Notas de estadística no parametrica

Pruebas de Bondad de Ajuste y Aleatoriedad

Humberto Vaquera Huerta

Colegio de Postgraduados

2022

Pruebas de Bondad de Ajuste

Una prueba de **bondad de ajuste** es una un tipo de prueba de hipótesis en la que la *hipótesis nula* es que la población tiene una distribución de probabilidad específica.

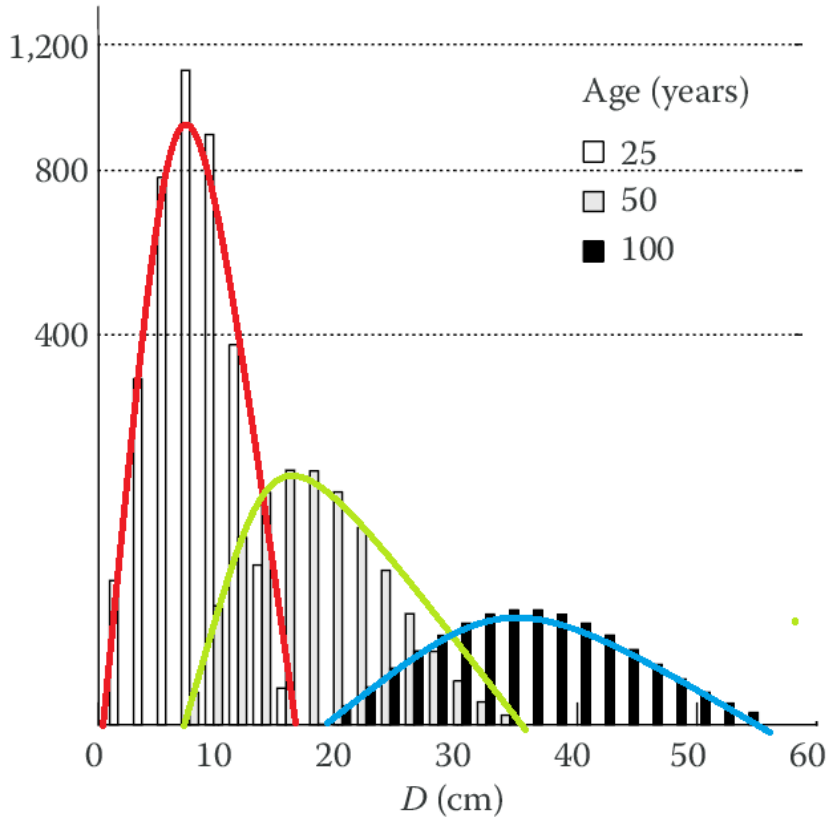
Como investigador, si se tiene un conjunto de datos le interesa saber cuál es la distribución probabilistica que gobierna el comportamiento de una variable aleatoria X de un conjunto de datos.

El objetivo es mostrar cómo podemos responder a esa pregunta usando algunas pruebas graficas y analíticas.

- 1. Bondad de ajuste: Chi-cuadrada de Pearson, Kolmogorov-Smirnov
- 2. Pruebas de Normalidad: Shapiro-Wilks y Jarque Vera
- 3. Aleatoriedad

Pruebas de Bondad de Ajuste

¿Qué tan bien se corresponden los datos observados con el modelo ajustado?



Histograma de frecuencia de los diámetros de los árboles(D) para un rodal ficticio de roble gestionado a los 25, 50 y 100 años de edad,

Pruebas de Bondad de Ajuste

Como investigador, si se tiene un conjunto de datos le interesa saber cuál es la distribución probabilistica que gobierna el comportamiento de una variable aleatoria X de un conjunto de datos.

Métodos gráficos

- Histograma,
- Gráfico de la densidad suavizada,
- Gráfico de cajas,
- Gráfico de la distribución empírica (o versión suavizada) y
- Gráficos P-P o Q-Q.

Pruebas de Bondad de Ajuste

Métodos gráficos

Histograma

Se agrupan los datos en intervalos $I_k = [L_{k-1}, L_k)$ con $k = 1, \ldots, K$ y a cada intervalo se le asocia un valor (altura de la barra) igual a la frecuencia absoluta de ese intervalo $n_k = \sum_{i=1}^{n} \mathbf{1}(X_i \in [L_{k-1}, L_k))$, si la longitud de los intervalos es constante, o proporcional a dicha frecuencia (de forma que el área coincida con la frecuencia relativa y pueda ser comparado con una función de densidad):

$$\hat{f}_n(x) = \frac{n_i}{n\,(L_k - L_{k-1})}$$

Como ya se ha visto anteriormente, en `R` podemos generar este gráfico con la función `hist()` del paquete base.

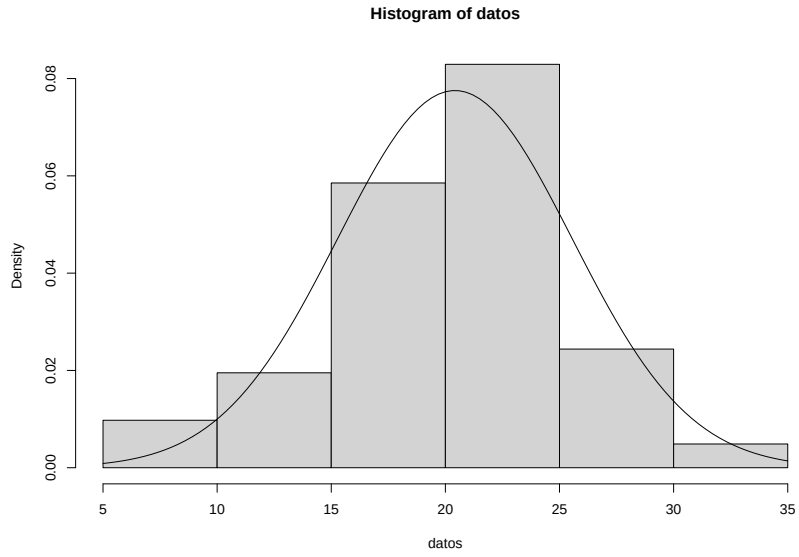
Pruebas de Bondad de Ajuste

Métodos gráficos

Histograma

Ejemplo:

```
datos <- c(22.56,22.33,24.58,23.14,19.03,26.76,18.33,23.10,
  21.53,9.06,16.75,23.29,22.14,16.28,18.89,27.48,10.44,
  26.86,27.27,18.74,19.88,15.76,30.77,21.16,24.26,22.90,
  27.14,18.02,21.53,24.99,19.81,11.88,24.01,22.11,21.91,
  14.35,11.14,9.93,20.22,17.73,19.05)
hist(datos, freq = FALSE)
curve(dnorm(x, mean(datos), sd(datos)), add = TRUE)
```



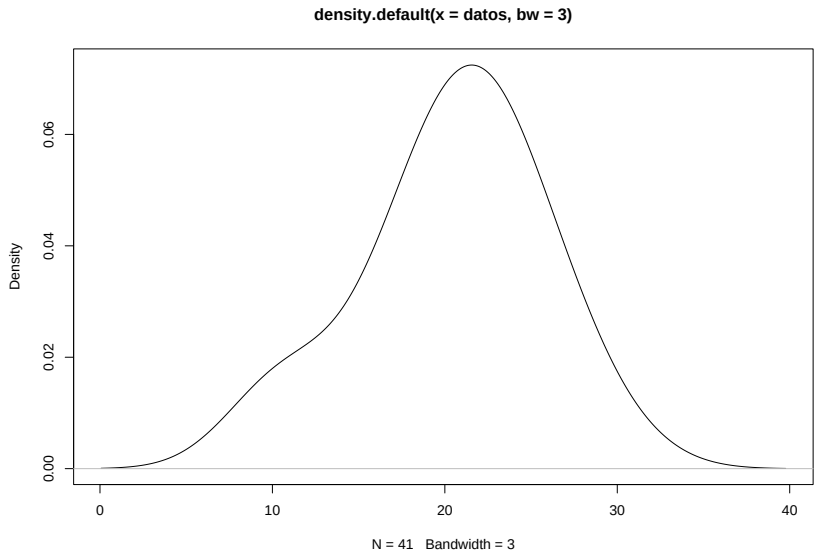
Pruebas de Bondad de Ajuste

Métodos gráficos

Densidad

Alternativamente se podría considerar una estimación suave de la densidad, por ejemplo empleando la estimación tipo núcleo implementada en la función `density()`.

```
den=density(datos,bw=3)
plot(den)
```



Pruebas de Bondad de Ajuste

Métodos gráficos

Función de distribución empírica

La función de distribución empírica $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i \leq x)$ asigna a cada número real x la frecuencia relativa de observaciones menores o iguales que x. Para obtener las frecuencias relativas acumuladas, se ordena la muestra $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ y:

$$F_n(x) = \begin{cases} 0 & \text{si} \quad x < X_{(1)} \\ \frac{i}{n} & \text{si} \quad X_{(i)} \leq x < X_{(i+1)} \\ 1 & \text{si} \quad X_{(n)} \leq x \end{cases}$$

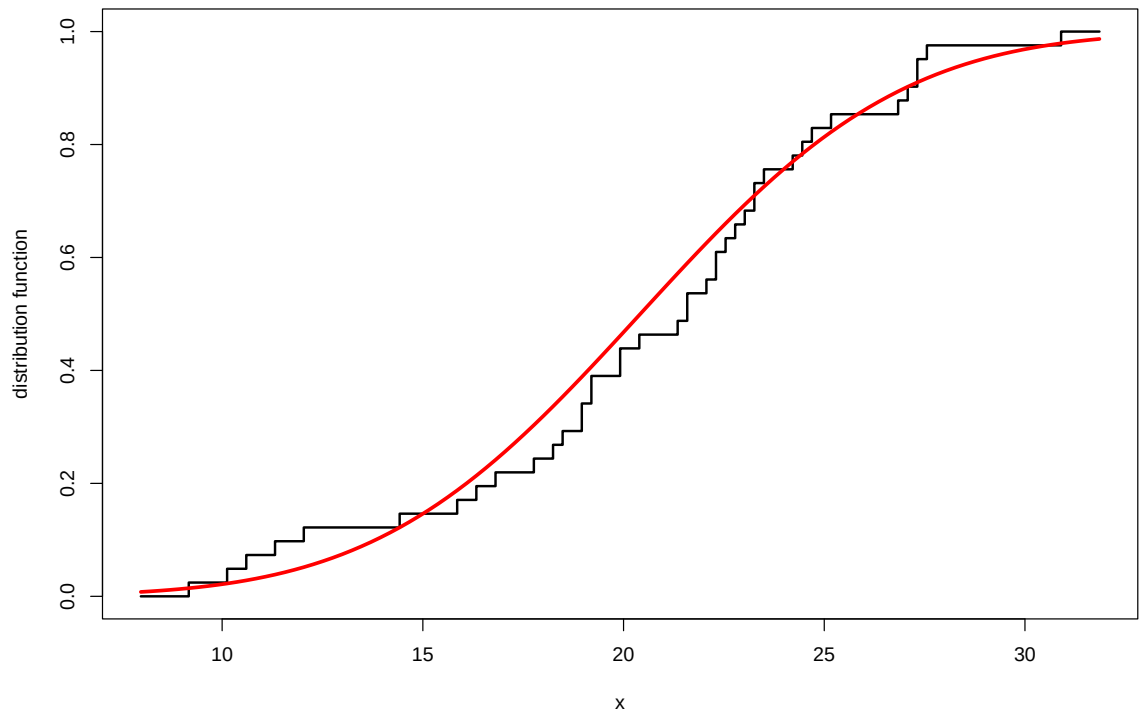
Pruebas de Bondad de Ajuste

Métodos gráficos

Función de distribución empírica

Ejemplo:

```
fn <- ecdf(datos)
curve(ecdf(datos)(x), xlim = extendrange(datos), type = 's',
      ylab = 'distribution function', lwd = 2)
curve(pnorm(x, mean(datos), sd(datos)), add = TRUE,col="red",lwd=3)
```



Pruebas de Bondad de Ajuste

Métodos gráficos

Gráficos P-P y Q-Q

El gráfico de probabilidad (o de probabilidad-probabilidad) es el gráfico de dispersión de:

$$\{(F_0(x_i), F_n(x_i)) : i = 1, \cdots, n\}$$

siendo F_n la función de distribución empírica y F_0 la función de distribución bajo H_0 (con la que desea comparar, si la hipótesis nula es simple) o una estimación bajo H_0 (si la hipótesis nula es compuesta; e.g. si $H_0 : F = \mathcal{N}(\mu, \sigma^2)$, $\hat{F}_0$ función de distribución de $\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$). Si H_0 es cierta, la nube de puntos estará en torno a la recta $y = x$ (probabilidades observadas próximas a las esperadas bajo H_0).

