Análisis Estadístico de Datos Climáticos

Análisis exploratorio de datos univariados (I)

(Wilks, Cap. 3)

2015

Datos univariados

Análisis exploratorio de datos

(para tener una “primera impresión” de los datos)

Datos climáticos

- Observaciones (datos medidos; datos interpolados) :

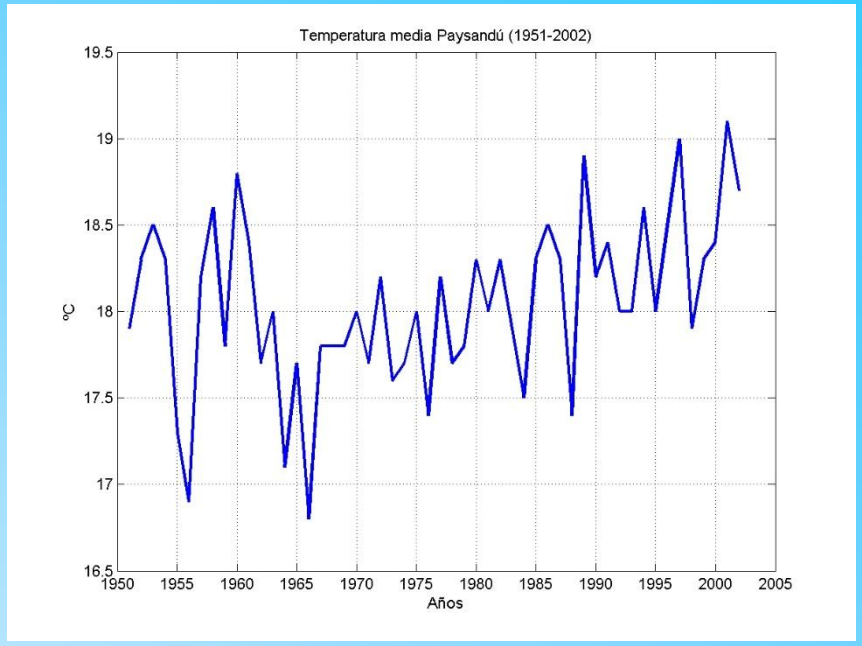
  Pueden ser in situ u obtenidas por sensoriamiento remoto (satélites, radares)

- Salidas de modelos numéricos o estadísticos:

  Simulaciones o pronósticos
  (posibilidad de variar condiciones iniciales o de borde)

La inspección visual de los datos es esencial!

Una simple gráfica puede mostrar características muy relevantes del conjunto de datos en cuestión.



También existen técnicas gráficas más sofisticadas para mostrar los datos, que permiten destacar algunos aspectos específicos de los mismos.

¿Hay tendencias?
¿Datos faltantes?
¿Outliers? (Datos atípicos)
¿Saltos?

Análisis exploratorio de datos univariados
(Wilks, Cap. 3)

- Robustez y resistencia
- Cuantiles (percentiles)
- Medidas numéricas de resumen
- Técnicas gráficas de resumen

Robustez y resistencia

Puede ser deseable que un método de análisis de datos sea **poco sensible** a suposiciones sobre la naturaleza de los datos.

P. ej., que los resultados no dependan esencialmente de que los datos sigan una distribución gaussiana o normal.

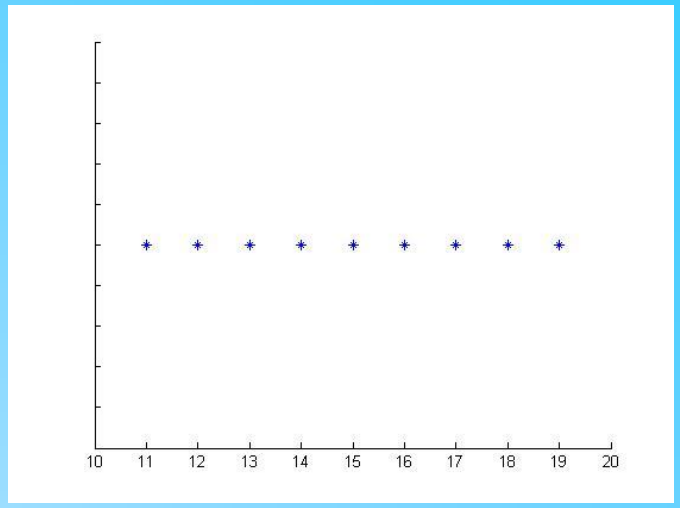
Un método es **robusto** cuando sus resultados no dependen esencialmente de cuál sea la distribución de probabilidades de los datos.

Un método es **resistente** si no es influido considerablemente por unos pocos datos atípicos (“outliers”)

Ejemplo:

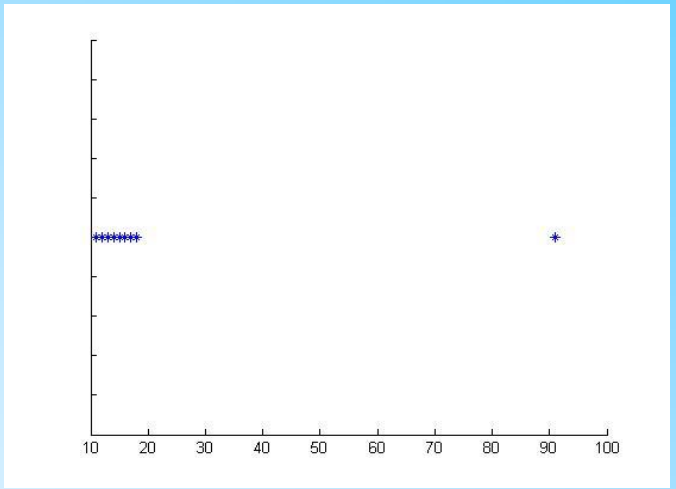
dado el conjunto

{11 12 13 14 15 16 17 18 19}



y el conjunto

{11 12 13 14 15 16 17 18 91}



Distintas medidas de “tendencia central”:
En ambos casos, el valor central es 15, pero los promedios son 15 y 23 respectivamente.

