Estadística Descriptiva

Edwin Santiago Alférez Baquero

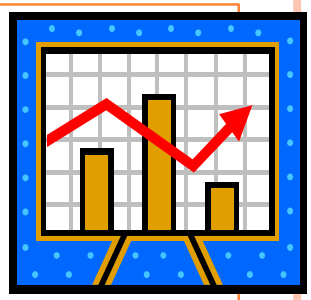
Contenido

Estadística descriptiva

- Conceptos básicos
- Organizar datos
- Resumir datos

Introducción

Estadística descriptiva

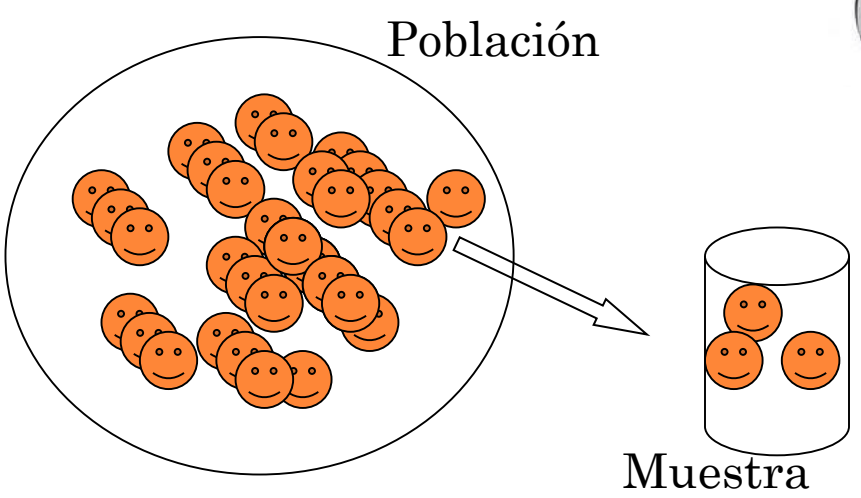


- Es la disciplina de resumir información cuantitativamente para describir las características principales de una colección de datos.
  - Tablas & Gráficas
  - Medidas de Tendencia Central
  - Medidas de Variabilidad

- Ejemplos:
  - Precipitación media en Bogotá el año pasado
  - Número de robos de automóviles en el año pasado
  - Resultados de prueba
  - Porcentaje de hombres en nuestra clase

Términos estadísticos (I)

Representación de la realidad



Población

- Conjunto completo de individuos, objetos o medidas

Muestra

- Un subconjunto de una población

Términos estadísticos (II)

Parámetro

- Una característica de una población
  - e.j., El *promedio* de la altura de todos los hombres de Hollywood.

Estadístico

- Una característica de una muestra
  - e.j., La *media* de la altura de una muestra de hombres de Hollywood.

Solo en estadística descriptiva

Términos estadísticos (III)

Variable

- Una característica que puede tomar diferentes valores

Datos

- Números o mediciones recolectadas

Variable: Altura

Información recolectada

Manejo de datos: Definiciones

Variable cualitativa

- Es una característica que expresa distintas cualidades o modalidades
- Dicotómicas (si o no)
- Politómicas
  - Ordinal, p.e. Bueno, regular, malo
  - Nominal, p.e. colores

Variable cuantitativa

- Es una característica que solo puede expresarse numéricamente
- Variables discretas
  - Toman valores únicamente enteros, p.e. # veces que ocurre un suceso
- Variables continuas
  - Toman cualquier valor real en un intervalo, p.e. Magnitudes reales

Estadística descriptiva

Organizar datos

Resumen de datos

Organizar datos

- Tablas
  - Distribuciones de frecuencia
- Gráficas
  - Gráfico de tallo y hoja - Gráfico de barras - Histograma - Polígono de frecuencia - Gráfico circular o por sectores -

Resumir datos

- Tendencia Central
  - Media– Mediana - Moda
- Variabilidad / Dispersión
  - Desviación Típica - Variance - Rango – Cuartil – Rango Intercuartílico
- Distancias relativas
  - Percentiles- Rango intercuartílico – Diagrama de cajas

Estadística descriptiva

Distribuciones de frecuencia

Frecuencia Absoluta

- Número de veces que aparece un cierto valor de la variable en el estudio

$$n_i \leq n, \quad \sum_{i=1}^{k} n_i = n$$

Frecuencia Relativa

- Número de veces que aparece un cierto valor de la variable dividido por todos los resultados

$$f_i = \frac{n_i}{n}$$

Frecuencia Acumulativa

- Es la suma de todas las frecuencias que se encuentran debajo de un valor particular
  - Frecuencia acumulativa absoluta N_i
  - Frecuencia acumulativa relativa F_i

ESTADÍSTICA DESCRIPTIVA

DISTRIBUCIONES DE FRECUENCIA

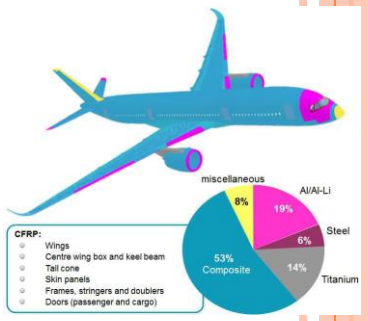
Ejemplo I

- El conjunto de datos para el control de calidad del agua de diferentes reactores es el siguiente:
  1, 5, 3, 1, 2, 3, 4, 5, 1, 4, 2, 4, 4, 5, 1, 4, 2, 4, 2, 2
- Donde cada número representa el reactor elegido como el mejor

Reactor	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia Absoluta Acum.	Frecuencia Relativa Acum.
1	4	4/20 = 0.2	4	0.2
2	5	5/20 = 0.25	9	0.45
3	2	2/20 = 0.1	11	0.55
4	6	6/20 = 0.3	17	0.85
5	3	3/20 = 0.15	20	1

ESTADÍSTICA DESCRIPTIVA

DISTRIBUCIONES DE FRECUENCIA



Ejemplo II

- Las resistencias a la compresión de la aleación en libras por pulgada cuadrada (psi) de 80 muestras de una nueva aleación de aluminio y litio en evaluación como posible material para elementos estructurales de aeronaves.