Pruebas de Bondad de Ajuste

Métodos gráficos

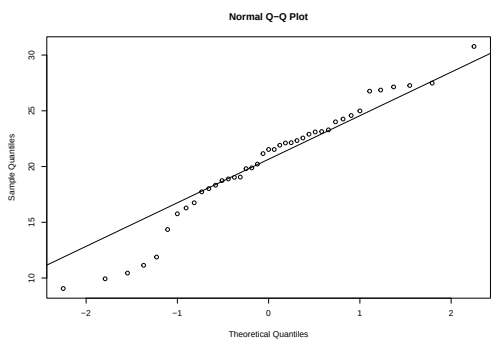
Gráficos P-P y Q-Q

El gráfico Q-Q (cuantil-cuantil) es equivalente al anterior pero en la escala de la variable:

$$\left\{\left(q_i, x_{(i)}\right) : i = 1, \cdots, n\right\}$$

siendo $x_{(i)}$ los cuantiles observados y $q_i = F_0^{-1}(p_i)$ los esperados[1] bajo H_0.
Ejemplo:

```
qqnorm(datos);qqline(datos)
```

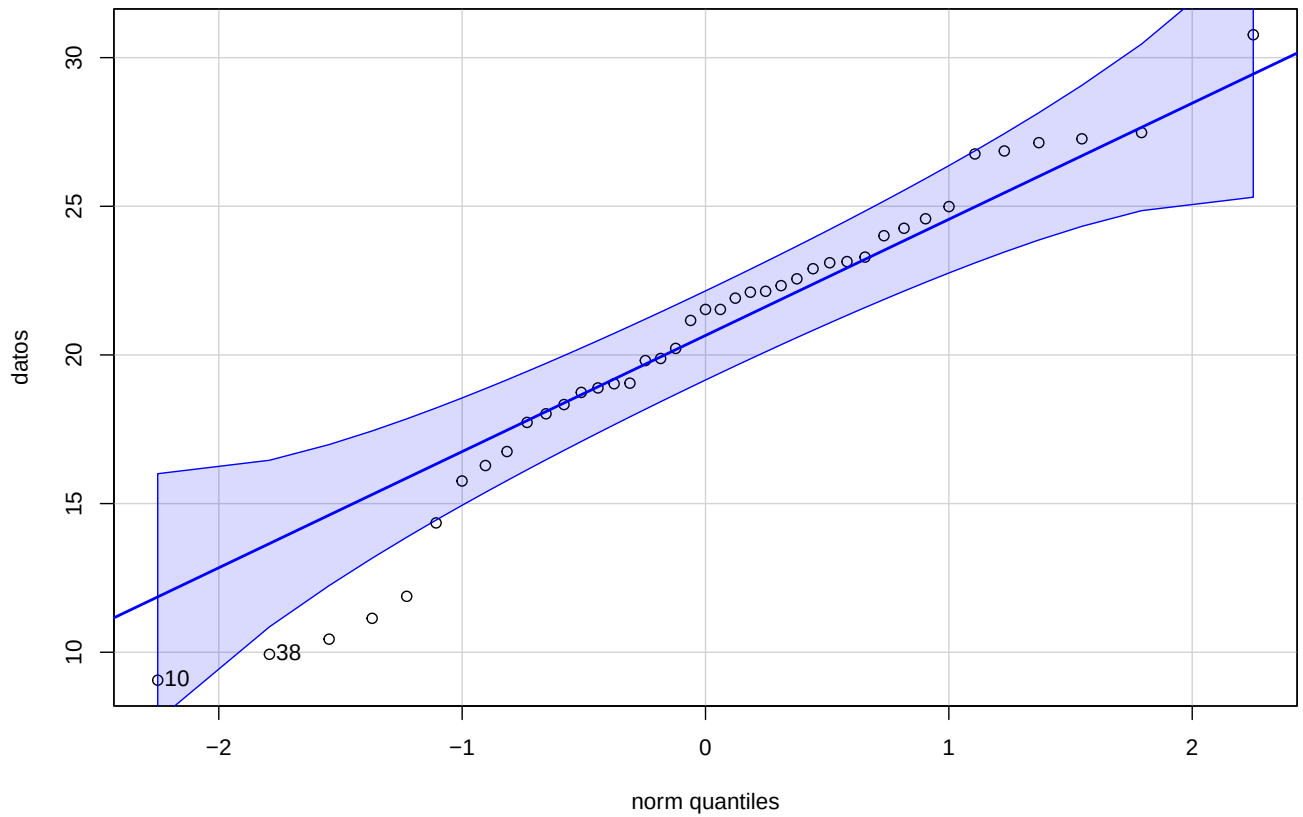


Pruebas de Bondad de Ajuste

Métodos gráficos

Gráficos P-P y Q-Q

```
require(car)
qqPlot(datos, "norm")
```



Pruebas de Bondad de Ajuste

Ejemplo: Datos precipitación Texcoco

Base de datos Climatológica Nacional (CLICOM)

http://clicom-mex.cicese.mx/

Texcoco precipitación promedio anual (mm).

Fecha	Precipitación	Fecha	Precipitación	Fecha	Precipitación
01/01/1952	24.36	01/01/1985	24.55	01/01/1992	25.07
01/01/1953		01/01/1986	25.49	01/01/1993	26.25
01/01/1954	23.93	01/01/1987	25.69	01/01/1994	26.53
01/01/1955	23.99	01/01/1988	26.07	01/01/1995	26.85
01/01/1956	23.61	01/01/1989	25.98	01/01/1996	26.88
01/01/1957	25.51	01/01/1990	25.65	01/01/1997	27.17
01/01/1958	23.69	01/01/1991	26.12	01/01/1998	28.11
01/01/1959	24.36	01/01/1992	25.07	01/01/1999	26.76
01/01/1960	25.47	01/01/1993	26.25	01/01/2000	26.64
01/01/1961	24.54	01/01/1994	26.53	01/01/2001	26.55
01/01/1962	24.06	01/01/1995	26.85	01/01/2002	26.98
01/01/1963	23.25	01/01/1996	26.88	01/01/2003	26.62
01/01/1964	23.33	01/01/1997	27.17	01/01/2004	26.45
01/01/1965	23.26	01/01/1998	28.11	01/01/2005	26.69
01/01/1966	23.01	01/01/1999	26.76	01/01/2006	26.37
01/01/1967	23.21	01/01/2000	26.64	01/01/2007	25.98
01/01/1968	23.05	01/01/2001	26.55	01/01/2008	26.29
01/01/1969	24.11	01/01/2002	26.98	01/01/2009	26.38
01/01/1970	24	01/01/2003	26.62	01/01/2010	25.23
01/01/1971	23.72	01/01/2004	26.45	01/01/2011	
01/01/1972	24.17	01/01/2005	26.69	01/01/2012	
01/01/1973	23.98	01/01/2006	26.37	01/01/2013	24.65
01/01/1974	23.73	01/01/2007	25.98	01/01/2014	24.69
01/01/1975	23.65	01/01/2008	26.29	01/01/2015	24.86
01/01/1976	23.27	01/01/2009	26.38		
01/01/1977	25.25	01/01/2010	25.23		
01/01/1978	24.89	01/01/2011			
01/01/1979	25.62	01/01/2012			
01/01/1980	25.38	01/01/2013	24.65		
01/01/1981	24.75	01/01/2014	24.69		
01/01/1982	26.09	01/01/2015	24.86		
01/01/1983	25.47	01/01/1990	25.65		
01/01/1984	24.78	01/01/1991	26.12		

Pruebas de Bondad de Ajuste

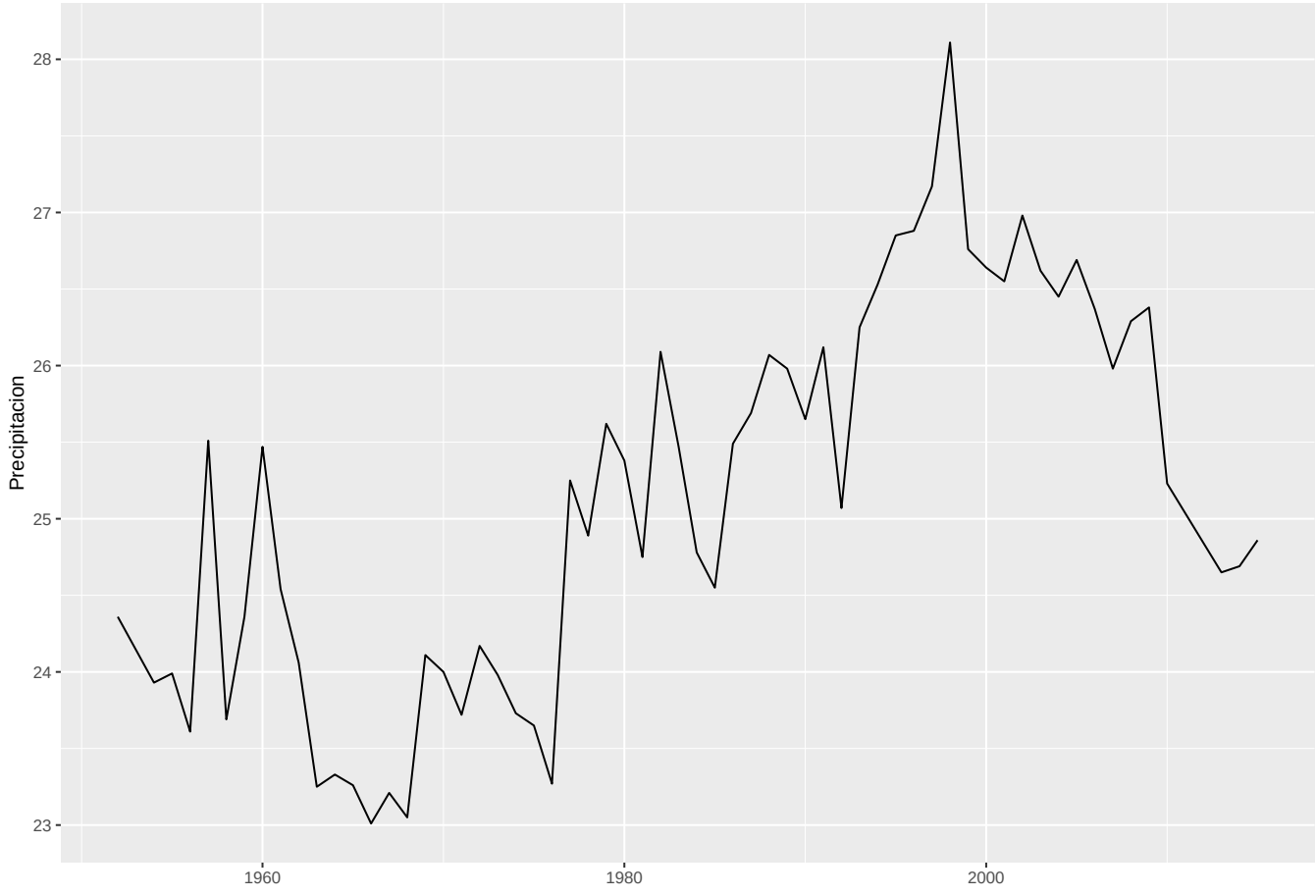
¿Como es la distribución de la la precipitacion de Texcoco?

```
precip <- na.omit(read.csv("D:/cursos/no param/Curso Noparametrica/precip.csv"))
precipitacion=na.omit(precip[,2])
pander::pander(head(precip,10))
```

	Fecha	Precipitacion	year
1	1952-01-01	24.36	1952
3	1954-01-01	23.93	1954
4	1955-01-01	23.99	1955
5	1956-01-01	23.61	1956
6	1957-01-01	25.51	1957
7	1958-01-01	23.69	1958
8	1959-01-01	24.36	1959
9	1960-01-01	25.47	1960
10	1961-01-01	24.54	1961
11	1962-01-01	24.06	1962

