



PABLO FIGUEROA

MINERÍA DE DATOS

ESTADÍSTICA DESCRIPTIVA

FRECUENCIAS

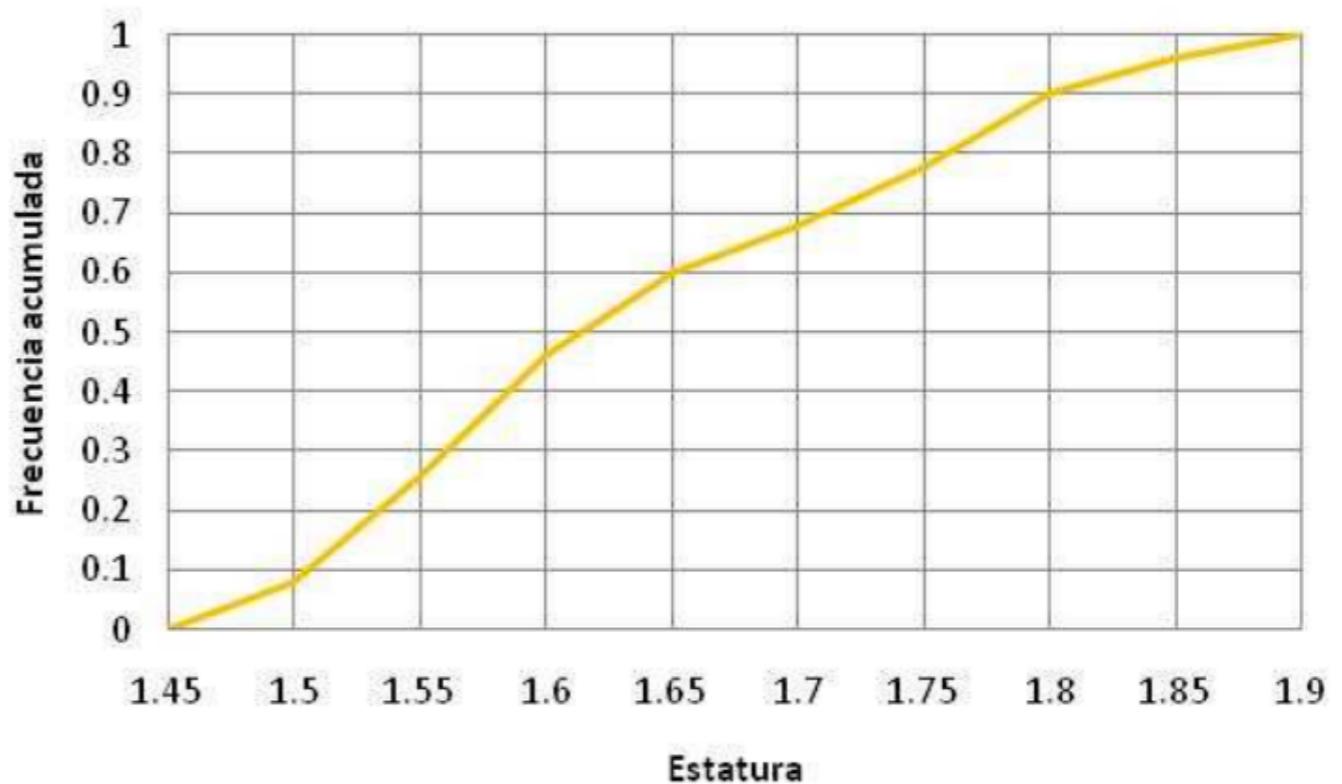
x	f	f_r	%	F	F_r	% acumulado
1	6	0.0400	4.00	6	0.0400	4.00
2	11	0.0733	7.33	17	0.1133	11.33
3	12	0.0800	8.00	29	0.1933	19.33
4	30	0.2000	20.00	59	0.3933	39.33
5	40	0.2667	26.67	99	0.6600	66.00
6	25	0.1667	16.67	124	0.8267	82.67
7	14	0.0933	9.33	138	0.9200	92.00
8	9	0.0600	6.00	147	0.9800	98.00
9	3	0.0200	2.00	150	1.0000	100.00
Total:	150	1.0000	100.00			

GRÁFICAS

Tipo de variable	Diagrama o gráfico
Cualitativa	Barras, sectores, pictogramas
Cuantitativa (discreta)	Diferencial (barras) Integral (escalera)
Cuantitativa (continua)	Diferencial (histograma, polígono de frecuencias) Integral (diagramas acumulativos)

GRÁFICAS

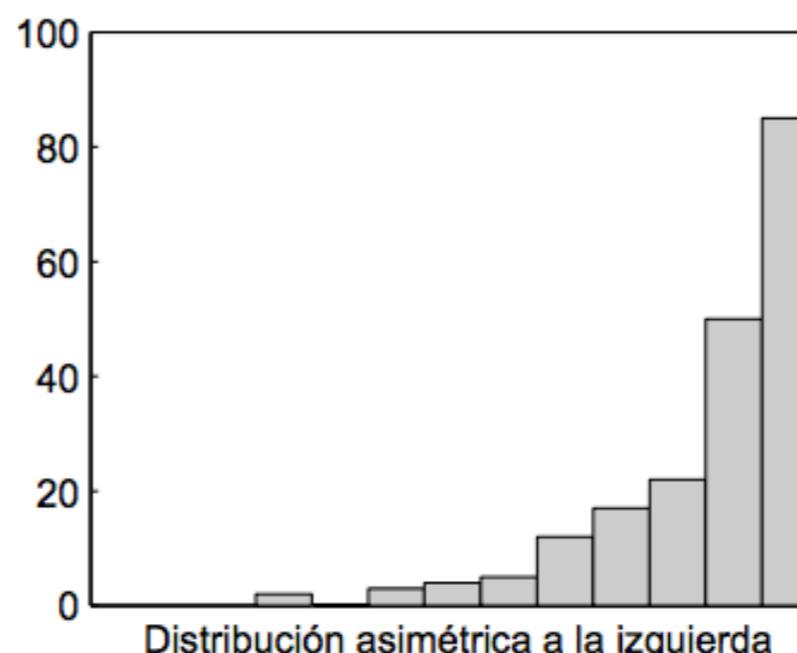
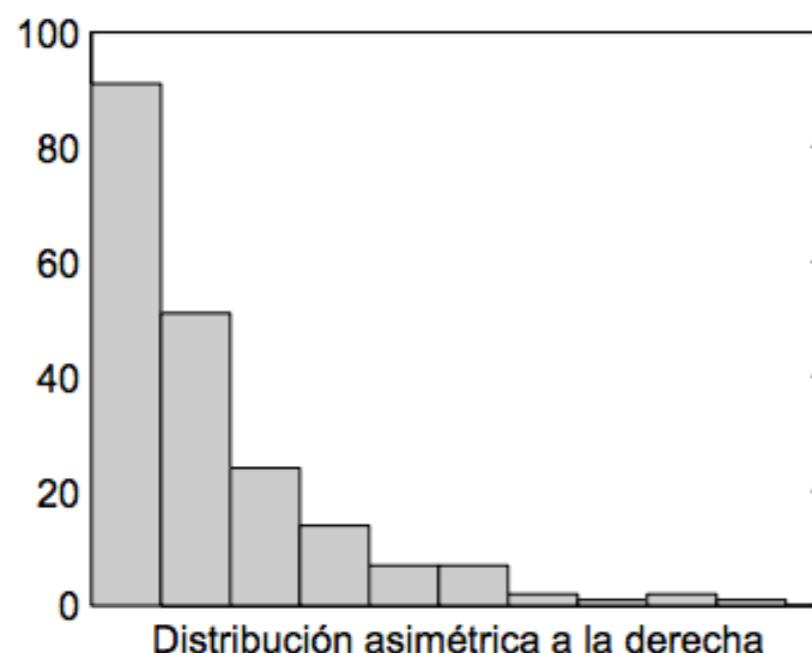
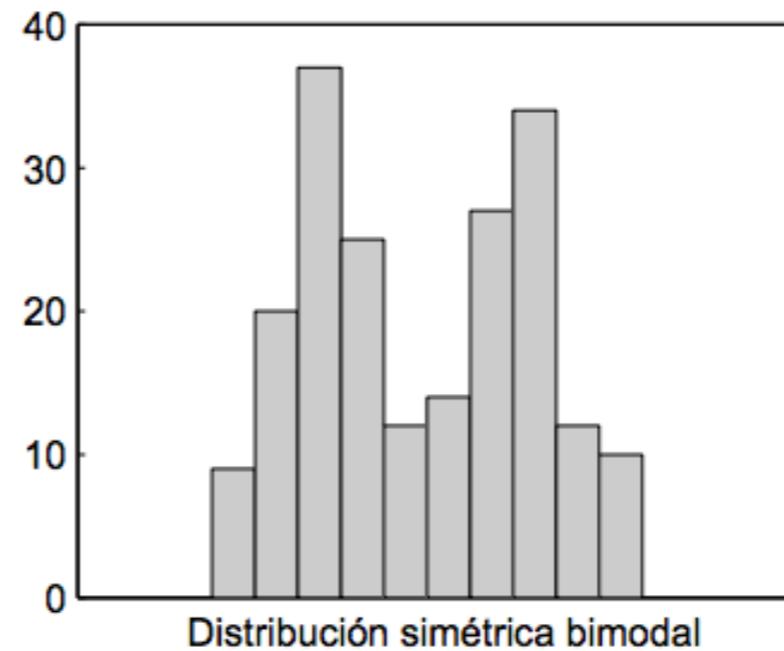
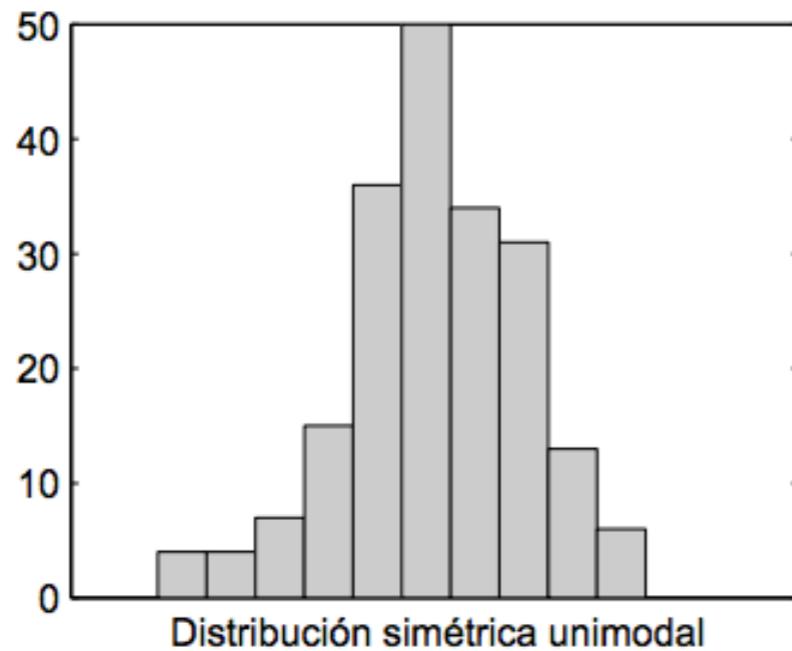
Altura (m) menor que	Número de estudiantes	Frecuencia acumulada
1.45	0	0
1.50	4	0.08
1.55	13	0.26
1.60	23	0.46
1.65	30	0.6
1.70	34	0.68
1.75	39	0.78
1.80	45	0.9
1.85	48	0.96
1.90	50	1