Estadísticos de orden de una muestra aleatoria

Sea $\{ x_1, x_2, ..., x_n \}$ una muestra aleatoria de datos

Se ordenan en forma ascendente:

$\{ x_{(1)}, x_{(2)}, ..., x_{(n)} \}$ son los estadísticos de orden

(cumpliéndose que $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$)

Ej: {7 -2 1 7 -3 4 0}

$\Rightarrow$ **{-3 -2 0 1 4 7 7}**

Cuantiles de una muestra aleatoria
(percentiles, cuartiles, quintiles, deciles, etc)

Ej.: 1) Sea la muestra aleatoria **{7 -2 2 7 -3 4 0}**

¿Cómo podemos estimar un valor central que, en sentido amplio, deje probabilidad ½ a ambos lados?

ordenamos

$\Rightarrow$ **{-3 -2 0 2 4 7 7}**

Parece natural tomar un valor que deje la misma cantidad de datos a cada lado, en este caso el 2**:**

{-3 -2 0 2 4 7 7}. Se dice que la **mediana** de la muestra es 2. **(“percentil 50”)** **($q_{0.5} = 2$)**

Cuantiles…

Ej. 2) Sea ahora la muestra **{7 1 7 -3 4 0}** (tiene un número par de datos)

¿Cuál será la mediana?

$\Rightarrow$ **{-3 0 1 4 7 7}**

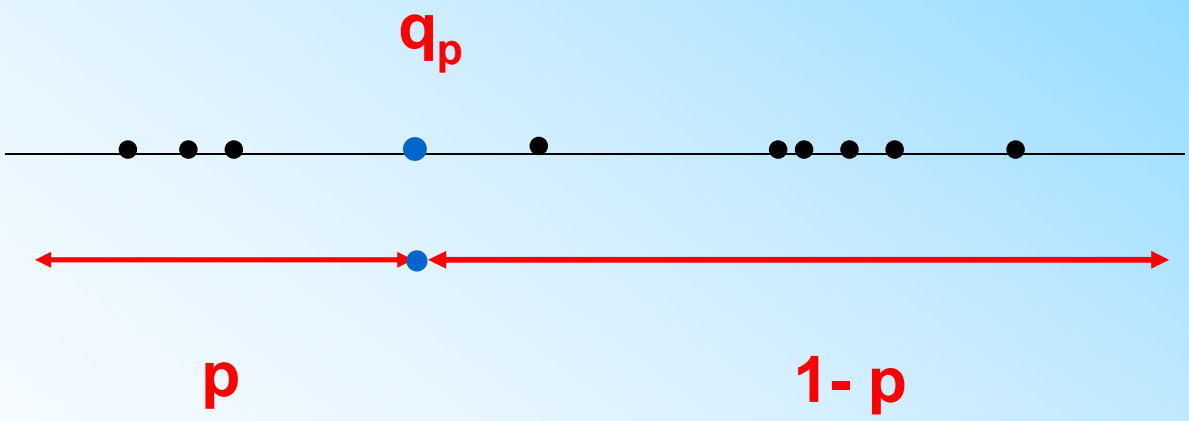
Convencionalmente, se suele tomar el promedio entre los dos valores centrales, o sea

(1 + 4) /2 = 2.5. **(un cuantil no tiene por qué ser un elemento de la muestra!)**

Pero, si no se tiene más información, podría elegirse cualquier valor en ese intervalo (1,4)

Generalizando, sea p tal que $0 < p < 1$.

Los **p-quantiles** (q_p) (o percentiles) son valores que dejan, en cierto sentido, probabilidad p a su izquierda, y probabilidad 1-p a su derecha. (Es decir que exceden la proporción de los datos dados por el subíndice p, con $0 < p < 1$).



$P(X \leq q_p) = p$ $\qquad$ $P(X > q_p) = 1 - p$

- Los cuantiles más usados:

- **Mediana q0.5**

- **Terciles, $q_{0.33}$, $q_{0.66}$**
- **Cuartiles, $q_{0.25}$, $q_{0.75}$**
- **Quintiles, deciles,**
- **$q_{0.05}$ $q_{0.95}$**

Estimación de los cuantiles

En general, los percentiles no son únicos y por lo tanto, no hay una única forma de estimarlos.

Una forma posible para una muestra aleatoria de tamaño n es:

1)tomar los estadísticos de orden como los cuantiles (0.5/n), (1.5/n), ..., ([n-0.5]/n) respectivamente

2) para los cuantiles con probabilidades entre (0.5/n) y ([n-0.5]/n), se interpola linealmente.

3) los valores mínimo o máximo de la muestra se asignan a los cuantiles para probabilidades fuera de ese rango.

Por ejemplo:

Para un vector de 5 elementos: {2, 10, 5, 9, 13}, los elementos del vector ordenado {2, 5, 9, 10, 13} corresponden respectivamente a los cuantiles 0.1, 0.3, 0.5, 0.7 y 0.9.

P. ej., el cuantil 0.6 se obtiene interpolando linealmente entre los de 0.5 y 0.7, obteniéndose $q_{0.6}$= 9.5.

Para cuantiles con probabilidades menores que 0.1 se asigna el valor mínimo (2) y para cuantiles con probabilidades mayores que 0.9, se asigna el valor máximo (13).