Pruebas de Bondad de Ajuste

Precipitación promedio anual en Texcoco

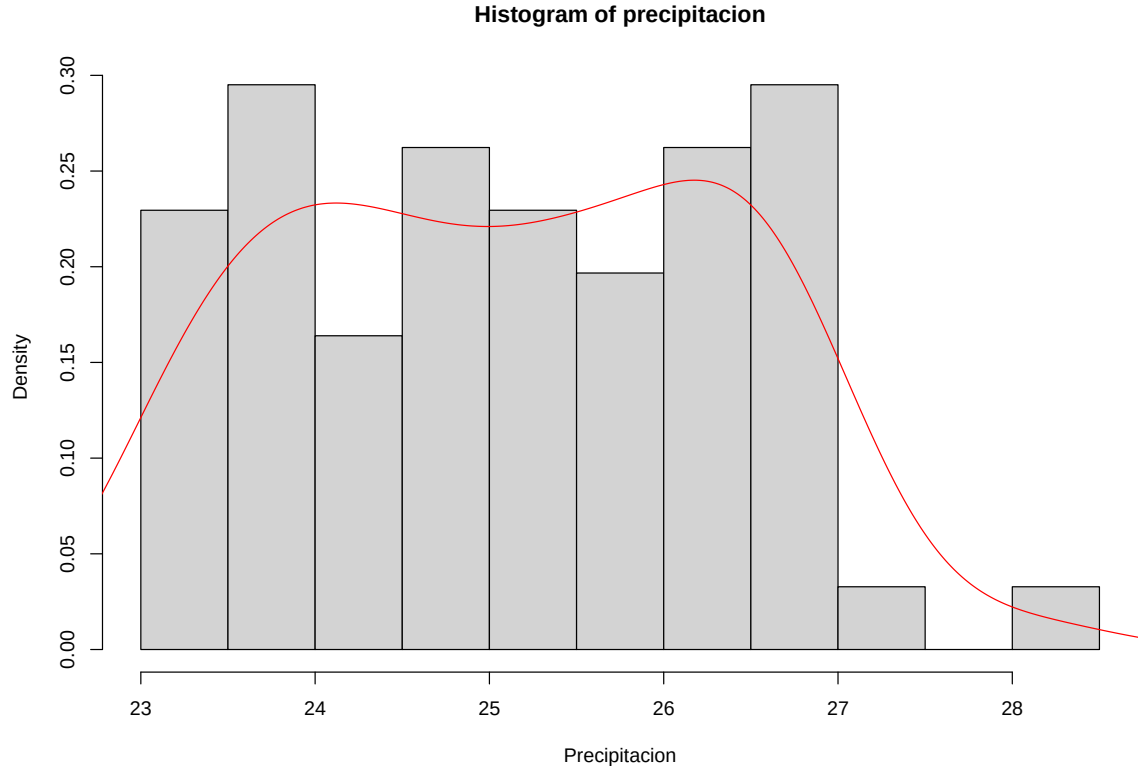


Pruebas de Bondad de Ajuste

¿Como es la distribución de la la precipitacion de Texcoco?

Histograma y grafica de densidad

```
hist(precipitacion,15,prob = TRUE,xlab ="Precipitacion")
lines(density(precipitacion),col="red")
```

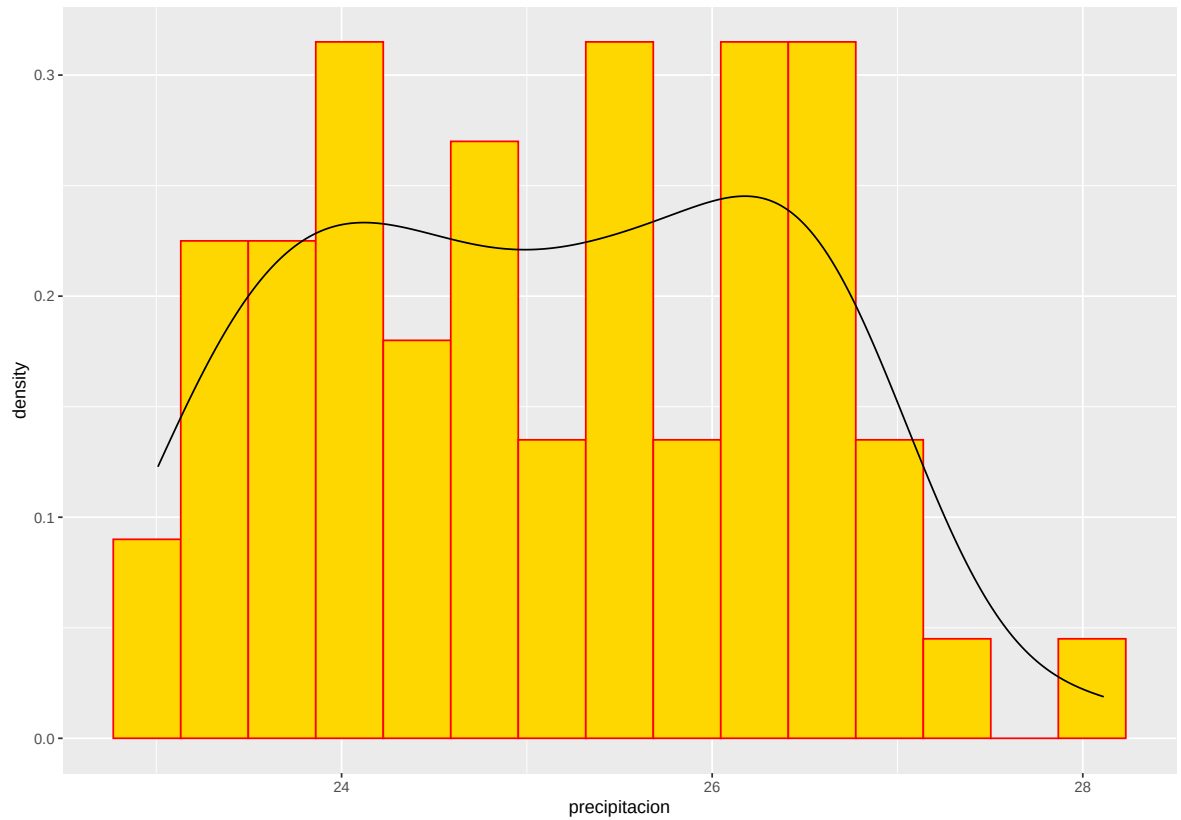


Pruebas de Bondad de Ajuste

¿Como es la distribución de la la precipitacion de Texcoco?

Histograma y grafica de densidad

```
library(ggplot2)
ggplot(precip, aes(precipitacion, y=..density.. ))+
geom_histogram(bins=15,color="red",fill="gold")+
geom_density(alpha=0.6)
```

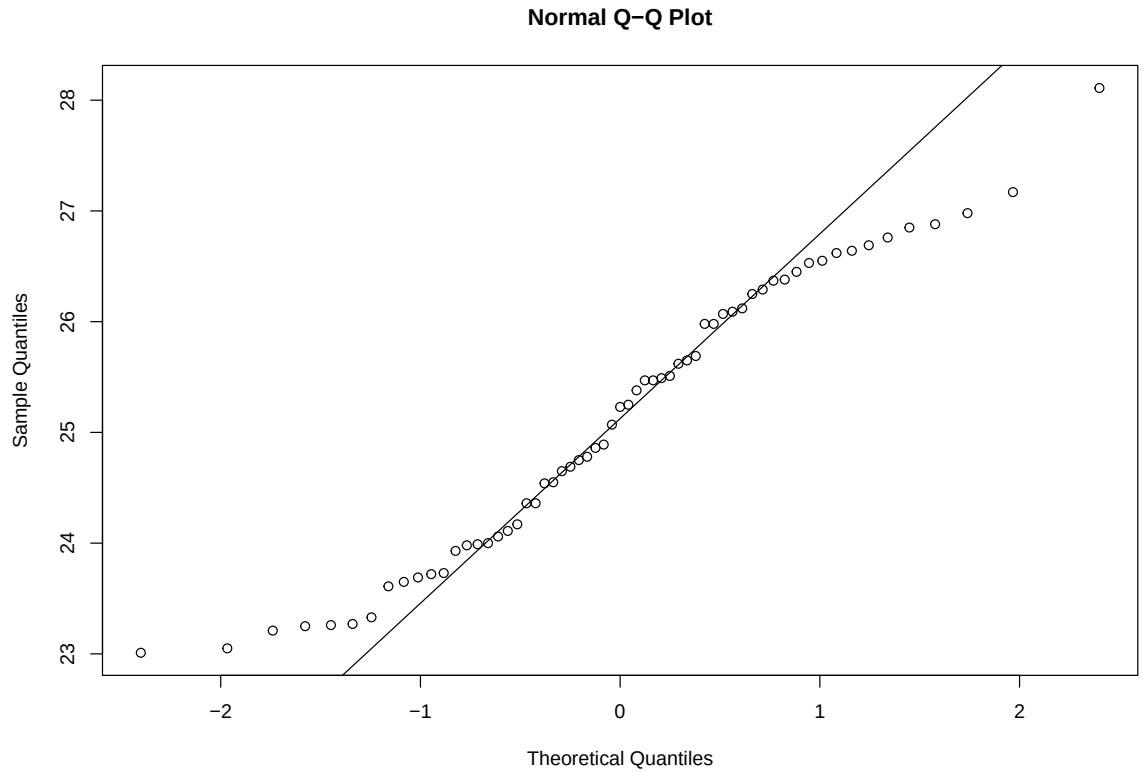


Pruebas de Bondad de Ajuste

¿Como es la distribución de la la precipitacion de Texcoco?

Grafica Q-Q

```
qqnorm(precipitacion)
qqline(precipitacion)
```

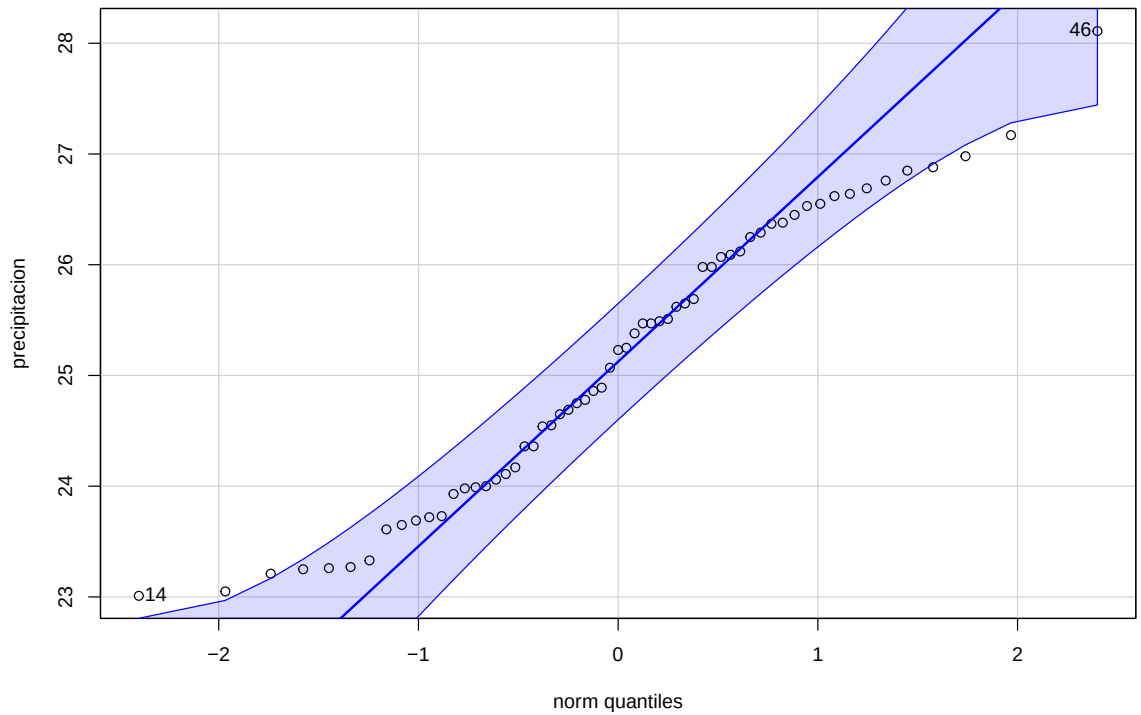


Pruebas de Bondad de Ajuste

¿Como es la distribución de la la precipitacion de Texcoco?