GRÁFICAS

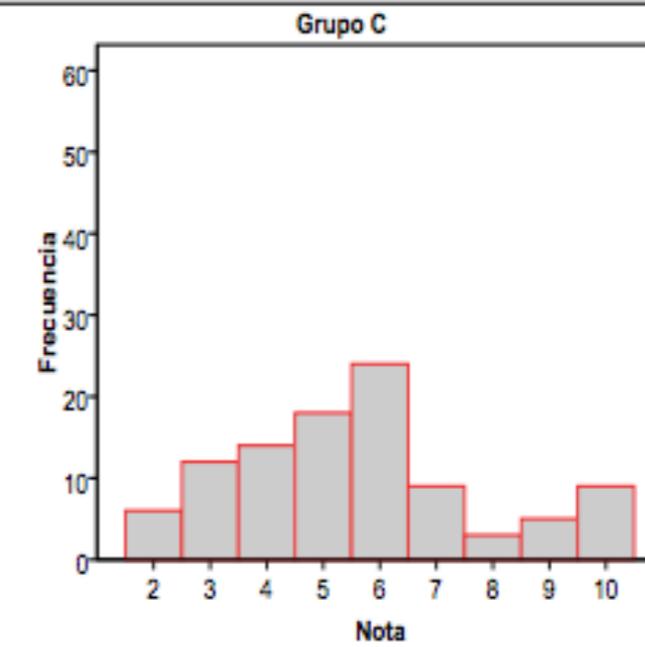
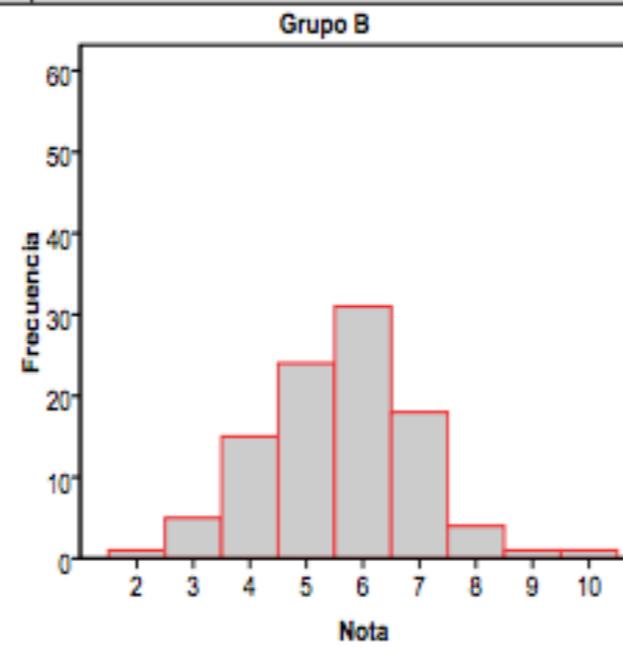
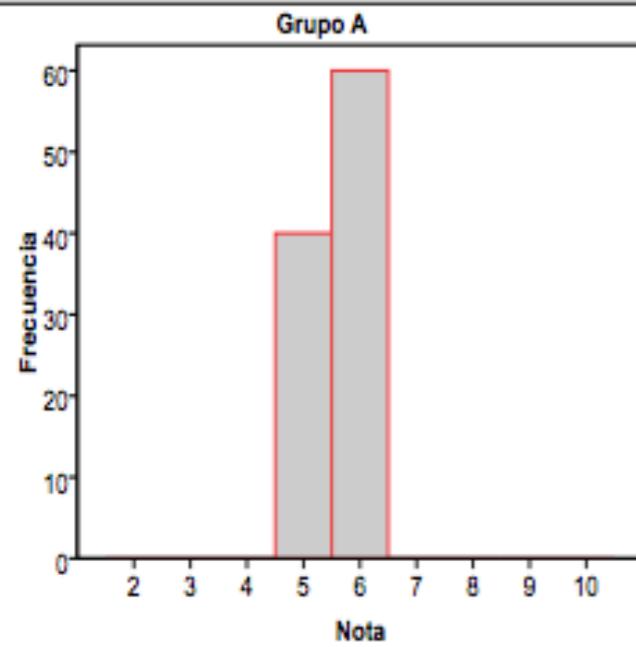
- Normalmente la base de todos los rectángulos es la misma por lo que la altura es proporcional a la frecuencia.
- Identificar si se han usado frecuencias absolutas o relativas.
- ¿Cuántas modas hay?
- ¿Hay algún dato atípico en relación al resto?
- ¿Es simétrica la distribución?
- En caso de asimetría, ¿es asimétrica a la izquierda o a la derecha
- ¿En torno a qué valor aproximado están centrados los datos?
- ¿Están muy dispersos los datos en torno a este centro o muy concentrados?

GRÁFICAS



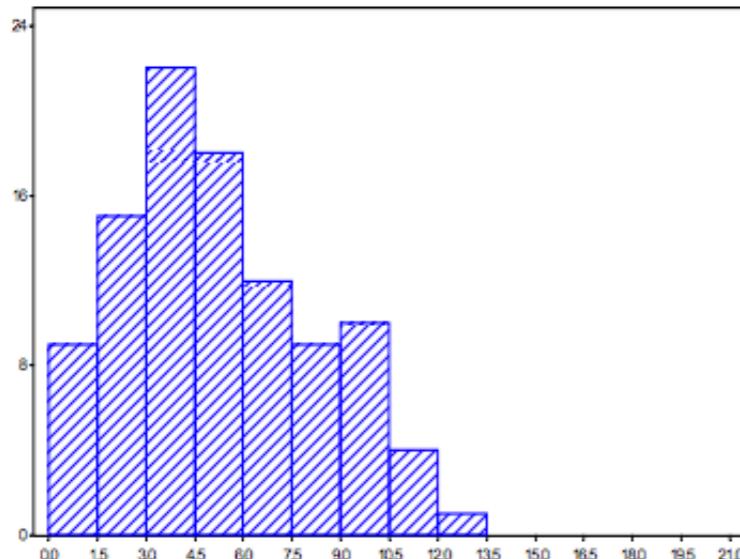
GRÁFICAS

	Nota obtenida								
	2	3	4	5	6	7	8	9	10
Nº alumnos grupo A	0	0	0	40	60	0	0	0	0
Nº alumnos grupo B	1	5	15	24	31	18	4	1	1
Nº alumnos grupo C	6	12	14	18	24	9	3	5	9

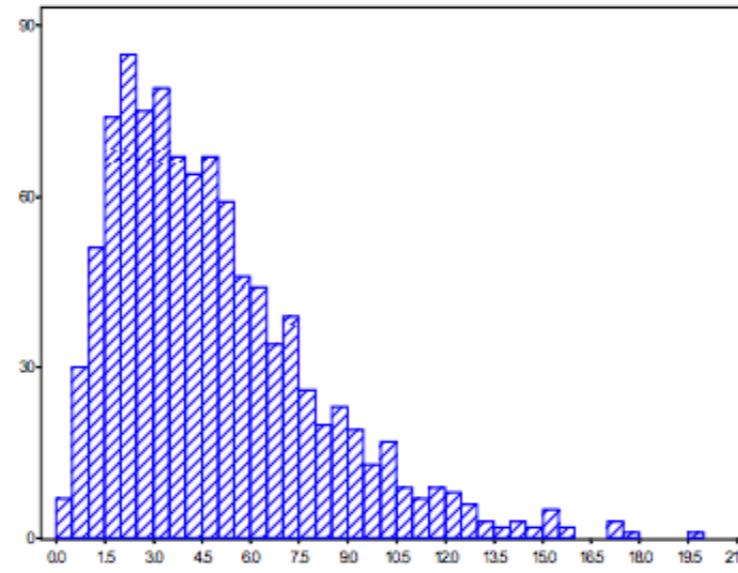


GRÁFICAS

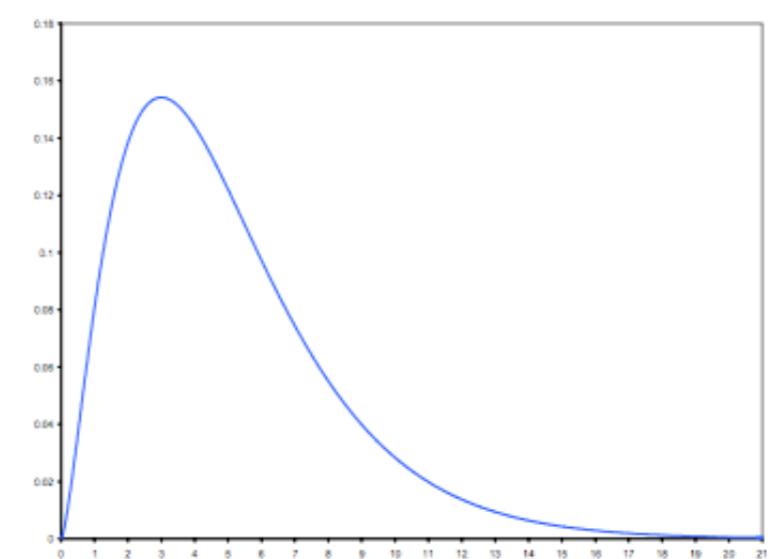
Figura 12. Histogramas para variables continuas.