105	221	183	186	121	181	180	143
97	154	153	174	120	168	167	141
245	228	174	199	181	158	176	110
163	131	154	115	160	208	158	133
207	180	190	193	194	133	156	123
134	178	76	167	184	135	229	146
218	157	101	171	165	172	158	169
199	151	142	163	145	171	148	158
160	175	149	87	160	237	150	135
196	201	200	176	150	170	118	149

ESTADÍSTICA DESCRIPTIVA

DISTRIBUCIONES DE FRECUENCIA

Ejemplo II (cont)

- La distribución de frecuencia se debe agrupar
- Divida el rango de los datos en intervalos (intervalos de clases, celdas o contenedores)
- Si es posible, los contenedores deben tener el mismo ancho.
- El número de contenedores depende del número de observaciones y la cantidad de dispersión o dispersión en los datos (generalmente de 5 a 20 contenedores).

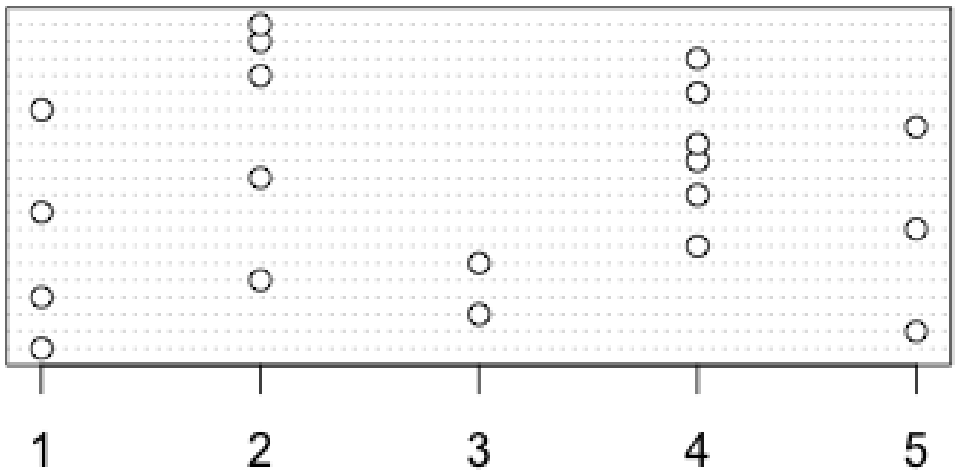
Class	$70 \le x < 90$	$90 \le x < 110$	$110 \le x < 130$	$130 \le x < 150$	$150 \le x < 170$	$170 \le x < 190$	$190 \le x < 210$	$210 \le x < 230$	$230 \le x < 250$
Frequency	2	3	6	14	22	17	10	4	2
Relative frequency	0.0250	0.0375	0.0750	0.1750	0.2750	0.2125	0.1250	0.0500	0.0250
Cumulative relative frequency	0.0250	0.0625	0.1375	0.3125	0.5875	0.8000	0.9250	0.9750	1.0000

Estadística descriptiva

Representaciones de distribuciones de frecuencia

Gráfica de puntos

- Es un buen resumen de datos numéricos cuando el conjunto de datos es razonablemente pequeño o hay relativamente pocos valores de datos distintos.
- Cada observación está representada por un punto sobre la ubicación correspondiente en una escala de medición horizontal.
- Cuando un valor ocurre más de una vez, hay un punto para cada ocurrencia, y estos puntos se apilan verticalmente



Ejemplo I

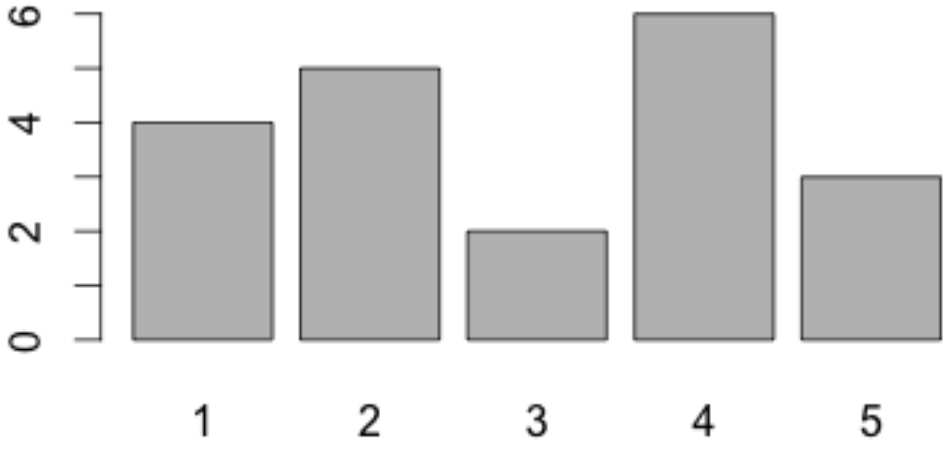
Reactor	Absolute Frequency
1	4
2	5
3	2
4	6
5	3

Estadística descriptiva

Representaciones de distribuciones de frecuencia

Gráfica de barras

- Es un gráfico con barras rectangulares con longitudes proporcionales a la frecuencia de cada valor.