Grafica Q-Q

```
library(car)
qqPlot(precipitacion)
```



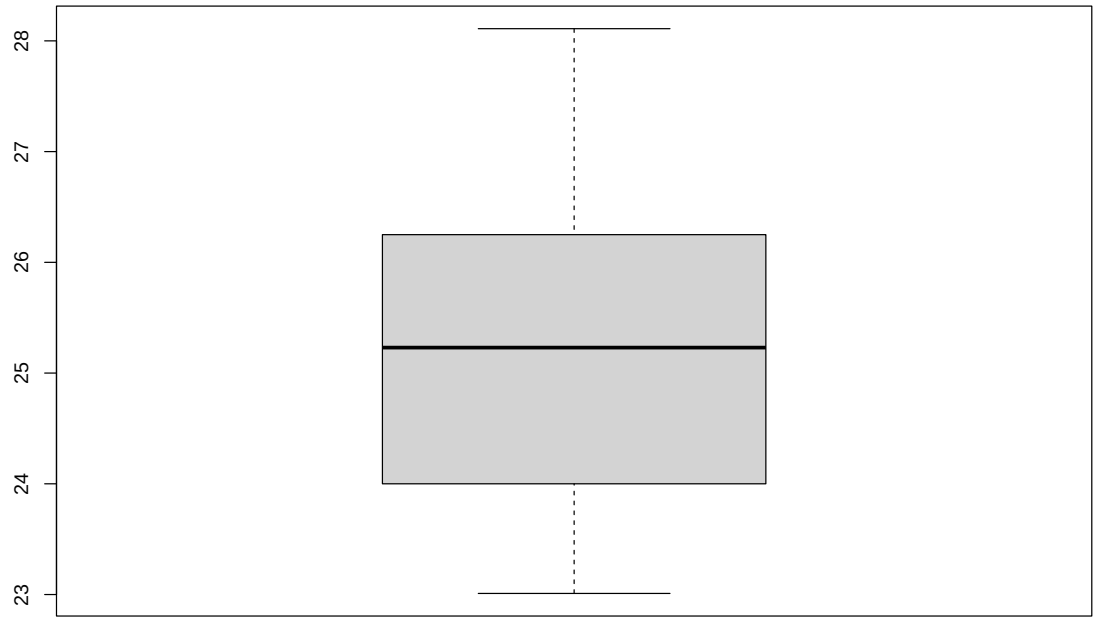

```
## [1] 46 14
```

Pruebas de Bondad de Ajuste

¿Como es la distribución de la la precipitacion de Texcoco?

Grafica de caja

```
boxplot(precip$Precipitacion)
```

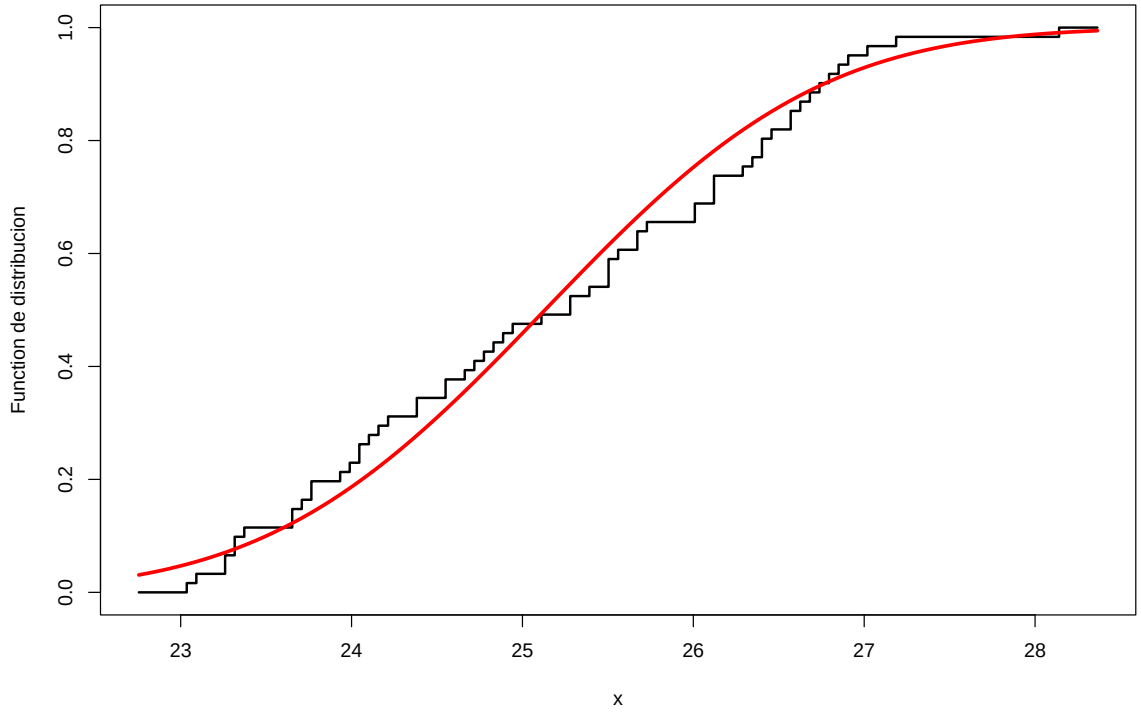


Pruebas de Bondad de Ajuste

¿Como es la distribución de la la precipitacion de Texcoco?

Función de distribución empírica

```
fn <- ecdf(precipitacion)
curve(ecdf(precipitacion)(x), xlim = extendrange(precipitacion),
type = 's', ylab = 'Function de distribucion', lwd = 2)
curve(pnorm(x, mean(precipitacion), sd(precipitacion)), add = TRUE,col="red",lwd=3)
```



Pruebas de Bondad de Ajuste

- **Objetivo :** Comprobar si una variable aleatoria *sigue* (o *se ajusta*) a un modelo propuesto (o no).
- **Juego de hipotesis:**

$$\begin{cases} H_0: & X \text{ se ajusta al modelo propuesto} F(x) \\ H_1: & \text{X no se ajusta a } F(x) \end{cases}$$

Regla:

"RECHAZAR H_0 si p-valor $< \alpha$" es un procedimiento de contraste con nivel de significación (menor o igual a) α.

Pruebas de Bondad de Ajuste

Prueba de Chi-cuadrada de Pearson:

Se trata de una prueba de hipotesis de bondad de ajuste:

$$\begin{cases} H_0 : F = F_0 \\ H_1 : F \neq F_0 \end{cases}$$

desarrollada inicialmente para variables categóricas. En el caso general, podemos pensar que los datos están agrupados en k clases: $C_1, \cdots, C_k$. Por ejemplo, si la variable es categórica o discreta, cada clase se puede corresponder con una modalidad. Si la variable es continua habrá que categorizarla en intervalos.

Pruebas de Bondad de Ajuste

Prueba de Chi-cuadrada de Pearson:

Clases	Discreta	Continua	H_0 simple	H_0 compuesta
C_1	x_1	$[L_0, L_1)$	p_1	$\hat{p}_1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_k	x_k	$[L_{k-1}, L_k)$	p_k	$\hat{p}_k$
			$\sum_i p_i = 1$	$\sum_i \hat{p}_i = 1$

Se realizará un contraste equivalente:

$$\begin{cases} H_0 : \text{Las probabilidades son correctas} \\ H_1 : \text{Las probabilidades no son correctas} \end{cases}$$

Pruebas de Bondad de Ajuste

Prueba de Chi-cuadrada de Pearson:

Si H_0 es cierta, la frecuencia relativa f_i de la clase C_i es una aproximación de la probabilidad teórica, $f_i \approx p_i$. Equivalentemente, las frecuencias observadas $n_i = n \cdot f_i$ deberían ser próximas a las esperadas $e_i = n \cdot p_i$ bajo H_0, sugiriendo el estadístico del contraste (Pearson, 1900):

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i} \underset{aprox.}{\sim} \chi^2_{k-r-1}, \text{ si } H_0 \text{ cierta}$$

siendo k el número de clases y r el número de parámetros estimados (para aproximar las probabilidades bajo H_0).

Pruebas de Bondad de Ajuste

Prueba de Chi-cuadrada de Pearson:

Clases	n_i observadas	p_i bajo H_0	e_i bajo H_0	$\frac{(n_i-e_i)^2}{e_i}$
C_1	n_1	p_1	e_1	$\frac{(n_1-e_1)^2}{e_1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_k	n_k	p_k	e_k	$\frac{(n_k-e_k)^2}{e_k}$
Total	$\sum_i n_i = n$	$\sum_i p_i = 1$	$\sum_i e_i = n$	$\chi^2 = \sum_{i=1}^{k} \frac{(n_i-e_i)^2}{e_i}$

Cuando H_0 es cierta el estadístico tiende a tomar valores pequeños y grandes cuando es falsa.

Pruebas de Bondad de Ajuste

Prueba de Chi-cuadrada de Pearson:

Por tanto se rechaza H_0, para un nivel de significación α, si:

$$\sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i} \geq \chi^2_{k-r-1,1-\alpha}$$

Si realizamos el contraste a partir del p-valor o nivel crítico:

$$p = P\left(\chi^2_{k-r-1} \geq \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i}\right)$$

rechazaremos H_0 si $p \leq \alpha$ (y cuanto menor sea se rechazará con mayor seguridad) y aceptaremos H_0 si $p > \alpha$ (con mayor seguridad cuanto mayor sea).

Pruebas de Bondad de Ajuste

Prueba de Chi-cuadrada de Pearson:

Se aplica a variables discretas o variables continuas agrupadas. el modelo propuesto $F(x)$ se expresa como tabla de probabilidades:

x_i	x_1	$\cdots$	x_k
p_i	p_1	$\cdots$	p_k

y la muestra como tabla de frecuencias

x_i	x_1	$\cdots$	x_k
n_i	n_1	$\cdots$	n_k

- Se puede aplicar a variable continua si se corta antes en intervalos y se calcula la probabilidad de cada intervalo.

Pruebas de Bondad de Ajuste

Prueba de Chi-cuadrada de Pearson:

Ejemplo: Poisson con Chi-Cuadrada

Tabla de datos de conteos de insectos

```
### Ho: Datos son Poisson vs Ha: No son Poisson
data <- data.frame(x=c(0:4), frec=c(2962,382,47,25,4))
names(data) <- c('x', 'frecuencia' )
pander::pander(data)
```

x	frecuencia
0	2962
1	382
2	47
3	25
4	4

```
valores <- rep(data$x, data$frec)
```

Pruebas de Bondad de Ajuste

Prueba de Chi-cuadrada de Pearson:

Ejemplo: Poisson con Chi-Cuadrada

```
x = c(0:4)
observed = c(2962,382,47,25,4)
total = sum(observed)
lambda = t(x)%*%observed/total
expected = total*dpois(x,lambda =lambda )
chi.squared = (observed - expected)^2/expected
expected = round(expected, 2)
chi.squared = round(chi.squared, 3)
goodness.of.fit = cbind(x, observed, expected)
colnames(goodness.of.fit) = c('x', 'Observed Counts', 'Expected Counts')
goodness.of.fit
```

```
##      x Observed Counts Expected Counts
## [1,] 0            2962         2897.51
## [2,] 1             382          480.38
## [3,] 2              47           39.82
## [4,] 3              25            2.20
## [5,] 4               4            0.09
```

```
chi.squared.statistic = sum(chi.squared)
p.value = pchisq(chi.squared.statistic, length(observed)-2, lower.tail = F)
lambda
```

```
##           [,1]
## [1,] 0.1657895
```

```
chi.squared.statistic
```

Ejemplo Poisson con Chi-Cuadrada

```
lambda
```

```
##             [,1]
## [1,] 0.1657895
```

```
chi.squared.statistic
```

```
## [1] 426.599
```

```
p.value
```

```
## [1] 3.829734e-92
```

Pruebas de Bondad de Ajuste

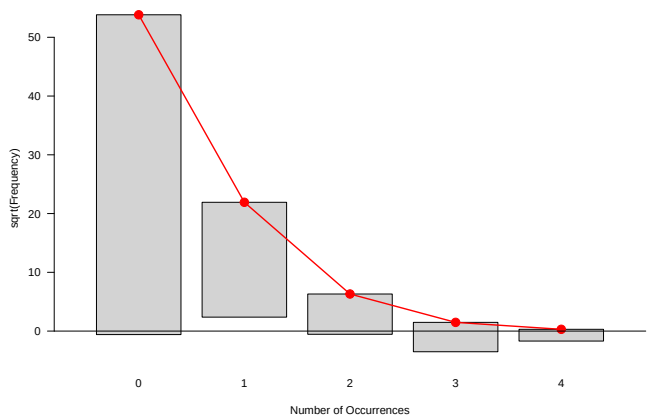
Prueba de Chi-cuadrada de Pearson:

Ejemplo: Poisson con Chi-Cuadrada

```
library("data.table")
counts <- data.frame(Freq=c(2962,382,47,25,4),x=c(0:4))
setDT(counts)
library(vcd)
HK.fit <- goodfit(counts)
summary(HK.fit)
```

```
##
##   Goodness-of-fit test for poisson distribution
##
##                       X^2 df     P(> X^2)
## Likelihood Ratio 122.6744  3 2.048368e-26
```

```
plot(HK.fit)
```