Muestra $n = 100$

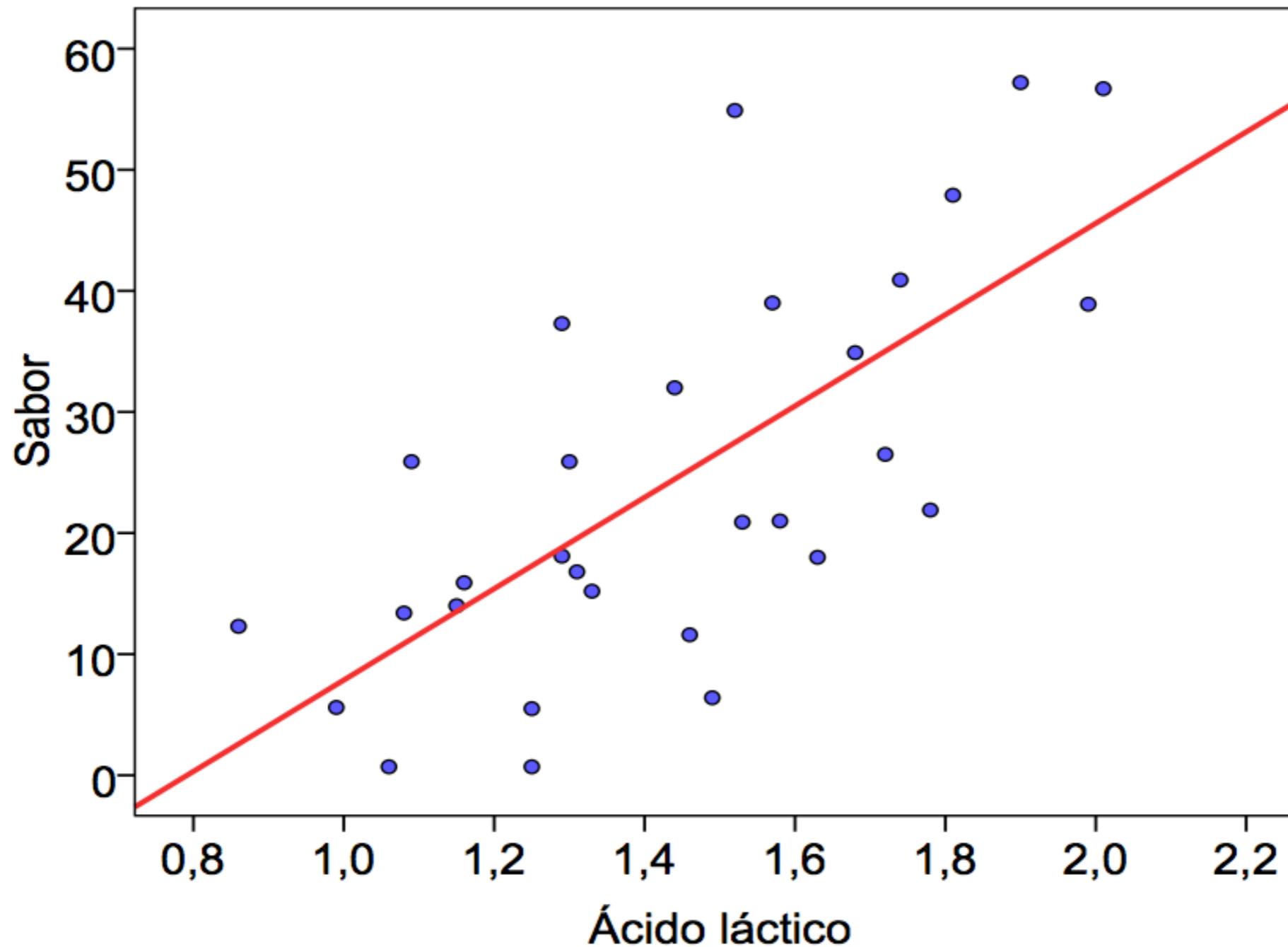


Muestra $n = 1000$

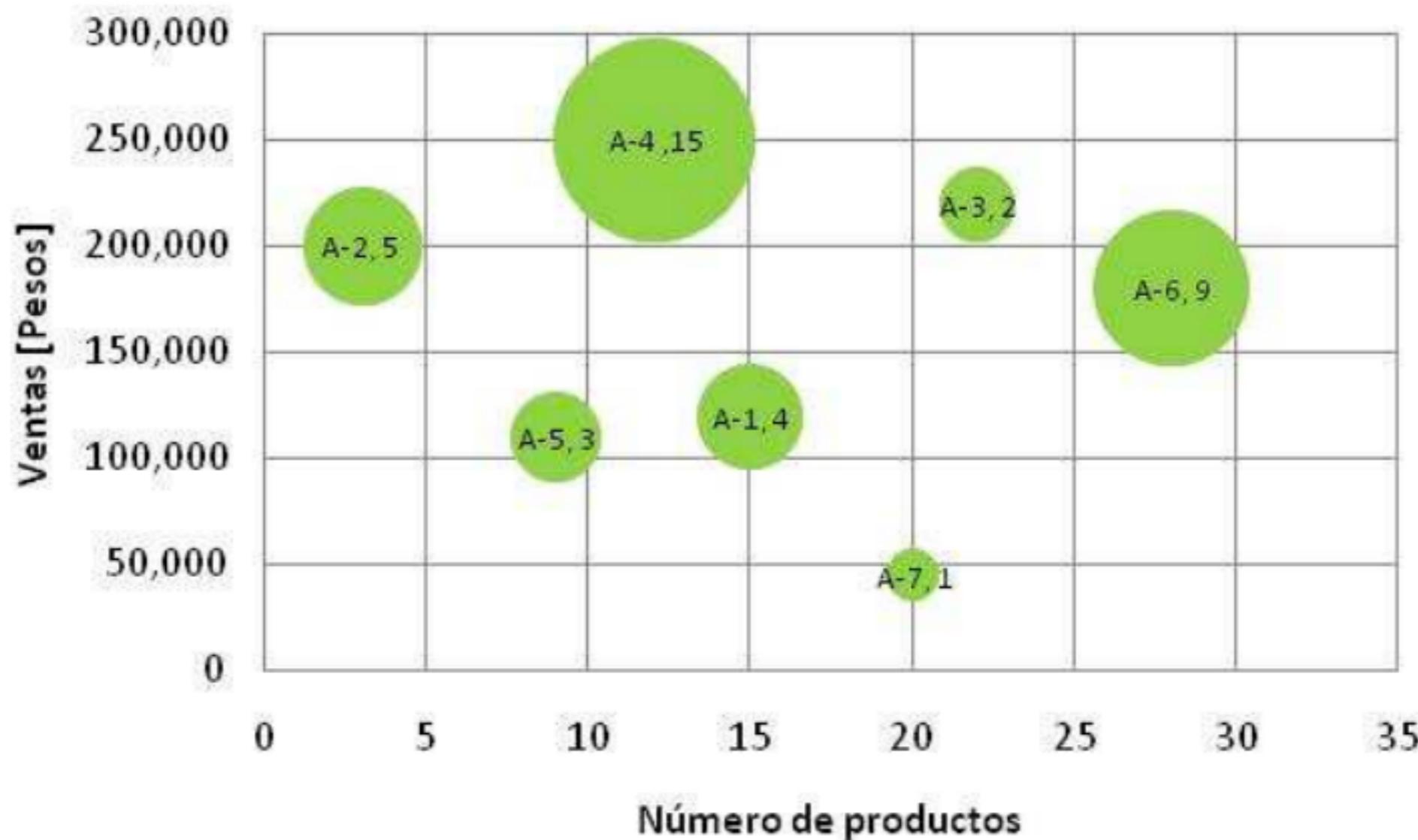


Población

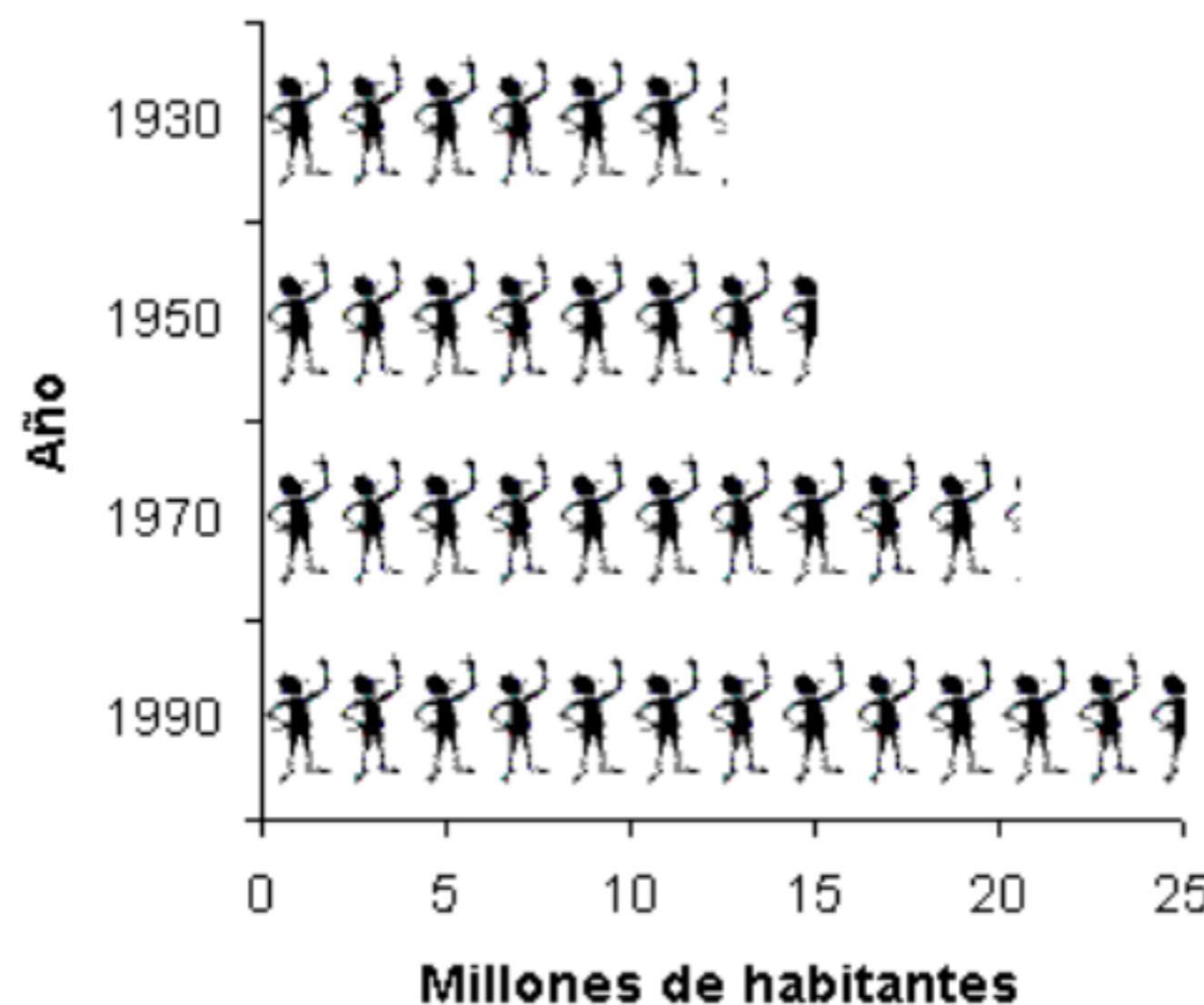
GRÁFICAS



GRÁFICAS



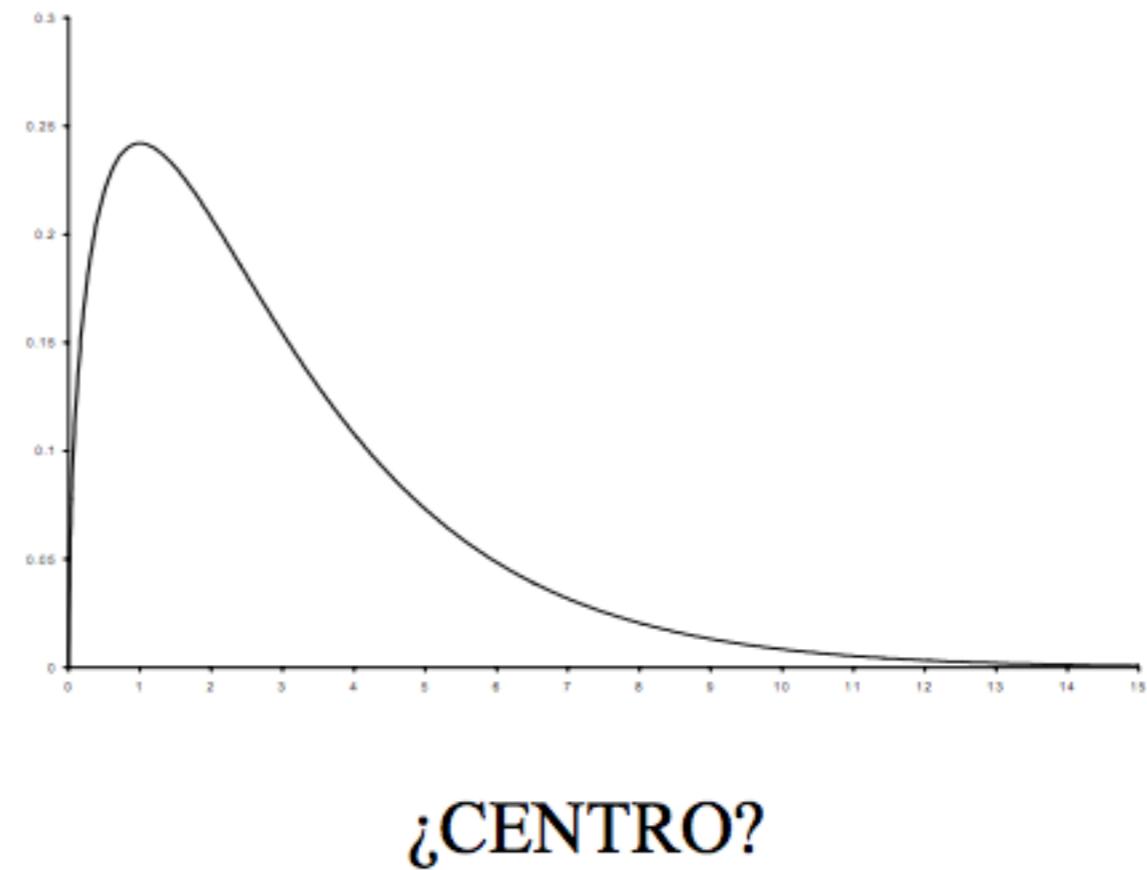
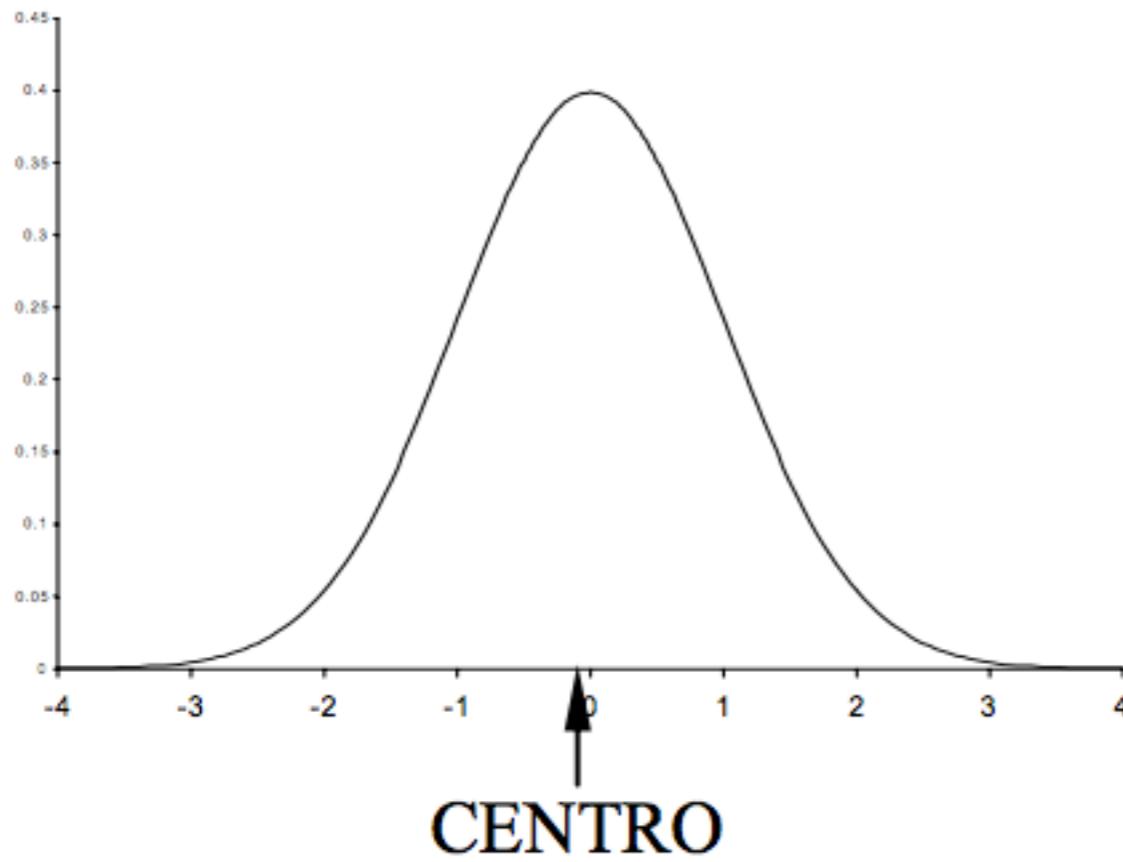
GRÁFICAS



MEDIDAS RESUMEN

Medidas de posición o localización \Rightarrow describen un valor alrededor del cual se encuentran las observaciones

Medidas de dispersión o escala \Rightarrow pretenden expresar cuan variable es un conjunto de datos.



