurtosis(Datos.Andalucia$Tasa.actividad.2001 , na.rm=TRUE)
names(kurt.Tasa)<-("Coeficiente de kurtosis")

kurt.Tasa
```

```
## Coeficiente de kurtosis
##               0.1241295
```

Histograma

```r
hist(Datos.Andalucia$Tasa.actividad.2001, breaks = 16, freq = TRUE, main =
"Histograma de la tasa de actividad en el año 2001",xlab="Actividad",ylab="Frecuencias", col="lightblue
border="blue")
```

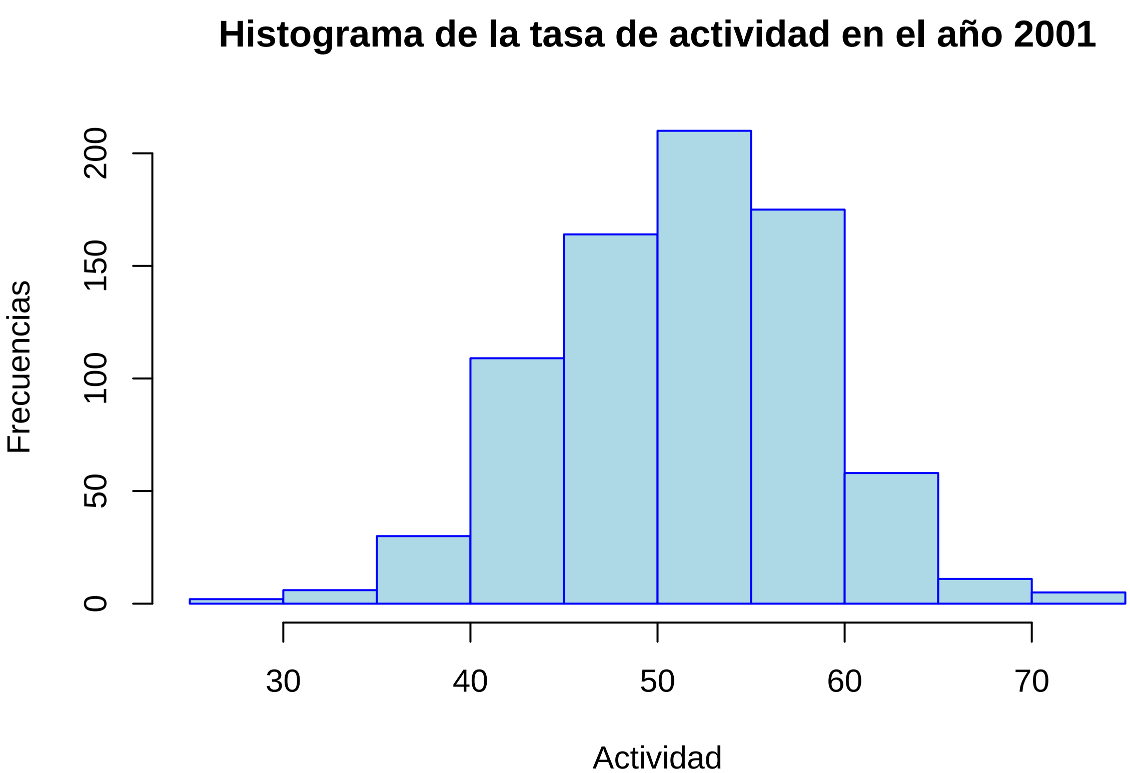
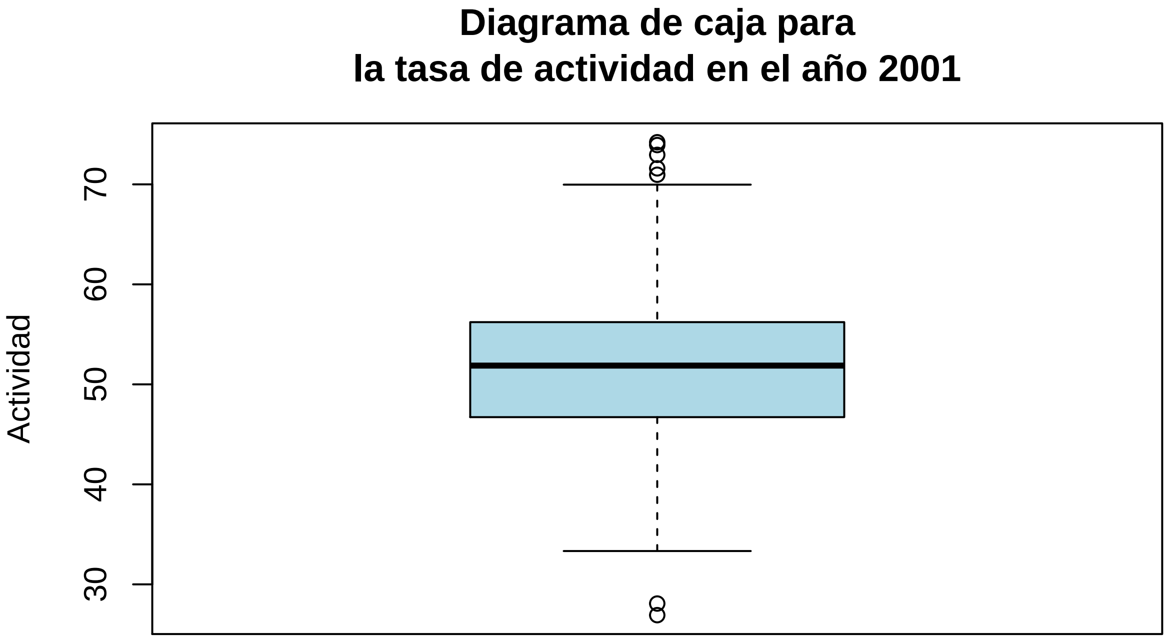


Diagrama de caja