Pruebas de Bondad de Ajuste

Prueba de Chi-cuadrada de Pearson:

Ejemplo: Poisson con Chi-Cuadrada

```
library(fitdistrplus)
ajuste=fitdist(valores, "pois")
pander::pander(gofstat(ajuste))
```

- **chisq**: *48.84*
- **chisqbreaks**: *0* and *1*
- **chisqpvalue**: *2.772e-12*
- **chisqdf**: *1*
- **chisqtable**:

	obscounts	theocounts
<= 0	2962	2898
<= 1	382	480.4
> 1	76	42.12

- **aic**:

1-mle-pois
3354

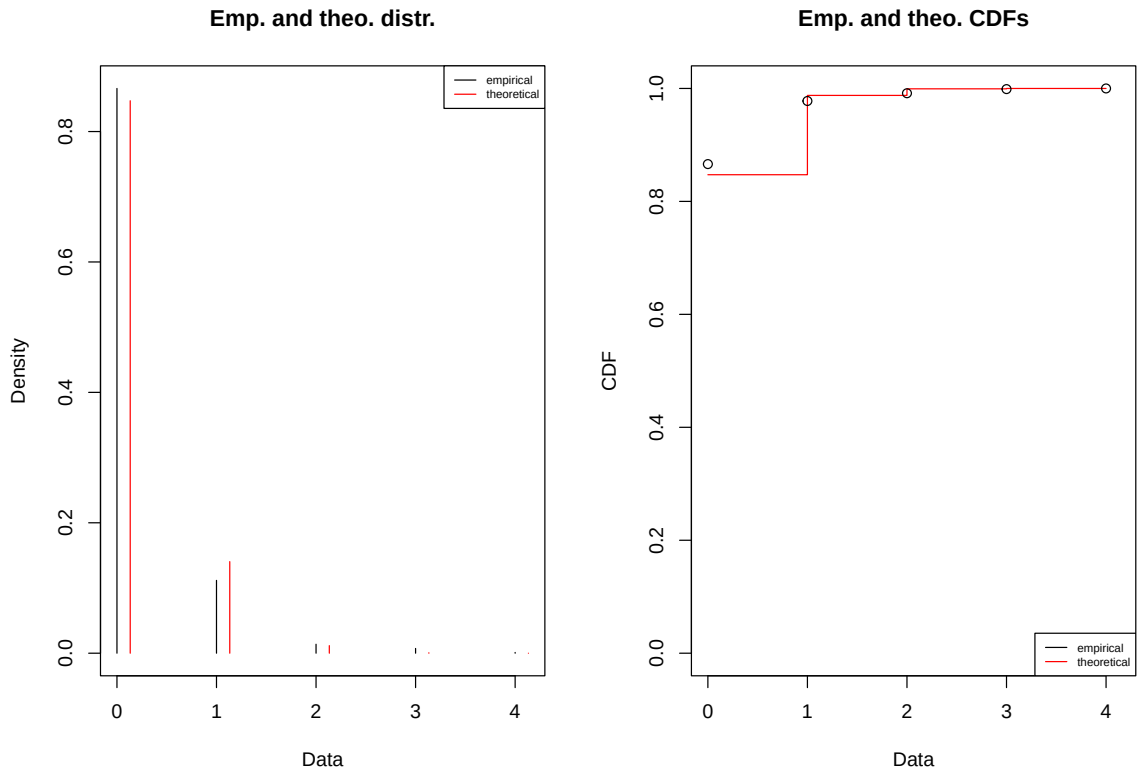
Ejemplo Poisson con Chi-Cuadrada

```
library(fitdistrplus)
ajuste=fitdist(valores, "pois")
ajuste
```

```
## Fitting of the distribution ' pois ' by maximum likelihood
## Parameters:
##         estimate Std. Error
## lambda 0.1657895 0.00696225
```

Ejemplo Poisson con Chi-Cuadrada

```
plot(ajuste)
```



Pruebas de Bondad de Ajuste

Prueba de Chi-cuadrada de Pearson:

Ejemplo Normal con Chi-Cuadrada

```
rend=c(1903,1935, 1910, 2496, 2108, 1961, 2060, 1444,
       1612, 1316, 1511, 2009, 1915, 2011, 2463, 2180,
       1925, 2122, 1482, 1542, 1443, 1535)
library(nortest)
pander::pander(pearson.test(rend))
```

Table 8: Pearson chi-square normality test: `rend`

Test statistic	P value
5.364	0.252

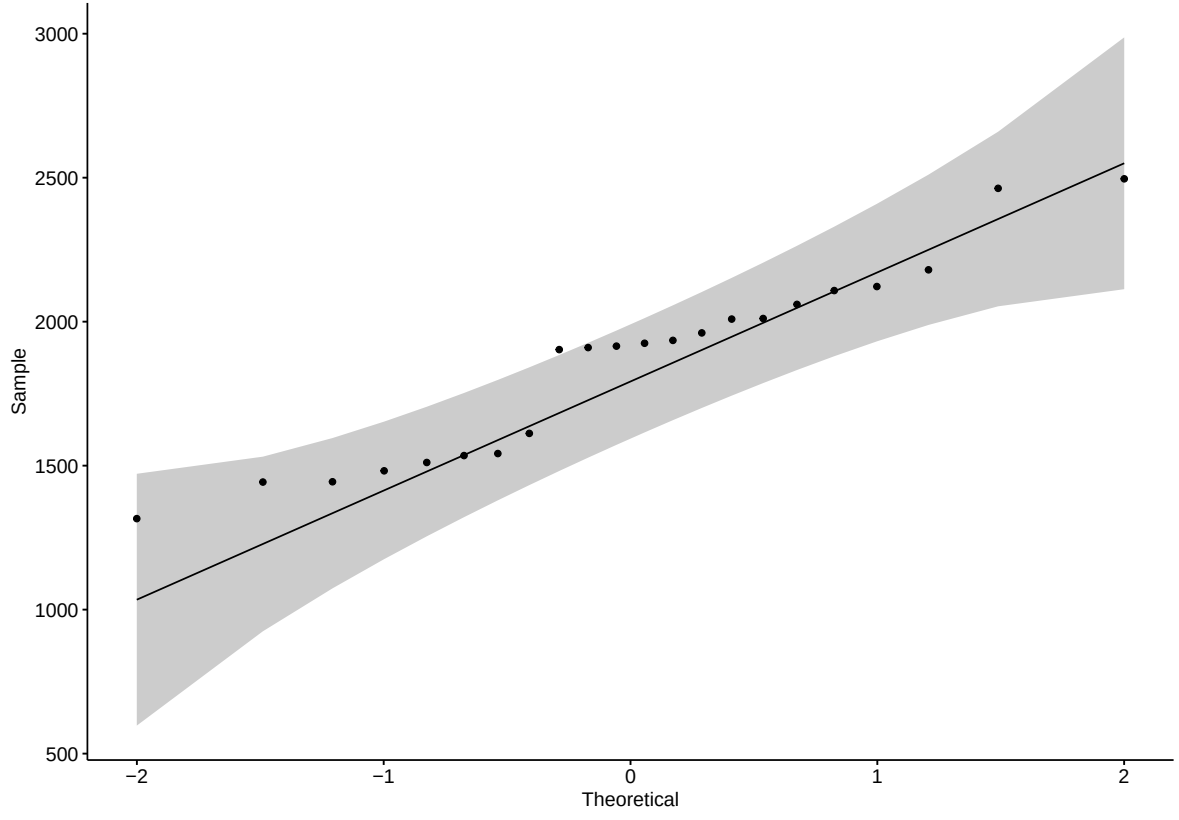
Pruebas de Bondad de Ajuste

Prueba de Chi-cuadrada de Pearson:

Ejemplo Normal con Chi-Cuadrada

Grafica de normalidad

```
library(ggpubr)
ggqqplot(rend)
```



- Prueba de Chi-cuadrada de Pearson:

```
- **En R:** función "chisq.test(x,p)" on:
```

- `x`: frecuencias de la muestra. Es decir `c(`n_1`,`n_2`,...`n_k`)`.
- `p`: probabilidades del modelo. Es decir `c(`p_1`,`p_2`,...`p_k`)`.
- Devuelve una lista con valores, entre los que destaca el `p.value`, base para la decisión del contraste.
- Devuelve un `Warning` si no se cumplen las condiciones teóricas de buena aproximación de la distribución ji-cuadrado. Hay que advertirlo si ocurre.

Pruebas de Bondad de Ajuste

Prueba de Kolmogorov-Smirnov:

Se trata de una prueba de bondad de ajuste para distribuciones continuas. Se basa en comparar la función de distribución F_0 bajo H_0 con la función de distribución empírica F_n:

$$\begin{aligned} D_n &= \sup_x |F_n(x) - F_0(x)|, \\ &= \max_{1\leq i\leq n} \left\{ |F_n(X_{(i)}) - F_0(X_{(i)})|, |F_n(X_{(i-1)}) - F_0(X_{(i)})| \right\} \end{aligned}$$

Teniendo en cuenta que $F_n\left(X_{(i)}\right) = \frac{i}{n}$:

$$\begin{aligned} D_n &= \max_{1\leq i\leq n} \left\{ \frac{i}{n} - F_0(X_{(i)}), F_0(X_{(i)}) - \frac{i-1}{n} \right\} \\ &= \max_{1\leq i\leq n} \left\{ D_{n,i}^+, D_{n,i}^- \right\} \end{aligned}$$

Pruebas de Bondad de Ajuste

Prueba de Kolmogorov-Smirnov:

Si H_0 es simple y F_0 es continua, la distribución del estadístico D_n bajo H_0 no depende F_0 (es de distribución libre). Esta distribución está tabulada (para tamaños muestrales grandes se utiliza la aproximación asintótica). Se rechaza H_0 si el valor observado d del estadístico es significativamente grande:

$$p = P\left(D_n \geq d\right) \leq \alpha.$$

Este método está implementado en la función `ks.test()` del paquete base de R:

```
ks.test(x, y, ...)
```

donde `x` es un vector que contiene los datos, `y` es una función de distribución (o una cadena de texto que la especifica; también puede ser otro vector de datos para el contraste de dos muestras) y `...` representa los parámetros de la distribución.

Pruebas de Bondad de Ajuste

Prueba de Kolmogorov-Smirnov:

Para contrastar $H_0 : F = \mathcal{N}(20, 5^2)$ podríamos emplear:

```
ks.test(datos, pnorm, mean = 20, sd = 5) # One-sample
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  datos
## D = 0.13239, p-value = 0.4688
## alternative hypothesis: two-sided
```

Pruebas de Bondad de Ajuste

Prueba de Kolmogorov-Smirnov:

Si H_0 es compuesta, el procedimiento habitual es estimar los parámetros desconocidos por máxima verosimilitud y emplear $\hat{F}_0$ en lugar de F_0. Para este caso para contrastar normalidad se desarrolló el test de Lilliefors, implementado en la función `lillie.test()` del paquete `nortest`. Por ejemplo:

```
ks.test(datos, pnorm, mean(datos), sd(datos)) # One-sample Kolmogorov-Smirnov test
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  datos
## D = 0.097809, p-value = 0.8277
## alternative hypothesis: two-sided
```

```
library(nortest)
lillie.test(datos)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  datos
## D = 0.097809, p-value = 0.4162
```

Pruebas de Bondad de Ajuste

Prueba de Kolmogorov-Smirnov:

```
library(nortest)
lillie.test(datos)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  datos
## D = 0.097809, p-value = 0.4162
```

Pruebas de Bondad de Ajuste

Prueba de Kolmogorov-Smirnov:

El estadístico de Kolmogorov–Smirnov, se define como la distancia vertical máxima entre una función de distribución empírica y una función de distribución acumulada teórica. La ventaja principal de este estadístico es que es sensible a diferencias tanto en la localización como en la forma de la función de distribución acumulada.