MEDIDAS RESUMEN - POSICIÓN

Si tenemos una muestra de n observaciones y denotadas por X_1, X_2, \dots, X_n , definimos la *media muestral* \bar{X} del siguiente modo:

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{\sum_{i=1}^n X_i}{n}$$

El símbolo $\sum_{i=1}^n X_i$ indica la suma de todos los valores obesrvados de la variable desde el primero ($i = 1$) hasta el último ($i = n$).

Ejemplo.

$$X_1 = 10 \quad X_2 = 14 \quad X_3 = 12 \quad X_4 = 11 \quad X_5 = 12 \quad X_6 = 13$$

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_6}{n} = \frac{10+14+12+11+12+13}{6} = \frac{72}{6} = 12$$

MEDIDAS RESUMEN - POSICIÓN

Media poblacional

Si se dispone de la información de una variable X para las N unidades de análisis de la población, es posible calcular la *media poblacional* a la que denotaremos con la letra griega μ (mu), para distinguirla de la media obtenida en una muestra de n

$$\mu = \frac{X_1 + X_2 + \dots + X_N}{N} = \frac{\sum_{i=1}^N X_i}{N}.$$

Media de datos agrupados

Supongamos que se dispone de dos conjuntos de datos en los que se conoce la media y el número de datos de cada uno de ellos (\bar{X}_1, n_1 y \bar{X}_2, n_2). Calculamos la media de los $n_1 + n_2$ datos como el *promedio pesado*

$$\bar{X} = \frac{n_1 \bar{X}_1 + n_2 \bar{X}_2}{n_1 + n_2}$$

CARACTERÍSTICAS Y PROPIEDADES DE LA MEDIA

- ▶ Se usa para datos numéricos.
- ▶ Representa el centro de gravedad o punto de equilibrio de los datos.
- ▶ La suma de las distancias de los datos a la media es cero.
- ▶ Es muy sensible a la presencia de valores atípicos (Outliers).

MEDIANA MUESTRAL

La *mediana* es el dato que ocupa la posición central en la muestra ordenada de menor a mayor.

¿Cómo calculamos la mediana de una muestra de n observaciones?

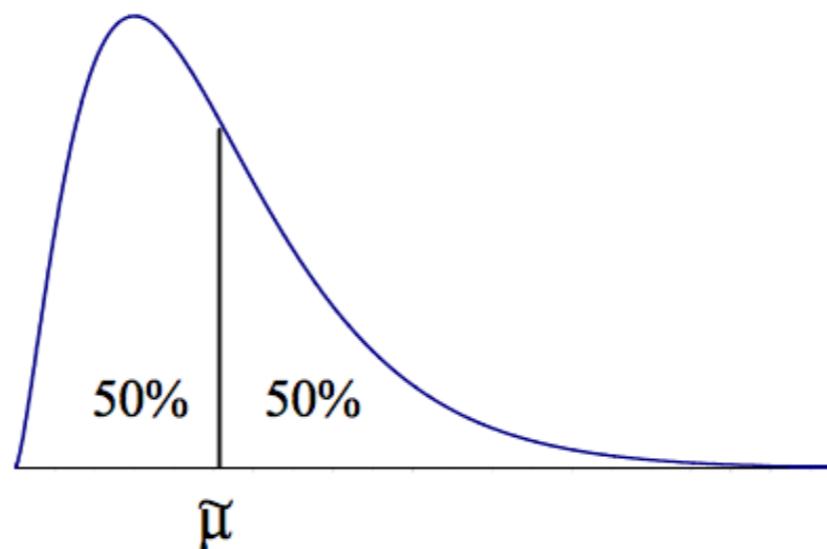
1. Ordenamos los datos de menor a mayor.
2. La mediana es el dato que ocupa la posición $\left(\frac{n+1}{2}\right)$ en la lista ordenada.

Si el número de datos es *ímpar*, la mediana \tilde{X} es el dato que ocupa la posición central.

Si el número de datos es *par*, la mediana \tilde{X} es el promedio de los dos datos centrales.

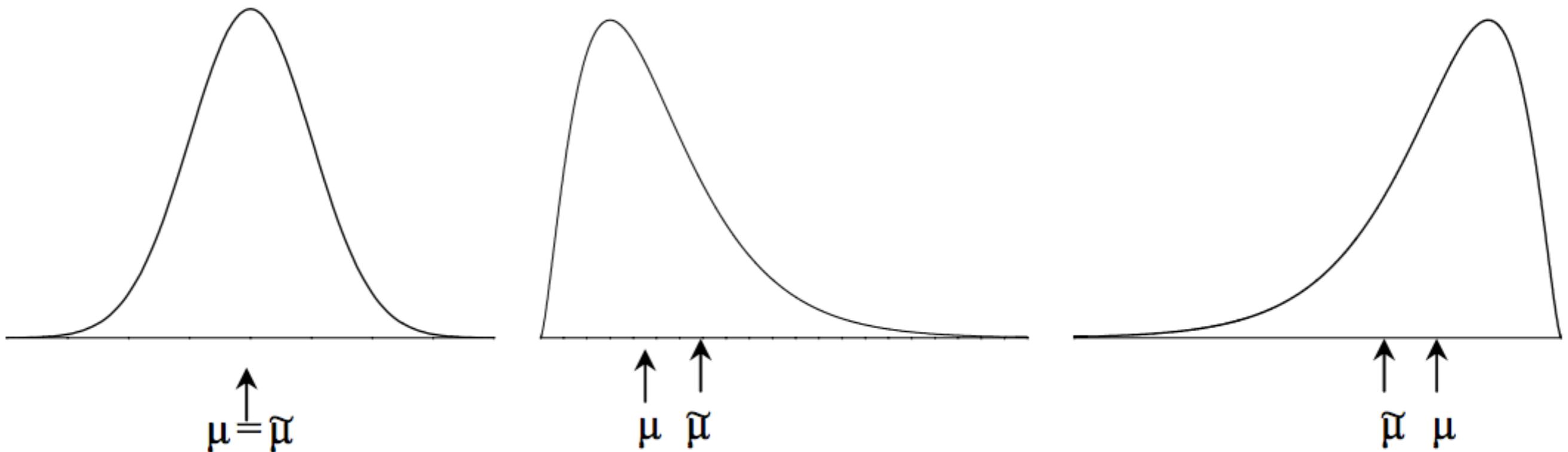
MEDIANA POBLACIONAL

La *mediana poblacional* se define de modo equivalente a la mediana muestral y es el valor de la variable por debajo del cual se encuentra a lo sumo el 50% de la población y por encima del cual se encuentra a lo sumo el 50% de la población. La denotamos como $\tilde{\mu}$.



PROPIEDADES DE LA MEDIANA

- ▶ La mediana puede ser usada no sólo para datos numéricos sino además para datos ordinales, ya que para calcularla solo es necesario establecer un orden de los datos.
- ▶ La media y la mediana difieren según la distribución de los datos:



PROPIEDADES DE LA MEDIANA

- ▶ La mediana es una medida de posición robusta. No se afecta por la presencia de datos outliers, salvo que modifiquemos casi el 50% de los datos menores o mayores de la muestra (la proporción de datos que debemos modificar para afectar la mediana depende del número de datos de la muestra).

Ejemplo

I)	10	11	12	12	13	14
II)	10	11	12	12	13	26

$$\begin{array}{ll} \bar{x} = 12 & \tilde{x} = 12 \\ \bar{x} = 14 & \tilde{x} = 12 \end{array}$$

PROPIEDADES DE LA MEDIANA

- ▶ La mediana es insensible a la distancia de las observaciones al centro, ya que solamente depende del orden de los datos. Esta característica que la hace robusta, es una desventaja de la mediana.

Ejemplo. Todos los conjuntos de datos siguientes tienen mediana 12

I)	10	11	12	13	14
II)	10	11	12	13	100
III)	0	11	12	12	12
IV)	10	11	12	100	100

PROPIEDADES DE LA MEDIANA

- ▶ Si hay datos censurados no es posible calcular la media, sin embargo, eventualmente puede calcularse la mediana.

Comparación de la media y la mediana

	MEDIA	MEDIANA
VENTAJAS	Usa toda la información que proveen los datos. Es de manejo algebraico simple.	Representa el centro de la distribución (en un sentido claramente definido). Robusta a la presencia de outliers. Útil para datos ordinales.
DESVENTAJAS	Muy sensible a la presencia de datos outliers.	Usa muy poca información de los datos.

MEDIA ALPHA PODADA

La media α -podada es un compromiso entre las dos medidas de posición presentadas. Es una medida más robusta que la media, pero que usa más información que la mediana.

La media α -podada se calcula despreciando $n.\alpha$ datos de cada extremo y promediando las observaciones centrales del conjunto ordenado de datos.

¿Cómo calculamos la media α -podada de una muestra de n observaciones?

1. Ordenamos los datos de menor a mayor.
2. Excluimos los $n.\alpha$ datos más pequeños y los $n.\alpha$ datos más grandes.
3. Calculamos el promedio de los datos restantes y lo denominamos \bar{X}_α .

MEDIA ALPHA PODADA

¿Cómo elegimos α ?

Depende de cuantos outliers se pretende excluir y de cuán robusta queremos que sea la medida de posición. Cuando seleccionamos $\alpha = 0$ tenemos la media, si elegimos el máximo valor posible para α (lo más cercano posible a 0.5) tenemos la mediana. Cualquier poda intermedia representa un compromiso entre ambas.

Una elección bastante común es $\alpha = 0.10$, que excluye un 20% de los datos.

¿Cuándo usar esta medida?

Cuando se sospecha que hay errores groseros en los datos, pero no tenemos modo de decidir si el dato es erróneo. Esto permite excluir datos aberrantes de un modo menos sesgado, porque estamos excluyendo datos de ambos extremos.

MEDIA ALPHA PODADA

Ejemplo

Calculamos la media 20% podada para los datos siguientes que corresponden a los puntajes asignados a una gimnasta por 5 jueces durante una competencia olímpica.

$$X_1 = 85 \quad X_2 = 98 \quad X_3 = 99 \quad X_4 = 95 \quad X_5 = 98$$

1. Ordenamos los datos: 85 95 98 98 99
2. Calculamos el número de datos que podaremos en cada extremo

$$n \cdot \alpha = 5 \cdot 0.20 = 1$$

Excluimos el primer y el último dato de la muestra ordenada.

3. Promediamos los datos restantes

$$\bar{X}_{0.20} = \frac{95 + 98 + 98}{3} = 97.$$

Para estos datos el promedio y la mediana resulta ser $\bar{X} = 95$, $\tilde{X} = 98$.

