El fichero Harwell.dat contiene datos sobre contaminantes atmosféricos y otras variables meterológicas como la velocidad del viento, la dirección y la temperatura. En la práctica solo se utiliza el mes y el ozono.

Al importar con read.table, lo hace directamente en formato data.frame. Es necesario incluir " header = TRUE " para que la primera fila la lea como el nombre de cada columna.

In [1]:
```
data = read.table(file = "Harwell.dat",   # Archivo de datos TXT/.dat indicado com
           header = TRUE,       # Si se muestra el encabezado (TRUE) o no (FALSE)
           sep = " ",           # Separador de las columnas del archivo
           dec = ".")           # Separador decimal

head(data)#; tail(data) #Vistazo rápido a las 5 primeras/últimas columnas (ya apare
length(data$year) #Se compueba que la longitud es 366 (año bisiesto) sin mirar dat
class(data) #Directamente, se importa como data frame
#str(data) #Otra forma de obtener el tipo de objeto con más información
```

year	month	day	hour	no2	no	o3	pm2.5	wspeed	windir	
2012	1	1	15	0.9052472	0.0009380291	32.00099	8.0007038945	7.201140	230.4011	1
2012	1	2	15	0.3202708	0.0010601799	35.00088	2.0010315493	8.100879	258.0010	
2012	1	3	15	0.5862021	0.0010596504	36.00099	3.0010828993	15.300838	281.7009	
2012	1	4	15	1.2243510	0.0008866445	33.00107	0.0009418681	13.500844	267.3011	
2012	1	5	15	0.9052161	0.0009445587	33.00102	NA	15.300812	322.2011	
2012	1	6	15	6.3839255	0.0007616141	22.00090	8.0010079721	5.800830	264.4010	

366

'data.frame'

1) Análisis exploratorio inicial.

1a) Concentraciones mensuales medias de ozono

In [2]:
```
tapply(data$o3, data$month, mean, na.rm = TRUE) #Tabla con la media: los meses com
```

1	26.572456888328
2	27.7596248041674
3	34.2677012284012
4	41.2152988620905
5	44.1300176093667
6	35.9676606958249
7	37.1300221286703
8	37.393861526205
9	32.5009809738516
10	24.9365051517852
11	25.8343526952344
12	26.9042473610645

1b) Diagrama "box and whisker para comparar las distribuciones de ozono en función del mes.

Para ello, hay que categorizar los datos de o3 para cada mes. La forma más sencilla es mediante split.

Al no tener cada mes el mismo número de días, se utiliza el formato lista.

Se podría hacer lo mismo con **data.frame** donde cada columna es un mes, pero las longitudes han de ser iguales, por ello añadiendo **rep("NA",31-length(data$o3[i])** para los meses con menos días y posteriormente omitir esos datos al representar, es otra opción.

In [3]:
```
ldatos = split(data$o3, data$month) #Crea una lista de los valores del ozono categ

mes = format(ISOdatetime(2012,1:12,1,0,0,0),"%b") # Con %B% meses sin abreviar
names(ldatos) = mes #Asigna el nombre del mes a cada vector de la lista

#Para hacer la media con listas (mismo resultado que antes)
#sapply(ldatos, function(x) mean(x, na.rm = TRUE));  #mismo resultado que antes
#media1 = sapply(ldatos, mean, na.rm=TRUE); media1 #Otra forma para hacer el 1a
```

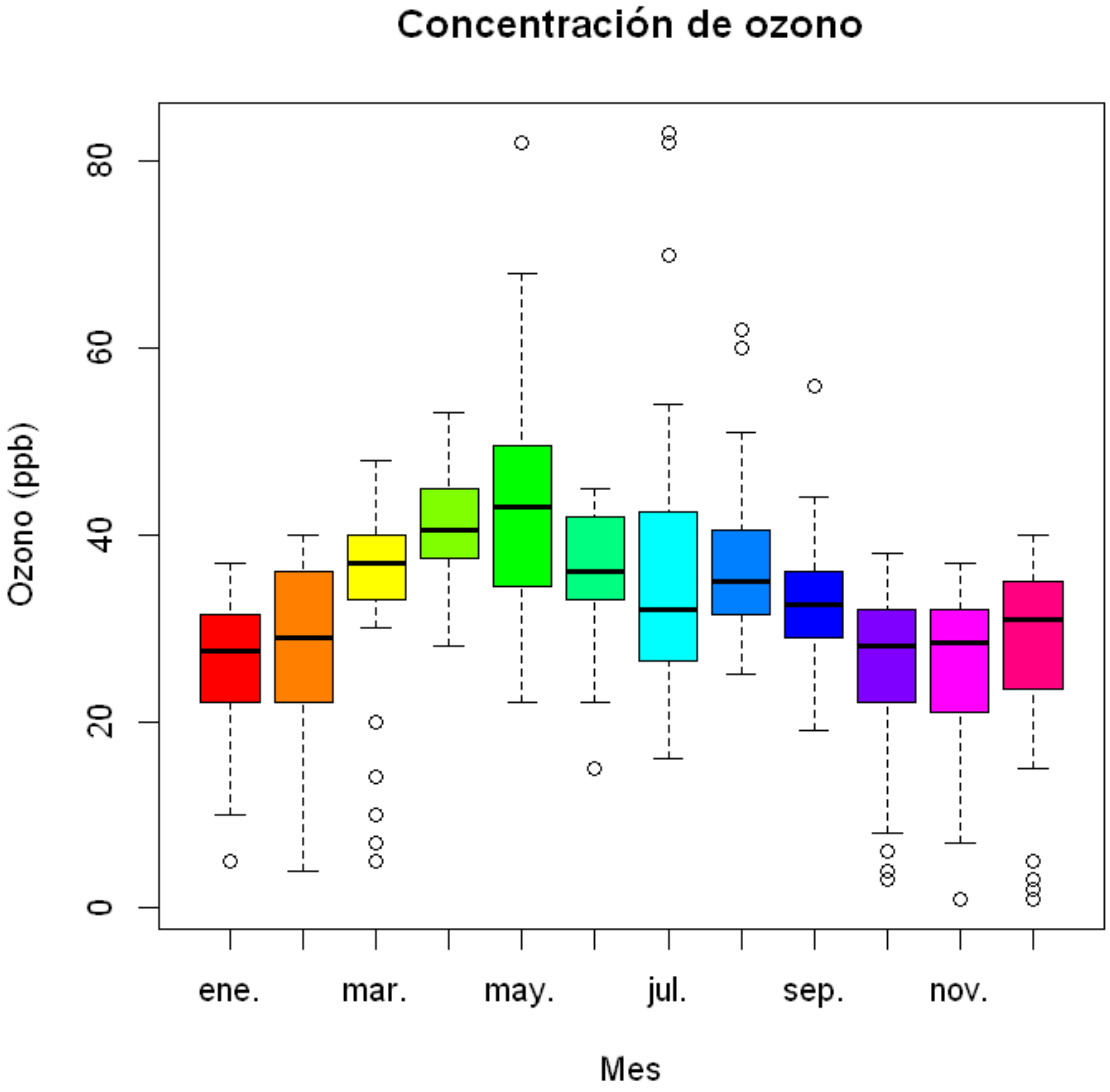
In [8]:
```
#par(mfrow=c(1,2)) #Parámetros gráficos. Para representar varios gráficos en una m
options(repr.plot.height = 6, repr.plot.width = 6) #Se ajusta el tamaño

b = boxplot(ldatos, horizontal = FALSE, main = "Concentración de ozono",
        ylab = "Ozono (ppb)", na.rm = TRUE, notch = FALSE,  col = rainbow(12), xlab

names(b)
b$names
```