Una vez calculada la distancia de Kolmogorov–Smirnov, hay que determinar si el valor de esta distancia es suficientemente grande (p-value). Hipótesis:$H_0 : F(x)$ es normal para toda x $H_0 : F(x)$ no es normal

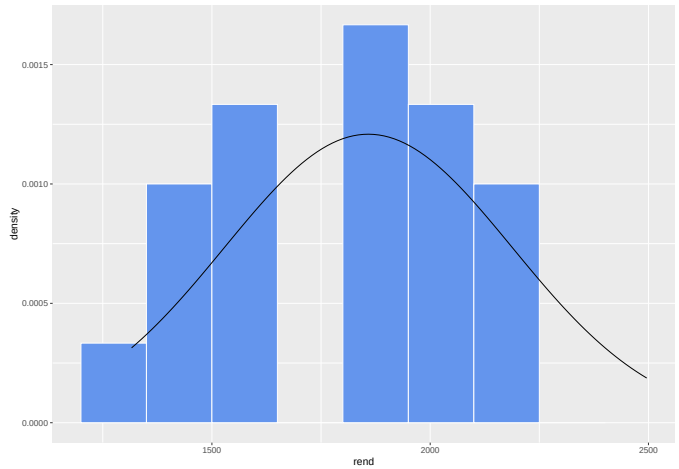
Pruebas de Bondad de Ajuste

Prueba de Kolmogorov-Smirnov:

normalidad

Grafica de normalidad

```
library(ggplot2)
rendi=data.frame(rend)
df <- rendi
ggplot(df, aes(x = rend)) +
  geom_histogram(aes(y =..density..),
      breaks = seq(1200, 2500, by = 150),
        colour = "white",
    fill = "cornflowerblue", size = 0.1)  +
  stat_function(fun = dnorm,
    args = list(mean = mean(df$rend),
      sd = sd(df$rend)))
```



Pruebas de Bondad de Ajuste

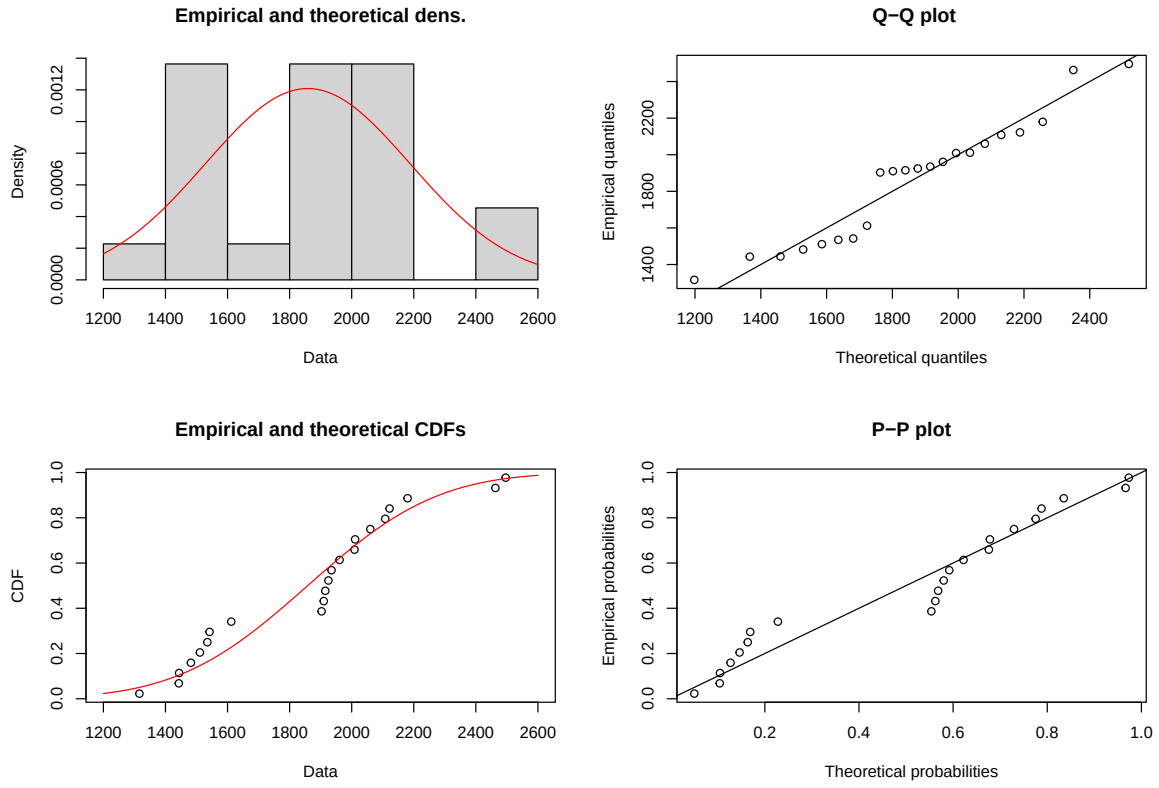
Prueba de Kolmogorov-Smirnov:

Pruebas de Bondad de Ajuste

Prueba de Kolmogorov-Smirnov:

normalidad

```
library(fitdistrplus)
plotdist(rend, "norm",para=list(mean=mean(rend),
                                sd=sd(rend)))
```



Pruebas de Bondad de Ajuste

Prueba de Kolmogorov-Smirnov:

Prueba de Kolmogorov en R

Grafica de normalidad

```
ks.test(x = rend,"pnorm", mean(rend), sd(rend))
```

```
##
##  Exact one-sample Kolmogorov-Smirnov test
##
## data:  rend
## D = 0.19019, p-value = 0.3578
## alternative hypothesis: two-sided
```

Pruebas de Bondad de Ajuste

Prueba de Kolmogorov-Smirnov:

el presunto modelo $\mathcal{M}$ es de variable (numérica) continua y se conoce la función de distribución $F_0(x)$

- **En R:** función `ks.test(x,y,...)` donde:
- `x`: muestra de datos
- `y`: nombre entrecomillado de la función F_0 (tal qual esté programada en R, por ejemplo `"punif"`, `"pnorm"`, `"pexp"`, etc.).
- `...`: parámetros adicionales de la función F_0 (por ejemplo, `mean` y `sd` si se trata de la normal, o `rate` si se trata de la exponencial, etc.)

Pruebas:

- Prueba de Kolmogorov-Smirnov:

- Devuelve una lista con valores, entre los que destaca el `p.value`, base para la decisión del contraste.
- Devuelve un `Warning` si la muestra contiene datos repetidos. Hay que advertirlo si ocurre.
- **Contraste de Kolmogorov-Smirnov-Lilliefors:** el presunto modelo $\mathcal{M}$ es la família normal completa
- **En R:** función `lillie.test(x)` (**atención!** cargar el *package* `nortest`, y instalarlo antes si no está disponible):
- `x`: muestra de datos
- Devuelve una lista con valores, entre los que destaca el `p.value`, base para la decisión del contraste. - Devuelve un `Warning` si la muestra contiene datos repetidos. Hay que advertirlo si ocurre.

Prueba de Shapiro-Wilk

Este test se emplea para contrastar normalidad cuando el tamaño de la muestra es menor de 50. Para muestras grandes es equivalente al test de kolmogorov-Smirnov.

Prueba de Shapiro-Wilk

```
shapiro.test(x = rend)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rend
## W = 0.93573, p-value = 0.1615
```

Prueba de normalidad de Jarque-Bera

Jarque-Bera no requiere estimaciones de los parámetros que caracterizan la normal. El estadístico de Jarque-Bera cuantifica que tanto se desvían los coeficientes de asimetría y curtosis de los esperados en una distribución normal. Puede calcularse mediante la función jarque.bera.test() del paquete tseries.

Prueba de normalidad de Jarque-Bera

```
   library("tseries")
jarque.bera.test(x = rend)
```

```
##
##  Jarque Bera Test
##
## data:  rend
## X-squared = 0.62282, df = 2, p-value = 0.7324
```

Prueba de aleatoriedad o independencia

- **Objetivo :** Comprobar si un proceso de muestreo "produce" datos independentes entre sí, o no.
- **Contraste:**
$$\begin{cases} H_0: & X \text{ produce datos independientes entre sí} \\ H_1: & \text{no } H_0 \end{cases}$$
- **En R:** función `runs.test(x)` (**atención!** cargar el *package* `tseries`, y instalarlo antes si no está disponible)::
  - `x`: vector de signos (o con dos únicas categoríes o números). Si no está disponible, hay que fabricarlo a partir de la muestra original (bien comparando cada dato con la mediana de todos ellos, o bien comparando cada dato con el anterior, si sube o baja)
  - Devuelve una lista con valores, entre los que destaca el `p.value`, base para la decisión del contraste.
  - **¡ATENCIÓN!:** `runs.test()`usa siempre la aproximación normal, incluso cuando no es aceptable (muestras con más de 20 signos de cada tipo). Tendrás que alertar si ocurre ese caso.**

Prueba de aleatoriedad o independencia

```
library("snpar")
#devtools::install_github("debinqiu/snpar")
runs.test(x = rend)
```

```
##
##  Approximate runs test
##
## data:  rend
## Runs = 9, p-value = 0.1899
## alternative hypothesis: two.sided
```

Ejemplo Gastos Maximos de rio Tempoal

Cuenca del Panuco: Sistema del Río Tempoal

El sistema del Río Tempoal, de la Región Hidrológica Núm. 26 (Pánuco) en cinco estaciones hidrométricas: Tempoal, El Cardón, Platón Sánchez, Los Hules y Terrerillos.

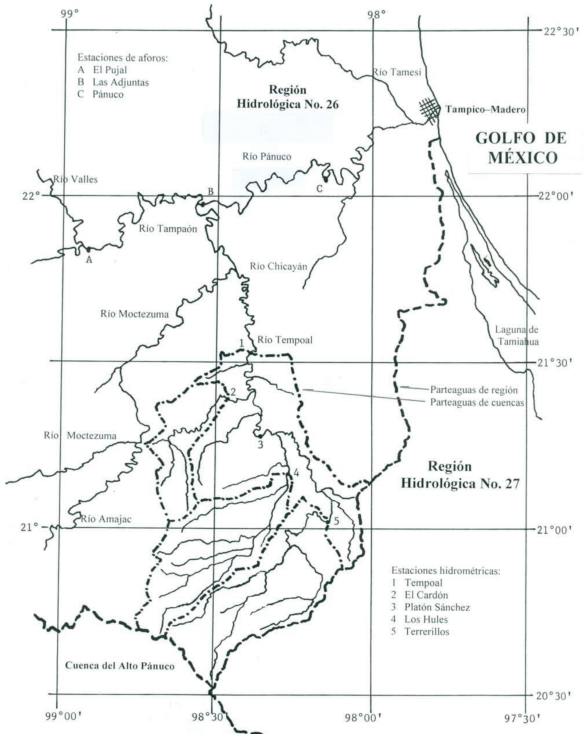


Figure 1: Rio Tempoal.

Gastos máximos anuales (m^3/s) en las estaciones hidrométricas del Río Tempoal en el periodo común de 1960 a 2002.

Referencia de datos y figura:

> *Campos-Aranda, Daniel Francisco. (2015). Estimación simultánea de datos hidrológicos anuales faltantes en múltiples sitios. Ingeniería, investigación y tecnología, 16(2), 295-306.*

Datos

Num	periodo	Tempoal	Cardon	P_Sanchez	Hules	Terrerillos
1	1960	1277.0	1080.0	NA	452.6	314.0
2	1961	852.9	303.5	NA	434.5	525.0
3	1962	739.2	262.0	NA	457.5	565.9
4	1963	1800.0	481.0	NA	947.4	895.9
5	1964	748.0	188.6	NA	258.0	397.1
6	1965	792.7	338.0	NA	414.9	659.4
7	1966	1778.0	287.0	NA	742.2	1121.7
8	1967	2245.0	854.2	NA	1009.4	1153.0
9	1968	1145.0	476.0	NA	1096.0	611.2
10	1969	1948.0	555.8	NA	825.0	2224.2
11	1970	1418.0	560.0	NA	800.0	1420.0
12	1971	1630.0	720.4	NA	1064.0	1488.5
13	1972	989.0	320.0	NA	1110.0	529.0
14	1973	1668.0	392.0	NA	749.0	1740.0
15	1974	4950.0	1198.3	NA	1950.0	3187.8
16	1975	4040.0	1204.2	NA	2470.0	2085.0
17	1976	1275.0	419.7	NA	937.7	1000.5
18	1977	514.0	179.1	NA	559.0	291.2
19	1978	3725.0	1390.0	2898.0	2874.0	2152.3
20	1979	1655.9	667.0	1040.0	1082.0	659.1
21	1980	1162.0	357.0	976.0	583.2	994.1
22	1981	2020.0	765.2	1940.0	1650.3	NA
23	1982	539.6	182.3	589.8	340.0	491.4
24	1983	868.0	269.8	827.3	544.0	768.4
25	1984	4030.0	572.0	4530.0	2834.9	2981.0
26	1985	1882.0	457.0	1608.0	938.4	1487.7
27	1986	476.0	192.0	462.0	308.0	434.0
28	1987	1765.0	346.8	1773.0	1440.0	2635.0
29	1988	3265.0	356.0	3653.0	4350.0	3710.0
30	1989	649.0	306.0	653.0	644.0	2100.0
31	1990	1611.0	306.0	4115.0	NA	702.0
32	1991	3532.0	1248.0	1916.0	NA	2860.0
33	1992	2291.0	790.0	1494.9	762.8	1607.5
34	1993	6120.0	865.5	4380.0	1684.1	3422.5
35	1994	1133.0	412.0	1153.8	723.8	1237.9
36	1995	741.9	412.2	537.0	568.0	531.0
37	1996	683.0	218.0	758.0	804.0	507.6
38	1997	905.0	348.2	1217.5	428.4	362.5
39	1998	1266.9	NA	1259.3	260.9	1605.9
40	1999	2693.7	602.9	2776.6	630.9	3328.3
41	2000	641.2	NA	580.4	84.9	753.4
42	2001	1847.9	498.3	1201.3	278.5	1512.2
43	2002	926.4	134.0	774.8	496.7	822.2