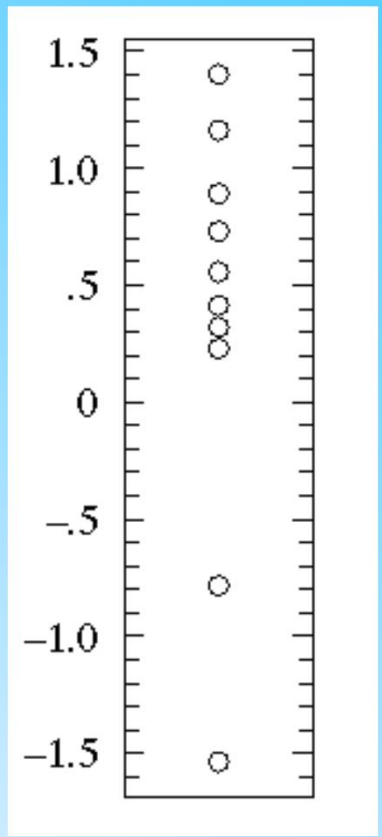
Principales medidas numéricas de resumen de un conjunto de datos

1) **Localización**: ej. valor de "tendencia central" del conjunto

2) **Dispersión**: alrededor del valor central

3) **Simetría**: cómo están distribuidos los datos respecto del valor central

4)…



Localización

Media $\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$

Valores de "tendencia central"

Mediana $q_{0.50}$

Ambas están comprendidas entre el mínimo y el máximo de la muestra.

La mediana "divide el conjunto de datos en dos subconjuntos ordenados con igual cantidad de datos" .

Es importante que los cuantiles (en particular la mediana) permiten trabajar con estimaciones de probabilidades

Localización

Ejemplo: (con muy pocos datos!!)

2 4 9 11 14 $\bar{x} = 8$

2 4 9 11 7004 $\bar{x} = 1406$

(outlier) ??

La media no es robusta ni resistente

Se puede estimar que P (X ≥ 9) ~ 0.5 ~ P(X ≤ 9)

Localización

Los cuantiles más usados…

- **Mediana $q_{0.5}$**
- **Terciles, $q_{0.33}$, $q_{0.66}$**
- **Cuartiles, $q_{0.25}$, $q_{0.75}$**
- **Quintiles, deciles,**
- **$q_{0.05}$ $q_{0.95}$**

$$\text{Trimedia} = \frac{q_{0.25} + 2q_{0.5} + q_{0.75}}{4}$$

Robustez vs. Eficiencia

¿Por qué se usa más la media que la mediana?

Porque en el caso (“muy frecuente”) de una distribución gaussiana es un estimador más eficiente que la mediana:

con menos valores (o sea, una muestra más pequeña) se obtiene la misma dispersión del estimador.

Además, la media es más fácil de tratar matemáticamente, y es única para una muestra dada.

Matlab

Variable	Comando
media	mean
cuantil	quantile
percentil	prctile
mediana	median

Ejemplo: si **X** es un vector de datos, y **p** está entre 0 y 1, el comando:

Y=quantile(X,p);

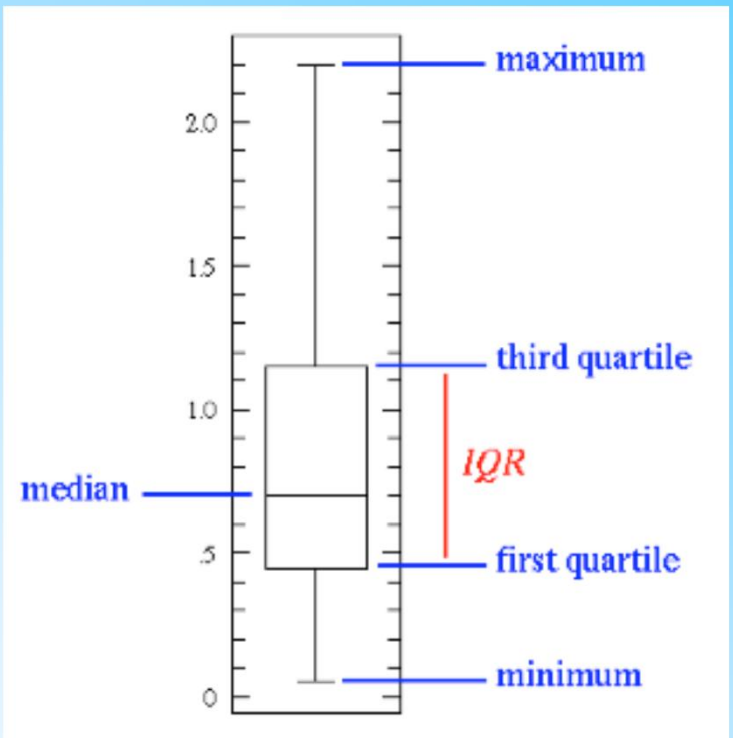
nos da $\mathbf{q_p}$ para el vector X.

Dispersión

- **Intervalo intercuartil** **(Robusto y resistente)**

$$IQR = q_{0.75} - q_{0.25}$$

"No usa" el 25% superior e inferior de los datos



Dispersión

- **Desviación estándar muestral** **(Ni robusta ni resistente)**

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2} \approx \sigma$$ **(σ^2 = varianza de la población)**

- **Desviación absoluta de la mediana**

MAD = median $\{|x_i - q_{0.5}|\}$ **Es robusto y resistente**

Simetría

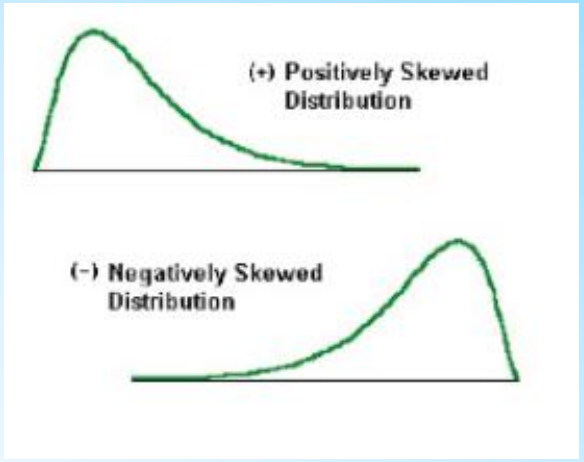
Coeficiente de asimetría de la muestra (ni robusto ni resistente)

$$\gamma = \frac{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^3}{s^3}$$

Indice de Yule-Kendall

$$\gamma_{YK} = \frac{\left(q_{0.75} - q_{0.5}\right) - \left(q_{0.5} - q_{0.25}\right)}{IQR}$$