Ejemplo I

Reactor	Absolute Frequency
1	4
2	5
3	2
4	6
5	3

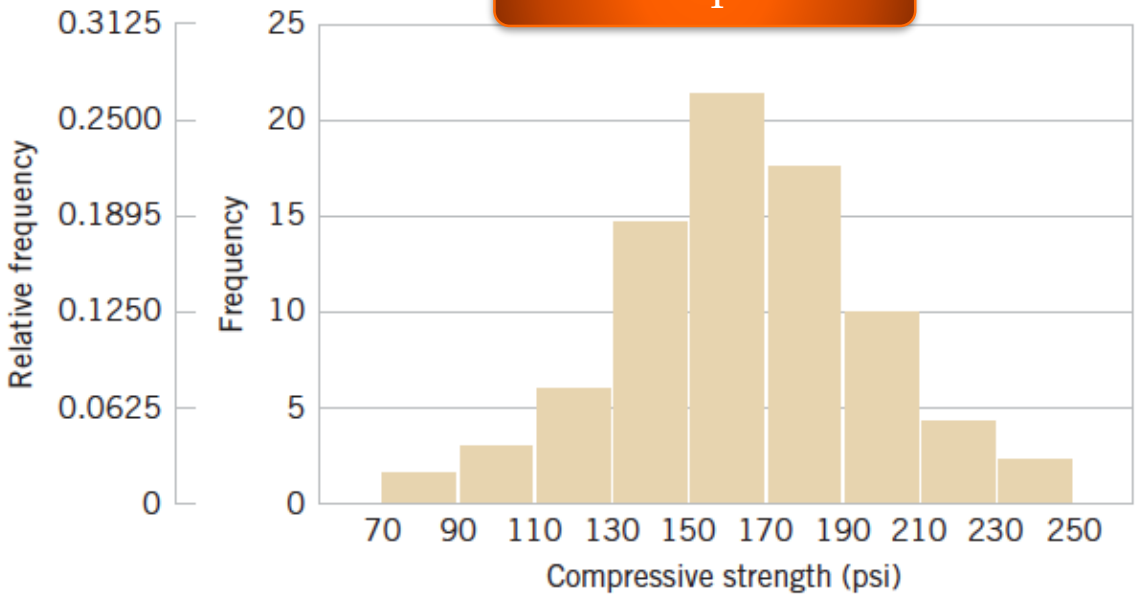
Estadística descriptiva

Representaciones de distribuciones de frecuencia

Histograma

- Es una visualización de la distribución de frecuencia
  - X valor (o puntos medios de intervalos de clase) en el eje x
  - Dibuja cada $f(x)$ con una barra, del mismo tamaño, tocando
  - Sin espacios entre barras

