Estadística Descriptiva

Estadística descriptiva

El objetivo de la **Estadística** es extraer conocimiento a partir de un conjunto de datos. En **Estadística Descriptiva** se exploran los datos a fin de identificar sus principales características mediante un número reducido de gráficos y/o números.

Los conjuntos de datos pueden provenir de medir una o más variables en un conjunto de individuos.

Para describir un conjunto de datos o muestra se comienza con un análisis individual de cada variable y posteriormente se estudian las relaciones entre variables medidas.

Suele comenzarse con **representaciones gráficas** y después se calculan las **medidas numéricas o de resumen**.

Apuntes: Notas de Liliana Orellana

Clases de A. M. Bianco-Daniela Rodríguez

POBLACIÓN: total de sujetos o unidades de análisis de interés en el estudio

Ej.: Todos los niños sanos con edad entre 0 y 5 años.

MUESTRA: cualquier subconjunto de los sujetos o unidades de análisis de la población, en el cual se recolectarán los datos.

Usamos una muestra para conocer o estimar características de la población, denominamos:

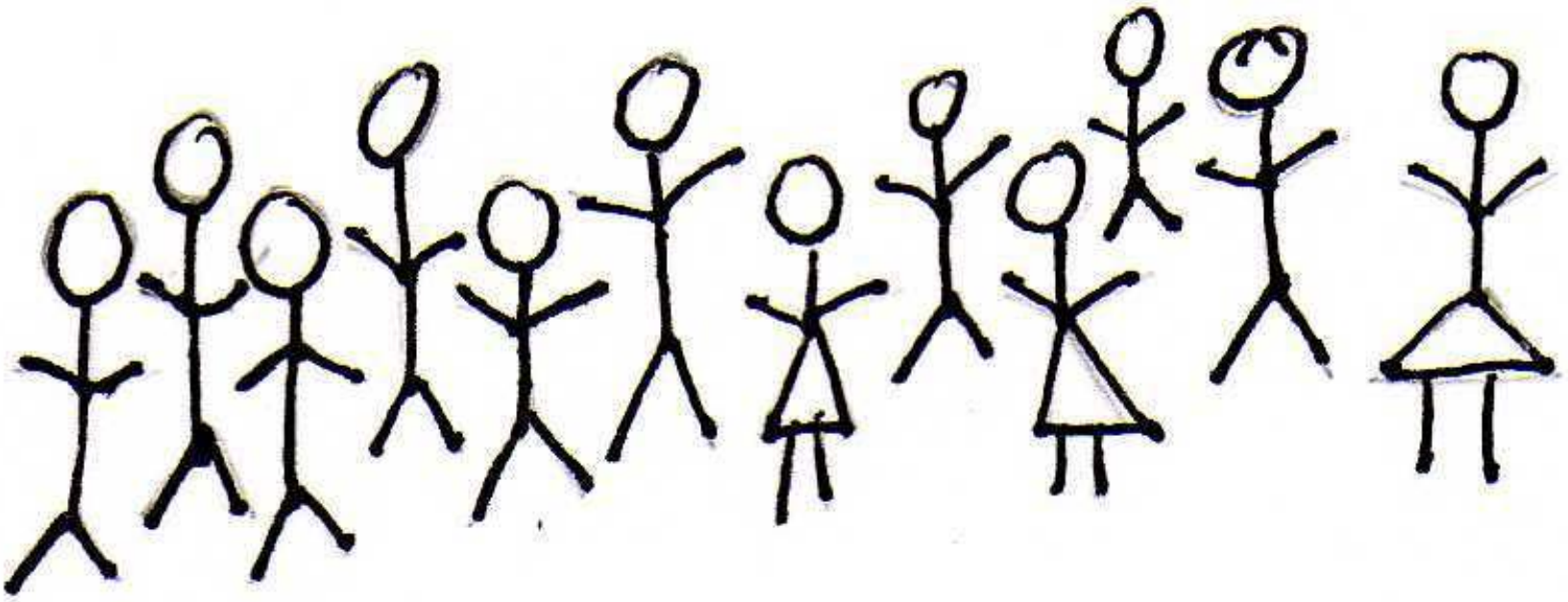
PARÁMETRO: una medida resumen calculada sobre la población: media, varianza, proporción.

ESTADÍSTICO : una medida resumen calculada sobre la muestra.

Cuando existen datos para toda la población (**CENSO**), en principio, no habría necesidad de usar métodos estadísticos, ya que sería posible calcular exactamente los parámetros de interés.

Ejemplo: en el censo poblacional, se registra el sexo de todas las personas censadas, que son prácticamente toda la población, así que es posible conocer exactamente la proporción de habitantes de los dos sexos.