Ambos son adimensionados



γ > 0

γ < 0

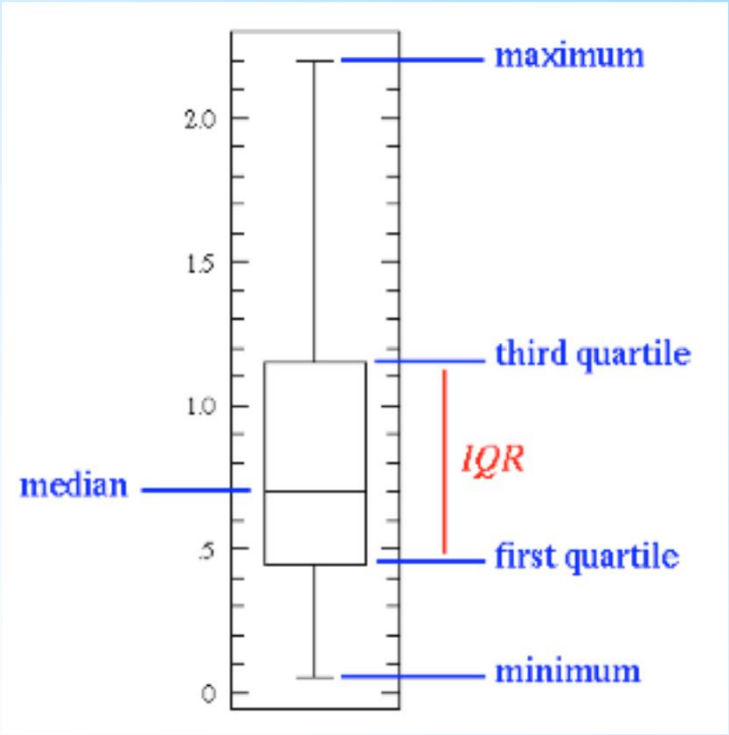
Matlab

Variable	Comando
intervalo intercuartil	iqr
desviación estándar	std
desviación absoluta de la mediana	mad
asimetría	skewness

Algunas técnicas gráficas de resumen

- Boxplots
- Histogramas
- Distribuciones de frecuencia acumulada

Boxplots (“barritas”)

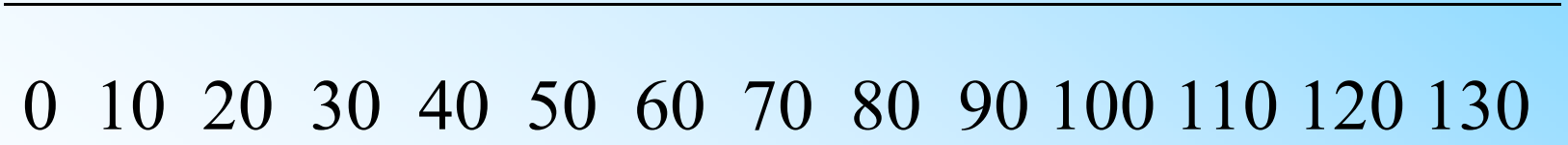


Usa 5 valores para describir un conjunto de datos: la mediana, el primer y tercer cuartiles, y el máximo y el mínimo (eventualmente con alguna restricción)

Boxplots (“barritas”)

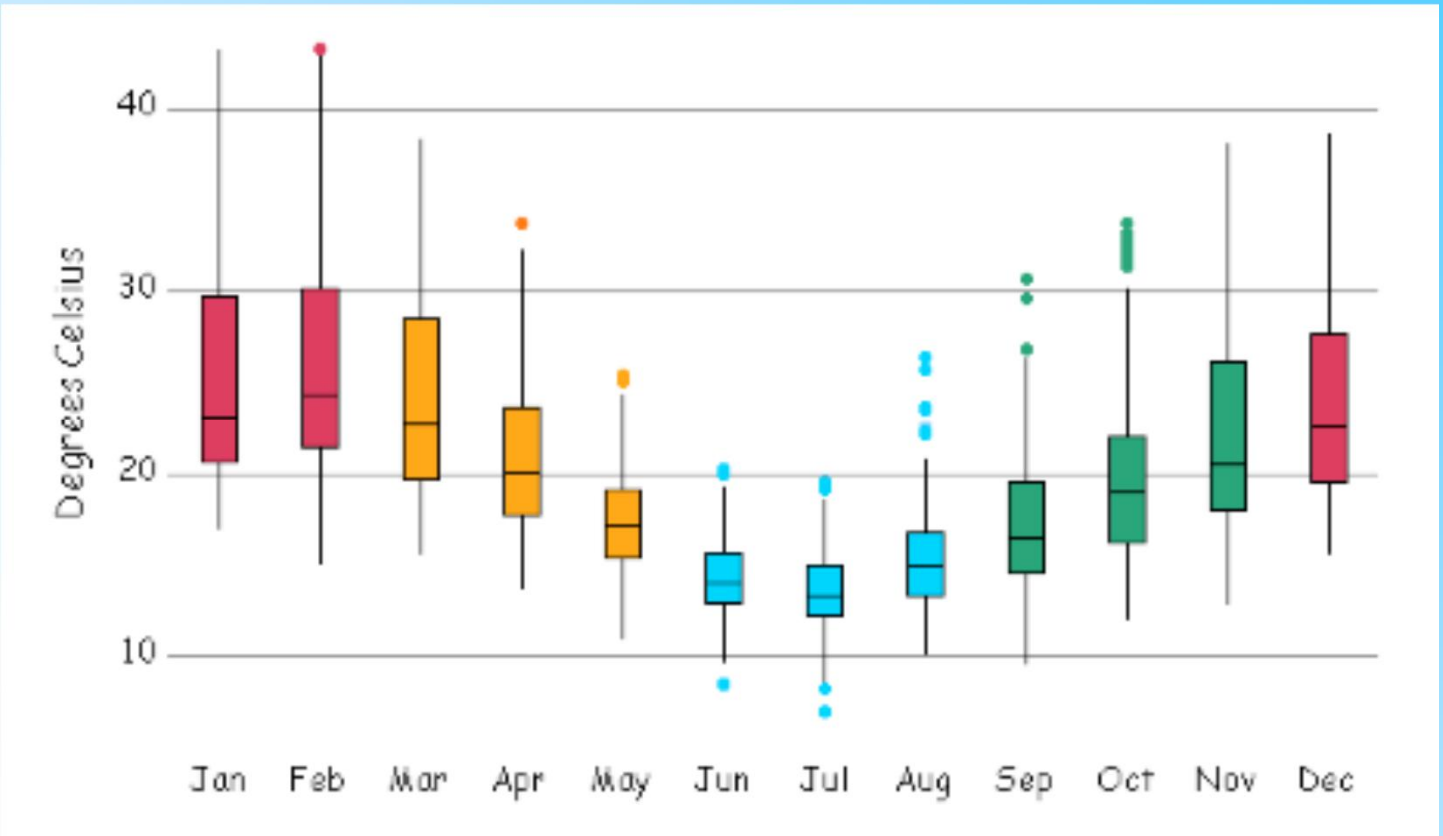
Min = 3.20 $q_{0.50} = 60.345$ Max = 124.27