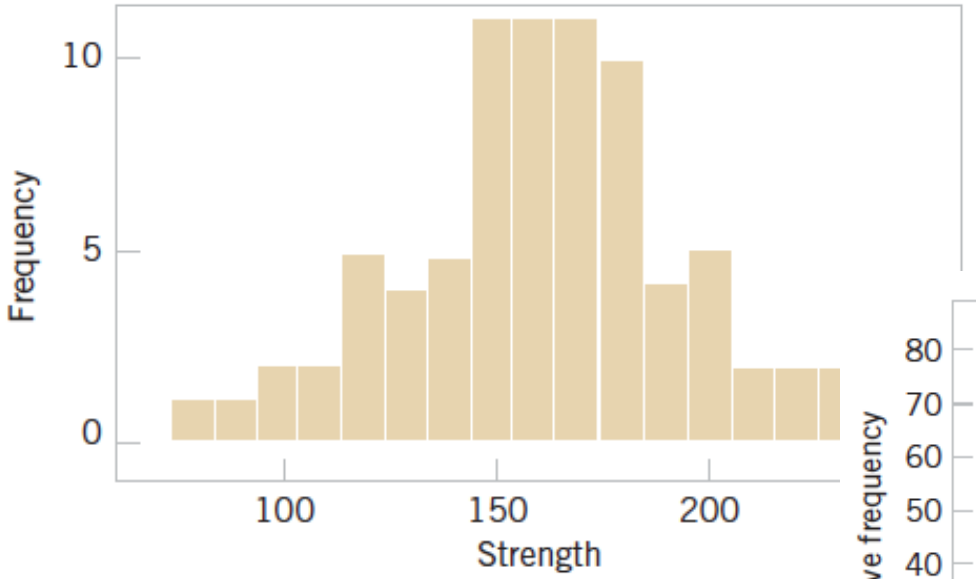
Example II



ESTADÍSTICA DESCRIPTIVA

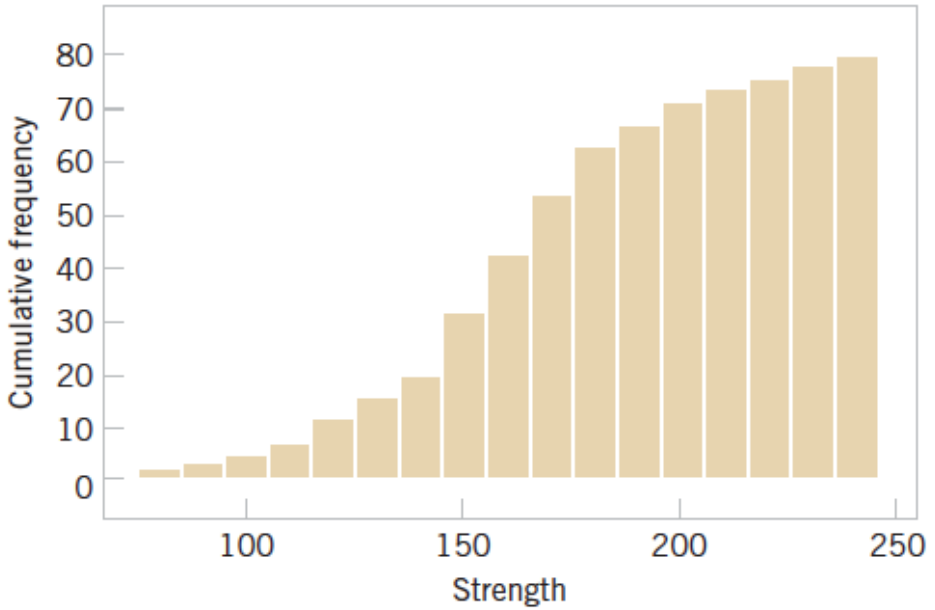
REPRESENTACIONES DE DISTRIBUCIONES DE FRECUENCIA

Histograma



Ejemplo II

Histograma acumulativo



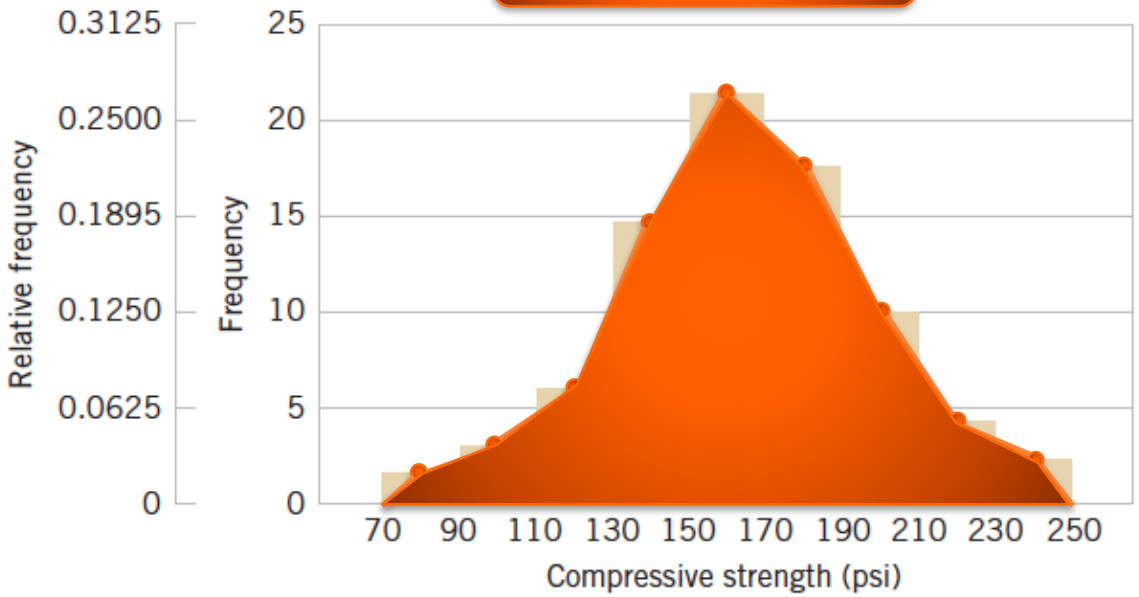
Estadística descriptiva

Representaciones de distribuciones de frecuencia

Polígonos de frecuencia

- Representa la información de una tabla de frecuencias o una tabla de frecuencias agrupadas como un **gráfico de líneas**

Ejemplo II

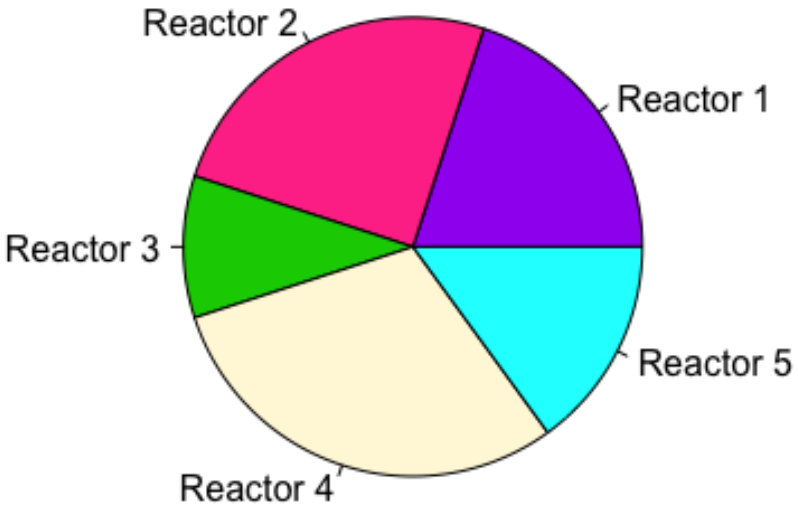


Estadística descriptiva

Representaciones de distribuciones de frecuencia

Diagrama por sectores

- Es una torta divida en sectores, ilustrando numéricamente una proporción
- La longitud del arco de cada sector (y consecuentemente su ángulo y área central) es proporcional a la frecuencia de cada valor



Ejemplo I

Reactor	Absolute Frequency
1	4
2	5
3	2
4	6
5	3

Estadística descriptiva

Resumen de datos

Medidas de Tendencia Central

- Se calculan para dar un **"centro"** alrededor del cual se distribuyen las mediciones en los datos

Medidas de Variación o Variabilidad

- Describen la **"dispersión de datos"** o cuán lejos están las mediciones del centro.

Medidas relativas

- Describa la **"posición relativa"** de las medidas específicas en los datos

Estadística descriptiva

Tendencia Central

Media

- Más comúnmente llamado el “promedio”.
- Es el “punto de equilibrio”.
- Sume los valores para cada caso y divida por el número total de casos.

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_N}{N} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

- Crucial para estadísticas inferenciales
- No es muy resistente a los valores atípicos
- Una “media recortada” puede ser mejor para fines descriptivos

Estadística descriptiva

Tendencia Central

Media (Ejemplo)

rim diameter (cm)

unit 1	unit 2
12.6	16.2
11.6	16.4
16.3	13.8
13.1	13.2
12.1	11.3
26.9	14.0
9.7	9.0
11.5	12.5
14.8	15.6
13.5	11.2
12.4	12.2
13.6	15.5
	11.7

n	12	13
total	168.1	172.6
total/n	14.0	13.3

rim diameter (cm)

unit 1	unit 2
~~9.7~~	~~9.0~~
11.5	11.2
11.6	11.3
12.1	11.7
12.4	12.2
12.6	12.5
13.1	13.2
13.5	13.8
13.6	14.0
14.8	15.5
16.3	15.6
~~26.9~~	16.2
	~~16.4~~

n	10	11
total	131.5	147.2
total/n	13.2	13.4

Estadística descriptiva

Tendencia Central

Media

- Si los datos están agrupados en una tabla de frecuencias, entonces

$$\overline{x} = \frac{x_1 . n_1 + x_2 . n_2 + x_3 . n_3 + \ldots + x_N . n_N}{N} = \frac{1}{N} \sum_{i=1}^{k} x_i . n_i$$

donde $n_1 + n_2 + n_3 + \ldots + n_k = N$

Estadística descriptiva

Tendencia Central

Mediana

- El elemento más central o más central en el conjunto de números ordenados; separa la distribución en dos mitades iguale
- Si n es impar, es el valor de la mitad de toda la secuencia
  - Si X = [1,2,4,6,**9**,10,12,14,17]
  - Entonces **9** es la mediana
- Si n es par, es el promedio de los 2 valores del medio
  - Si X = [1,2,4,6,**9**,**10**,11,12,14,17]
  - Entonces **9.5** es la mediana; i.e., (9+10)/2

- La Mediana no se ase afecta por los valores extremos

Estadística descriptiva

Tendencia Central

Mediana (Ejemplo)

rim diameter (cm)

unit 1		unit 2	
9.7		9.0	
11.5		11.2	
11.6		11.3	
12.1		11.7	
12.4		12.2	
12.6		12.5	
12.9	<--	13.2	13.2
13.1		13.8	
13.5		14.0	
13.6		15.5	
14.8		15.6	
16.3		16.2	
26.9		16.4	