```r
boxplot(Datos.Andalucia$Tasa.actividad.2001,main="Diagrama de caja para
la tasa de actividad en el año 2001",ylab="Actividad", col="lightblue")
```



3.1. ¿Cuál es la tasa media de actividad de los municipios andaluces?

51.44

¿Crees que este valor es adecuado para representar la Tasa de Actividad de los municipios andaluces durante 2001?

Sí, ya que la *desviación típica* no es muy pronunciada, al ser 6.99. También lo podemos apreciar en el *recorrido intercuartílico* que es 9.49,es decir el 50 % de los valores centrados en la *mediana* tienen un rango de variación de 9.49 esto unido al hecho de que el *mínimo* es 26.92 y el *máximo* es 74.21, o lo que es lo mismo que el *rango* sea de 47.29. Lo cual nos refleja que no hay una gran dispersión de los datos, situación en la cual es idoneo usar la *media* para representar a la actividad de los municipios andaluces. Como comprobación si miramos el valor de la *mediana* 51.88 y la *moda* 55.09 vemos que no existe una deferencia notable, ya que la única diferencia sería en la *moda* y al ser un rango continuó sobre el que se ha recogido la actividad de los municipios, no nos aporta mucha fiabilidad o seguridad el usar la moda como parámetro representativo.

3.2. ¿Cómo valoras la homogeneidad de los valores de la tasa de actividad en los municipios andaluces?

Los valores son bastante homogéneos como se ha comentado en el apartado anterior para justificar la idoneidad de usar el parámetro de la *media* como representativo de la muestra.

¿Qué parámetro elegirías para representar la dispersión de la Tasa de Actividad de 2001?

Elegiría la desviación típica, ya que tiene en cuenta la cercanía o lejanía de cada una de los valores recogidos a la *media*, la cual ha sido justificada como un parámetro que representa muy bien a los valores recogidos.

3.3. ¿En ese sentido, qué municipios andaluces destacan significativamente del resto (como atípicos) por su alta tasa de actividad y por su baja tasa de actividad?

```r
Datos.Andalucia$Municipio[which.min(Datos.Andalucia$Tasa.actividad.2001)]
```

```
## [1] Benitagla
## 770 Levels: Abla Abrucena Adamuz Adra Agrón ... Zurgena
```

Por su baja tasa de actividad destaca *Benitagla*, lo cual tiene sentido porque es un municipio de *Almería* con tan solo 69 habitantes.