$q_{0.25} = 43.645$ $q_{0.75} = 84.96$



Matlab: boxplot (X)

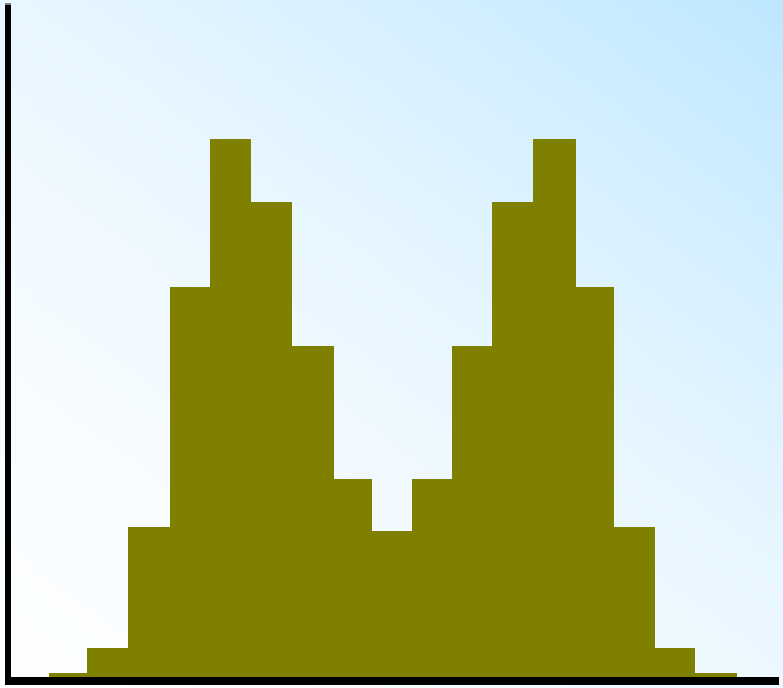
Temperatura diaria máxima en Melbourne



Se destacan valores extremos inusuales

Histogramas

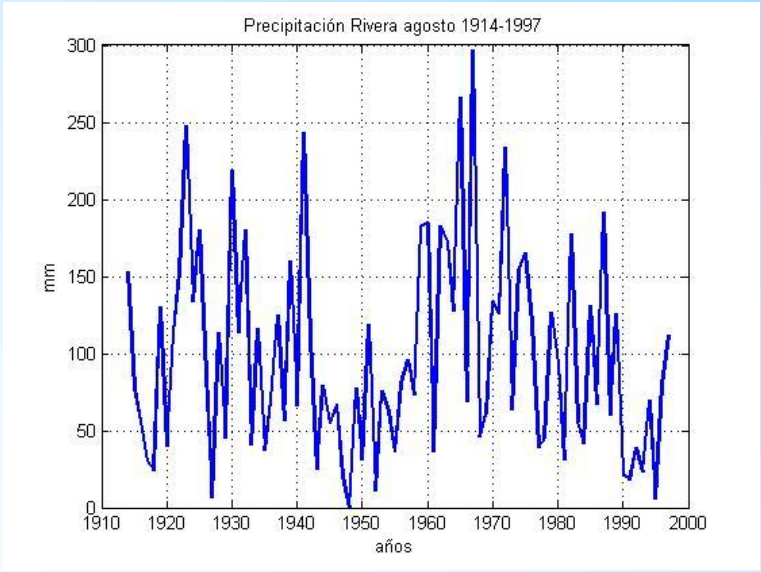
Además de dar idea sobre la localización, la dispersión, y la simetría, también muestran si los datos son multimodales