Estadística descriptiva

Tendencia Central

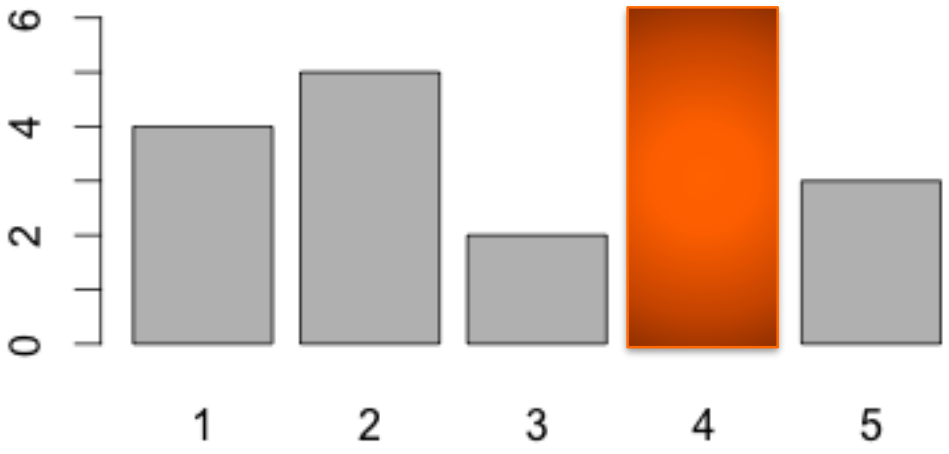
Moda

- La moda es el número que ocurre con mayor frecuencia en una distribución
  - Si X = [1,2,4,**7,7,7**,8,10,12,14,17]
  - Entonces 7 es la moda
- Fácil de ver en una distribución de frecuencia simple
- Es posible no tener moda o que existan más de una moda
  - Bimodal y multimodal
- No tiene que ser exactamente la misma frecuencia
  - mayor moda, menor moda
- La moda no se ve afectada por valores extremos

Estadística descriptiva

Tendencia Central

Moda (Ejemplo)



Example I

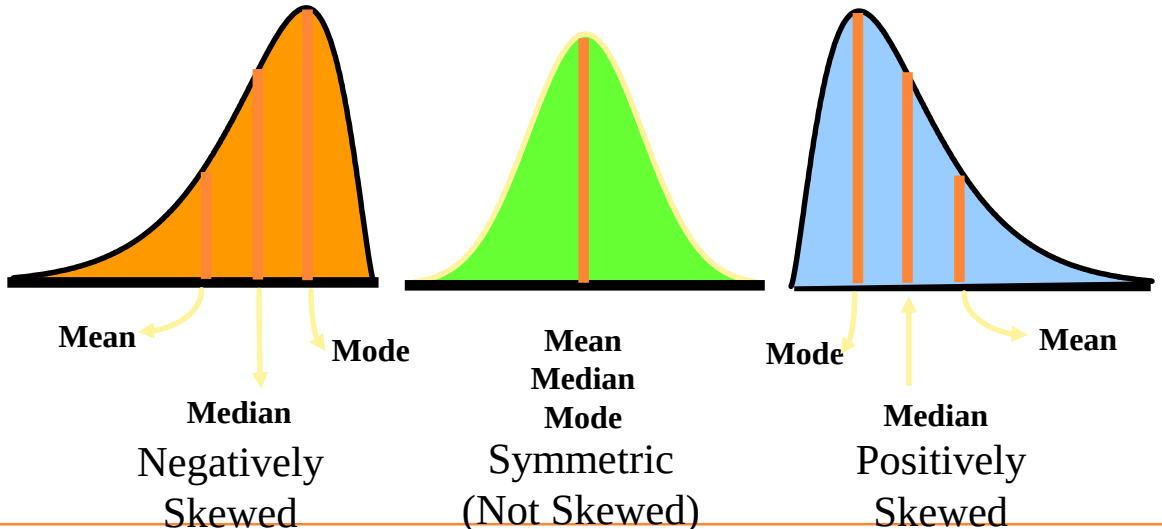
Reactor	Absolute Frequency
1	4
2	5
3	2
4	**6**
5	3

Estadística descriptiva

Tendencia Central

Cuándo usar y qué usar?

- La media es una gran medida. Pero, hay momentos en que su uso es inapropiado o imposible.
  - Datos Nominales: Moda
  - La distribución es bimodal: Moda
  - Datos ordinales: Mediana o moda
  - Son algunos puntajes extremos: Mediana



ESTADÍSTICA DESCRIPTIVA

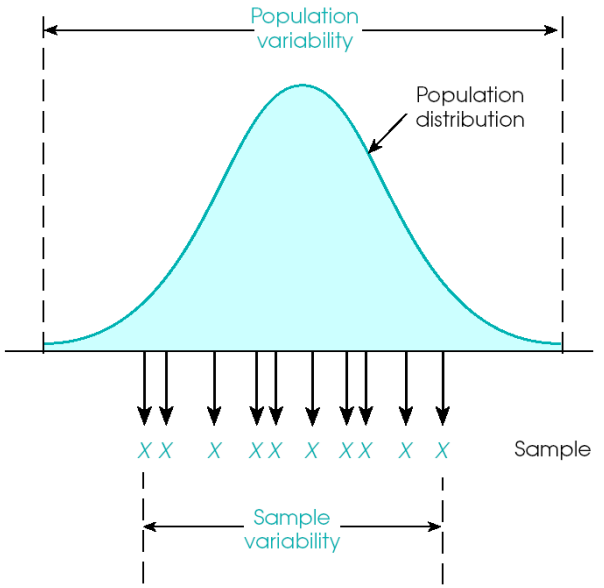
VARIABILIDAD

Dispersión

- Qué tan estrechamente agrupados o qué tan variables son los valores en un conjunto de datos.

- Ejemplo
  - Data set 1: [0,25,50,75,100]
  - Data set 2: [48,49,50,51,52]
  - Ambos tienen una media de 50, pero el conjunto de datos 1 claramente tiene una **mayor variabilidad** que el conjunto de datos 2.