Grafica de gastos maximos

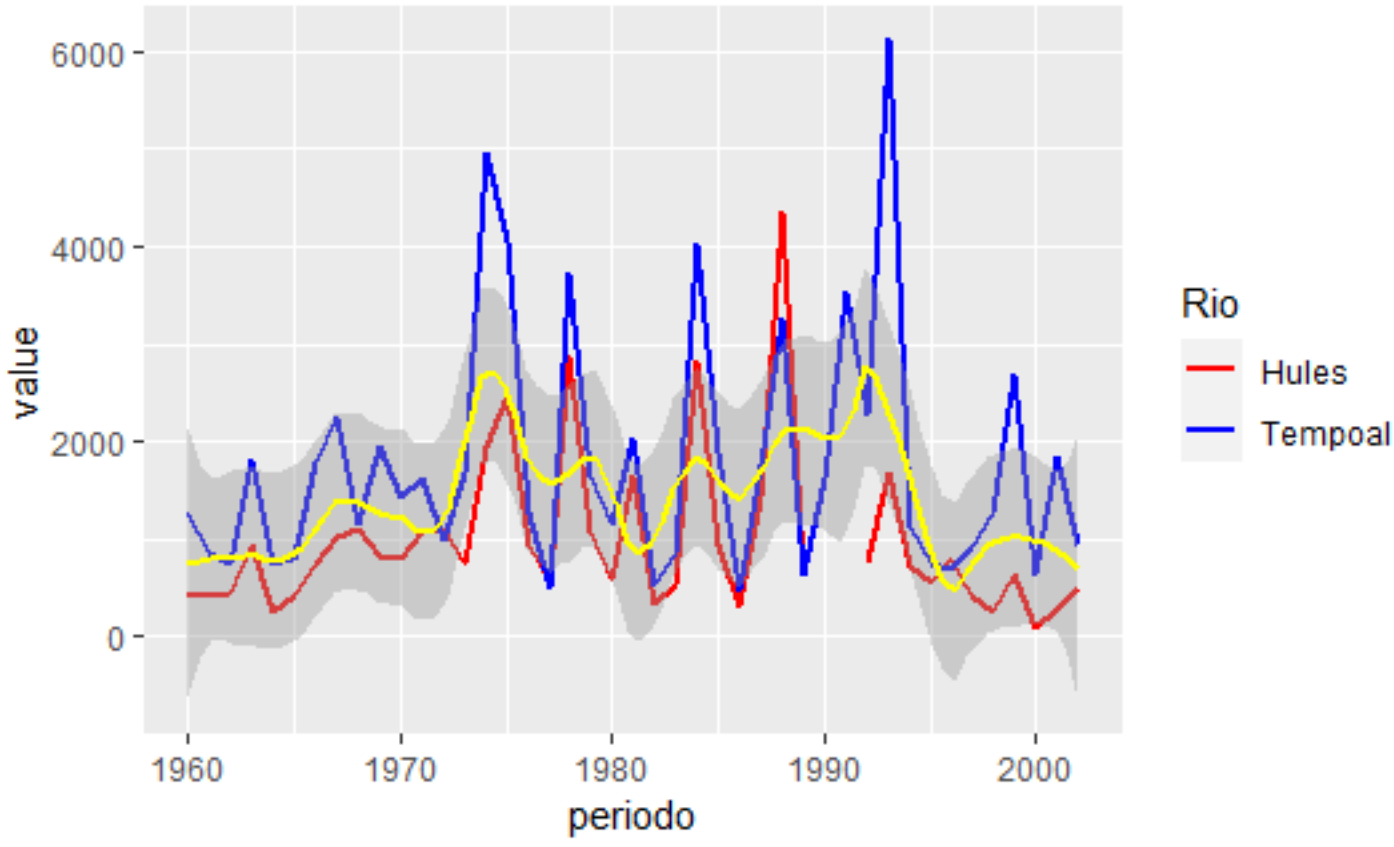
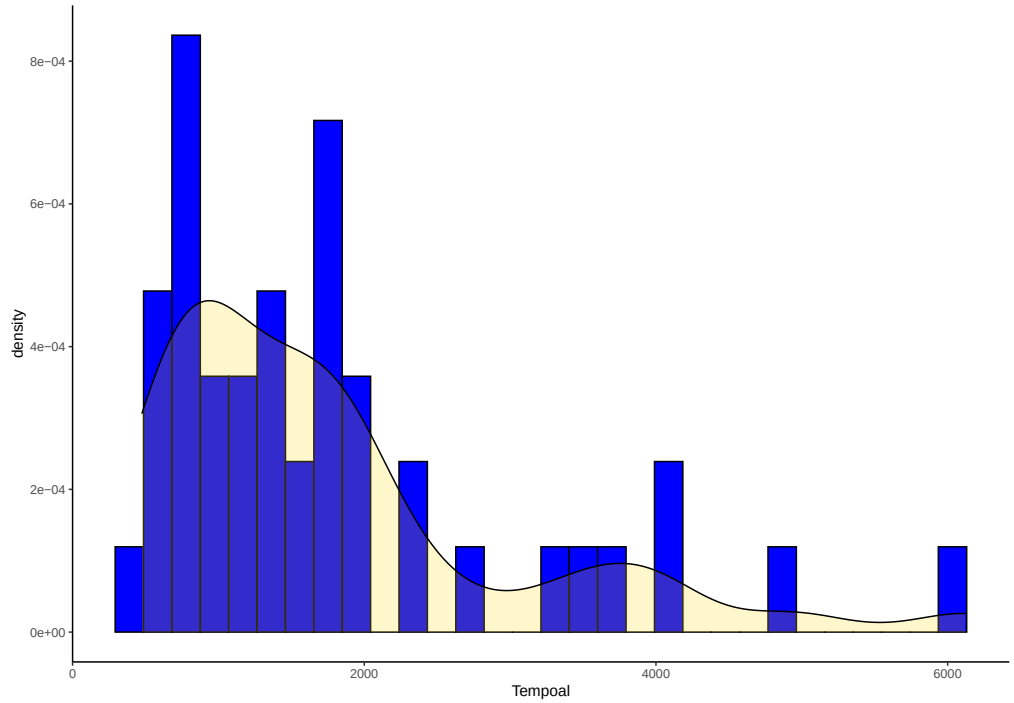


Figure 2: Rio Tempoal.

Histograma para gastos maximos de Tempoal

```
library(ggplot2)
theme_set(  theme_classic() +
theme(legend.position = "top"))
ggplot(gastos, aes(x = Tempoal))+
geom_histogram(aes(y = stat(density)), colour="black", fill="blue") +
geom_density(alpha = 0.2, fill = "gold")
```



Distribucion Gumbel para maximos

Distribucion Gumbel para modelar maximos.

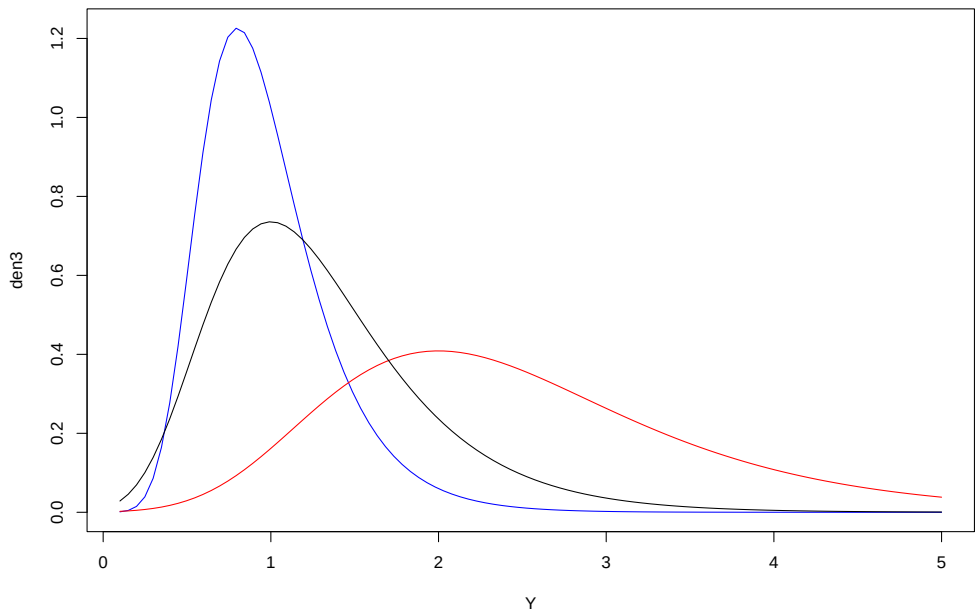
La distribución de Gumbel (Gumbel, 1958) se ha propuesto y utilizado para describir fenómenos extremos en diversas disciplinas, incluida la hidrología. Se ha encontrado que la distribución de Gumbel proporciona una excelente bondad de ajuste a variables extremas en muchos casos.

$$f(x) = \frac{1}{\beta} e^{-\frac{x-\mu}{\beta}} e^{-e^{-\frac{x-\mu}{\beta}}}, -\infty < x < \infty, \beta > 0$$

donde μ es el parámetro de ubicación y β es el parámetro de escala. El caso en el que $\mu = 0$ y $\beta = 1$ se denomina distribución *Gumbel estándar*.

Graficas de la densidad Gumbel

```
library(ExtDist)
Y=seq(.1,5,length=100)
den <- dGumbel(Y, location = 1, scale= .5)
den2=dGumbel(Y, location = 2, scale= .9)
den3=dGumbel(Y, location = .8, scale= .3)
plot(Y, den3,type="l", col="blue")
lines(Y, den2,type="l", col="red")
lines(Y, den,type="l", col="black")
```