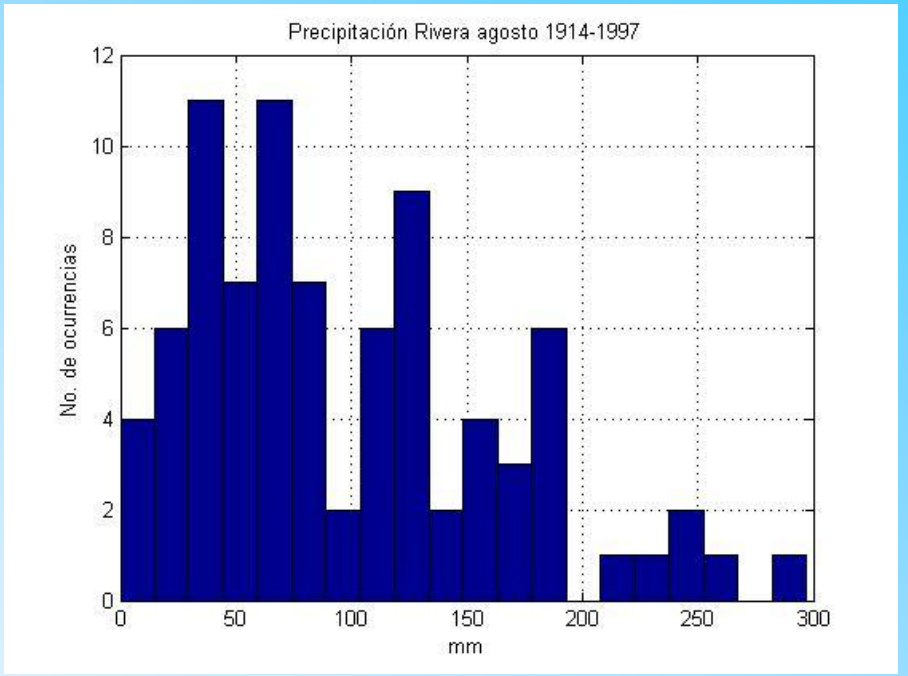


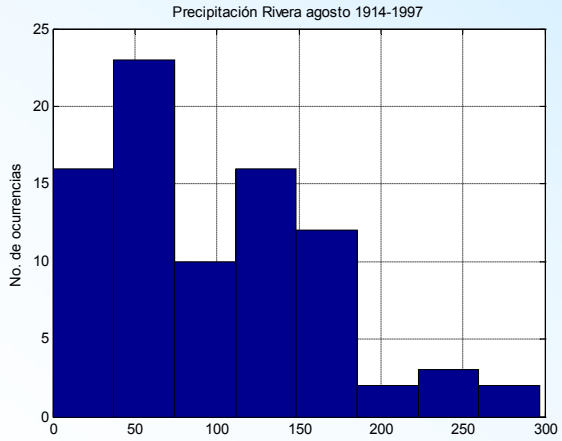
Distribución bimodal

Histogramas

Precipitación Rivera agosto 1914-1997







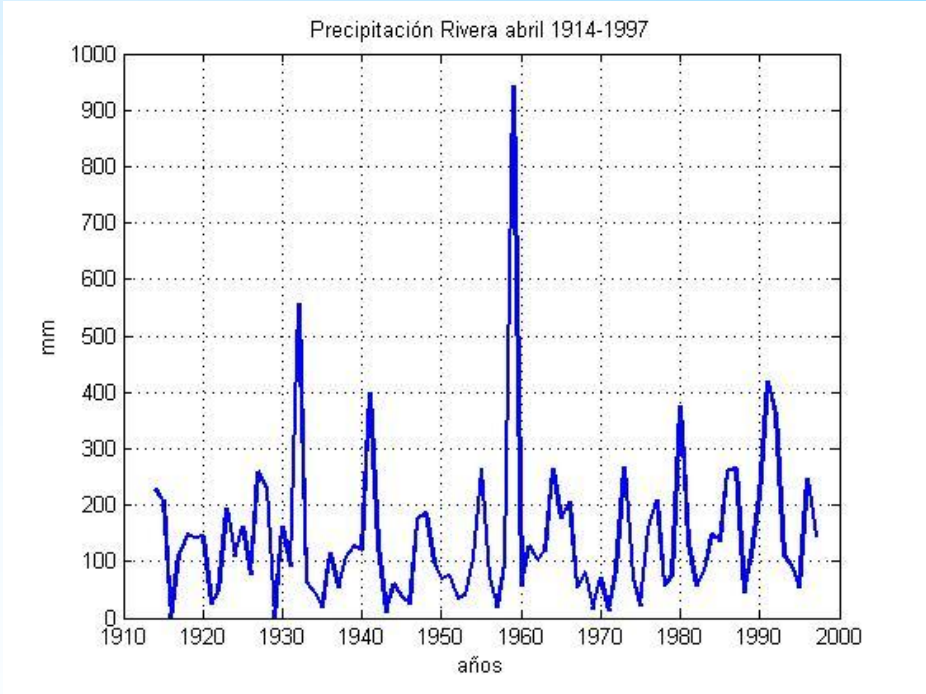
mediana=78.5 mm

media = 97.9 mm

Matlab: hist (X,n)