Estamos interesados en estudiar un fenómeno de una población



CENSO

~~CENSO~~

Limitaciones

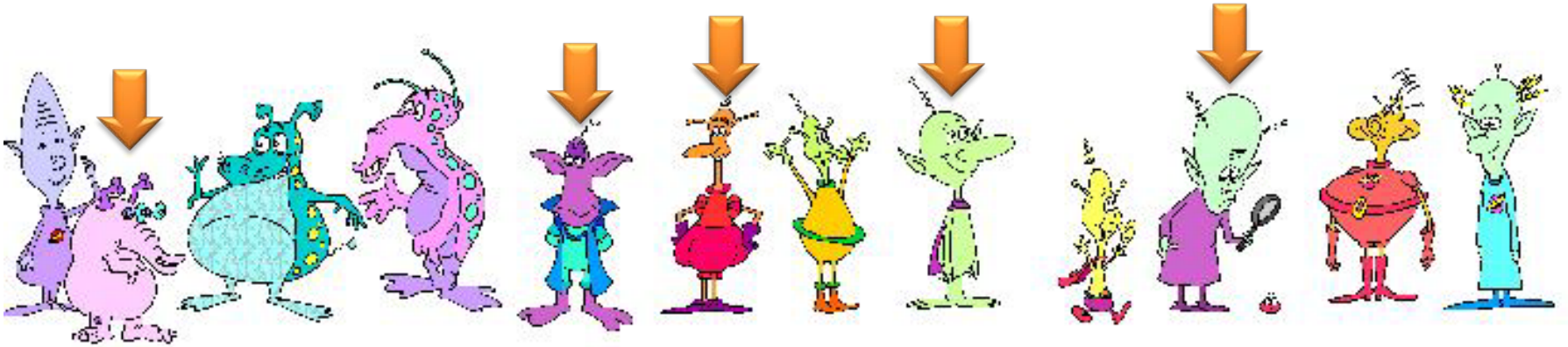
Imposibilidad

Población

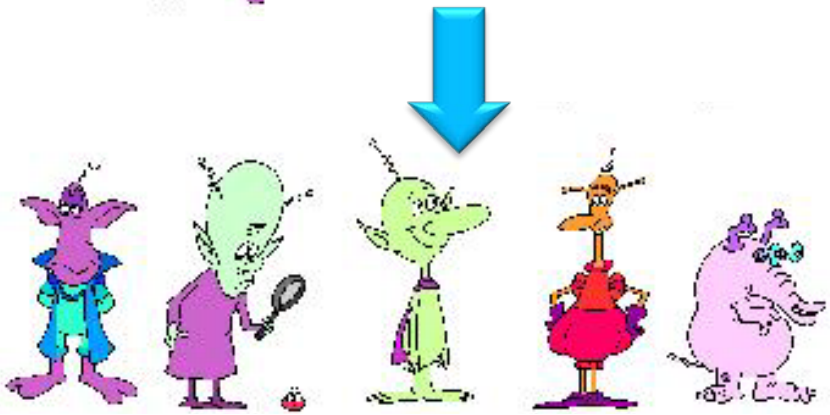
Población

Población

Población



Población



Muestra

VARIABLE: Una variable es una característica que varía de individuo en individuo.

(edad, peso, altura, género, concentración de colesterol en sangre, club de fútbol preferido, etc.)

DATOS: son los valores de la variable en estudio.

Los datos disponibles se obtienen a partir de una muestra de la población de interés, como los valores observados de la o las variables de interés.

- Los datos guardan información, pero será necesario analizarlos o procesarlos para obtener respuestas a algunas preguntas y llegar a conclusiones.

Métodos Gráficos

REPRESENTACIÓN DE DATOS NUMERICOS

Trataremos de responder a preguntas tales como:

- **¿Son los valores medidos casi todos iguales?**
- **¿Son muy diferentes unos de otros?**
- **¿En qué sentido difieren?**
- **¿Cómo podemos describir cualquier patrón o tendencia?**
- **¿Son un único grupo? ¿Hay varios grupos?**
- **¿Difieren algunos pocos datos notablemente del resto?**

TIPOS DE DATOS

- 1. **Variables cualitativas**: Describen cualidades o atributos

  (ej.: género, color del ojos, estado civil, fuma no fuma, severidad de la patología: Ausente/leve/moderado/severo).

- 2. **Variables cuantitativas discretas**: Toman un cierto número de valores posibles. En general, aparecen por conteo.

  (ej.: número de miembros del hogar, número de hijos, número de intervenciones quirúrgicas, número de casos notificados de una cierta patología)

- 3. **Variables cuantitativas continuas**: Toman valores en un intervalo (ej.: altura, peso, pH, nivel de colesterol en sangre, tiempo hasta que llega un tren).

El tipo de dato nos permite decidir qué análisis estadístico utilizar.

Ejemplo: Edad es continua, pero si se la registra en años resulta ser discreta. En estudios con adultos, en que la edad va de 20 a 70 años, por ejemplo, no hay problemas en tratarla como continua, ya que el número de valores posibles es muy grande. Pero en el caso de niños en edad preescolar, si la edad se registra en años debe tratarse como discreta, en tanto que si se la registra en meses puede tratarse como continua.

Los datos numéricos (discretos o continuos) pueden ser transformados en categóricos y ser tratados como tales.

Aunque esto es correcto no necesariamente es eficiente y *siempre es* preferible registrar el valor numérico de la medición.

¿Por qué es importante identificar el tipo de datos?

Porque el tipo de datos DETERMINA el método de análisis apropiado y válido y cada método de análisis estadístico es específico para un cierto tipo de datos.

La distinción más importante es entre datos numéricos y categóricos.

Métodos Gráficos:

REPRESENTACIÓN DE DATOS CATEGÓRICOS

TABLA DE FRECUENCIA

El modo más simple de presentar datos categóricos es por medio de una tabla de frecuencias que indica el número observaciones que caen en cada una de las clases de la variable.

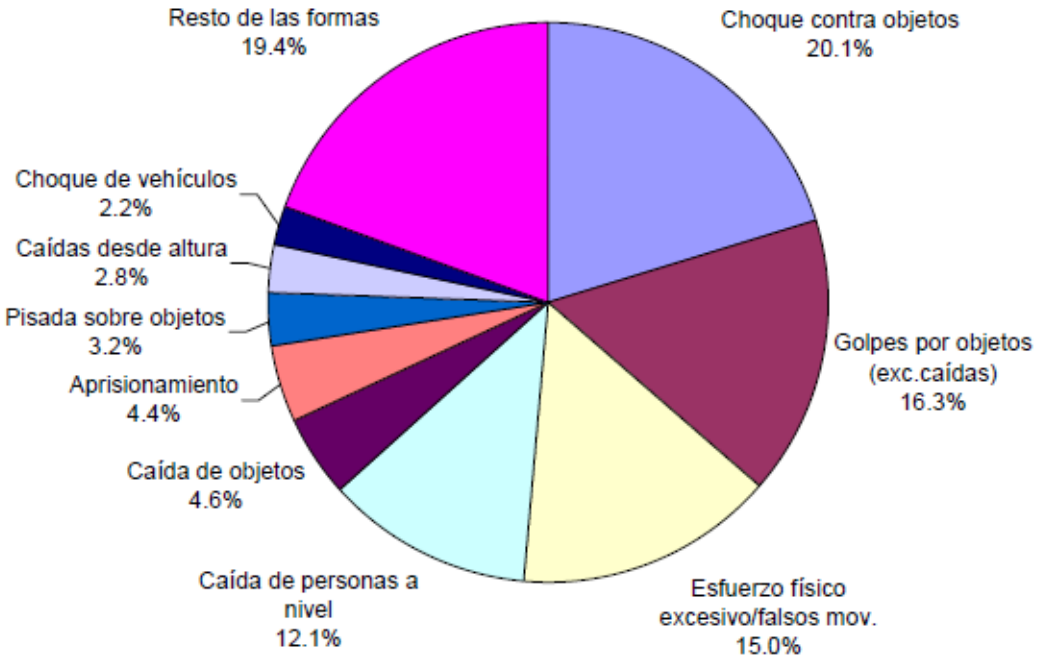
GRÁFICO DE BARRAS

A cada categoría o clase de la variable se le asocia una barra cuya *altura representa la frecuencia* o *la frecuencia relativa* de esa clase. Las barras difieren sólo en altura, no en ancho.