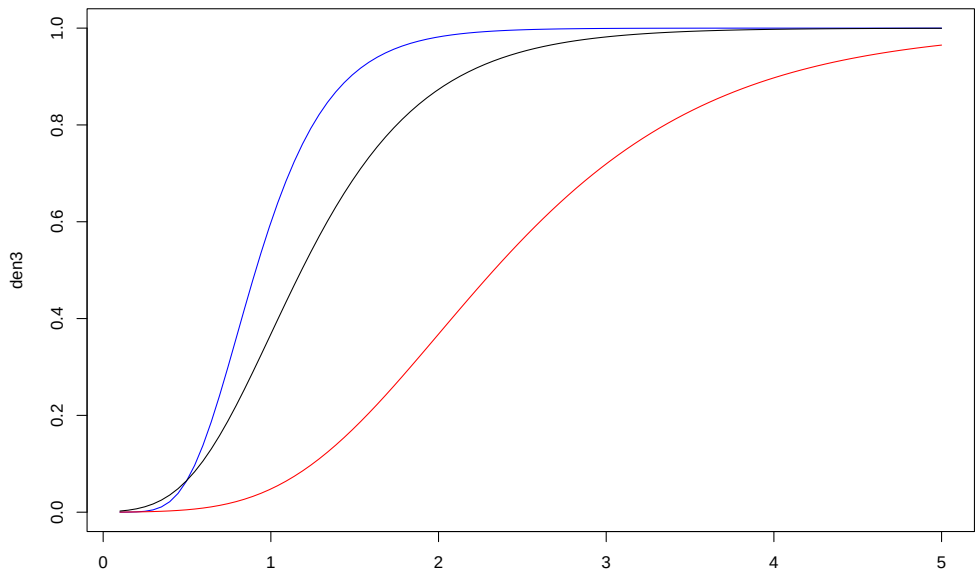


La funcion acumulativa de probalidades de la Gumbel

$$F_X(x) = 1 - e^{-e^{-\frac{x-\mu}{\beta}}}, -\infty < x < \infty, \beta > 0$$

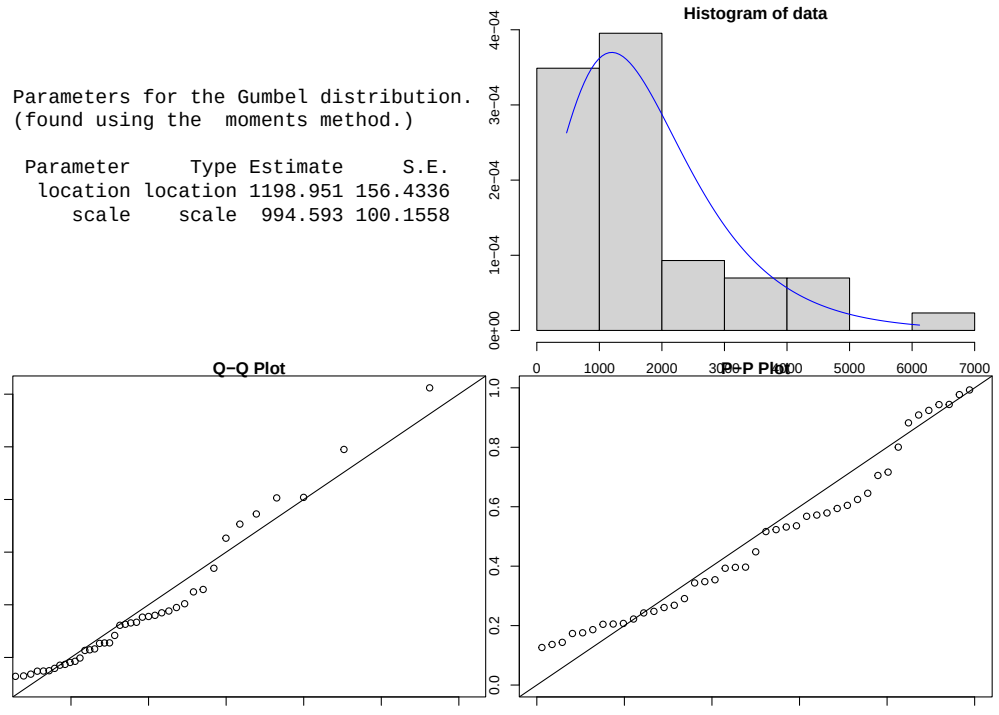
Graficas de la funcion acumulativa Gumbel

```
library(ExtDist)
Y=seq(.1,5,length=100)
den <- pGumbel(Y, location = 1, scale= .5)
den2=pGumbel(Y, location = 2, scale= .9)
den3=pGumbel(Y, location = .8, scale= .3)
plot(Y, den3,type="l", col="blue")
lines(Y, den2,type="l", col="red")
lines(Y, den,type="l", col="black")
```



Ajuste de datos de Tempoal a Gumbel

```
library(ExtDist)
gasto_tempoal=gastos[,3]
est.par <- eGumbel(gasto_tempoal, method="moments")
par(mar=c(1,1,1,1))
plot(est.par)
```



parametros estimados de localidad y escala

```
est.par[attributes(est.par)$par.type=="location"]
```

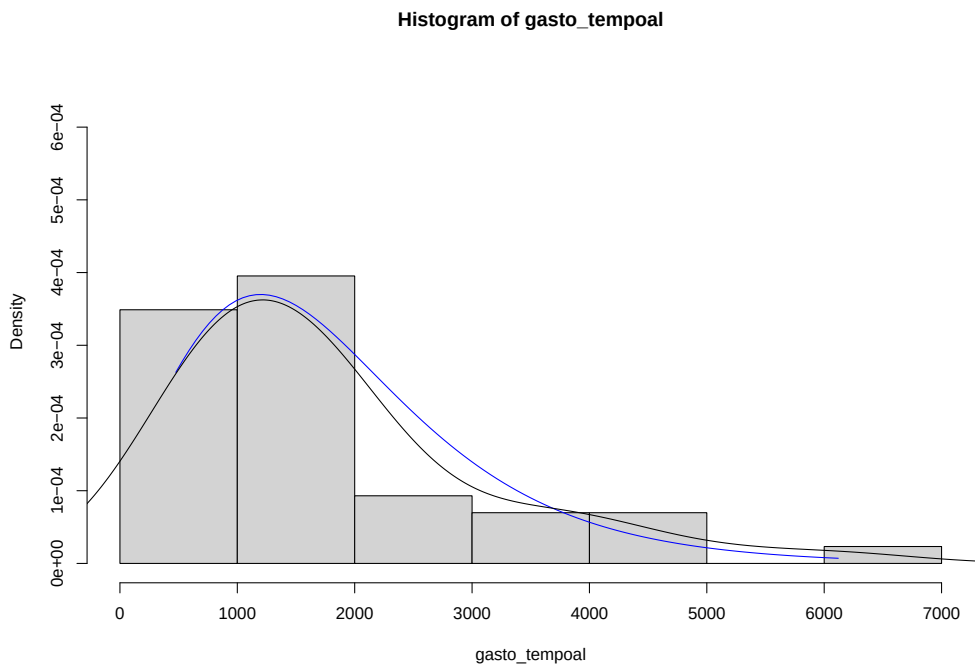
```
## $location
## [1] 1198.951
```

```
est.par[attributes(est.par)$par.type=="scale"]
```

```
## $scale
## [1] 994.593
```

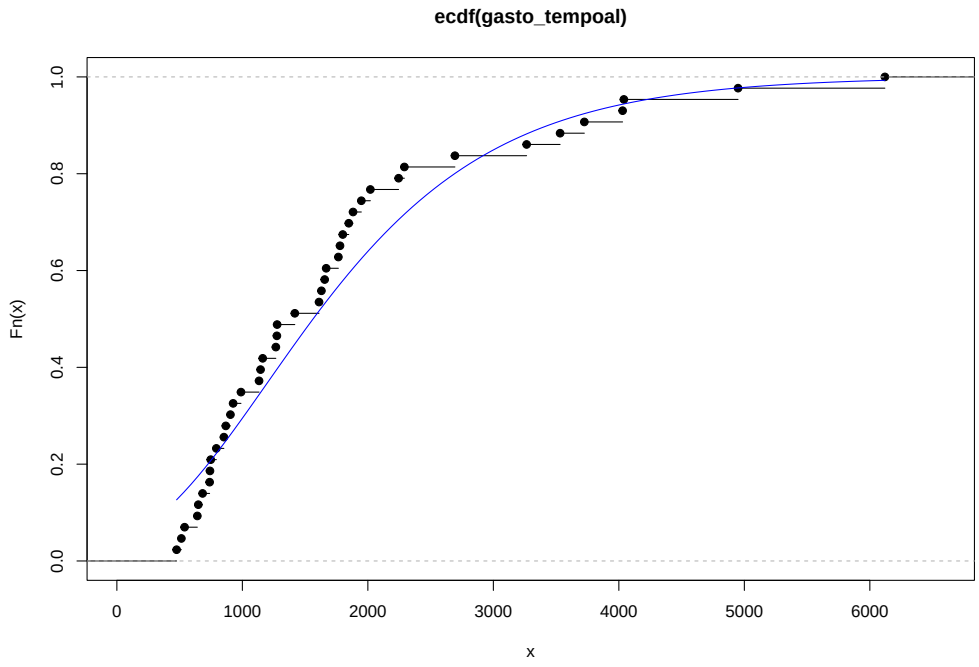
Datos ajustados a Gumbel y estimador de densidad

```
#  Fitted density curve and histogram
den.x <- seq(min(gasto_tempoal),max(gasto_tempoal),length=100)
den.y <- dGumbel(den.x, location = est.par$location, scale= est.par$scale)
hist(gasto_tempoal, breaks=8, probability=TRUE, ylim = c(0,1.8*max(den.y)))
lines(den.x, den.y, col="blue")
lines(density(gasto_tempoal,bw = 700))
```



Datos ajustados a Gumbel funcion acumulativa

```
den.x <- seq(min(gasto_tempoal),max(gasto_tempoal),length=100)
p.y <- pGumbel(den.x, location = est.par$location, scale= est.par$scale)
plot(ecdf(gasto_tempoal))
lines(den.x, p.y, col="blue")
```



Prueba de Kolmogorov Smirnov

Queremos probar si los datos de gastos maximos de tempoal se ajustan a la Gumbel

$$H_0 : F(x) \quad Gumbel$$

vs

$$H_0 : F(x) \quad No\ es\ Gumbel$$

```
#remotes::install_github("NVE/fitdistrib")
library(fitdistrib)
estimate=gumbel_mle(gasto_tempoal)
mu=estimate$mle[1]
mu
```

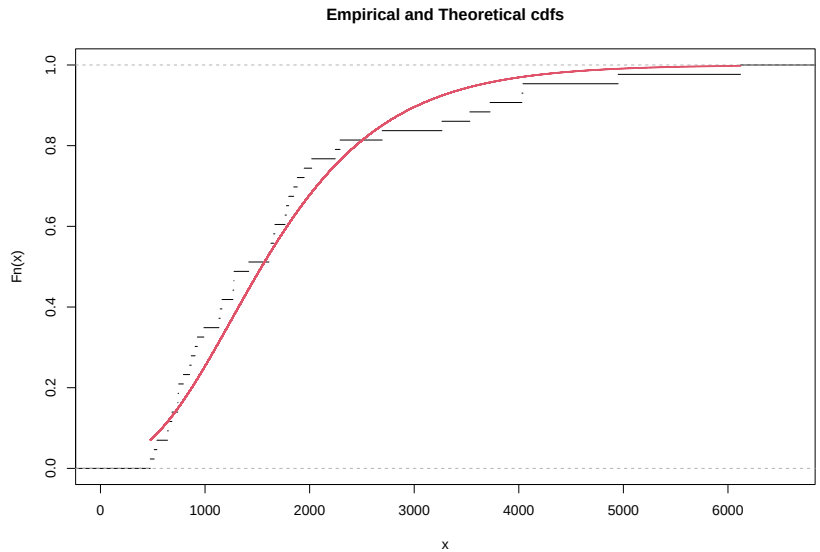
```
## [1] 1251.263
```

```
beta=estimate$mle[2]
beta
```

```
## [1] 793.0382
```

Prueba de Kolmogorov Smirnov

```
library(reliaR)
gastot=sort(gasto_tempoal)
ks.gumbel(gastot,1251.2634,793.0382, alternative="two.sided",plot=TRUE)
```




```
##
##  Exact one-sample Kolmogorov-Smirnov test
##
## data:  x
## D = 0.10856, p-value = 0.6516
## alternative hypothesis: two-sided
```

Conclusión : con un 95% de confianza los gastos maximos de Tempoal se ajustan al la distribución *Gumbel*.

periodos de retorno

```
library(ExtDist)
Prob=1-pGumbel(4000, location = 1251.2634, scale= 793.0382)
Prob
```

```
## [1] 0.03075624
```

```
1/Prob
```

```
## [1] 32.51373
```

La probabilidad que se presente un evento superior al gasto maximo anual 4000 mts^3/s en el rio Tempoal bajo el modelo Gumbel es de **0.03075624** y tiene un periodo de retorno de **32** años.

Ajuste a Gumbel en SAS

```
data panuco;
input periodo Tempoal Cardon P_Sanchez Hules Terrerillos;
cards;
1960 1277 1080 . 452.6 314
1961 852.9 303.5 . 434.5 525
1962 739.2 262 . 457.5 565.9
1963 1800 481 . 947.4 895.9
.
.
2000 641.2 . 580.4 84.9 753.4
2001 1847.9 498.3 1201.3 278.5 1512.2
2002 926.4 134 774.8 496.7 822.2
;
proc sgplot data=panuco;
    histogram Tempoal;
    *density Tempoal/ type=normal;
    density Tempoal/TYPE=KERNEL(c=1.1) ;
run;
proc univariate data=panuco noprint;
   qqplot Tempoal / gumbel(mu=est sigma=est) square;
run;
* CDF y Gumbel ;
proc univariate data=panuco noprint;
   cdf Tempoal / gumbel odstitle = title;
   inset gumbel(mu sigma);
run;
proc univariate data=panuco noprint;
   qqplot Tempoal / gumbel(mu=est sigma=est);
run;
proc univariate data=panuco;
   histogram tempoal/ gumbel(mu=est sigma=est);
run;
```