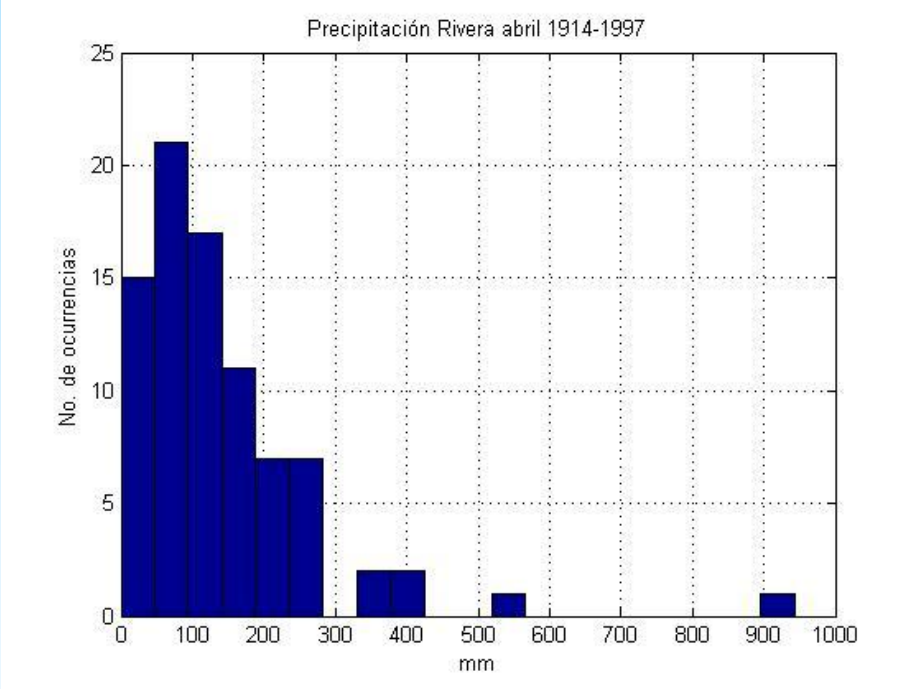
Histogramas

Precipitación Rivera abril 1914-1997

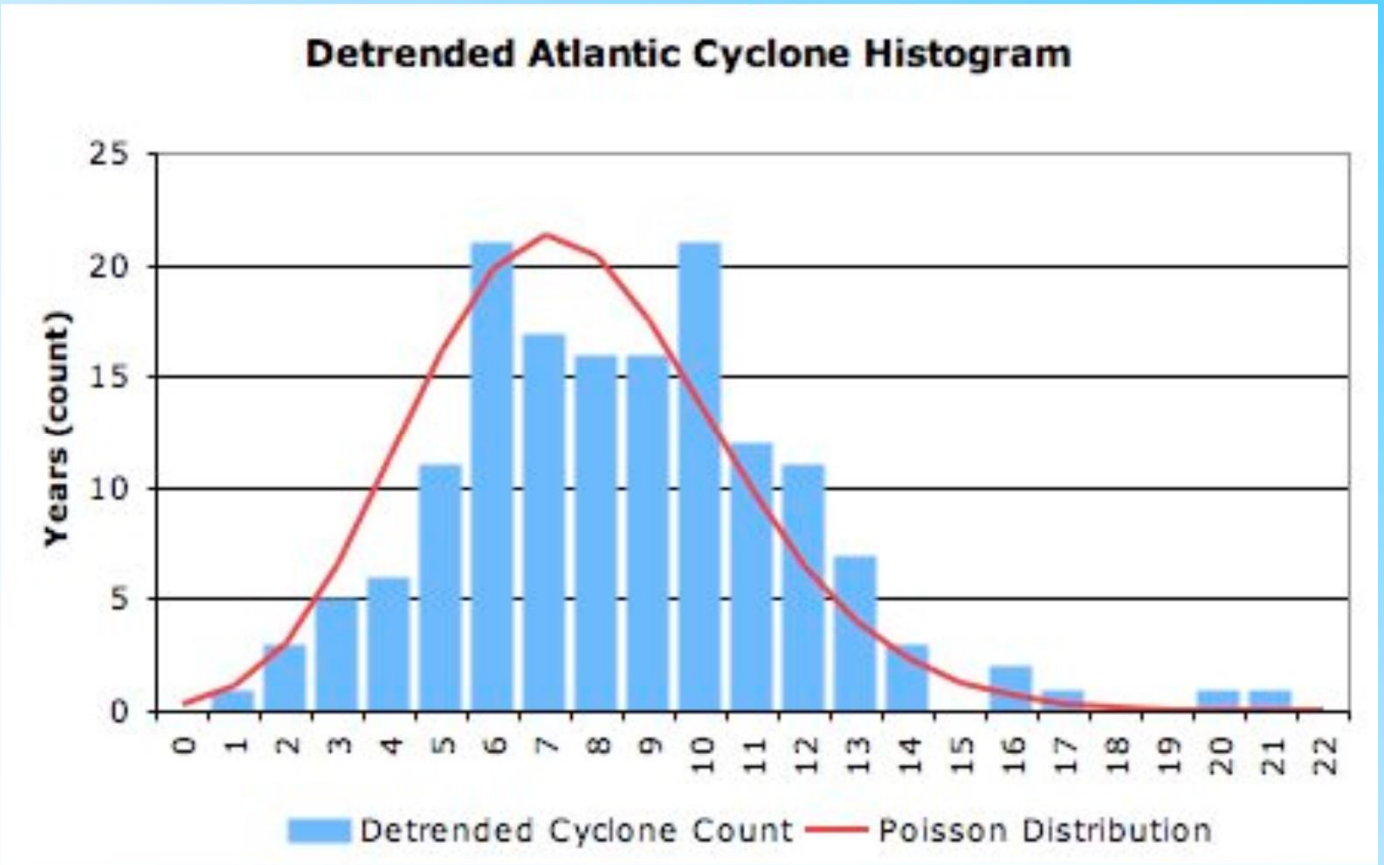


mediana=110.5 mm

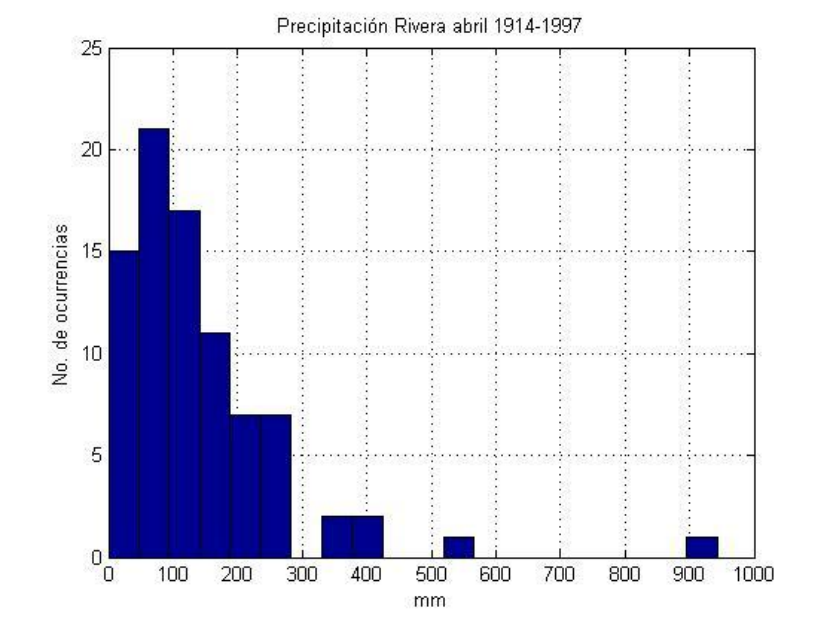
media = 141.7 mm



Histogramas



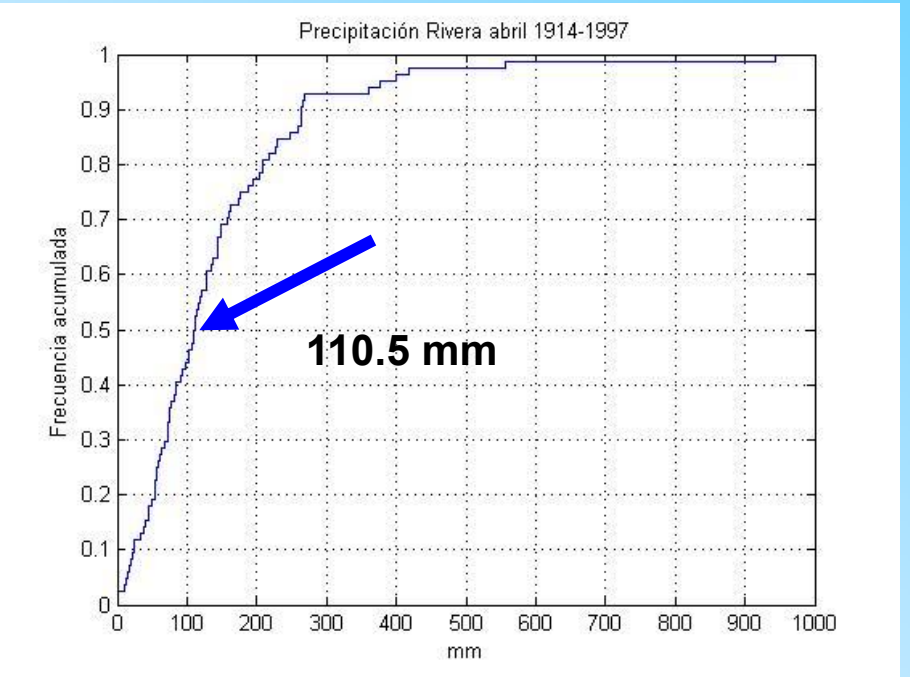
Distribuciones empíricas de frecuencia acumulada



Rivera abril 1914-1997

Interesa P (X ≤ x), probabilidad de no excedencia

P. ej. se puede estimar así:
$P(X \leq x_{(i)}) = (i - ½) / n$,
o de otras formas



Mediana ~ 110.5 mm

P(X≤110.5) ~ 0.5