- Rango– Varianza – Desviación Típica



ESTADÍSTICA DESCRIPTIVA

VARIABILIDAD

Rango

- La propagación entre los valores más bajos y más altos de una variable.
- Muy sensible a los valores atípicos, insensible a la forma.
- Ignora cómo se distribuyen los datos y solo toma en cuenta los puntajes extremos
- Muy sensible a los valores atípicos, insensible a la forma.

$$Range\_1 = 26.9 - 9.7 = 17.2$$

$$Range\_2 = 16.4 - 9 = 7.4$$

unit 1	unit 2
9.7	9.0
11.5	11.2
11.6	11.3
12.1	11.7
12.4	12.2
12.6	12.5
13.1	13.2
13.5	13.8
13.6	14.0
14.8	15.5
16.3	15.6
26.9	16.2
	16.4

Estadística descriptiva

Variabilidad

Varianza

- Mide cuántos cada número en el conjunto es de la media
- El promedio de las diferencias al cuadrado de la media

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

- Nota: las unidades de la varianza son cuadradas, y hace que la difícil de interpretar.
- Ex.: muestra de punto proyectil:
  - Media= 22.6 mm
  - Varianza = 38 mm^2
- **Qué significa esto???**

ESTADÍSTICA DESCRIPTIVA

VARIABILIDAD

Desviación Típica

- Raíz cuadrada de la Varianza

$$s = \sqrt{s^2} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

- Las unidades están en las mismas unidades que las mediciones de base

- Ex.: muestra de punto proyectil:
  - media= 22.6 mm
  - Desviación típica= 6.2 mm

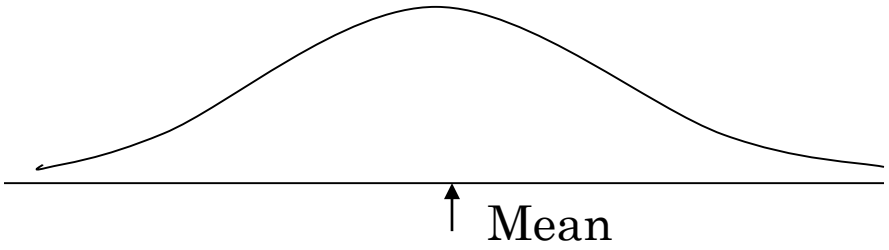
- Mean +/- sd (16.4—28.8 mm)
  - debería dar al menos un sentido intuitivo de dónde yacen la mayoría de los casos, salvo los efectos principales de los valores atípicos

Estadística descriptiva

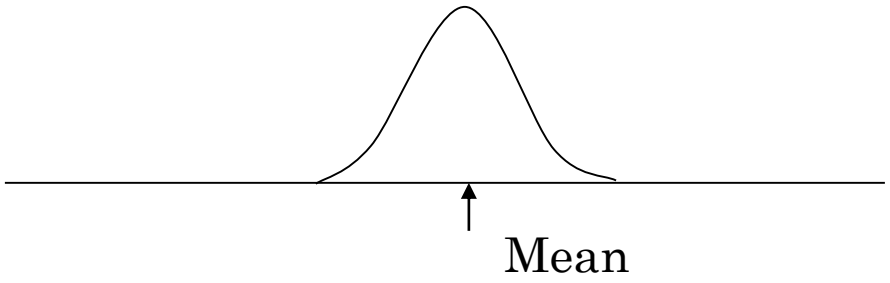
Variabilidad

Varianza y Desviación Típica

- Cuanto mayor es la varianza, más lejos están los casos individuales de la media.



- Cuanto menor es la varianza, más cercanos son los puntajes individuales a la media.

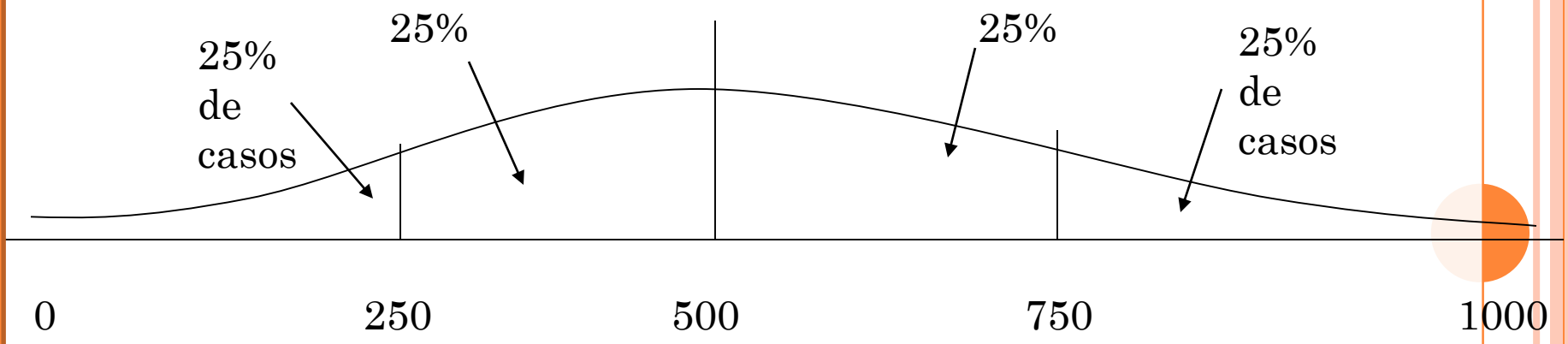


Estadística descriptiva

Distancia relativa

Percentiles

- El percentil p-ésimo es un número tal que, como máximo, el p% de las mediciones están por debajo de él y, como máximo, el 100-p por ciento de los datos están por encima de él.
  - $P_{25} = Q_1$ => Primer cuartil
  - $P_{50} = Q_2$ => Segundo cuartil => Mediana
  - $P_{75} = Q_3$ => Tercer cuartil