1. 'stats'
2. 'n'
3. 'conf'
4. 'out'
5. 'group'
6. 'names'

1. 'ene.'
2. 'feb.'
3. 'mar.'
4. 'abr.'
5. 'may.'
6. 'jun.'
7. 'jul.'
8. 'ago.'
9. 'sep.'
10. 'oct.'
11. 'nov.'
12. 'dic.'



1C) Describir muy brevemente cómo es el ciclo anual de las concentraciones de ozono a partir de los anteriores resultados.

Los meses desde marzo a septiembre presentan concentraciones más elevadas y la mediana varía bastante entre ellos. De octubre a febrero la mediana prácticamente es la misma para todos (cambia un poco en diciembre). Marzo (inicio primavera) y septiembre (fin de verano) presentan una menor dispersión (rango intercuartílico pequeño). En general, no hay simetría en las distribuciones, con alguna excepción como febrero y septiembre.

Al haber "outliers" una medida robusta para comparar es la mediana (máxima en el mes de mayo), ya que la media se ve más afectada por los valores extremos. Además, se da el caso, que en los meses fríos los outliers están por debajo y en los cálidos por encima.

2. Concentraciones de ozono para abril, mayo y junio.

Selecciona las concentraciones de ozono para todos los días de los meses de abril, mayo y junio. Guarda los datos en una nueva variable llamada o3.spring

In [209…

```
# ----------------- Varias formas de utilizar lista ------- util para el examen ---

#o3.spring = c(ldatos[4:6]); o3.spring; #Guarda en una lista de longitud 3
#o3.spring = ldatos[c("Abril","Mayo","Junio")]; o3.spring #Cuidado si están los mes
#o3.spring = df[c("Abril","Mayo","Junio")]; o3.spring

o3.spring = c(ldatos[[4]], ldatos[[5]], ldatos[[6]]); #o3.spring; #length(o3.spring
```

¿Es el valor medio de o3.spring significativamente diferente de 38.5 ppb? Usar un nivel de significación de 0.05.

Suponiendo que las medidas siguen una distribución normal, con varianza poblacional desconocida y un número de medidas grande, se utiliza una distribución normal tipificada sustituyendo σ por s.

La hipótesis nula y alternativa son:

$$H_0 : \mu = \mu_0 \equiv 38.5$$

$$H_1 : \mu \neq \mu_0 \equiv 38.5$$

En el caso de que la hipótesis nula fuera cierta, $\bar{X}$ seguiría una distribución $N(\mu_0, \sigma/\sqrt{n})$.
El eltadístico de prueba es: $z = \frac{|\bar{x}-\mu_0|}{s/\sqrt{n}}$

El contraste es bilateral, para conservar el área el valor crítico de la abcisa se obtiene para $\alpha/2$

In [210…

```
#head(o3.spring) #util si se pone en formato data.frame, para listas NO

mo3 = mean(o3.spring, na.rm = TRUE); sdo3 = sd(o3.spring, na.rm = TRUE) #Media y d
mu0 =  38.5; #Media con la que se quiere comparar

cat("Media en primavera: ", mo3)
cat("\nDesviación en primavera: ", sdo3)
```

```
Media en primavera:  40.46167
Desviación en primavera:  10.1007
```

In [211…

```
#Como hay más de 30 datos y la varianza es desconocida, se utiliza la distribución
alpha = 0.05 #Nivel de significación.
n = length(o3.spring)
#p-value valor critico
z.alpha = qnorm(alpha/2, lower.tail = FALSE); z.alpha #Contraste bilateral alpha/2
                                                      #Valor crítico de la abcisa

z = abs(mo3 - mu0)/(sdo3/sqrt(n)); z #Valor del estadístico

if (z <= z.alpha){
    cat("Se acepta la hipótesis nula con un nivel de significación alpha = 0.05")
}else{
```

```
    cat("Se rechaza la hipótesis nula con un nivel de significación alpha = 0.05")
}
```

1.95996398454005

1.85265864810611

```
Se acepta la hipótesis nula con un nivel de significación alpha = 0.05
```

Al aceptar la hipótesis de que las medias son iguales, se rechaza que la media muestral sea significativamente distinta del valor dado para alpha = 0.05.