```r
Datos.Andalucia$Municipio[which.max(Datos.Andalucia$Tasa.actividad.2001)]
```

```
## [1] Mojonera (La)
## 770 Levels: Abla Abrucena Adamuz Adra Agrón ... Zurgena
```

Por su alta tasa de actividad destaca *Mojonera*, tiene sentido porque es un municipio de *Almería* con 8740 habitantes.

¿Se te ocurre alguna explicación al respecto?

Sí, parece debido a la extensión y número de habitantes.

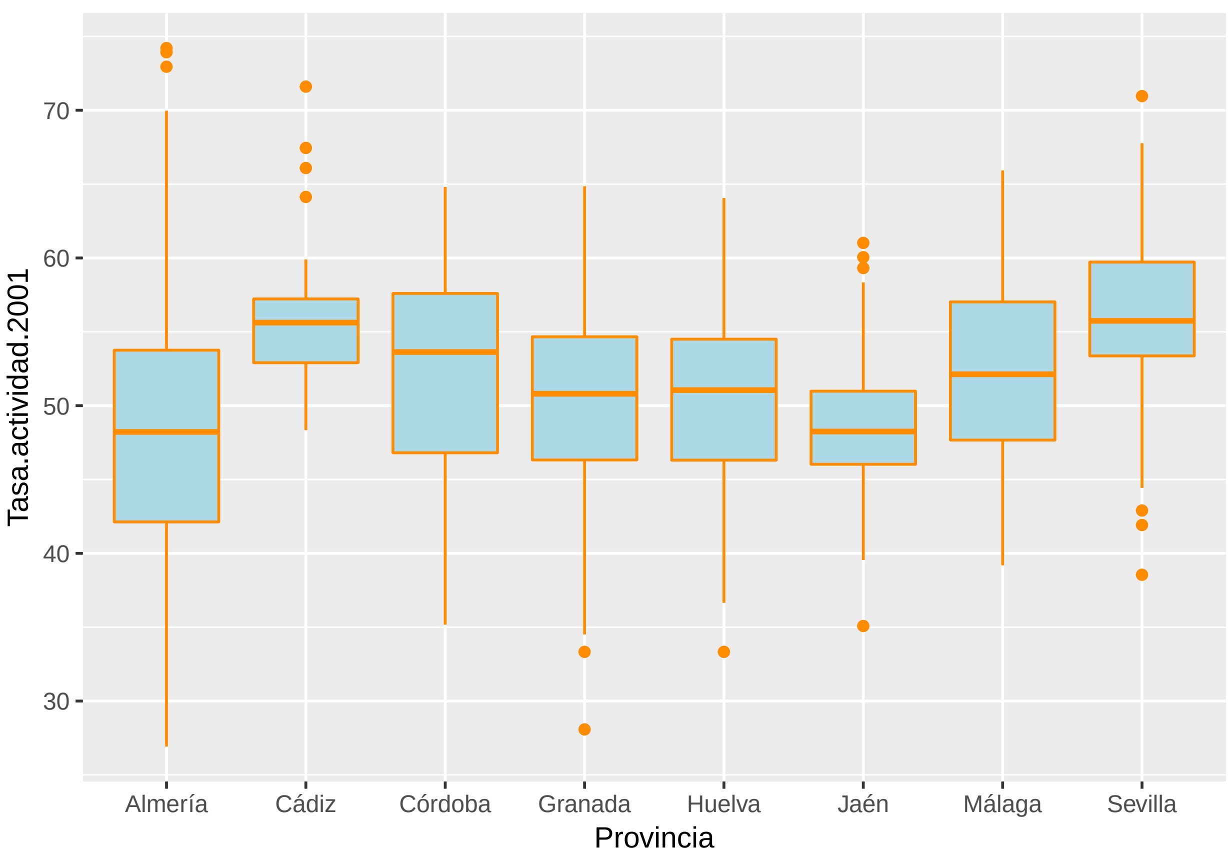
3.4. ¿Cómo valoras la simetría de la distribución de frecuencias?

El *coeficiente de asimetría* es de -0.09917887 lo cual indica que la distribución de frecuencias es muy simétrica. Esto además se aprecia a simple vista observando el histograma representado.

4. Obtener un gráfico de caja de la Tasa de actividad en 2001 en función de la provincia y describe brevemente la información que contienen los datos a partir del gráfico.