Es la "función inversa" del cuantil

Matlab:

cdfplot(X)

Distribuciones empíricas de frecuencia acumulada

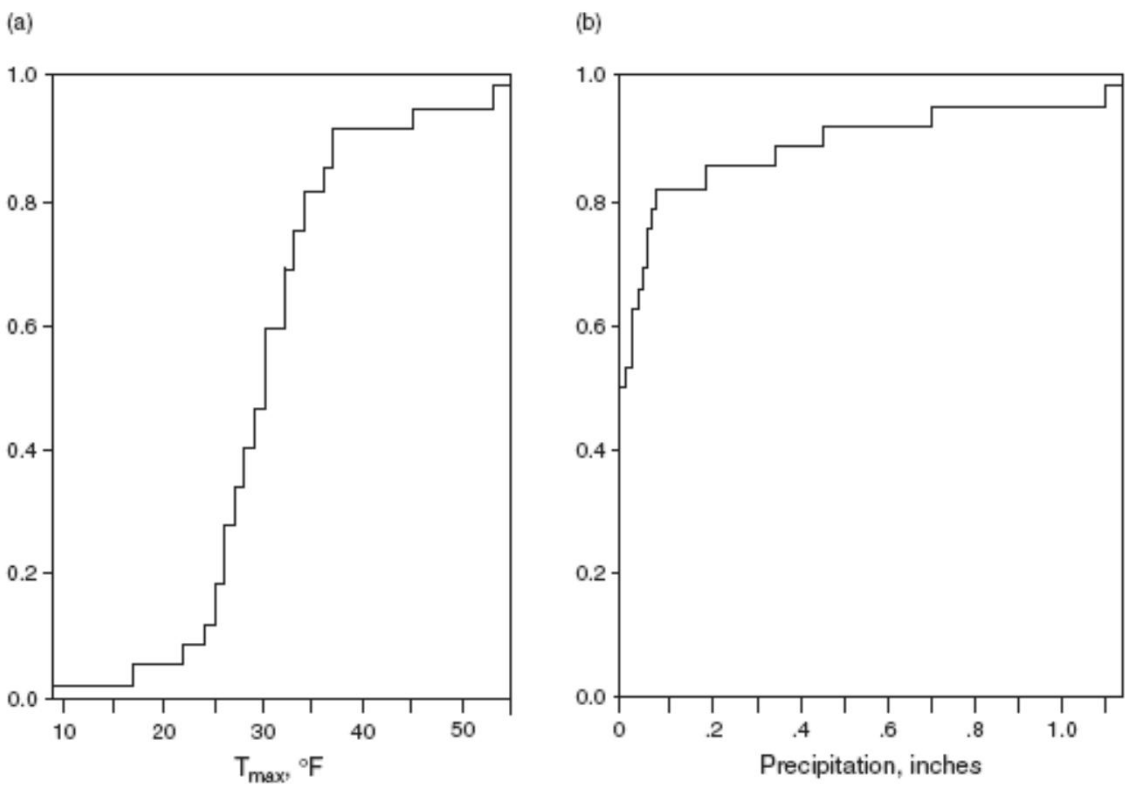


FIGURE 3.10 Empirical cumulative frequency distribution functions for the January 1987 Ithaca maximum temperature data (a), and precipitation data (b). The S shape exhibited by the temperature data is characteristic of reasonably symmetrical data, and the concave downward look exhibited by the precipitation data is characteristic of data that are skewed to the right.