GRÁFICO DE TORTAS

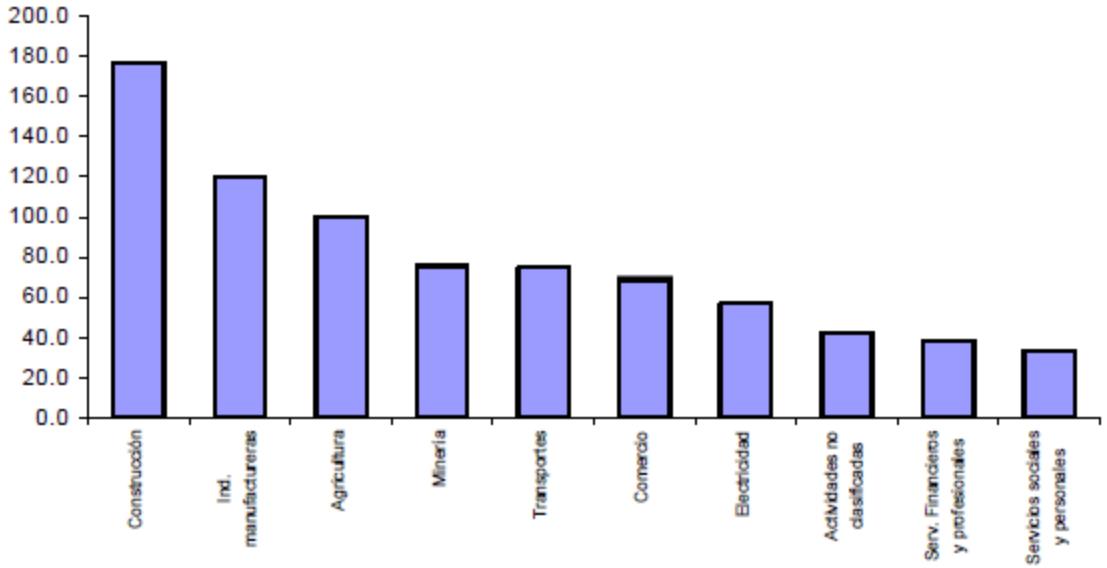
Se representa la frecuencia relativa de cada categoría como una porción de un círculo, en la que el ángulo se corresponde con la frecuencia relativa correspondiente.

GRÁFICO DE TORTAS



Fuente: http://www.srt.gob.ar/estadisticas/anuario/1999.pdf

GRÁFICO DE BARRAS



Métodos Gráficos

REPRESENTACIÓN DE DATOS NUMERICOS

HISTOGRAMAS

El histograma es el más conocido de los gráficos para resumir un conjunto de datos Numéricos.

Para construir un histograma es necesario previamente construir una *tabla de frecuencias.*

Métodos Gráficos

REPRESENTACIÓN DE DATOS NUMERICOS

HISTOGRAMAS

Dividimos el rango de los **n datos en intervalos o clases, que no se** superponen. Las clases deben ser **excluyentes y exhaustivas.**

Contamos la cantidad de datos en cada intervalo o clase, es decir la **frecuencia.**

También podemos usar para cada intervalo la **frecuencia relativa**

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

Graficamos el histograma en un par de ejes coordenados representando en las abscisas los intervalos y sobre cada uno de ellos un rectángulo cuya área es proporcional a la frecuencia relativa (o frecuencia) de dicho intervalo.

HISTOGRAMAS

Ejemplo: Porcentajes de octanos para mezclas de naftas.

85.3	87.5	87.8	88.5	89.9	90.4	91.8	92.7
86.7	87.8	88.2	88.6	90.3	91.0	91.8	93.2
88.3	88.3	89.0	89.2	90.4	91.0	92.3	93.3
89.9	90.1	90.1	90.8	90.9	91.1	92.7	93.4
91.2	91.5	92.6	92.7	93.3	94.2	94.7	94.2
95.6	96.1						

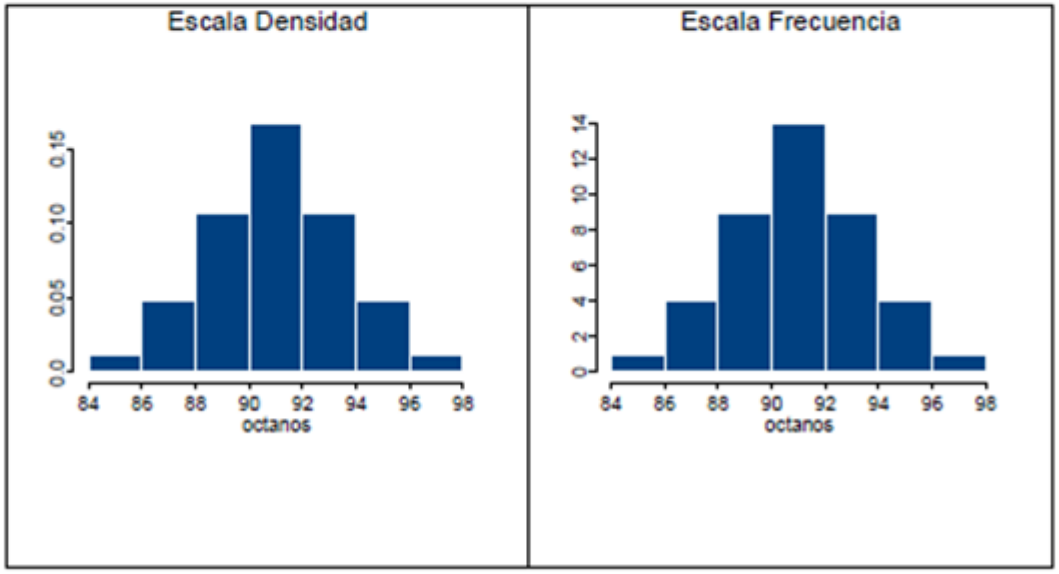
Clase	Frecuencia f_i	Frecuencia relativa fr_i
[84, 86]	1	0.02380952
(86, 88]	4	0.09523810
(88, 90]	9	0.21428571
(90,92]	14	0.33333333
(92,94]	9	0.21428571
(94,96]	4	0.09523810
(96,98]	1	0.02380952
Total	**42**	**1**