LA MODA

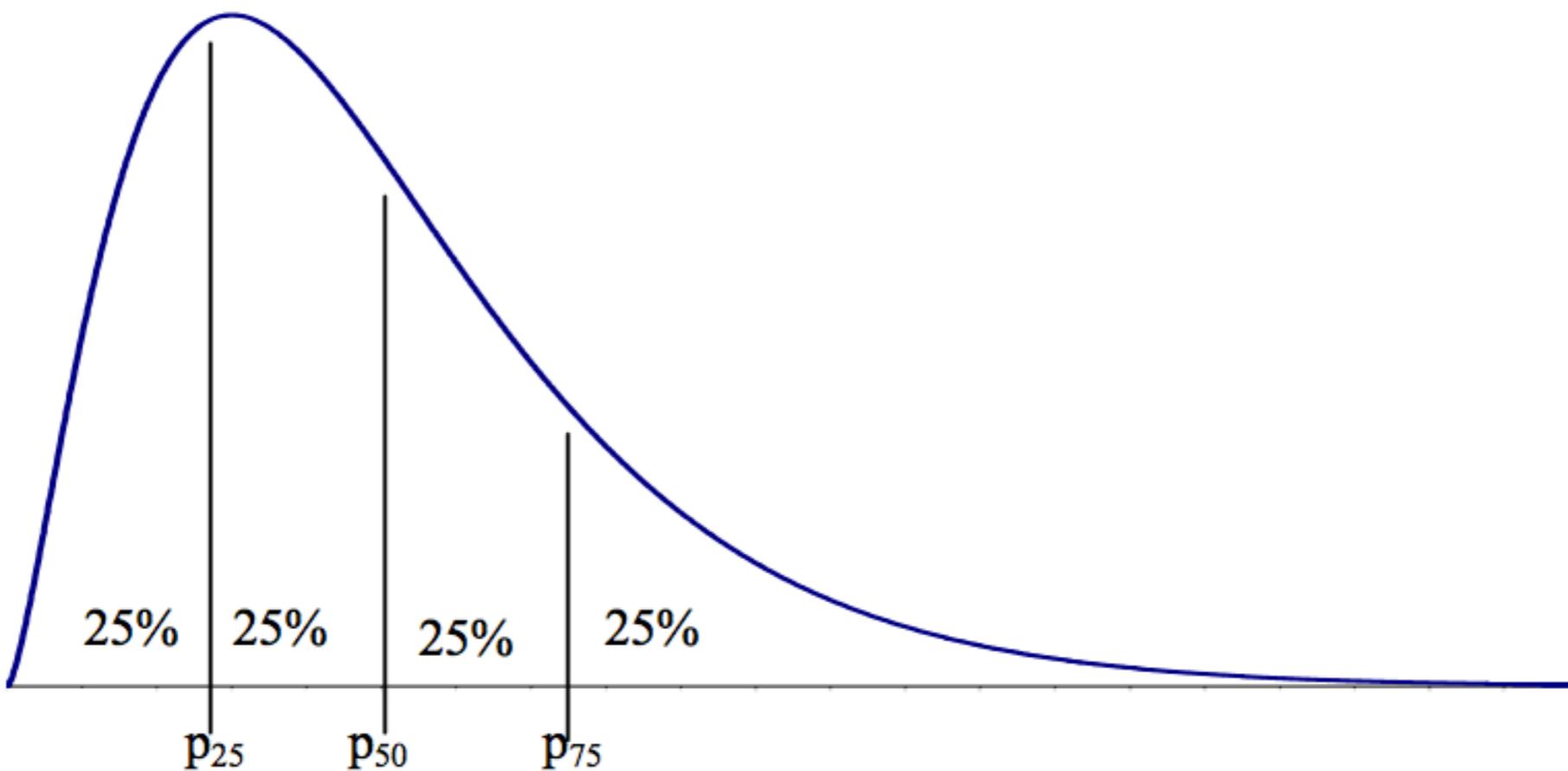
La moda es el dato que ocurre con mayor frecuencia en el conjunto.

Es una medida de poca utilidad salvo para datos categóricos en los que suele interesar identificar la categoría con mayor cantidad de datos. En una muestra de datos numéricos, puede ocurrir que la moda sea un valor que se repite un cierto número de veces, pero que no es típico.

Cuando se considera la distribución poblacional de una variable continua, decimos que esta es **UNIMODAL** si presenta un pico y **BIMODAL** si aparecen dos picos claros.

CUARTILES Y PERCENTILES

La mediana es el percentil 50%. Otros percentiles con nombre propio son el percentil 25% y el percentil 75% que se denominan *cuartil inferior* y *superior* respectivamente, ya que juntamente con la mediana dividen a la distribución en 4 porciones iguales.



MEDIDAS RESUMEN

¿Cómo se calculan los cuartiles de una muestra de n observaciones?

1. Ordenar los datos de menor a mayor.
2. El cuartil inferior es el dato que ocupa la posición $(n+1)/4$ en la muestra ordenada.
3. El cuartil superior es el dato que ocupa la posición $3(n+1)/4$ en la muestra ordenada.

Si la posición resulta ser un número decimal, promediamos los datos que se encuentran a izquierda y derecha de la posición obtenida.

Ejemplo

Consideremos los siguientes datos ordenados ($n = 13$).

Posición	1	2	3	4	5	6	7	8	9	10	11	12	13
Datos	104	112	134	146	155	168	170	195	246	302	338	412	678

$$\text{Posición del Cuartil Inferior} = (13+1)/4 = 3.5 \quad \Rightarrow C_I = \frac{134 + 146}{2} = 140$$

$$\text{Posición de la mediana} = (13+1)/2 = 7 \quad \Rightarrow \tilde{X} = 170$$

$$\text{Posición del Cuartil Superior} = 3.(13+1)/4 = 10.5 \quad \Rightarrow C_S = \frac{302 + 338}{2} = 320$$

MEDIDAS RESUMEN - DISPERSIÓN

Las medidas de posición dan una idea de dónde se encuentra el centro de la distribución, pero no nos dicen cuán disperso es el conjunto de datos. Consideremos los siguientes conjuntos de datos:

Muestra A:	55	55	55	55	55	55	55
Muestra B:	47	51	53	55	57	59	63
Muestra C:	39	47	53	55	57	63	71

En todos ellos $\bar{X} = \tilde{X} = 55$, pero las muestras difieren notablemente.

Las *medidas de dispersión o variabilidad* describen cuán cercanos se encuentran los datos entre ellos, o cuán cerca se encuentran de alguna medida de posición. Introduciremos a continuación algunos estadísticos que miden variabilidad del conjunto de datos.

RANGO

El *rango* de n observaciones X_1, X_2, \dots, X_n es la diferencia entre la observación más grande y la más pequeña,

$$\text{Rango} = \max(X_i) - \min(X_i)$$

Ejemplo

Muestra A: 55 55 55 55 55 55 55 Rango = $55 - 55 = 0$

Muestra B: 47 51 53 55 57 59 63 Rango = $63 - 47 = 16$

Muestra C: 39 47 53 55 57 63 71 Rango = $71 - 39 = 32$

Características y propiedades

- Es muy simple de obtener.
- Es extremadamente sensible a la presencia de datos atípicos. Si hay datos outliers, estos estarán en los extremos, que son los datos que se usan para calcular el rango.
- Ignora la mayoría de los datos.
- En general aumenta cuando aumenta el tamaño de la muestra (las observaciones atípicas tienen más chance de aparecer en una muestra con muchas observaciones).

En consecuencia, reportar el rango o el máximo y el mínimo de un conjunto de datos, no informa demasiado sobre las características de los datos. A pesar de esto es frecuente encontrar en las publicaciones científicas datos numéricos resumidos a través de una medida de posición acompañada por los valores mínimo y máximo.

DESVIACIÓN ESTÁNDAR Y VARIANZA MUESTRAL

La *desviación estándar* mide cuan lejos se encuentran los datos de la media muestral.

Definimos la *varianza* de una muestra de observaciones X_1, X_2, \dots, X_n , cuya media es \bar{X} , como

$$s^2 = \frac{(X_1 - \bar{X})^2 + \dots + (X_n - \bar{X})^2}{n-1} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}.$$

La varianza muestral puede pensarse como “promedio” de las distancias a la media al cuadrado.

Sin embargo, la varianza no tiene las mismas unidades que los datos. Para salvar este inconveniente, definimos la *desviación estándar muestral* como la raíz cuadrada positiva de la varianza

$$s = \sqrt{s^2}.$$

DESVIACIÓN ESTÁNDAR Y VARIANZA POBLACIONAL

Si se dispone de la información de una variable X para las N unidades de análisis de la población, denotamos con σ^2 y σ (sigma) *la varianza y la desviación estandar de la población* respectivamente y las definimos del siguiente modo:

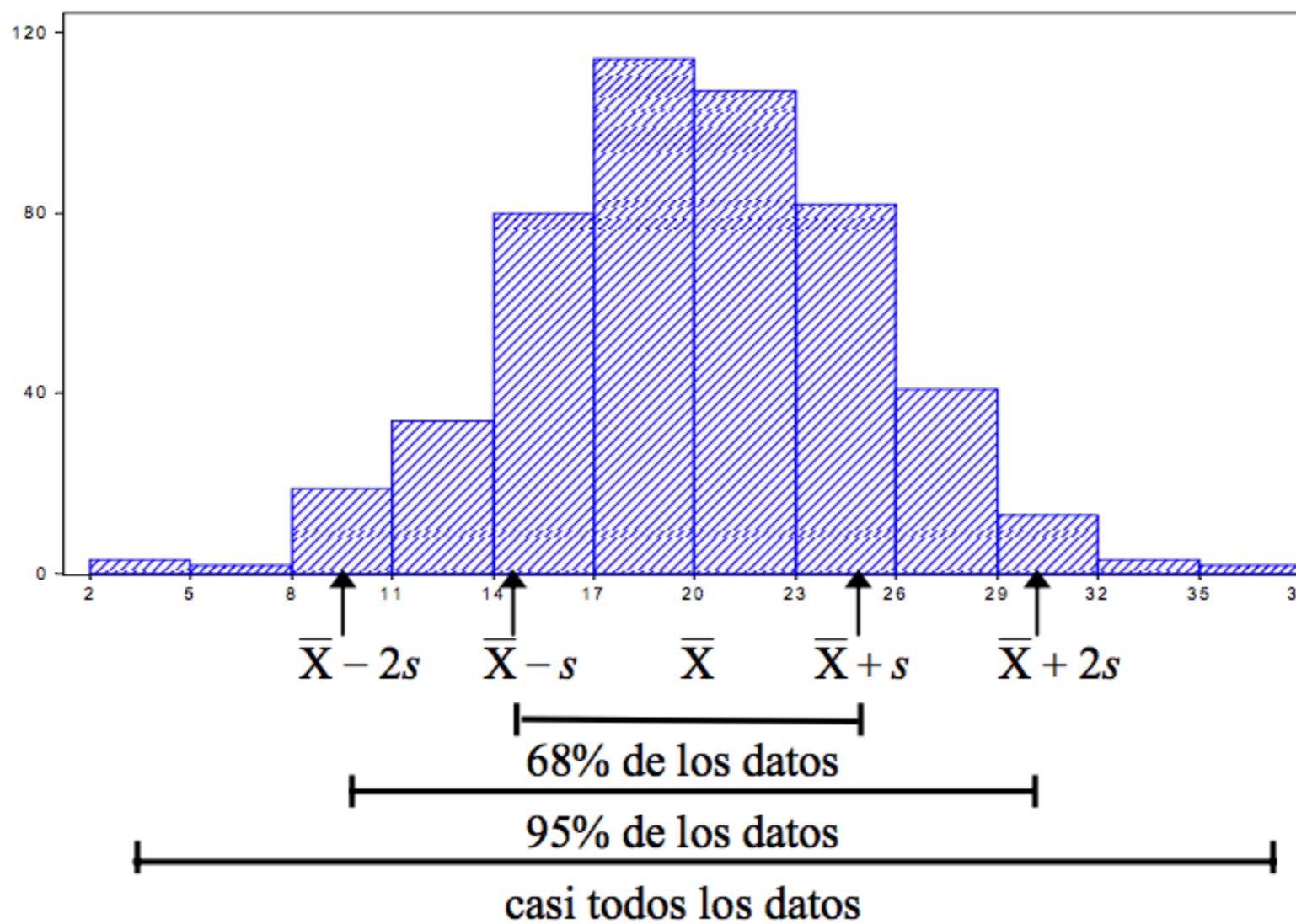
$$\sigma^2 = \frac{(X_1 - \mu)^2 + \dots + (X_n - \mu)^2}{N} = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N} \quad \sigma = \sqrt{\sigma^2}$$

La razón para usar $(n - 1)$ y no n en el denominador de la varianza muestral tiene que ver con el hecho de que el valor de s^2 obtenido en una muestra, se usa para estimar la varianza poblacional σ^2 . Definida con $(n - 1)$ en el denominador la varianza muestral posee una propiedad deseable, resulta ser *insesgado*, esto es, en promedio no subestima ni sobreestima el valor de la varianza poblacional.