Estadística descriptiva

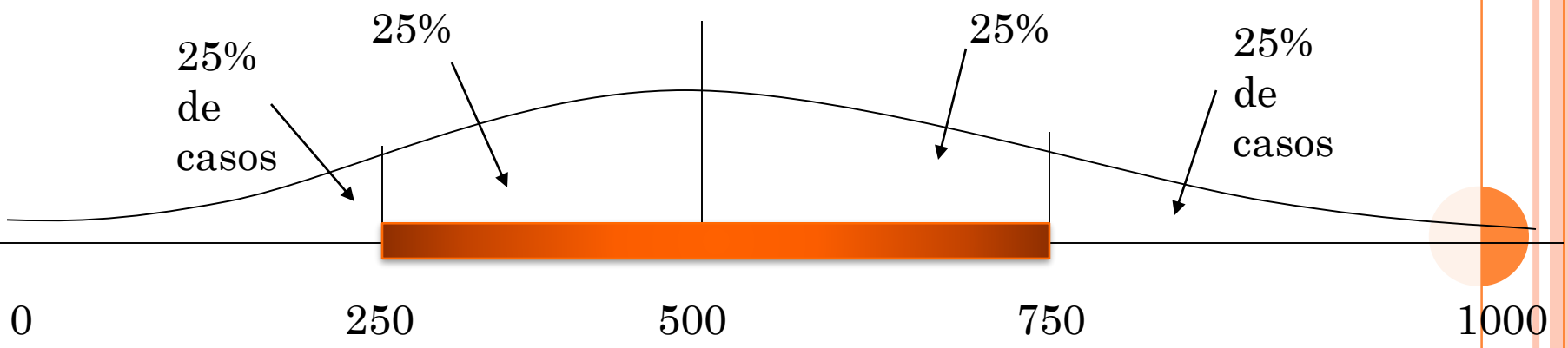
Distancia relativa

Rango intercuartílico

- Diferencia entre tercer y primer cuartil

$$IQR = Q_3 - Q_1$$

- Contiene el 50% de la información
- No se ve afectado por valores extremos

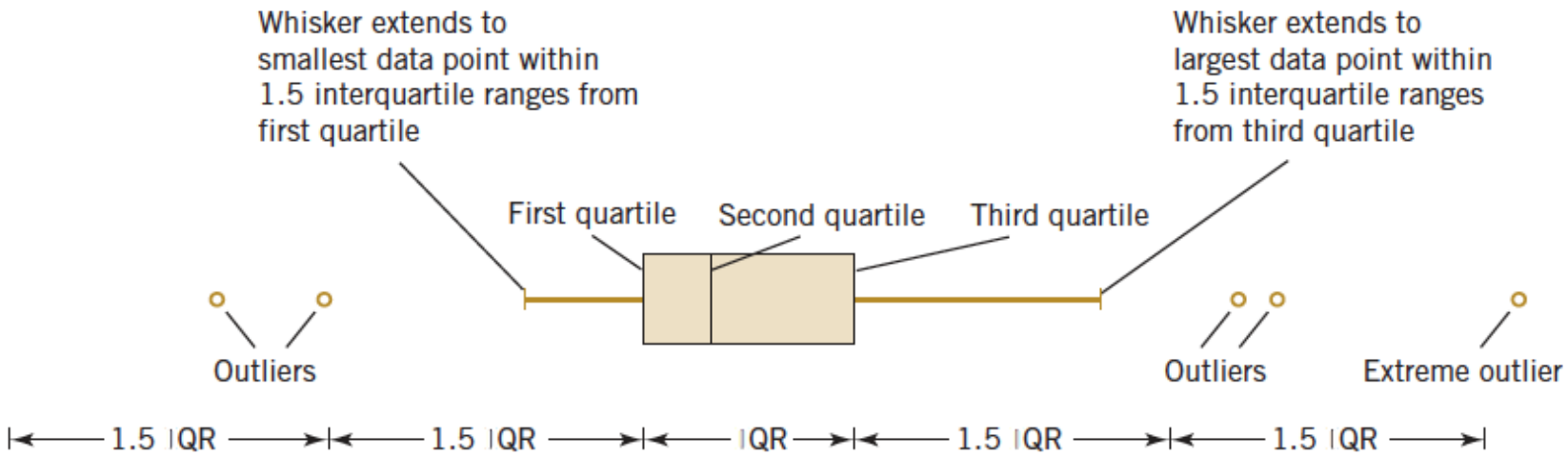


ESTADÍSTICA DESCRIPTIVA

REPRESENTACIÓN DE VARIABILIDAD Y DISTANCIAS

Diagrama de cajas

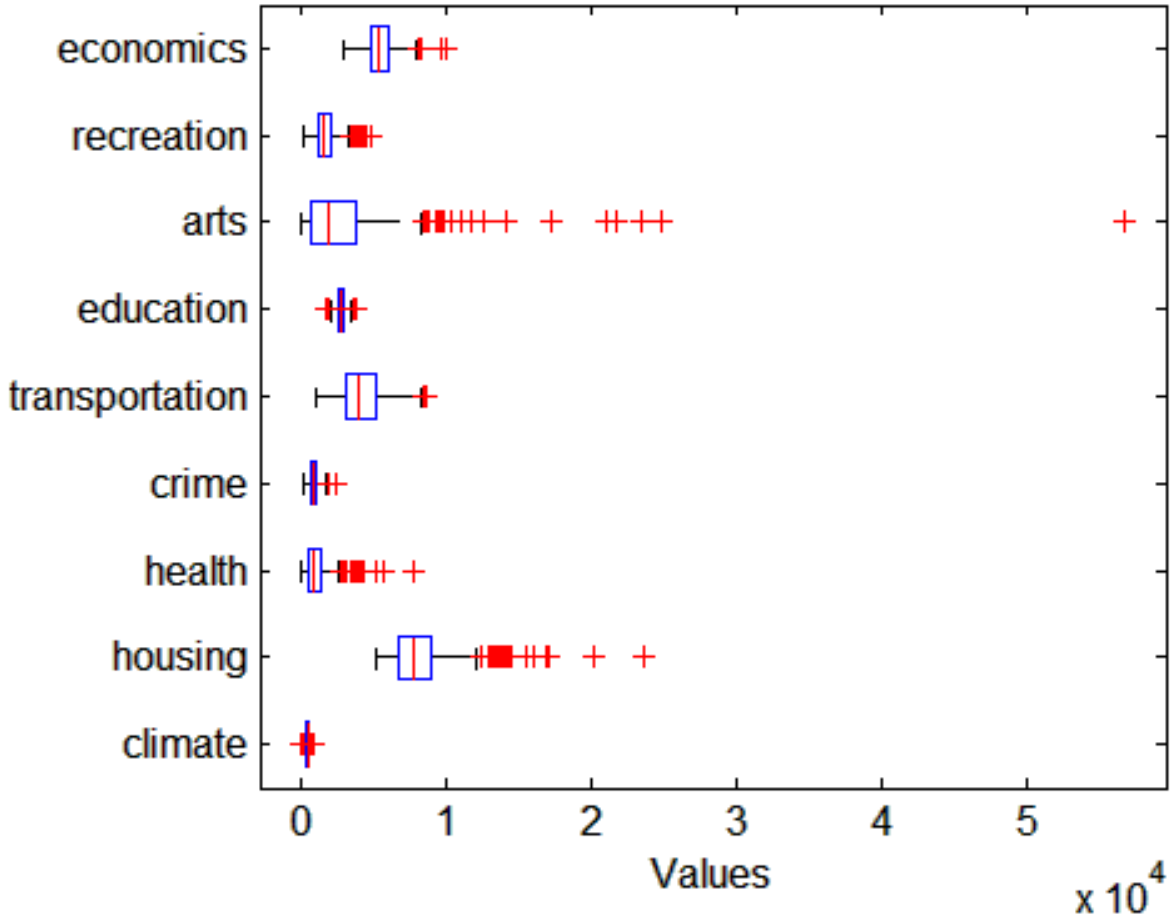
- Un gráfico que tiene un recuadro de Q1 a Q3, y contiene también un número para resumir los datos: **Mínimo, Q1, Mediana, Q3, Máximo**
- También indica valores atípicos identificados por separado
- Extremo = observación extrema
- abajo LQ – 1.5(IQR)