HISTOGRAMAS

Los comandos son

hist(octanos.per,freq=T)
hist(octanos.per,freq=F) (para graficar escala densidad)

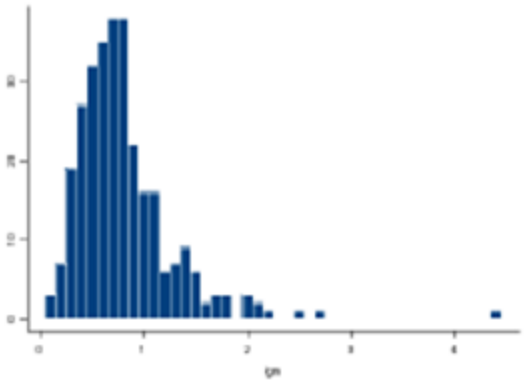
Histogramas para datos de OCTANOS



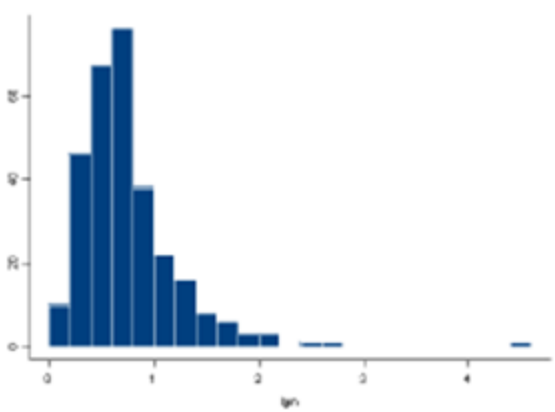
HISTOGRAMAS

Ejemplo: Concentración de Inmunoglobulina

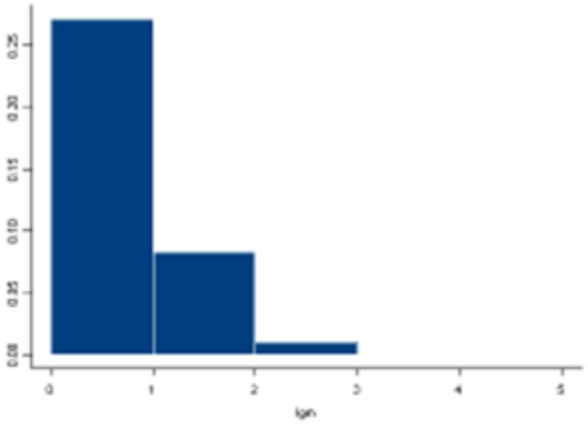
En este ejemplo vemos cómo la elección del ancho de las clases afecta el gráfico.



Longitud de Clase= 0.1 g/l

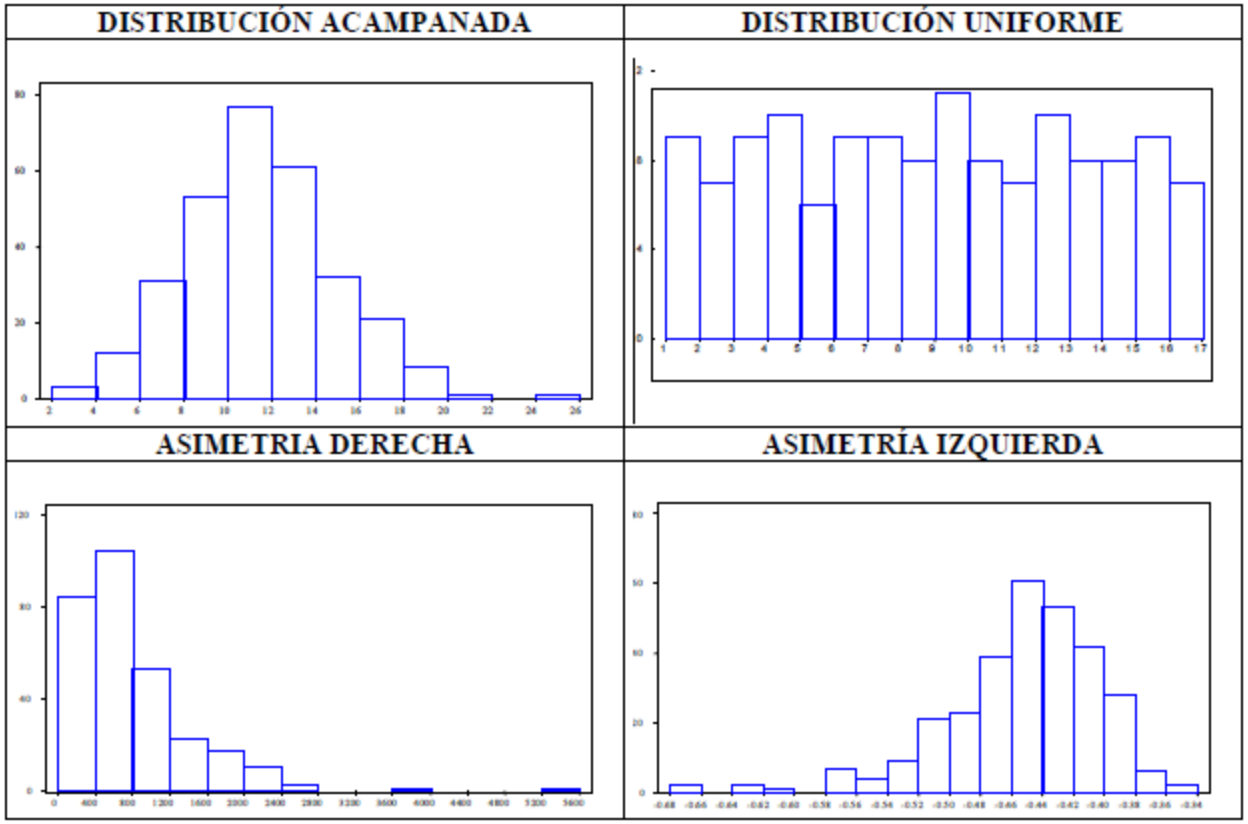


Longitud de Clase= 0.2 g/l



Longitud de Clase=1g/l

EJEMPLOS DE HISTOGRAMAS



Interpretación de un Histograma

- En general, los intervalos se toman de igual longitud y de esa manera la altura es proporcional a la frecuencia y se facilita la lectura.
- Es aconsejable identificar si se han usado frecuencias absolutas o relativas, sobre todo si se van a comparar distintos histogramas.
- Rango de variación de los datos (Mínimo – Máximo).
- Intervalos más frecuentes
- ¿La distribución es unimodal o hay más de una moda?
- ¿La distribución es simétrica?
- Si es asimétrica, ¿ la asimetría es a derecha o a izquierda?
- ¿En torno a qué valor están aproximadamente centrados los datos?
- ¿Cuán dispersos en torno a este centro están los datos ?
- ¿Hay datos atípico en relación a la mayoría de los datos?