INTERPRETACIÓN DE LA DESVIACIÓN ESTÁNDAR

Si el histograma de los datos es aproximadamente simétrico y acampanado entonces,

- Aproximadamente el 68% de las observaciones caen en el intervalo $\bar{X} - s$ y $\bar{X} + s$.
- Aproximadamente el 95% de las observaciones caen en el intervalo $\bar{X} - 2s$ y $\bar{X} + 2s$.
- Prácticamente todas las observaciones caen en el intervalo $\bar{X} - 3s$ y $\bar{X} + 3s$.



PROPIEDADES DE LA DESVIACIÓN ESTÁNDAR

- ▶ s mide la dispersión alrededor de la media, por lo tanto es natural elegir esta medida de dispersión cuando se usa la media como medida de posición.
- ▶ $s = 0$ solamente cuando todos los datos son iguales, de otro modo $s > 0$.
- ▶ s es una medida de dispersión muy sensible a la presencia de datos outliers. De hecho es más sensible que la media ya que están elevadas al cuadrado.

MEDIANA DE LAS DESVIACIONES ABSOLUTAS (MAD)

La MAD (median absolute deviations) es otra medida de dispersión que pretende dar una idea resumen de “distancias a un punto central” tal como ocurre con el desvío estándar. Pero, ¿en qué difiere del desvío estándar?

- Considera la mediana como punto central de la distribución para calcular las desviaciones.
- Toma el valor absoluto de las desviaciones para eliminar el signo (en vez de elevar al cuadrado como hacemos al calcular el desvío estándar).
- Toma la mediana de las distancias (en vez de promediar como hacemos con s).

Definimos la *MAD* de una muestra X_1, X_2, \dots, X_n como

$$MAD = \text{mediana}(|X_i - \tilde{X}|)$$

¿Cómo calculamos la MAD?

1. Ordenamos los datos de menor a mayor.
2. Calculamos la mediana.
3. Calculamos la distancia de cada dato a la mediana.
4. Despreciamos el signo de las distancias y las ordenamos de menor a mayor.
5. Buscamos la mediana de las distancias sin signo.

PROPIEDADES DE LA MAD

- Si la distribución es acampanada y simétrica la MAD y el desvío estándar s se relacionan del siguiente modo:

$$s \approx 1.48 \text{ MAD}$$

- La MAD es una medida de dispersión muy robusta a la presencia de datos outliers.

Ejemplo

Consideremos los siguientes datos ordenados ($n = 13$).

Posición	1	2	3	4	5	6	7	8	9	10	11	12	13
Datos	104	112	134	146	155	168	170	195	246	302	338	412	678

1. Como $n = 13$ la mediana es el dato que ocupa la posición $(13+1)/2 = 7 \Rightarrow \tilde{X} = 170$.
2. Calculamos las diferencias a la mediana
 $-66, -58, -36, -24, -15, -2, 0, 25, 76, 132, 168, 242, 508$
3. Despreciamos el signo de las distancias y las ordenamos de menor a mayor
 $0, 2, 15, 24, 25, 36, 58, 66, 76, 132, 168, 242, 508$
4. Tenemos $n = 13$ diferencias, por lo tanto la mediana es la diferencia que ocupa el séptimo lugar, en consecuencia

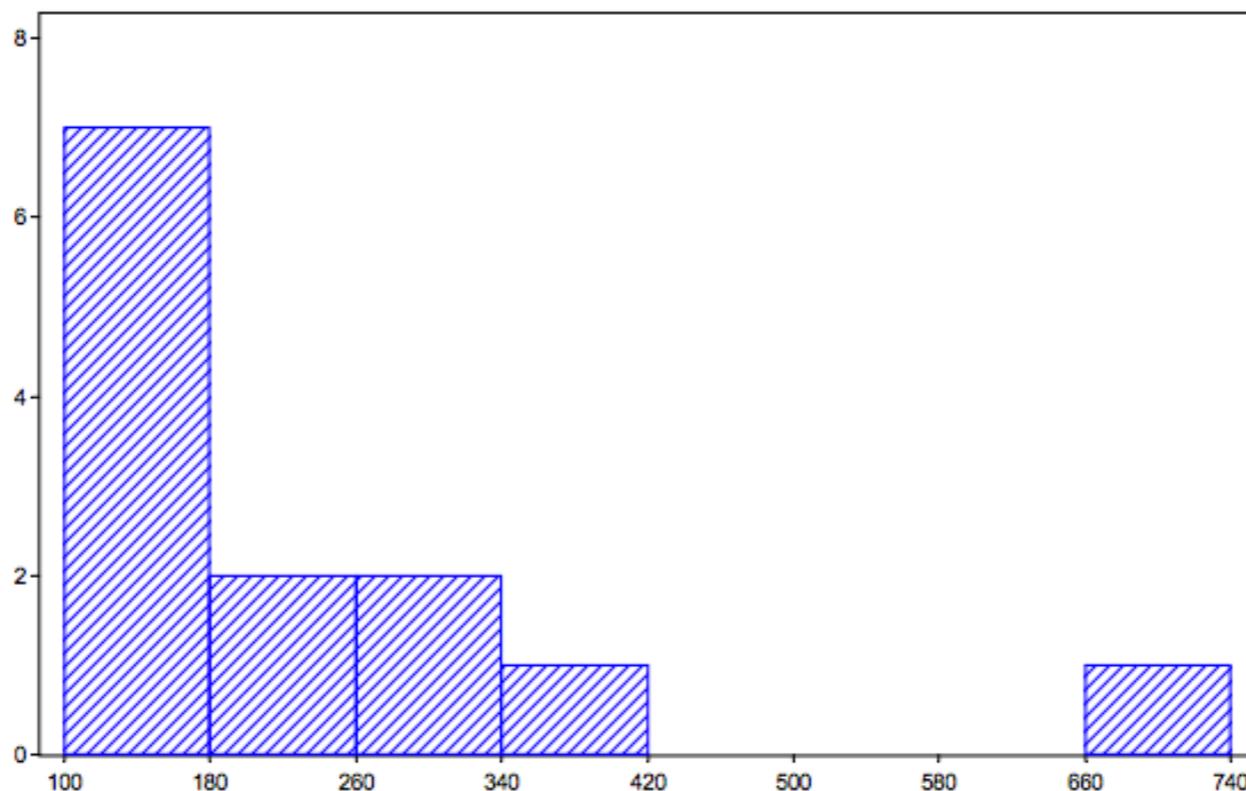
$$\text{MAD} = 58$$

PROPIEDADES DE LA MAD

Si la distribución fuera simétrica esperaríamos que el desvío estándar fuera

$$s \approx 1.48 \text{ MAD} = 1.48 \cdot 58 = 85.8$$

pero para estos datos $s = 160.48$. Esta gran diferencia nos dice que la distribución es asimétrica. El histograma de estos datos, que se presenta en la figura siguiente confirma este hecho.



DISTANCIA O RANGO INTERCUARTIL

El *rango intercuartil o distancia intercuartil* (D_I) de un conjunto de datos es la distancia entre los dos cuartiles:

$$D_I = C_S - C_I$$

Indica el rango donde se encuentra aproximadamente el 50% “central” de las observaciones.

Propiedades

- Si todos los datos son iguales $D_I = 0$. Pero D_I puedes ser igual a cero aún cuando no todos los datos sean iguales.

Ejemplo 5 12 12 12 12 12 20 $n = 7$ $C_I = 12$ $C_S = 12$ $D_I = 0$

- Es una medida robusta de dispersión.
- Cuando la distribución es simétrica y acampanada la relación entre la distancia intercuartil y el desvío estándar es la siguiente

$$D_I \cong \frac{4}{3} s$$

Para distribuciones muy asimétricas $s > D_I$

GRÁFICO DE CAJA BOX-PLOT

¿Cómo se construye un box-plot?

1. Ordenar los datos de menor a mayor
2. Calcular la mediana, el cuartil inferior, el cuartil superior y la distancia intercuartil.
3. Calcular cotas que nos permitirán decidir si un dato es outlier:

- 2^a cota inferior = $C_I - 3 D_I$
- 1^a cota inferior = $C_I - 1.5 D_I$
- 1^a cota superior = $C_S + 1.5 D_I$
- 2^a cota superior = $C_S + 3 D_I$

Cualquier dato que caiga entre la 1^a y 2^a cota inferior o entre la 1^a y 2^a cota superior será declarado *outlier*.

Cualquier dato que caiga por fuera de la 2^a cota inferior o la 2^a cota superior será declarado *outlier severo*.

4. Dibujar una escala que cubra el rango de variación de los datos y marcar la mediana y los cuartiles. Dibujar una caja que se extienda entre los cuartiles y marcar en ella la posición de la mediana.

GRÁFICO DE CAJA BOX-PLOT

5. Partiendo del cuartil inferior trazar una línea (bigote) que llegue hasta el último dato contenido “dentro” de la 1^a cota inferior.
Partiendo del cuartil superior trazar una línea (bigote) que llegue hasta el último dato contenido “dentro” de la 1^a cota superior.
6. Marcar la posición de los outliers con un símbolo (por ejemplo, *) y de los outliers severos con otro símbolo (por ejemplo, ○).

Ejemplo

Consideremos nuevamente los datos siguientes.

Posición	1	2	3	4	5	6	7	8	9	10	11	12	13
Datos	104	112	134	146	155	168	170	195	246	302	338	412	678

De los ejemplos anteriores sabemos que:

$$C_I = 140 \quad \tilde{X} = 170 \quad C_S = 320 \quad D_I = 320 - 140 = 80$$

Calculamos las cotas:

$$2^{\text{a}} \text{ cota inferior} = C_I - 3 D_I = 140 - 3 \cdot 80 = -100$$

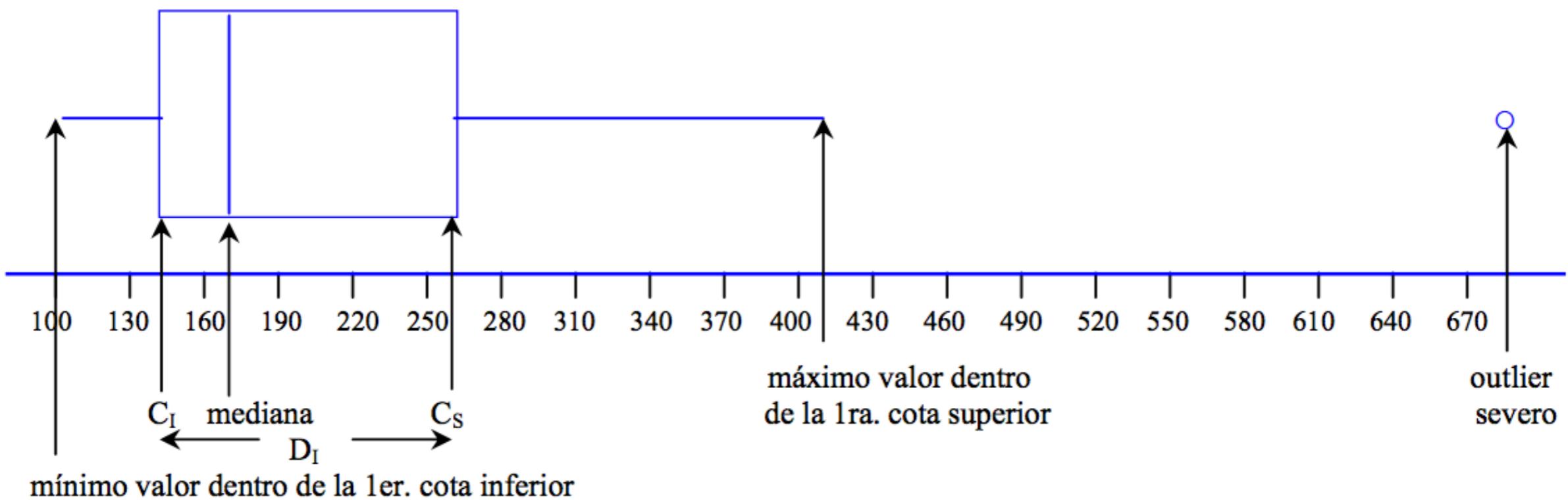
$$1^{\text{a}} \text{ cota inferior} = C_I - 1.5 D_I = 140 - 1.5 \cdot 80 = 20$$

$$1^{\text{a}} \text{ cota superior} = C_S + 1.5 D_I = 320 + 1.5 \cdot 80 = 440$$

$$2^{\text{a}} \text{ cota superior} = C_S + 3 D_I = 320 + 3 \cdot 80 = 580$$

GRÁFICO DE CAJA BOX-PLOT

El gráfico de caja resultante se muestra en la figura siguiente.