```r
#install.packages("ggplot2")
library(ggplot2)
```

```r
ggplot(Datos.Andalucia, aes(Provincia,Tasa.actividad.2001))+geom_boxplot(colour="darkorange",fill="ligh
```



Se ve que en Almería hay una gran dispersión de los datos, mucha variabilidad y que la *media* de la tasa de actividad está ligeramente por debajo de la *media* global de la tasa de actividad de los municipios. Además solo presenta dos casos atípicos.

Cádiz es la provincia con menos variabilidad en la tasa de actividad, además la *media* está pronunciadamente por encima de la *media* global por municipios de la comunidad. Aunque presenta cuatro datos atípicos.

Córdoba, Granada, Huelva y Málaga son muy parecidos en cuanto a tasa de variabilidad y *media* de tasa de actividad. Su *media* está muy proxima a la *media* global por municipios, y no presentan apenas ningún dato atípico.

Jaén presenta poca variabilidad y su *media* está ligeramente por debajo de la *media* global por municipios. Presentando cuatro datos atípicos.

Sevilla presenta también poca variabilidad pero en este caso la *media* de tasa de actividad está por encima de la *media* global por municipios, siendo esta junto a la provincia de Cádiz las más altas de todas las provincias.

5. Guardar la hoja de datos Datos.Andalucia con la nueva variable creada en los apartados anteriores junto con los parámetros obtenidos en un archivo de datos de R y llámalo Andalucia.RData.

```
save(Datos.Andalucia, Tabla, tabla.porcentaje, par.posicion, par.asimetria, akew.Tasa, kurt.Tasa, file =
```

3. Distribuciones de probabilidad

1. Consideremos una variable aleatoria que sigue una distribución B (15, 0.33). Se pide:

1.1. ¿Qué valor de la variable deja por debajo de sí el 75% de la probabilidad?

```
probs.acum<-cumsum(dbinom(0:15, 15, 0.33))
print("F(x) con x perteneciente a [0,15]")
```

```
## [1] "F(x) con x perteneciente a [0,15]"
```

```
names(probs.acum)<-0:15
probs.acum
```

```
##           0           1           2           3           4           5
## 0.002461059 0.020643511 0.083332261 0.217130639 0.414832720 0.629059154
##           6           7           8           9          10          11
## 0.804916674 0.916280605 0.971131497 0.992144027 0.998353700 0.999743926
##          12          13          14          15
## 0.999972172 0.999998115 0.999999940 1.000000000
```

El valor de la variable 6 deja po debajo el 80 % de la probabilidad, por tanto deja también el 75 %, y como el valor anterior 5 deja por debajo el 63 % de la probabilidad, no hay otro caso posible.

Si usamos directamente la función de R:

```
qbinom(0.75,15,0.33)
```

```
## [1] 6
```

1.2. Calcular el percentil 95% de la distribución.

Fijándonos en los resultados de la función de distribución para cada valor del rango de la variable, vemos que el valor de 8 es el que deja por debajo el 95 % de a probabilidad.

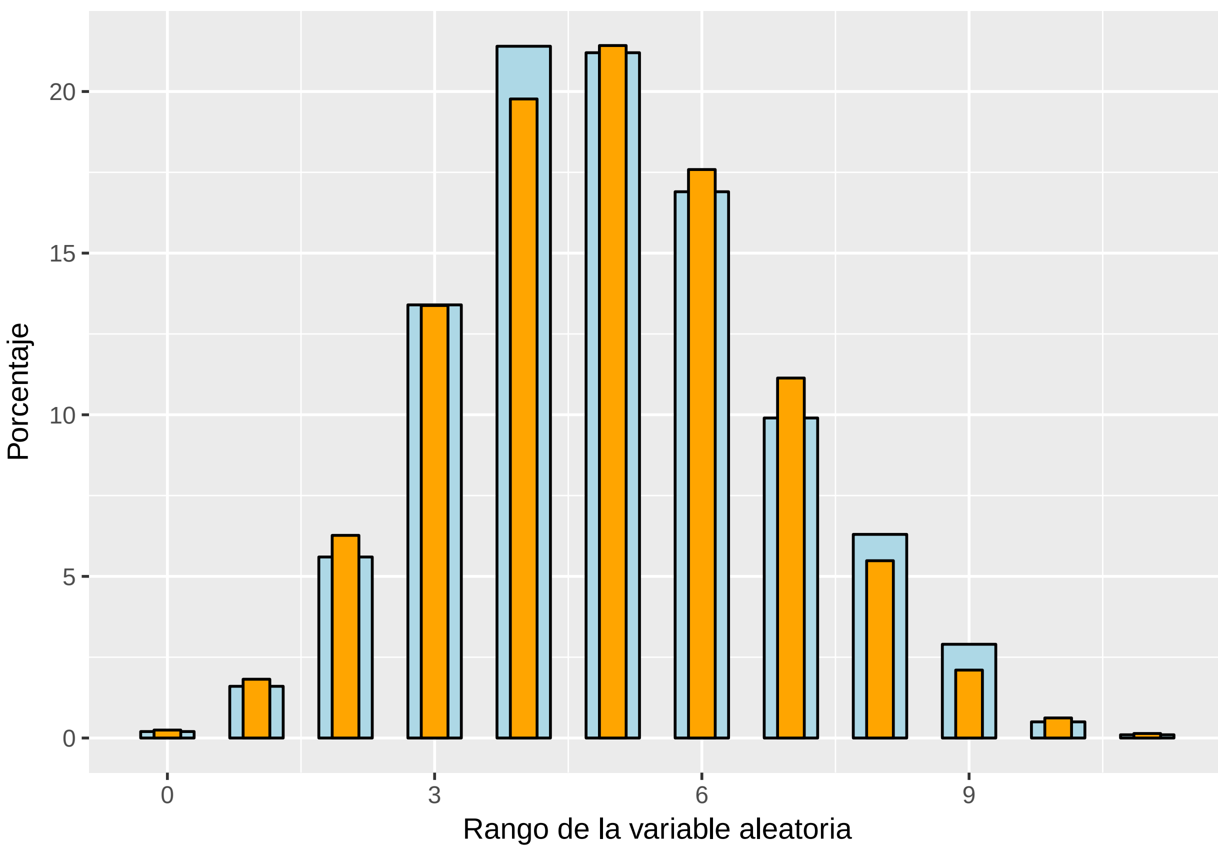
Si usamos directamente la función de R:

```
qbinom(0.95,15,0.33)
```

```
## [1] 8
```

1.3. Obtener una muestra de tamaño 1000 de esta distribución, representarla gráficamente las frecuencias observadas de cada valor de la distribución mediante un diagrama de barras y comparar éste con las frecuencias esperadas según el modelo que genera los datos.

```
muestra<-rbinom(1000,15,0.33)

tabla.muestra<-table(muestra)
tabla.muestra.prop<-prop.table(tabla.muestra)

datos.plot<-as.data.frame(tabla.muestra.prop*100)
names(datos.plot)<-c("Rango de la variable aleatoria","Porcentaje")
datos.plot$`Rango de la variable aleatoria`<-as.double(names(tabla.muestra))

x<-datos.plot$`Rango de la variable aleatoria`
y<-dbinom(x, 15, 0.33)
datos.plot1<-as.data.frame(cbind(x,y*100))
names(datos.plot1)<-c("Rango de la variable aleatoria","Porcentaje")

plot1<-geom_bar(data= datos.plot, aes(x=`Rango de la variable aleatoria`,y=Porcentaje), stat = "identity
plot2<-geom_bar(data = datos.plot1, aes(x=`Rango de la variable aleatoria`,y=Porcentaje), stat = "identi

ggplot()+plot1+plot2
```



En la figura vemos en color azul la frecuencia relativa en % de la muestra de 1000 elementos de la distribución binomial, y en naranja el valor teórico esperado. Vemos que con 1000 elementos de la muestra se obtiene una buena aproximación de la distribución real.

2. Consideremos una variable aleatoria W con distribución N (250, 13). Se pide:

2.1. P $[240 < W \leq 245.5]$

Como la variable es continua es lo mismo que calcular $P[240 \leq W \leq 245.5]$:

```
pnorm(245.5, mean=250, sd=13, lower.tail=T) - pnorm(240, mean=250, sd=13, lower.tail=T)
```

```
## [1] 0.1437354
```

2.2. P $[W \geq 256]$.

```
pnorm(256, mean = 250, sd = 13, lower.tail = F)
```

```
## [1] 0.3222062
```

2.3. Si queremos desechar el 5% de valores más altos de la distribución y el 5% de valores más bajos, ¿con qué intervalo de valores nos quedaremos?

```
x1<-qnorm(0.05, mean=250, sd=13, lower.tail=T)
x2<-qnorm(0.05, mean=250, sd=13, lower.tail=F)
x1
```

```
## [1] 228.6169
```

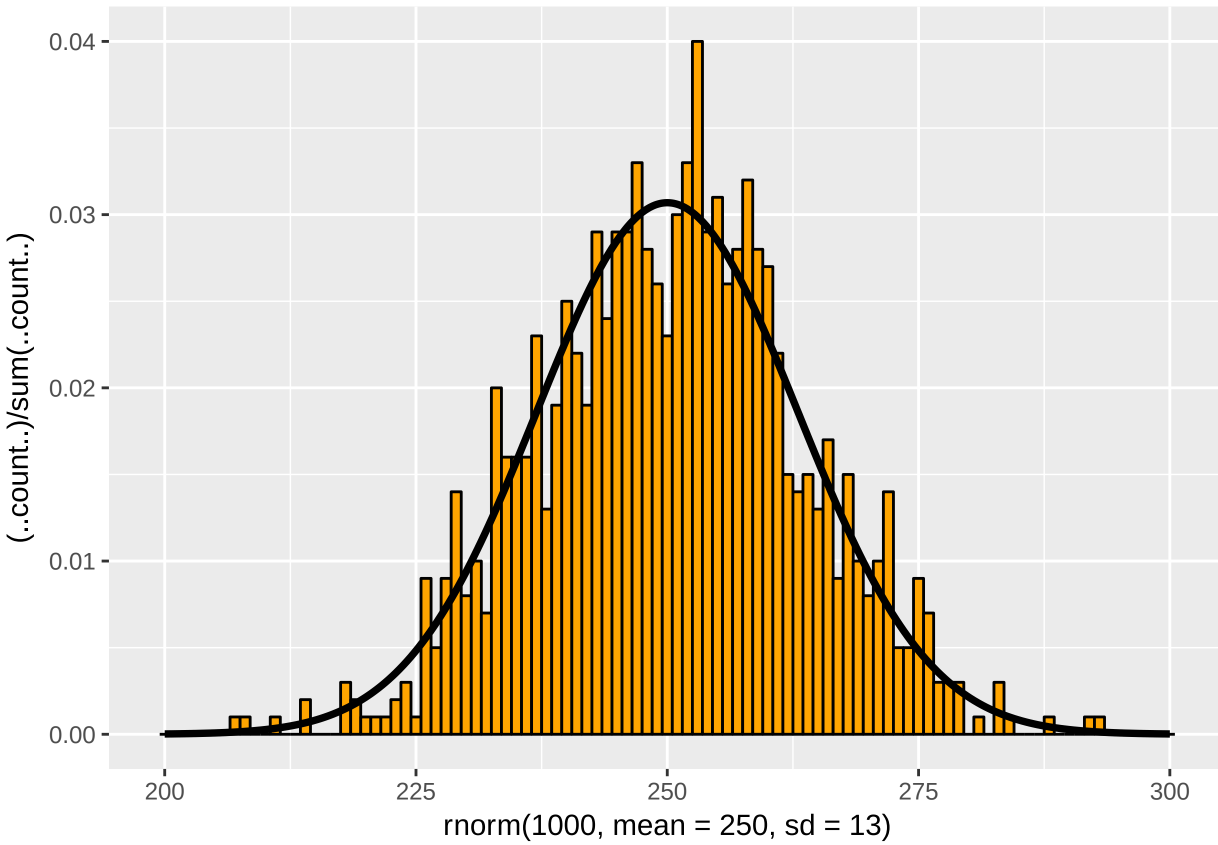
```
x2
```

```
## [1] 271.3831
```

El intervalo buscado sería [228.6169,271.3831].

2.4. Obtener una muestra de tamaño 1000 de la distribución, representar la función de densidad de esta distribución y compararla con el histograma de la muestra obtenida.

```
muestra<-as.data.frame(rnorm(1000,mean=250,sd=13))

x<-seq(200,300,length.out = 1000)
y<-dnorm(x,mean=250,sd=13)
datos.plot1<-as.data.frame(cbind(x,y))

plot1<-geom_histogram(data=muestra,aes(x=`rnorm(1000, mean = 250, sd = 13)`,(..count..)/sum(..count..))
plot2<-geom_line(data = datos.plot1, aes(x=x,y=y),colour="black",size=1.3)

ggplot()+plot1+plot2
```



4. Contrastes de Hipótesis e Intervalos de Confianza

Descripción del dataset

Mediante una red de sensores se han recogido datos sobre la temperatura media diaria (ºC) en dos estaciones A y B durante 52 días. Los valores recogidos de la temperatura se encuentran en la hoja de datos "Temper" incluida en el fichero Temperatura.RData.

Ejercicios

1. Cargar el fichero Temperatura.RData.

```
load("Temperatura.RData")
```

2. Crear dos nuevas variables, temp.A y temp.B, que contengan las temperaturas de las estaciones A y B, respectivamente.

```
temp.A<-Temper[which(Temper$Estacion=="A"),]

temp.B<-Temper[which(Temper$Estacion=="B"),]
```

3. Da un intervalo de confianza para la temperatura media diaria de la estación A, al 95%, y a partir de éste indica si se puede admitir, y por qué, que la temperatura media diaria en dicha estación sea de 19ºC, con ese mismo nivel de confianza.