2B. En el siguiente apartado, se pide el nivel de significación para el que se rechaza la hipótesis. Esto es, la probabilidad acumulada desde el valor de abcisa que proporciona el estadístico hasta el final de la distribución.

In [212…

```
#Se busca el área que queda fuera cuando se establece z como valor crítico:

#alpha.z = 2*pnorm(z, lower.tail = FALSE); alpha.z; #Bilateral, se tiene en cuenta
p2b = 2*pnorm(z, lower.tail = FALSE);

cat("La hipótesis se rechaza para un nivel de significación mínimo de", round(p2b,
```

```
La hipótesis se rechaza para un nivel de significación mínimo de 0.064
```

Para $\alpha = 0.064$ se rechaza la hipótesis nula y la media es significativamente diferente de 38.5.

2C) ¿Es el valor medio de o3.spring significativamente mayor que 38.5 ppb? Usar un nivel de significación de 0.05

Contraste unilateral: en este caso, tenemos una desigualdad. Hay que tener cuidado en el lado de la desigualdad, ya que dependiendo de esto la hipótesis se acepta/rechaza de forma distinta.

Como la media es 40.46, la hipótesis nula se establece como lo contrario a lo que dicen los datos (menor que 38.5).

$$H_0 : \mu \leq \mu_0 \equiv 38.5$$

$$H_1 : \mu > \mu_0$$

Si se rechaza la hipótesis nula, se acepta que la media muestral es mayor de 38.5

In [213…

```
z.alpha2 = qnorm(0.05, lower.tail = FALSE); z.alpha2 #z crítico
z2 = abs(mo3 - mu0)/(sdo3/sqrt(n)); z #valor del estadístico

if (z2 <= z.alpha2){
    cat("Se acepta la hipótesis nula con un nivel de significación del 0.05")
}else{
    cat("Se rechaza la hipótesis nula con un nivel de significación del 0.05")
}
```

1.64485362695147

1.85265864810611

```
Se rechaza la hipótesis nula con un nivel de significación del 0.05
```

2D) ¿Para qué nivel de significación mínimo (valor de p) es significativamente mayor que 38.5 ppb?

In [214…

```
#La hipótesis nula se rechaza cuando:

zr2 = pnorm(z2, lower.tail = FALSE); zr2 #Como es contraste unilateral, no hay que
cat("La hipótesis se rechaza para un nivel de significación de", round(zr2,3))
```

```
0.0319656502875287
La hipótesis se rechaza para un nivel de significación de 0.032
```

Cuando $\alpha \geq 0.032$, se rechaza al hipótesis nula y por tanto, la media no es menor de 38.5 para este nivel de significación o lo que es lo mismo, a partir de $\alpha = 0.032$, la media muestral es significativamente mayor de 38.5

In [215…

```
#utilizando la aproximación t-Student
t.test(o3.spring, alternative="greater", mu=mu0, conf.level = 1-zr2) #Hipótesis al
```

```
        One Sample t-test

data:  o3.spring
t = 1.8322, df = 88, p-value = 0.03515
alternative hypothesis: true mean is greater than 38.5
96.80343 percent confidence interval:
 38.45279      Inf
sample estimates:
mean of x
 40.46167
```

Como la muestra es grande, aunque sigue una distribución normal y la t-Student no es exactamente igual pero se le asemeja mucho, puede ser útil para comparar el resultado. Con t-Student, si alpha es igual o mayor de 0.035 la hipótesis se rechaza (resultado muy parecido). De la misma forma, para un nivel de confianza del 96.803%, la media está contenida en el intervalo (no tiene sentido rechazarla).

3) Continuando con las concentraciones de ozono y suponiendo que siguen una distribución normal.

3A) ¿Puede ser la desviación típica de la población igual a un valor de 9 ppb? Usar un nivel de significación de 0.05

Ahora, en vez de comparar la media, se compara la varianza y la distribución utilizada será χ^2 y el estadístico

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

$$H_0 : \sigma^2 = \sigma_0^2 \equiv 9^2$$

$$H_1 : \sigma^2 \neq \sigma_0^2 \equiv 9^2$$

Es un contraste bilateral, con nivel de significación $\alpha = 0.05$. A diferencia de los anteriores, la distribución no es simétrica y el rango de aceptación se calcula con las dos abcisas.

In [216…

```
sd0 = 9; sd3 = sd(o3.spring, na.rm=TRUE); n = length(o3.spring)

alpha = 0.05

chi = (n-1)*sdo3**2 / sd0**2; chi
chi.alpha1 = qchisq(alpha/2, df=n-1); chi.alpha1 #Cola izquierda, equivalente a (1
chi.alpha2 = qchisq(alpha/2, df=n-1, lower.tail=FALSE); chi.alpha2 #Cola derecha: (

chi > chi.alpha1 & chi < chi.alpha2
```

113.360170269418

65.6466175764689

118.135892560615

TRUE

Se acepta la hipótesis nula, para el nivel elegido, la desviación típica es significativamente igual a 9.

3B) ¿Para qué nivel de significación mínimo (valor de p) es la desviación típica significativamente diferente de un valor de 9 ppb?