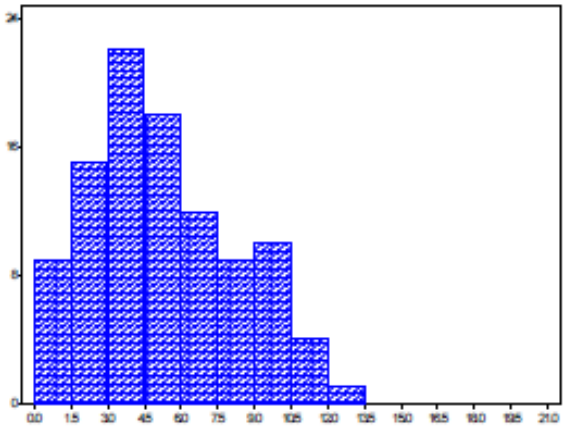
HISTOGRAMAS

¿En que difieren un gráfico de barras y un histograma?

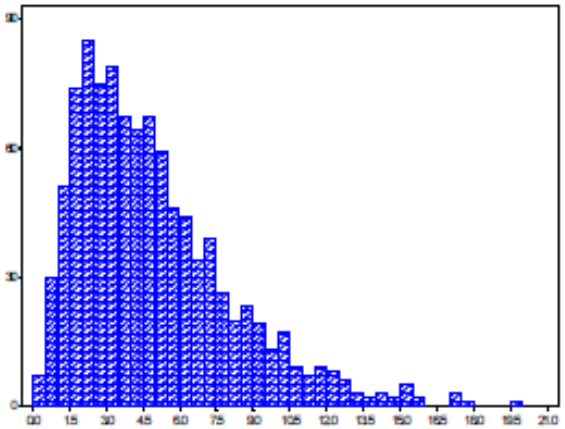
-El gráfico de barras representa el porcentaje en la altura de la barra. Mientras que en un histograma el porcentaje se representa en el área de la barra.

- En el gráfico de barras, las barras se representan separadas para indicar que no hay continuidad entre las categorías. En un histograma barras adyacentes *deben estar en* contacto indicando que la variable es continua.

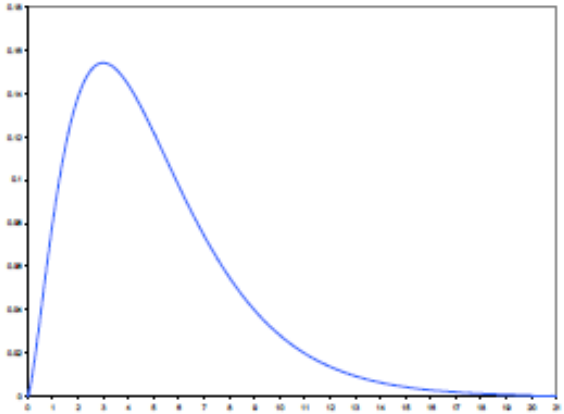
HISTOGRAMAS



Muestra n = 100



Muestra n = 1000



Población

Medidas de resumen

Resumiremos la información de los datos mediante medidas de fácil interpretación que reflejen sus características más relevantes. Las medidas de resumen son útiles para comparar conjuntos de datos y para presentar los resultados de un estudio.

Se clasifican en dos grupos principales:

Medidas de posición o localización: describen un valor alrededor del cual se encuentran las observaciones.

Medidas de dispersión o escala: pretenden expresar cuán variable es un conjunto de datos.

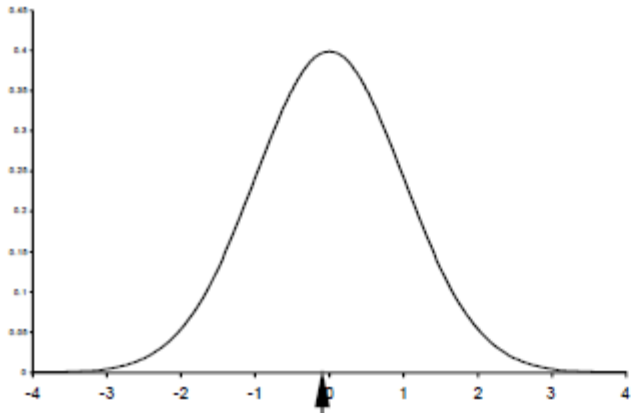
Medidas de Posición o Centrado

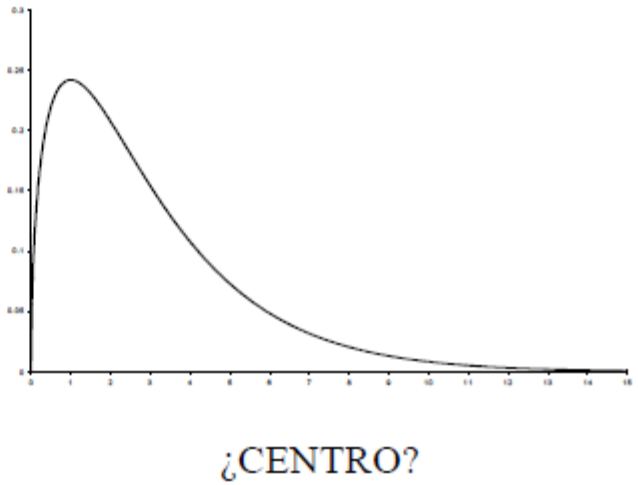
¿Cuál es el valor central o que mejor representa a los datos?

Buscamos un valor típico que represente a los datos.

Si la distribución es simétrica diferentes medidas darán resultados similares y hay un claro centro.

Si es asimétrica no existe un centro evidente y diferentes criterios para resumir los datos pueden diferir considerablemente.