```
t.test(temp.A$Temper, alternative = "two.sided", conf.level = 0.95)
```

```
##
```

```
##  One Sample t-test
##
## data:  temp.A$Temper
## t = 93.167, df = 76, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  19.39401 20.24131
## sample estimates:
## mean of x
##  19.81766
```

Intervalo de confianza al 95% es: [19.39401 , 20.24131] Dado que 19 ºC no está en el intervalo de confianza, se rechaza la hipótesis de que la media sea 19 ºC.

4. Plantea un test de hipótesis que refleje la pregunta del apartado anterior y resuélvelo sin usar el intervalo de confianza (riesgo de 1ª especie 5%).

Se plantea el siguiente test de hipótesis:

Contaste bilateral

H0: media = 19 ºC

H1: media $\neq$ 19 ºC

```
t.test(temp.A$Temper, alternative = "two.sided", mu=19, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  temp.A$Temper
## t = 3.844, df = 76, p-value = 0.0002496
## alternative hypothesis: true mean is not equal to 19
## 95 percent confidence interval:
##  19.39401 20.24131
## sample estimates:
## mean of x
##  19.81766
```

El p-valor obtenido es 0.0002496 y dado que hemos elegido el parámetro de que la probabilidad de cometer el error de tipo 1 sea inferior a 0.05 entonces comparando vemos que el p-valor $0.0002496 < 0.025$ que es la mitad de 0.05 ya que la t de student sigue una distribución simétrica y el p-valor nos muestra la probabilidad teniendo en cuenta solo la cola superior. Por tanto al ser el p-valor menor que el humbral fijado rechazamos la hipótesis nula y aceptamos la alternativa de que la media sea distinta de 19 ºC.

5. Determina si puede admitirse, con un riesgo de primera especie de 1%, que la temperatura media diaria es la misma en las dos estaciones. Plantea previamente el correspondiente contraste de hipótesis.

Para poder aplicar el test de comparación de medias antes debemos aplicar un test para ver si se puede suponer que tienen la misma varianza:

Test de hipótesis:

H0: $\sigma_A = \sigma_B \Leftrightarrow \sigma_A/\sigma_B = 1$

H1: $\sigma_A \neq \sigma_B \Leftrightarrow \sigma_A/\sigma_B \neq 1$