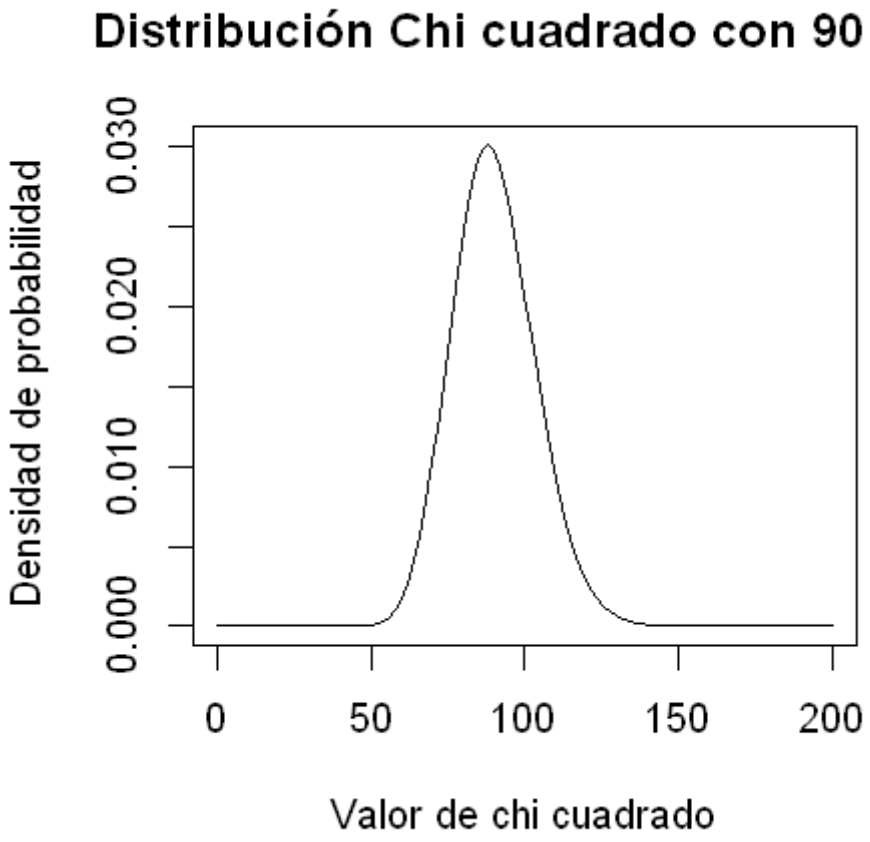
Se busca el p-value, es decir, el área que queda fuera para la abcisa crítica igual al estadístico. La distribución Chi cuadrado no es simétrica, pero al tener una muestra grande y 90 grados de libertad, se puede considerar "simétrica".

El valor se compara con la función de R var.Test que también proporciona el p-value.

In [217…

```
options(repr.plot.height = 4, repr.plot.width = 4) #Se ajusta el tamaño
curve(dchisq(x, df=n-1), from=0, to=200, main = "Distribución Chi cuadrado con 90 g

#xc = seq(0, 200, by = 0.1)
#yc = dchisq(xc, df) #densidad de probabilidad para cada valor de x
#plot(xc, yc, type = "l", xlab = "Valor de chi cuadrado", ylab = "Densidad de probc
```



In [192…
```r
#alpha.chi = 2 * (1 - chisq.cdf(chi, n - 1))
alpha.chi = 2*pchisq(chi, df = n - 1, lower.tail = FALSE); alpha.chi

cat("Para alpha =",round(alpha.chi,3),"la desviación típica es significativamente
```

0.0971948921434263

```
Para alpha = 0.097 la desviación típica es significativamente diferente de 9
```

In [193…
```r
install.packages("EnvStats") #Hace falta dejarlo puesto
library("EnvStats")
```

```
  There is a binary version available but the source version is later:
        binary source needs_compilation
EnvStats  2.4.0  2.7.0             FALSE

installing the source package 'EnvStats'
```

In [194…
```r
vT = varTest(o3.spring, conf.level=0.95, alternative = "two.sided", sigma.squared=
pv = vT$p.value; round(pv,3) #names(vT);
```

```
Warning message in is.not.finite.warning(x):
"There were 2 nonfinite values in x : 2 NA's"Warning message in varTest(o3.spring,
conf.level = 0.95, alternative = "two.sided", :
"2 observations with NA/NaN/Inf in 'x' removed."
```

0.101

Se obtiene un valor parecido, en el anterior hacíamos la suposición de que chi cuadrado es simétrica.

Con varTest, para un p-value (nivel de significación) de 0.101, la desviación típica significativamente diferente de 9.

4A) ¿Dirías que la concentración media de ozono del mes de abril es significativamente mayor que la del mes de junio? En caso afirmativo, ¿para qué nivel de significación?

Ahora, tenemos un contraste de igualdad de medias entre dos poblaciones normales.

In [218…
```
o3.a = ldatos[[4]]; o3.j = ldatos[[6]] # Con el doble [[]] se guardan como vectores
#o3.a; o3.j

mo3.a = mean(o3.a, na.rm = TRUE); mo3.j = mean(o3.j, na.rm = TRUE)
cat("Abril (ppm):",mo3.a,"\nJunio (ppm):",mo3.j)

sdo3.a = sd(o3.a, na.rm = TRUE); sdo3.j = sd(o3.j, na.rm = TRUE)
```

```
Abril (ppm): 41.2153
Junio (ppm): 35.96766
```

La concentración de abril es mayor que la de junio, por tanto se establece como hipótesis nula la contraria.

$$H_0 : \mu_1 \leq \mu_2$$

$$H_1 : \mu_1 > \mu_2$$

Con μ_1 para abril (mayor media) y μ_2 para junio. Al tener cada muestra tamaño 30 y ser las varianzas poblacionales desconocida, se utiliza la distribución normal.

El enunciado pide el nivel de significación para el cual la media de abril es mayor que la de junio.