Medidas de Posición o Centrado

Promedio o Media Muestral

Sumamos todas las observaciones y dividimos por el número total datos:

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Es el punto de equilibrio de los datos.
- Es una medida muy sensible a datos atípicos.
- La suma de los desvíos respecto del promedio es cero:

$$(x_1 - \bar{x}) + (x_2 - \bar{x}) + \cdots + (x_n - \bar{x}) = 0$$

Medidas de Posición o Centrado

Es el punto de equilibrio del conjunto de datos.

X´s: 1, 2, 2, 3

X´s: 1, 2, 2, 7

Es una medida muy sensible a la presencia de datos anómalos (outliers).

Medidas de Posición o Centrado

Mediana Muestral

Es una medida del centro de los datos en tanto divide a la muestra ordenada en dos partes de igual tamaño. "Deja la mitad de los datos a cada lado".
Sean los estadísticos de orden muestrales:

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

definamos como mediana

$$\tilde{x} = \begin{cases} x_{(k+1)} & \text{si} \quad n = 2k+1 \\ \dfrac{x_{(k)} + x_{(k+1)}}{2} & \text{si} \quad n = 2k \end{cases}$$

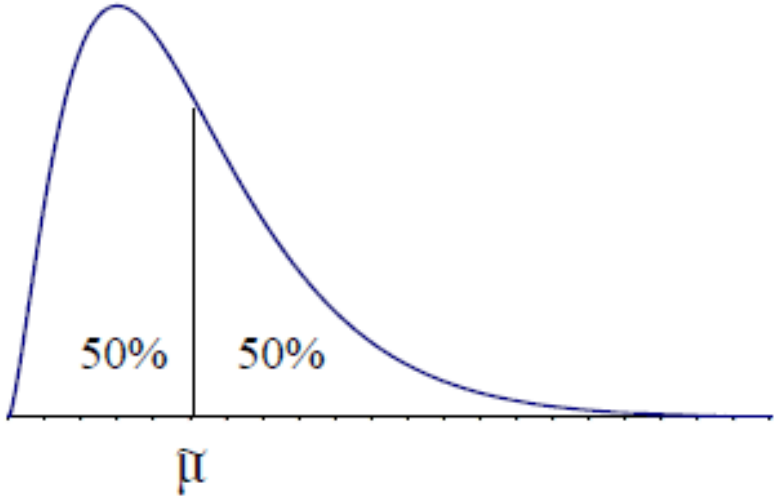
Si la distribución es simétrica la mediana y la media identifican al mismo punto.

La mediana es resistente a la presencia de datos atípicos.

Medidas de Posición o Centrado

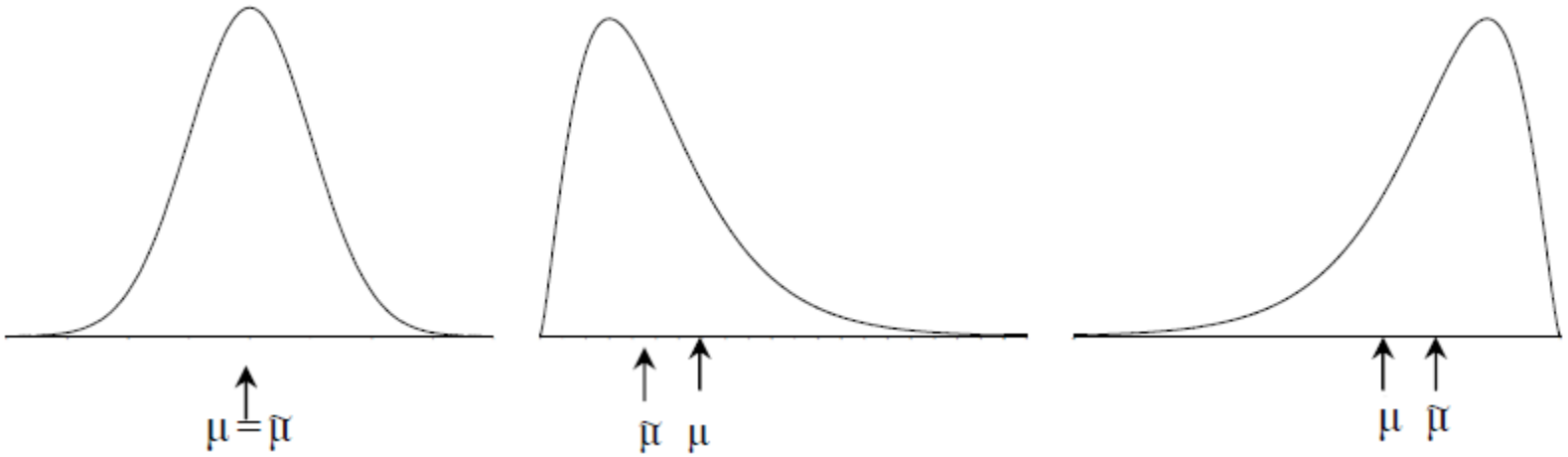
Mediana poblacional

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.



Medidas de Posición o Centrado

Relación entre mediana y media poblacionales



Medidas de Posición o Centrado

Si tenemos:

X´s: 1,2,2,3 $\quad \bar{x} = 2 \quad \tilde{x} = 2$

X´s: 1,2,2,7 $\quad \bar{x} = 3 \quad \tilde{x} = 2$

¿Qué pasa si tenemos un 70 en lugar de 7?

$$\bar{x} = 18.75 \qquad \tilde{x} = 2$$

Si tenemos una muestra de salarios de una población dada, ¿sería más adecuado tomar la media o la mediana muestral para representarlos?

Medidas de Posición o Centrado

Medias α-Podadas

Es un promedio calculado sobre los datos una vez que se han eliminado **α %** de los datos más pequeños y **un α %** de los datos más grandes. Formalmente podemos definirla como:

$$\bar{x}_{\alpha} = \frac{x_{([n\alpha]+1)} + \ldots + x_{(n-[n\alpha])}}{n - 2[n\alpha]}$$

Ejemplo: Sea el siguiente conjunto de 10 observaciones, ya ordenadas

X's: 2 5 8 10 14 17 21 25 28 40

y calculemos la media 0.10-podada. Como el 10% de 10 es 1, debemos podar 1 dato en cada extremo y calcular el promedio de los 8 datos restantes, es decir

$$\bar{x}_{0.10} = \frac{5+8+10+14+17+21+25+28}{8} = \frac{128}{8} = 16$$

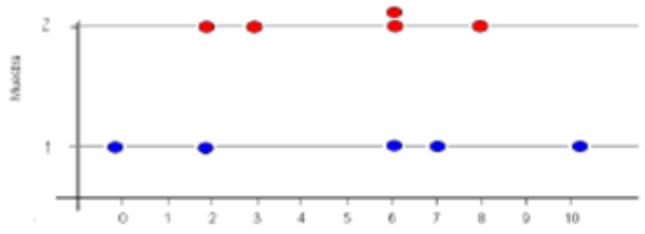
Medidas de Dispersión

Medidas de Dispersión o Variabilidad:

¿Cuán dispersos están los datos? ¿Cuán cercanos son los datos al valor típico?