```
var.test(temp.A$Temper, temp.B$Temper , ratio=1, alternative="two.sided", conf.level=0.99)
```

```
##
```

```
##  F test to compare two variances
##
## data:  temp.A$Temper and temp.B$Temper
## F = 1.0978, num df = 76, denom df = 78, p-value = 0.6825
## alternative hypothesis: true ratio of variances is not equal to 1
## 99 percent confidence interval:
##  0.6071355 1.9889245
## sample estimates:
## ratio of variances
##            1.0978
```

P-valor = 0.6825 > 0.025 por tanto aceptamos la hipótesis nula de que las varianzas sean iguales. Ahora planteamos el siguiente contraste:

Test de hipótesis:

H0: $\mu_A = \mu_B \Leftrightarrow \mu_A$ - $\mu_B = 0$

H1: $\mu_A \neq \mu_B \Leftrightarrow \mu_A$ - $\mu_B \neq 0$

```
t.test(temp.A$Temper, temp.B$Temper, alternative="two.sided", mu=0,var.equal=T, conflevel=0.99)
```

```
##
##  Two Sample t-test
##
## data:  temp.A$Temper and temp.B$Temper
## t = -0.64116, df = 154, p-value = 0.5224
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.7642880  0.3897392
## sample estimates:
## mean of x mean of y
##  19.81766  20.00494
```

Como el p-valor = 0.5224 > 0.005 (la mitad de 0.01 que el valor de error de tipo 1 prefijado) aceptamos la hipóteis nula de que las medias sean iguales.

6. Obtén un intervalo de confianza (99%) para la diferencia de temperaturas entre estaciones. ¿Aporta alguna información adicional al resultado obtenido en el apartado anterior?

Usando el test realizado en el apartado anterior tenemos que el intervalo de confianza es: [-0.7642880 , 0.3897392]. Vemos que el 0 está en el intervalo, por tanto coincide con el resultado anterior, por otro lado aporta información sobre lo ajustado de ese valor, en este caso el rango en el que oscila es bastante amplio por lo que también sería factible afirmar que las medias son ligeramente distintas, teniendo una mayor media el grupo *temp.B* ya que el intervalo tiene valores mayores en valor absoluto por el lado izquierdo del intervalo.

7. Se sabe que a lo largo de los 52 días, la estación A falló 5 días y la B 7 días. ¿Puede afirmarse con un nivel de confianza del 90% que la proporción de días fallados es la misma en las dos estaciones?

Aplicamos el siguiente test de hipótesis:

H0: Proporción_A = Proporción_B ⇔ Proporción_A - Proporción_B = 0

H1: Proporción_A ≠ Proporción_B ⇔ Proporción_A - Proporción_B ≠ 0

```
prop.test(c(5,7), c(52,52),alternative="two.sided", conf.level=0.9, correct=T)
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
```

```
##
## data:  c(5, 7) out of c(52, 52)
## X-squared = 0.094203, df = 1, p-value = 0.7589
## alternative hypothesis: two.sided
## 90 percent confidence interval:
##  -0.16056582  0.08364275
## sample estimates:
##     prop 1     prop 2
## 0.09615385 0.13461538
```

Se ve por un lado que el p-valor obtenido es 0.7589, y por otro lado se ve que el 0 entá en el itervalo de confianza al 90 % [-0.16056582 , 0.08364275], por tanto aceptado la hipótesis nula de que las dos proporciones sean iguales.