Se busca rechazar la hipótesis nula, que la media de junio no es mayor que la de abril. Esto se da cuando se toma como valor crítico el del estadístico.

In [219…
```
z3 = (mo3.a - mo3.j)  / sqrt(sdo3.a**2/n + sdo3.j**2/n ); #z3 #En este caso, tanto

p.aj = pnorm(z3, lower.tail = FALSE); p.aj
```

6.11347590305203e-08

Para el nivel de significación $\alpha = 6.11 \cdot 10^{-8}$ la concentración media de abril es mayor que la de junio.

Por ejemplo, si elegimos un nivel de 0.05, al ser mayor que el valor obtenido, se rechazará.

In [220…
```
alpha = 0.05
z3.cri = qnorm(alpha, lower.tail = FALSE); z3.cri; z3

z3 <= z3.cri
```

1.64485362695147

5.29007005592408

FALSE

4B) ¿Dirías que la concentración media de ozono en el mes de abril es significativamente diferente de la del mes de mayo?

En caso afirmativo, ¿para qué nivel de significación?

In [221…
```
o3.m = ldatos[[5]]; mo3.m = mean(o3.m, na.rm = TRUE); sdo3.m = sd(o3.m, na.rm = TRU

cat("MAYO. Media:",mo3.m,". Desviación típica:",sdo3.m)
cat("\nABRIL. Media:",mo3.a,". Desviación típica:",sdo3.a)
```

```
MAYO. Media: 44.13002 . Desviación típica: 13.43814
ABRIL. Media: 41.2153 . Desviación típica: 6.130413
```

Como hipótesis nula, se propone que la concentración media es igual.

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

El nivel de significación vendrá determinado por el valor del estadístico.

In [222…
```
z4 = abs(mo3.a - mo3.m)  / sqrt(sdo3.a**2/n + sdo3.m**2/n ); #z4 #En este caso, tal

#Se rechaza a partir de
alpha.4 = 2*pnorm(z4, lower.tail = FALSE); alpha.4
```

0.059774479542294

Por tanto, a partir de $\alpha = 0.597$ se rechaza la hipótesis nula y las medias son significativamente diferentes. Por debajo, se acepta la hipótesis nula ("son significativamente iguales")

In [223…
```
#Por ejemplo, fijamos alpha = 0.06 (se rechazaría la hipótesis nula)
z4.cri = qnorm(0.06/2, lower.tail = FALSE); z4.cri

z4 < z4.cri
```

1.88079360815125

FALSE

Si fijamos de antemano un nivel de 0.06, al ser mayor de 0.05, se rechaza la hipótesis.

5) Test de bondad de ajuste.

Hasta ahora, se ha supuesto que los datos siguen una distribución normal, pero ¿es una buena suposición?

Para ello se crea un histograma agrupando los datos y se guarda como parámetro para obtener la tabla de frecuencias.

Hacemos un contraste de hipótesis:

H_0 : los datos siguen una distribución normal

H_1 : los datos no siguen una distribución normal

In [224…
```
limits = seq(5, 85, 5) #Extremos de los intervalos

options(repr.plot.height = 6, repr.plot.width = 6) #Tamaño del gráfico
num_colores = length(limits) - 1 #Numero de colores distintos (numero de intervalos
```

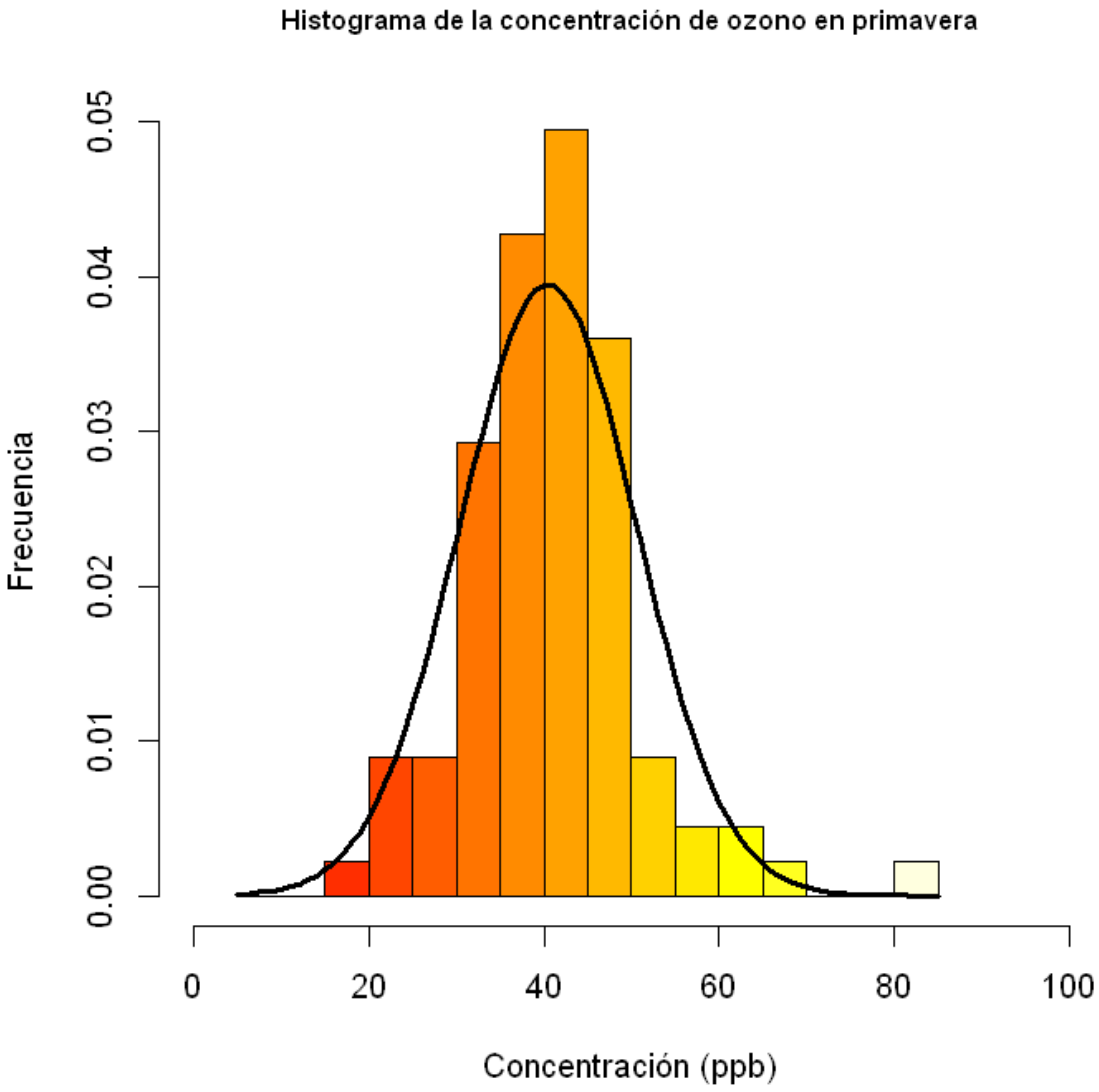
```
colores = heat.colors(length(limits)-1) #colores = rainbow(num_colores); #colores =

h = hist(o3.spring, breaks = limits, freq=FALSE, xlim = c(0,100), col = colores, ma
         cex.main = 0.8, ylab = "Frecuencia", xlab = "Concentración (ppb)")

#x = seq(5,85,length.out = 100)

x = seq(min(h$breaks), max(h$breaks), by=1)
y = dnorm(x,mean=mo3,sd=sdo3)
#curve(dnorm(x,mean=mo3,sd=sdo3), col="black", add=TRUE, lwd=2.5)

lines(x, y, type="l", col="black", lwd=3)
```