¿Qué se observa?

- Un dato outlier.
- La distribución de los datos es asimétrica hacia la derecha, la mitad inferior de los datos se distribuye en un rango mucho menor que la mitad superior.

GRÁFICO DE CAJA BOX-PLOT

¿Qué características de la distribución de los datos se manifiestan en un box-plot?

- Muestra los cinco números resúmenes
- Muestra una medida de posición robusta ⇒ MEDIANA
- Muestra una medida de dispersión robusta ⇒ DISTANCIA INTERCUARTIL
- Permite estudiar la simetría de la distribución
- Nos da un criterio de detección de datos outliers

GRÁFICO DE CAJA BOX-PLOT

¿Qué concluimos a partir de este gráfico?

La satisfacción de los habitantes de las distintas poblaciones difiere en posición (la mediana cambia notablemente) y en dispersión (la población 3 presenta mayor dispersión que las demás).

Las distribuciones tienen diferentes formas (Población 4 muy asimétrica, mucha gente está totalmente de acuerdo con el desempeño de sus gobernantes, mientras que en las demás la distribución es simétrica).

Podemos observar además que el cuartil inferior (percentil 25) del puntaje en la Población 3 es aproximadamente 63 y coincide con el cuartil superior de la Población 2, es decir, en la población 3 el 75% de los encuestados asignaron puntajes de 63 o más, en tanto que en la Población 2 sólo el 25% asignaron puntajes de 63 o más.

Del mismo modo, podemos observar que los encuestados de la Población 4 tienen un grado de satisfacción más alto que prácticamente todos los encuestados en las demás poblaciones.

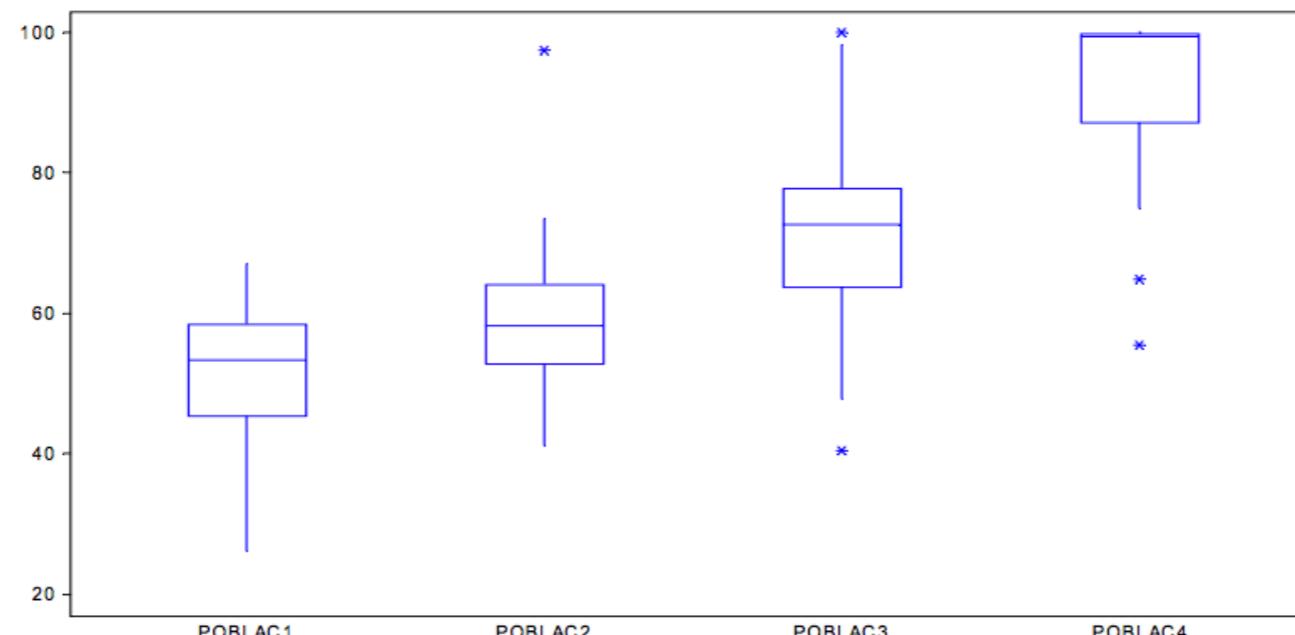
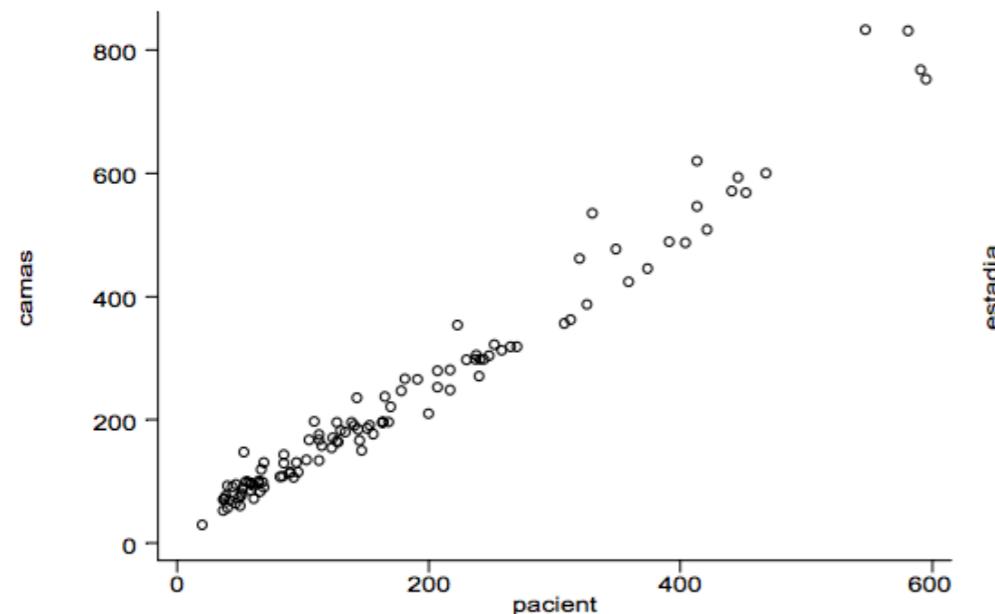
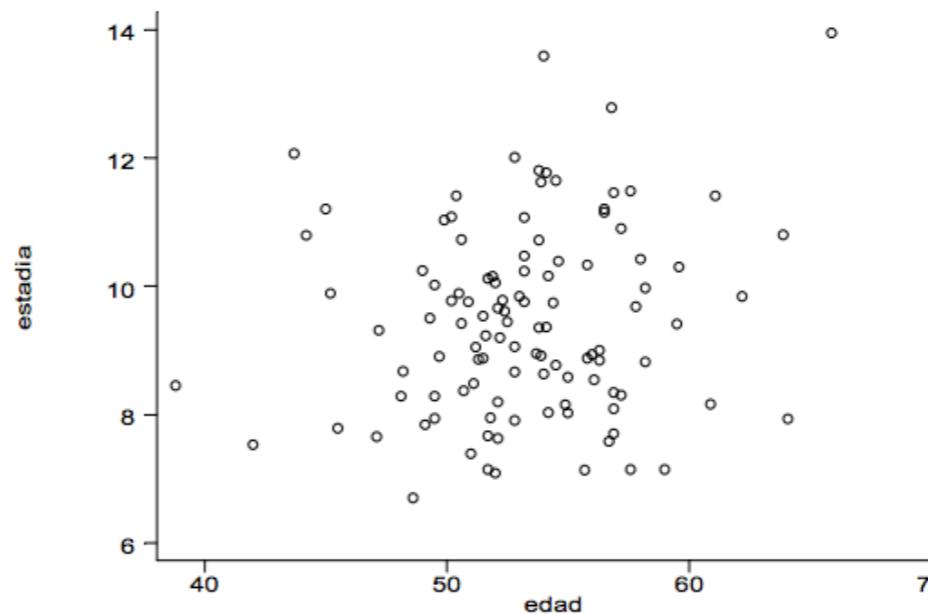


GRÁFICO DE DISPERSIÓN

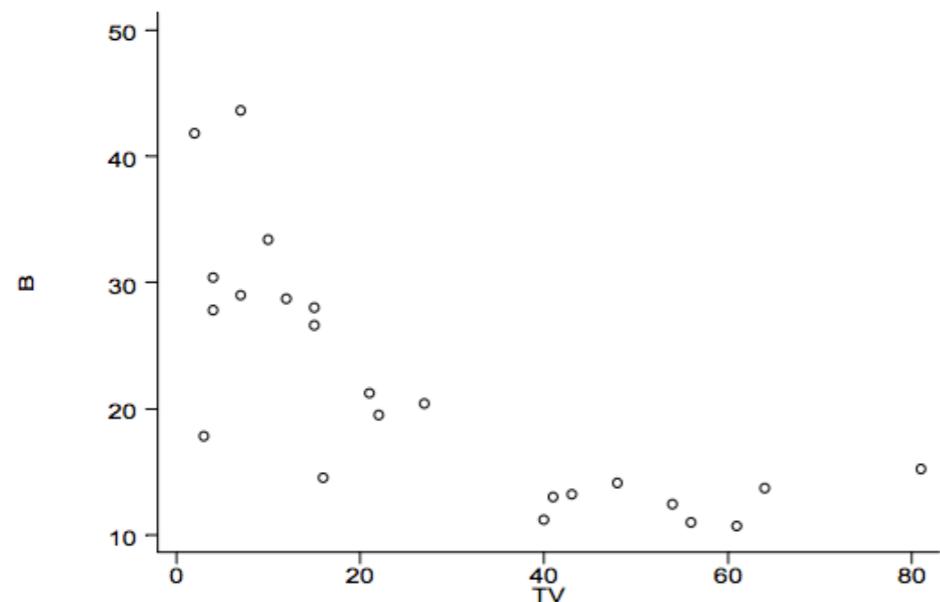
a) Número de pacientes versus número de camas en hospitales.



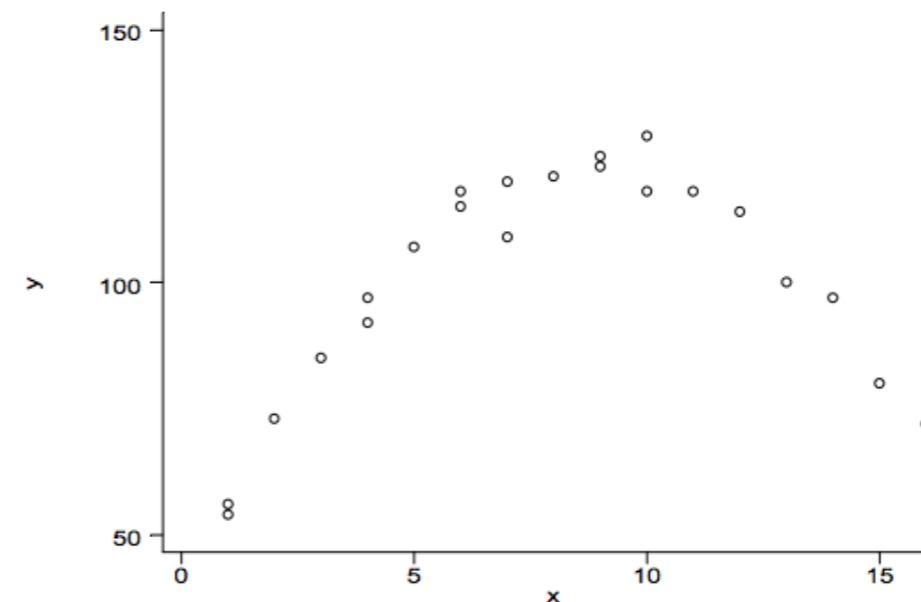
b) Tiempo de internación versus edad del paciente.



a) Tasa de natalidad versus número de aparatos de TV



b) Datos ficticios.