Supongamos que tenemos datos $x_1, x_2, \ldots\ldots, x_n$

X's: 0 2 6 7 10
Y's: 2 3 6 6 8



$$\overline{X} = \overline{Y} = 5$$
$$\widetilde{X} = \widetilde{Y} = 6$$

¿Cómo medir la diferencia que se observa entre ambas muestras?

Medidas de Dispersión

Rango Muestral

Se define como la diferencia entre el valor más grande y el pequeño de los datos:

$$\text{Rango} = \text{máx}(X_i) - \text{mín}(X_i)$$

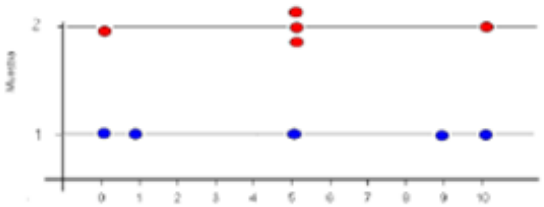
Ejemplo: en nuestros conjuntos de datos:

$$\text{Rango (X)} = 10 \qquad \text{Rango(Y)} = 6$$

- Esta medida es muy sensible a la presencia de outliers.

Veamos otro ejemplo:

X's: 0 1 5 9 10
Y's: 0 5 5 5 10



$\overline{X} = \overline{Y}$ $\qquad \tilde{X} = \tilde{Y}$ $\qquad Rango(X) = Rango(Y)$

Medidas de Dispersión

Varianza Muestral

Es una medida de la variabilidad de los datos alrededor de la media muestral.

Varianza muestral : $S^2 = \dfrac{\sum_{i-1}^{n} (x_i - \bar{x})^2}{n-1}$

Desvío estándar muestral : $S = \sqrt{S^2}$

Ejemplo:

$S^2_x = 20.5 \quad S_x = 4.258$

$S^2_y = 12.5 \quad S_y = 3.536$

Medidas de Dispersión

Distancia Intercuartil

Es una medida basada en el rango de los datos centrales de la muestra y más resistente que el desvío estándar.

Comenzaremos por definir los **percentiles.** El percentil $\alpha \cdot 100$ **%** de la muestra es el valor por debajo del cual se encuentra el $\alpha \cdot 100$ **%** de los datos en la muestra ordenada.

Para calcularlo:

- Ordenamos la muestra de menor a mayor
- Buscamos el dato que ocupa la posición $\alpha \cdot (n+1)$ en la muestra ordenada. Si este número no es entero se interpolan los dos adyacentes.

Medidas de Dispersión

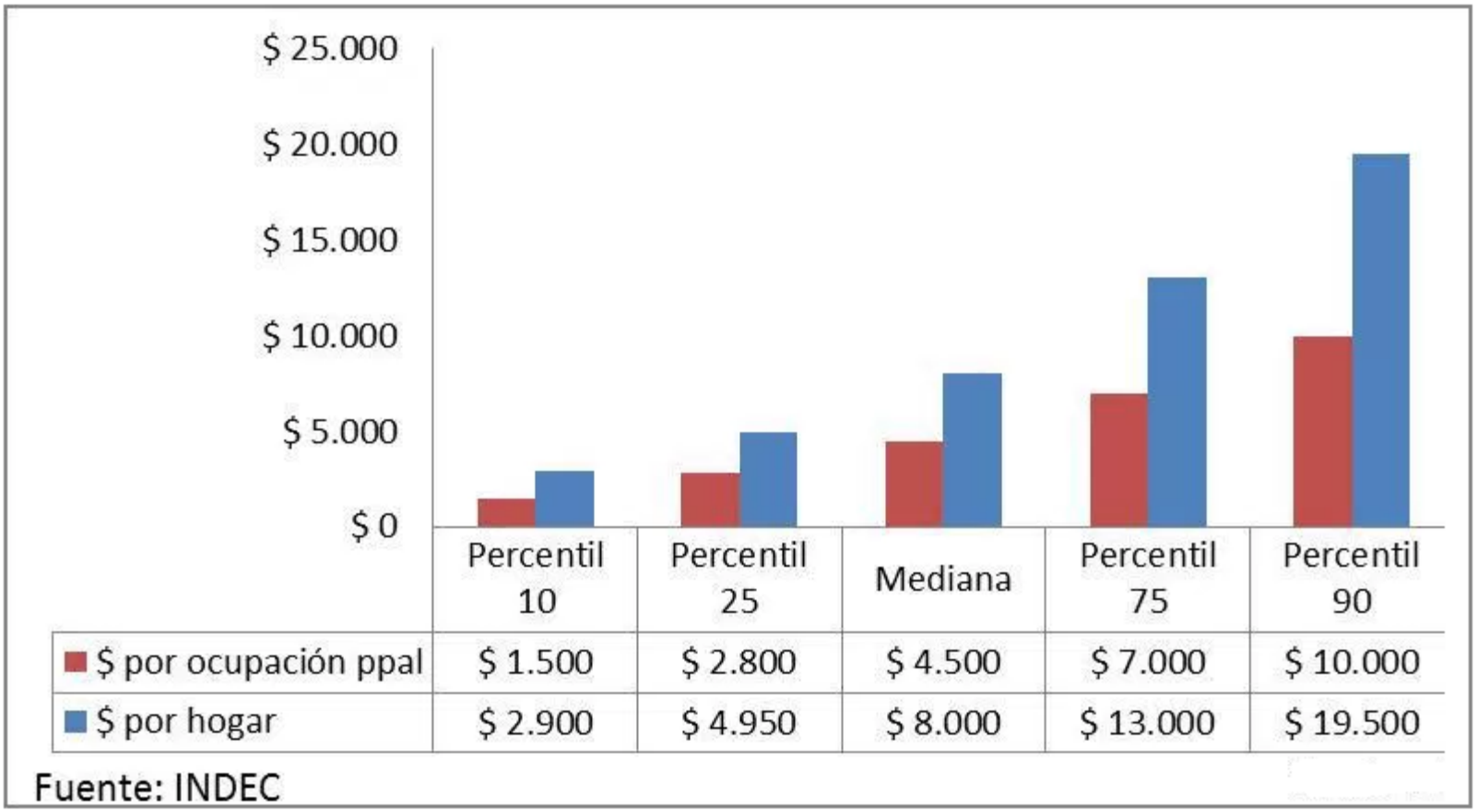
Ejemplo: Tenemos 19 datos que ordenados son