In [225…

```
N = length(h$breaks)
ind <- h$breaks[2]-h$breaks[1]

h$breaks

tfreq <- data.frame(h$breaks[1:N-1], h$breaks[2:N], h$mids,
h$counts,h$density*ind,cumsum(h$counts),cumsum(h$density)*ind)
names(tfreq) <-c("a_i","a_i+1","c_i","n_i","f_i","N_i","F_i")
tfreq
```

1. 5
2. 10
3. 15
4. 20
5. 25
6. 30
7. 35
8. 40
9. 45
10. 50
11. 55
12. 60
13. 65
14. 70
15. 75
16. 80
17. 85

a_i	a_i+1	c_i	n_i	f_i	N_i	F_i
5	10	7.5	0	0.00000000	0	0.00000000
10	15	12.5	0	0.00000000	0	0.00000000
15	20	17.5	1	0.01123596	1	0.01123596
20	25	22.5	4	0.04494382	5	0.05617978
25	30	27.5	4	0.04494382	9	0.10112360
30	35	32.5	13	0.14606742	22	0.24719101
35	40	37.5	19	0.21348315	41	0.46067416
40	45	42.5	22	0.24719101	63	0.70786517
45	50	47.5	16	0.17977528	79	0.88764045
50	55	52.5	4	0.04494382	83	0.93258427
55	60	57.5	2	0.02247191	85	0.95505618
60	65	62.5	2	0.02247191	87	0.97752809
65	70	67.5	1	0.01123596	88	0.98876404
70	75	72.5	0	0.00000000	88	0.98876404
75	80	77.5	0	0.00000000	88	0.98876404
80	85	82.5	1	0.01123596	89	1.00000000

Como algunos intervalos tienen una frecuencia menor a 5, hay que agrupar.

In [226…

```
# calculamos las frecuencias esperadas
limites = limits

# número total de intervalos
ninterv <- length(limites)-1
```

```r
# inicializamos vector de frecuencias relativas (probabilidades)
p <- numeric(ninterv)

# Rellenamos el vector para cada intervalo
for (i in 1:ninterv) {

     # último intervalo
     if (i == ninterv) {
         p[i] = pnorm(limites[i], mean=mo3, sd=sdo3, lower.tail=FALSE) # cola dere

     } else {
           # primer intervalo
           if (i == 1) {
              p[i] = pnorm(limites[i+1], mean=mo3, sd=sdo3)

            # ni primer ni último intervalo
            } else {
                p[i] = pnorm(limites[i+1], mean=mo3, sd=sdo3)-
                          pnorm(limites[i], mean=mo3, sd=sdo3)
             }

         }
   }

#Como hay algunos valores NA, no se tienen en cuenta para length

o = h$counts
e = p * 89
#df = data.frame(e, o); df

sum(o); sum(e)
```

89

89

Si los datos siguen una distribución normal, el valor esperado debe parecerse a la frecuencia observada.

In [227…

```r
#Al haber intervalos con menos de 5 datos, hace falta juntarlos. A cambio, se piera
N = length(h$breaks)

ind <- h$breaks[2]-h$breaks[1]
tfreq1 <- data.frame(h$breaks[1:N-1], h$breaks[2:N], h$mids,
h$counts,h$density*ind,cumsum(h$counts),cumsum(h$density)*ind, e)
names(tfreq1) <-c("a_i","a_i+1","c_i","n_i","f_i","N_i","F_i", "e_i")
tfreq1
```