RELACIÓN ENTRE DOS VARIABLES

Al estudiar la relación entre dos variables CUANTITATIVAS. En general interesa:

- ✓ Investigar *si existe asociación* entre las dos variables.
- ✓ Cuantificar la *fuerza de la asociación*, a través de una medida de asociación denominada *coeficiente de correlación*.
- ✓ Estudiar la *forma de la relación* y en lo posible proponer un *modelo matemático* para la relación.
- ✓ *Predecir* una variable a partir de la otra usando el modelo propuesto (REGRESIÓN)

CORRELACIÓN

El *grado de asociación* entre dos variables numéricas puede ser resumido en un estadístico denominado COEFICIENTE DE CORRELACIÓN.

Presentaremos en primer lugar el coeficiente de correlación de Pearson, que mide el grado de asociación lineal entre dos variables y posteriormente un estadístico basado en rangos que estima la correlación sin hacer supuestos sobre el tipo de relación entre las variables.

CORRELACIÓN DE PEARSON

Supongamos que tenemos dos variables (X , Y) registradas en cada una de los n sujetos de una muestra. Sean (X_i, Y_i) las observaciones realizadas para cada variable en el sujeto i -ésimo. Definimos la *covarianza muestral* entre X e Y como:

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

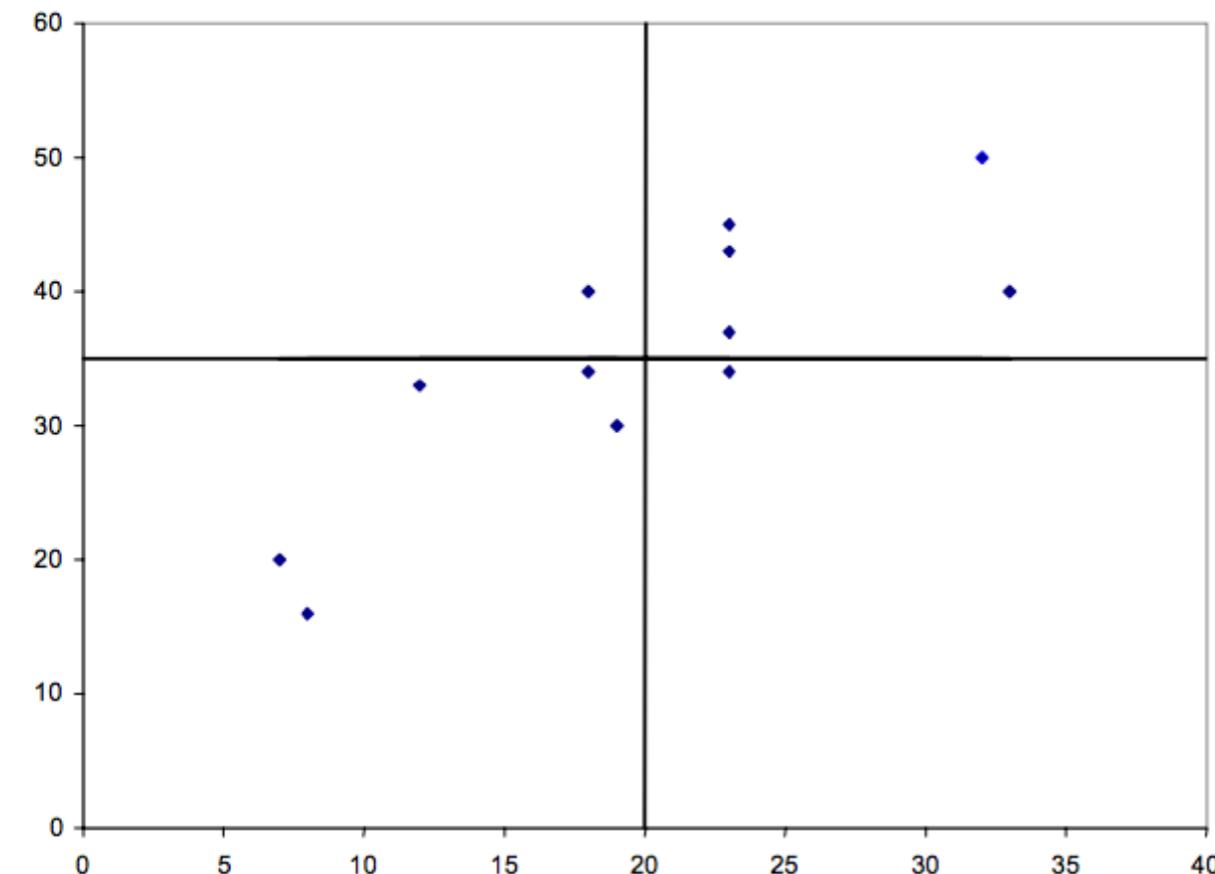
$$\text{donde } \bar{X} = \frac{\sum_{i=1}^n X_i}{n} \quad \text{e} \quad \bar{Y} = \frac{\sum_{i=1}^n Y_i}{n}.$$

La covarianza es el “promedio” de los productos de las desviaciones de las variables respecto de las correspondientes medias.

CORRELACIÓN DE PEARSON

IV

I



III

II

Cuadrante	$(X - \bar{X})$	$(Y - \bar{Y})$	$(X - \bar{X})(Y - \bar{Y})$
I	+	+	+
II	+	-	-
III	-	-	+
IV	-	+	-

CORRELACIÓN DE PEARSON

Definición

Sean (X_i, Y_i) las observaciones realizadas en cada uno de los n sujetos de una muestra de tamaño n . Definimos el *coeficiente de correlación muestral de Pearson* entre X e Y como:

$$r = \text{Corr}(X, Y) = \frac{\text{cov}(X, Y)}{s_X s_Y} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{(n-1) s_X s_Y}$$

donde s_x y s_y son los desvíos estandares muestrales de las variables X e Y respectivamente.

CORRELACIÓN DE PEARSON

Ejemplo

	X	Y	(X - \bar{X})	(Y - \bar{Y})	(X - \bar{X}) (Y - \bar{Y})
	3	10	-3.86	3.14	-12.12
	6	7	-0.86	0.14	-0.12
	5	9	-1.86	2.14	-3.98
	8	6	1.14	-0.86	-0.98
	9	8	2.14	1.14	2.45
	10	7	3.14	0.14	0.45
	7	8	0.14	1.14	0.16
Media	6.86	7.86		Suma =	-14.14
DS	2.41	1.35			

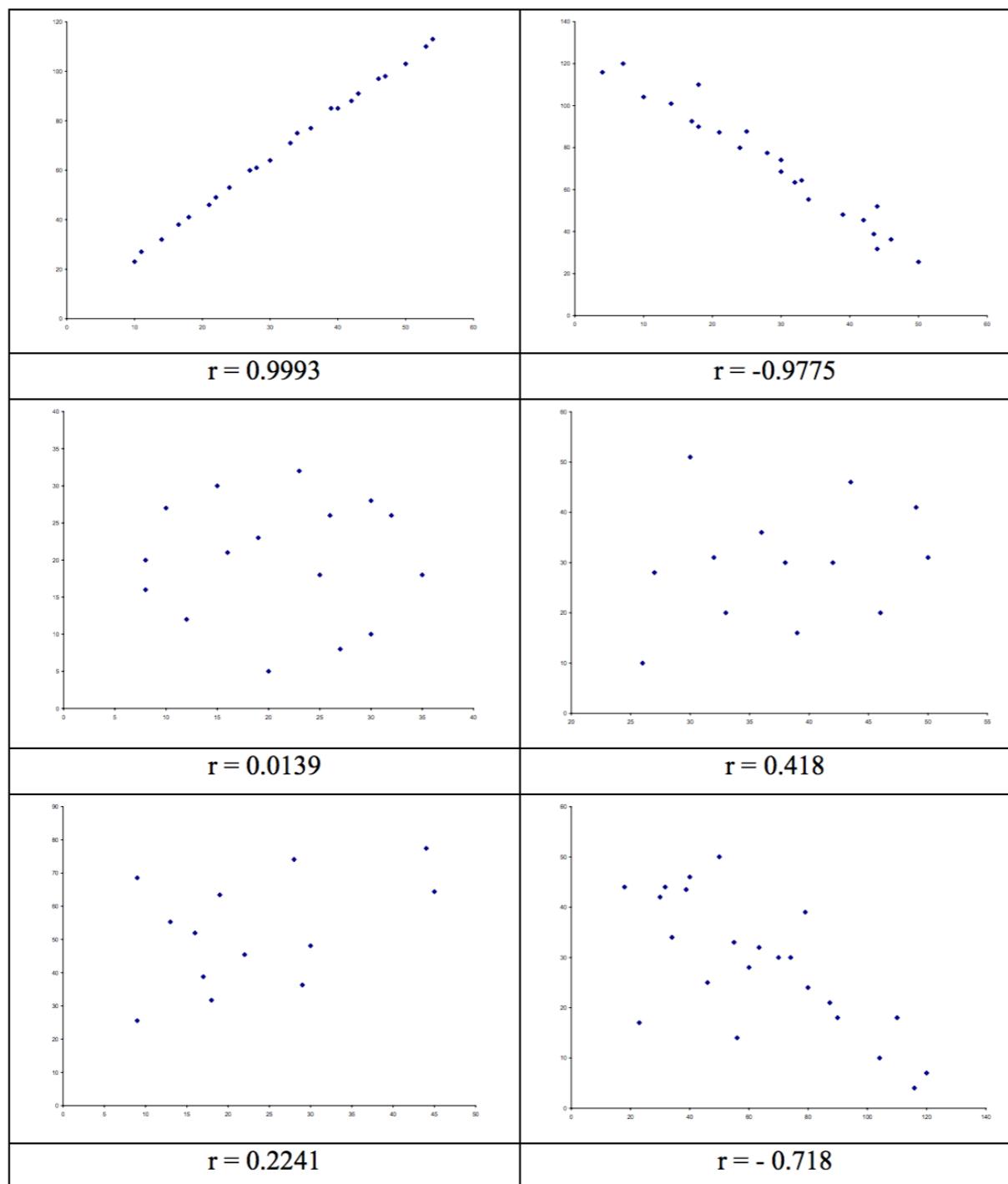
$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{(n-1) s_X s_Y} = \frac{-14.14}{(7-1) 2.41 1.35} = -0.73$$

Propiedades del coeficiente de correlación de Pearson

- r toma valores entre -1 y 1 ($-1 \leq r \leq 1$),
- r mide la *fuerza* de la asociación LINEAL entre X e Y,
- $r = 0$ implica que no hay relación lineal entre las variables,
- $r = + 1$ implica que todos los puntos caen sobre una recta de pendiente positiva (asociación positiva),
- $r = - 1$ implica que todos los puntos caen sobre una recta de pendiente negativa (asociación negativa),
- mientras mayor el valor absoluto de r mayor la fuerza de la asociación,
- el valor de r no depende de las unidades de medición,
- el coeficiente de correlación trata a X e Y simétricamente, no identifica cual es la variable dependiente y cual la independiente.

CORRELACIÓN DE PEARSON

Figura 5. Ejemplos de conjuntos de datos con diferente grado de correlación.



ESTUDIAR LA CORRELACIÓN DE SPEARMAN

Tarea

ENTORNO MINERÍA DE DATOS

INSTALACIÓN

- ▶ Instalar Python 2.7.9:
 - ▶ <https://www.python.org/downloads/>
- ▶ Agregarlo al PATH, C:\Python27:
 - ▶ Computer/Properties/Advanced system settings/
Environment Variables/
- ▶ Instalar Pip:
 - ▶ *python -m pip install -U pip setuptools*
 - ▶ Agregarlo al PATH, C:\Python27\Scripts

INSTALACIÓN

- ▶ Instalar Virtual Environment:
 - ▶ *pip install virtualenv*
- ▶ Crear y activar virtualenv:
 - ▶ `virtualenv <DIR>`
 - ▶ `<DIR>\Scripts\activate`
- ▶ Instalar Microsoft Visual C++ Compiler for Python 2.7
 - ▶ <https://www.microsoft.com/en-us/download/confirmation.aspx?id=44266>
- ▶ Instalar Jupyter Notebook:
 - ▶ *pip install jupyter*



PRIMEROS PASOS

Python

PROYECTOS