Estadística descriptiva

Representación de variabilidad y distancias

Diagrama de cajas



Ejercicio 1

De gran importancia para residentes de la región central de Florida es la cantidad de material radiactivo presente en el suelo de zonas recuperadas de la explotación minera de fosfatos. Las mediciones de la cantidad de ^{238}U en 25 muestras de suelo fueron como sigue (mediciones en picocurios por gramo):

.74	6.47	1.90	2.69	.75
.32	9.99	1.77	2.41	1.96
1.66	.70	2.42	.54	3.36
3.59	.37	1.09	8.32	4.06
4.55	.76	2.03	5.70	12.48

Construya un histograma de frecuencia relativa para estos datos.

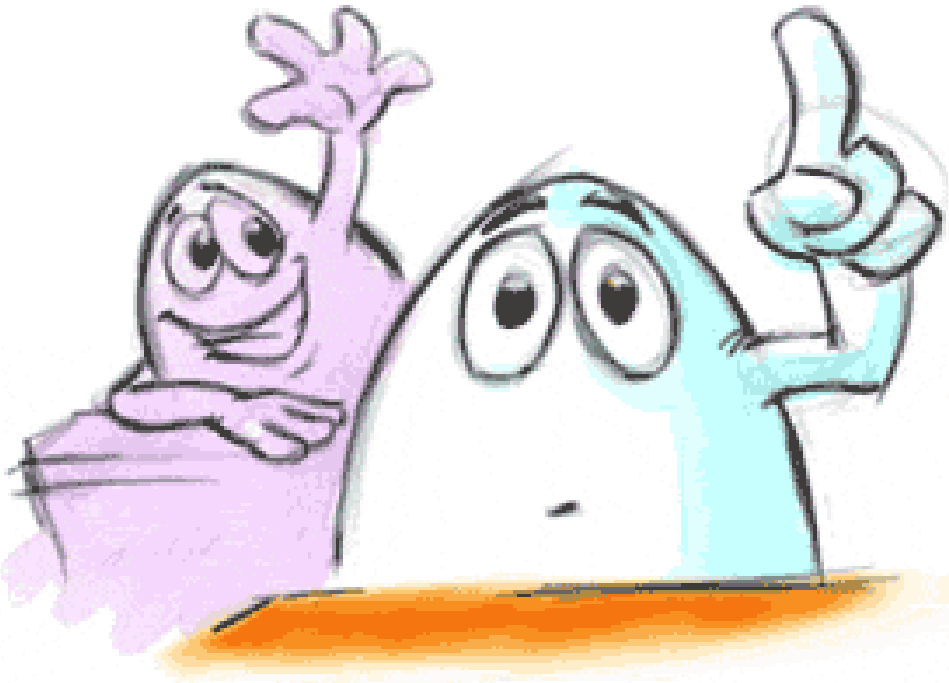
Ejercicio 2

Una compañía farmacéutica desea saber si un medicamento experimental tiene efecto sobre la presión sistólica de la sangre. A 15 pacientes seleccionados al azar se les aplicó el medicamento y, después de un tiempo suficiente para que el medicamento tuviera efecto, se registraron sus presiones sistólicas. Los datos aparecen a continuación:

172 140 123 130 115
148 108 129 137 161
123 152 133 128 142

Teorema de Tchebysheff
Para cualquier conjunto de n mediciones, la fracción incluida en el intervalo $\bar{y} - ks$ a $\bar{y} + ks$ es al menos $(1 - \frac{1}{k^2})$

a) Dibuje un histograma de frecuencias relativas
b) Calcule el valor de s usando la aproximación del rango.
c) Calcule los valores de $\bar{y}$ y s para las 15 lecturas de presión sanguínea.
d) Use el teorema de Tchebysheff para hallar valores de a y b tales que al menos 75% de las mediciones de presión sanguínea se encuentren entre a y b.
e) ¿Funcionó el teorema de Tchebysheff? Es decir, use la información para hallar el porcentaje real de lecturas de presión sanguínea que están entre los valores de a y b hallados en el inciso d. ¿Este porcentaje real es mayor que 75%?



ESTADÍSTICA DESCRIPTIVA

REPRESENTACIONES DE DISTRIBUCIONES DE FRECUENCIA

Ejemplo II

Diagramas de tallo y hojas

- Es un dispositivo para presentar datos cuantitativos en un formato gráfico para ayudar a visualizar la forma de una distribución.
- Es una tabla especial donde cada valor de datos se divide en una "hoja" (generalmente el último dígito) y un "tallo" (los otros dígitos).
- Da información sobre la ubicación, propagación, extremos y huecos.

Stem	Leaf	Frequency
7	6	1
8	7	1
9	7	1
10	5 1	2
11	5 8 0	3
12	1 0 3	3
13	4 1 3 5 3 5	6
14	2 9 5 8 3 1 6 9	8
15	4 7 1 3 4 0 8 8 6 8 0 8	12
16	3 0 7 3 0 5 0 8 7 9	10
17	8 5 4 4 1 6 2 1 0 6	10
18	0 3 6 1 4 1 0	7
19	9 6 0 9 3 4	6
20	7 1 0 8	4
21	8	1
22	1 8 9	3
23	7	1
24	5	1

Stem: Tens and hundreds digits (psi); Leaf: Ones digits (psi).