a_i	a_i+1	c_i	n_i	f_i	N_i	F_i	e_i
5	10	7.5	0	0.00000000	0	0.00000000	0.114055315
10	15	12.5	0	0.00000000	0	0.00000000	0.407013632
15	20	17.5	1	0.01123596	1	0.01123596	1.383024408
20	25	22.5	4	0.04494382	5	0.05617978	3.695368853
25	30	27.5	4	0.04494382	9	0.10112360	7.764990157
30	35	32.5	13	0.14606742	22	0.24719101	12.832662182
35	40	37.5	19	0.21348315	41	0.46067416	16.680589285
40	45	42.5	22	0.24719101	63	0.70786517	17.054484209
45	50	47.5	16	0.17977528	79	0.88764045	13.715117567
50	55	52.5	4	0.04494382	83	0.93258427	8.675264825
55	60	57.5	2	0.02247191	85	0.95505618	4.315830560
60	65	62.5	2	0.02247191	87	0.97752809	1.688530701
65	70	67.5	1	0.01123596	88	0.98876404	0.519481661
70	75	72.5	0	0.00000000	88	0.98876404	0.125658642
75	80	77.5	0	0.00000000	88	0.98876404	0.023895240
80	85	82.5	1	0.01123596	89	1.00000000	0.004032761

In [228…

```
# Tabla final (agrupando 5 primeros y 3 últimos intervalos)
o_new = c(o[1]+o[2]+o[3]+o[4]+o[5], o[6:9], o[10]+o[11]+o[12]+o[13]+o[14]+o[15]+o[
e_new = c(e[1]+e[2]+e[3]+e[4]+e[5], e[6:9], e[10]+e[11]+e[12]+e[13]+e[14]+e[15]+e[

hb = h$breaks
h_new = c(hb[1], hb[6:9], hb[10], hb[17]); h_new

sum(o_new); sum(e_new)
```

1. 5
2. 30
3. 35
4. 40
5. 45
6. 50
7. 85

89

89

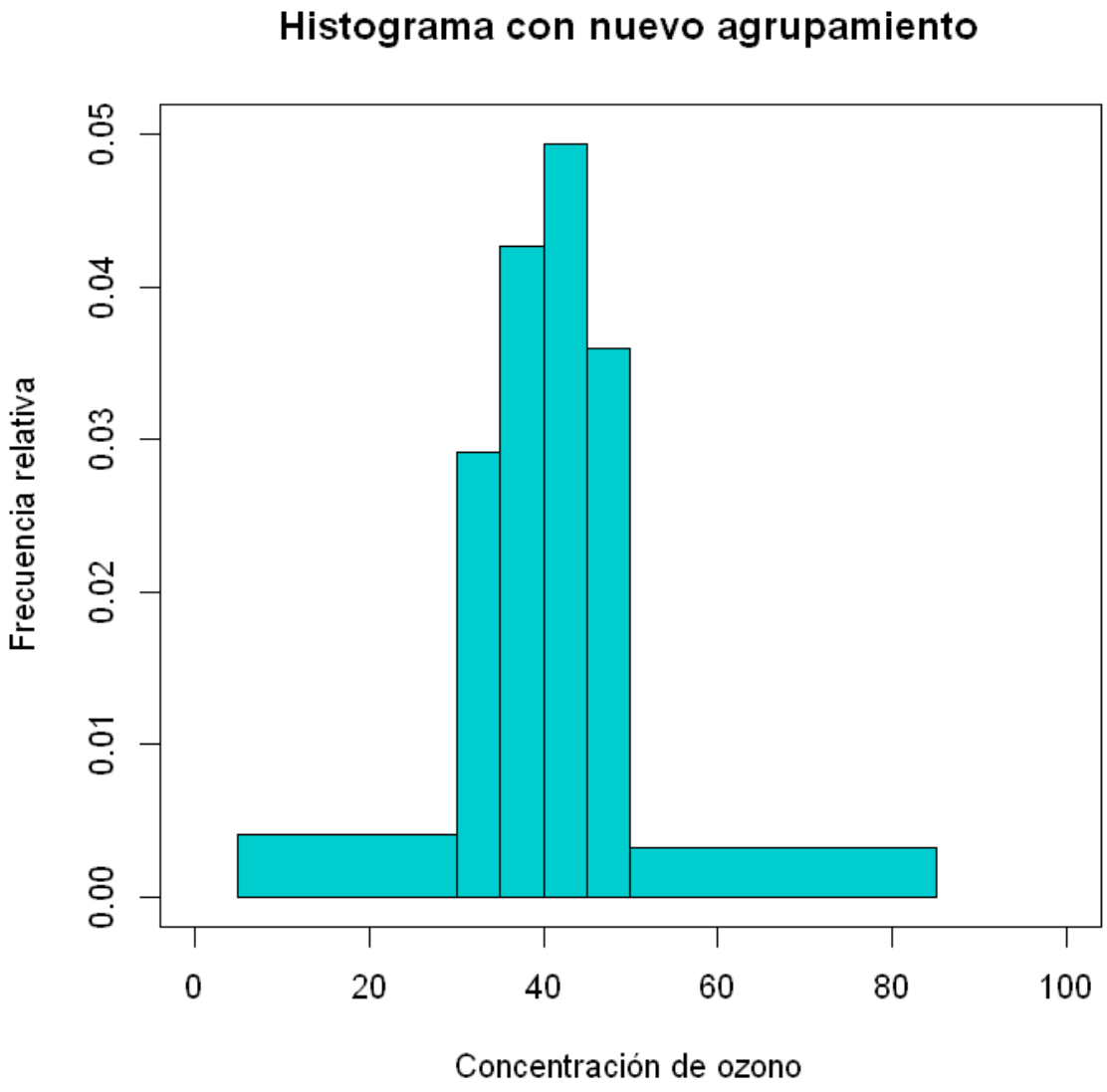
In [229…

```
NN = length(h_new)
tfreq2 = data.frame(h_new[1:NN-1], h_new[2:NN], o_new, e_new)
names(tfreq2) = c("a_i","a_i+1","n_i","e_i")

#Puesto que se ha realizado un agrupamiento, para el contraste de hipótesis se uti
tfreq2

new_h = hist(o3.spring, breaks = h_new, freq = FALSE, ylim = c(0,0.05), xlim = c(0
             xlab = "Concentración de ozono", ylab = "Frecuencia relativa")
box()
```

a_i	a_i+1	n_i	e_i
5	30	9	13.36445
30	35	13	12.83266
35	40	19	16.68059
40	45	22	17.05448
45	50	16	13.71512
50	85	10	15.35269



In [230…

```
# El estadístico de prueba será la suma de los elementos de:
#(o_new-e_new)**2/e_new

chi5 = sum((o_new-e_new)**2/e_new) ; chi5

length(o_new); length(e_new)
```

5.43097755848241

6

6

In [231…
```
# k  = 6: número de intervalos
# p  = 2: número de parámetros estimados (media & stdev)
# df = k - p - 1 = 6 - 2 - 1 = 3
#
# valor crítico (para alfa=0.05 y 3 grados de libertad)
alpha = 0.05
chi_01 = qchisq(alpha, df=3, lower.tail=FALSE)
chi_01
```

7.81472790325118

In [232…
```
# aceptamos H_0 para alfa=0.05 porque se cumple que

chi5 < chi_01
```

TRUE

Al reducir los grados de libertad por el agrupamiento, no se puede utilizar chisq.test

Se acepta la hipótesis de que se pueda aproximar por una distribución normal para el nivel de significación elegido ($\alpha = 0.05$).

Por tanto, tiene sentido la suposición de que los datos siguen una distribución normal realizada en los apartados 2, 3 y 4.

6) Volver a usar las concentraciones de ozono considerando todos los meses para los que hay datos. Hacer un análisis de varianza para determinar si hay relación entre la concentración media y el mes.

Con el análisis de varianza se pretende comprobar la igualdad de medias entre dos o más poblaciones, en este caso, entre los 12 meses del año.

H_0 : No hay diferencia significativa entre la concentración media de ozono en diferentes meses.

H_1 : Existe una diferencia significativa entre la concentración media de ozono en diferentes meses.

In [235…
```
ozono = data$o3
mes = data$month

df = data.frame(mes, ozono); #str(df)
newdf = na.omit(df); str(newdf) #Se crea un nuevo dataframe omitiendo las filas co

#head(df); tail(df)
#head(newdf); tail(newdf)
```

```
'data.frame':   357 obs. of  2 variables:
 $ mes  : int  1 1 1 1 1 1 1 1 1 ...
 $ ozono: num  32 35 36 33 33 ...
 - attr(*, "na.action")= 'omit' Named int  16 17 26 81 103 114 229 234 235
  ..- attr(*, "names")= chr  "16" "17" "26" "81" ...
```

In [236…
```
Type = factor(newdf$mes); #Type
p = length(levels(Type)); #p
nt = nrow(newdf); #nt
```

```
C = (sum(newdf$ozono,na.rm = TRUE))**2/nt; #C
VT = sum(newdf$ozono**2,na.rm = TRUE) - C; #VT
T = tapply(newdf$ozono, Type, sum, na.rm = TRUE); #T
n = tapply(newdf$ozono, Type, length);
VET = sum(T**2/n)-C
VDT = VT - VET
MT = VET/(p-1) ; #MT #Cuadrado medio de los tratamientos
ME = VDT/(nt-p) ; #ME #Cuadrado medio del azar
F = MT/ME ; #F # Estadistico

cat("El cuadrado medio de los tratamientos es", MT, "\n")
cat("El cuadrado medio del azar es", ME, "\n")

f.alpha = pf(F, df1=p-1, df2=nt-p, lower.tail=FALSE); #f.alpha
cat("La hipótesis nula de igualdad de medias se rechaza para un nivel se significac
cat("\nEl valor del estadístico es:",round(F,2))
```

```
El cuadrado medio de los tratamientos es 1252.172
El cuadrado medio del azar es 108.0686
La hipótesis nula de igualdad de medias se rechaza para un nivel se significación:
1.971117e-18
El valor del estadístico es: 11.59
```

El cuadrado medio de los tratamientos es mucho mayor que el del azar $F = \frac{MT}{ME} \approx 11$, hay una gran variación en al menos 2 medias de los tratamientos en comparación con la variación esperada por azar. Por ello, el nivel para el que se rechaza la hipótesis nula es tremendamente bajo 10^{-18}.

En el apartado (1), el boxplot muestra diferencias para los distintos meses con una cierta tendencia para primera-verano y otra para otoño-invierno. Con el análisis de varianzas, se ha demostrado que **al menos dos de las medias son diferentes**, lo que no sabemos es entre cuales, para ello habría que realizar un análisis 2 a 2, por meses. Otra opción sería comparar la media total de marzo-septiembre frente a la de octubre-febrero.

Si cogemos un nivel típico de 0.05 (mayor que el p-value), la hipótesis nula será rechazada:

In [237…
```
alpha = 0.05 # Nivel de significación
Fcrit = qf(alpha, df1=p-1, df2=nt-p, lower.tail=FALSE) ; #Fcrit

F < Fcrit
```

FALSE

Se rechaza la hipótesis nula para $\alpha = 0.05$, la evidencia sugiere que existe una diferencia significativa en los niveles de ozono de cada mes, por tanto, se puede concluir que las diferencias no son producto del azar, para al menos dos meses **hay relación entre los niveles de ozono y el mes**.

R dispone de una función propia para realizar un análisis de varianzas (ANOVA). Se obtiene los mismos resultados que antes.

In [239…
```
mianova = aov(newdf$ozono ~ Type)
summary(mianova)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
Type         11  13774  1252.2   11.59 <2e-16 ***
Residuals   345  37284   108.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```