1 1 2 2 3 4 4 5 5 6 7 7 8 8 9 9 10 10 11

Percentil	Posición	Valor	
10%	0.10 (19+1) = **2**	1	
25%	0.25 (19+1) = **5**	3	**Cuartil Inferior**
50%	0.50 (19+1) = **10**	6	**Mediana**
75%	0.75(19+1) = **15**	9	**Cuartil Superior**
95%	0.95(19+1) = **19**	11	

- Notemos que el percentil 50% o segundo cuartil coincide con la mediana. Denotaremos Q_1 al primer cuartil (25%) y Q_3 al tercer cuartil (75%).
- Los cuartiles y la mediana dividen a la muestra en cuatro partes igualmente pobladas: 25% de la muestra en cada una de ellas.
- Entre Q_1 y Q_3 se halla el 50% central de los datos y el rango de estos rango es:

Distancia Intercuartil: $d_I = Q_3 - Q_1$



	Percentil 10	Percentil 25	Mediana	Percentil 75	Percentil 90
$ por ocupación ppal	$ 1.500	$ 2.800	$ 4.500	$ 7.000	$ 10.000
$ por hogar	$ 2.900	$ 4.950	$ 8.000	$ 13.000	$ 19.500

Fuente: INDEC

Así en una nota de 2014 en http://fortunaweb.com.ar/ se resume la información sobre salarios brindada por el INDEC.

Números de resumen

- Observemos que porcentaje de datos hay
- ✓ a la izquierda de Q_1
- ✓ a la derecha de Q_3
- ✓ entre Q_1 y Q_3
- ✓ entre Q_1 y el máximo
- ✓ entre el mínimo y Q_3

- Resultan muy útiles para describir la muestra las siguientes medidas conocidos como **Números de resumen**
- Mínimo
- Q_1 : Cuartil Inferior (o Primer Cuartil)
- Mediana (o Segundo Cuartil)
- Q_3 : Cuartil Superior (o Tercer Cuartil)
- Máximo

Medidas de Dispersión

Desvío Absoluto Mediano (Desviación absoluta respecto de la Mediana) MAD

Es una versión robusta del desvío estándar basada en la mediana. Definimos la MAD como:

$$MAD = \text{mediana}\left(\left|x_i - \tilde{x}\right|\right)$$

¿Cómo calculamos la MAD?

- Ordenamos los datos de menor a mayor.
- Calculamos la mediana.
- Calculamos la distancia de cada dato a la mediana.
- Despreciamos el signo de las distancias y las ordenamos de menor a mayor.
- Buscamos la mediana de las distancias sin signo.

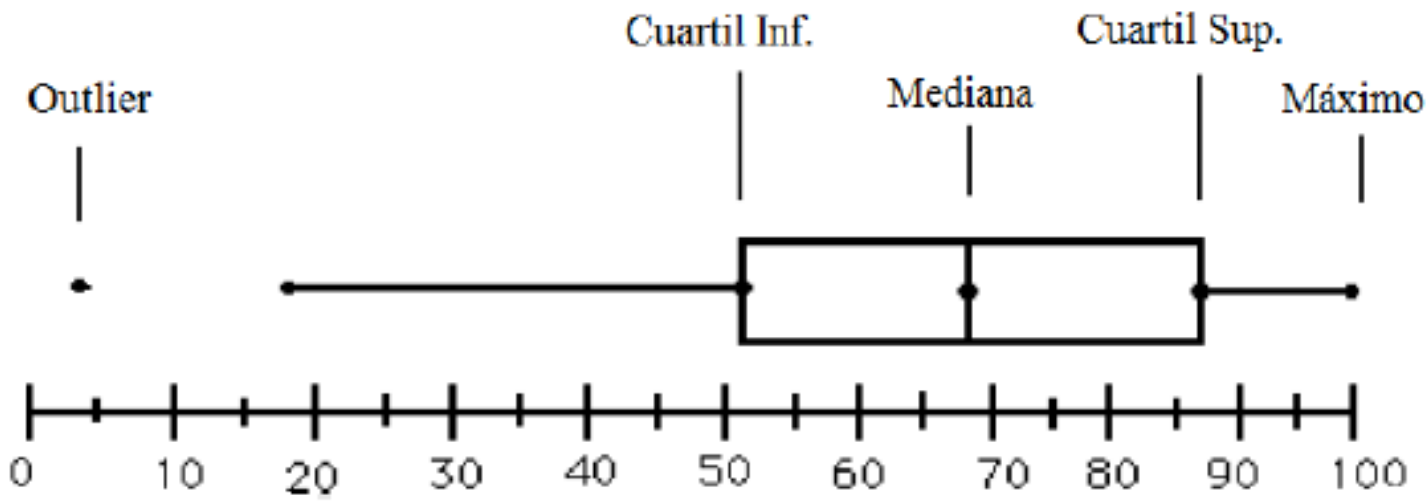
Observación: Si deseamos comparar la distancia intercuartil y la MAD con el desvío standard es conveniente dividirlas por constantes adecuadas. En ese caso se compara a S mediante

$$\frac{MAD}{0.675} \qquad \frac{d_I}{1.35}$$

Métodos Gráficos

REPRESENTACIÓN DE DATOS NUMERICOS

Boxplot



Boxplot

1. Representamos una escala vertical u horizontal

2. Dibujamos una caja cuyos extremos son los cuartiles y dentro de ella un segmento que corresponde a la mediana.

3. A partir de cada extremo dibujamos un segmento hasta el dato más alejado que está a lo sumo 1.5 d_I del extremo de la caja. Estos segmentos se llaman bigotes.

4. Marcamos con * a aquellos datos que están a más de 1.5 d_I de cada extremo de la caja.

Boxplot

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

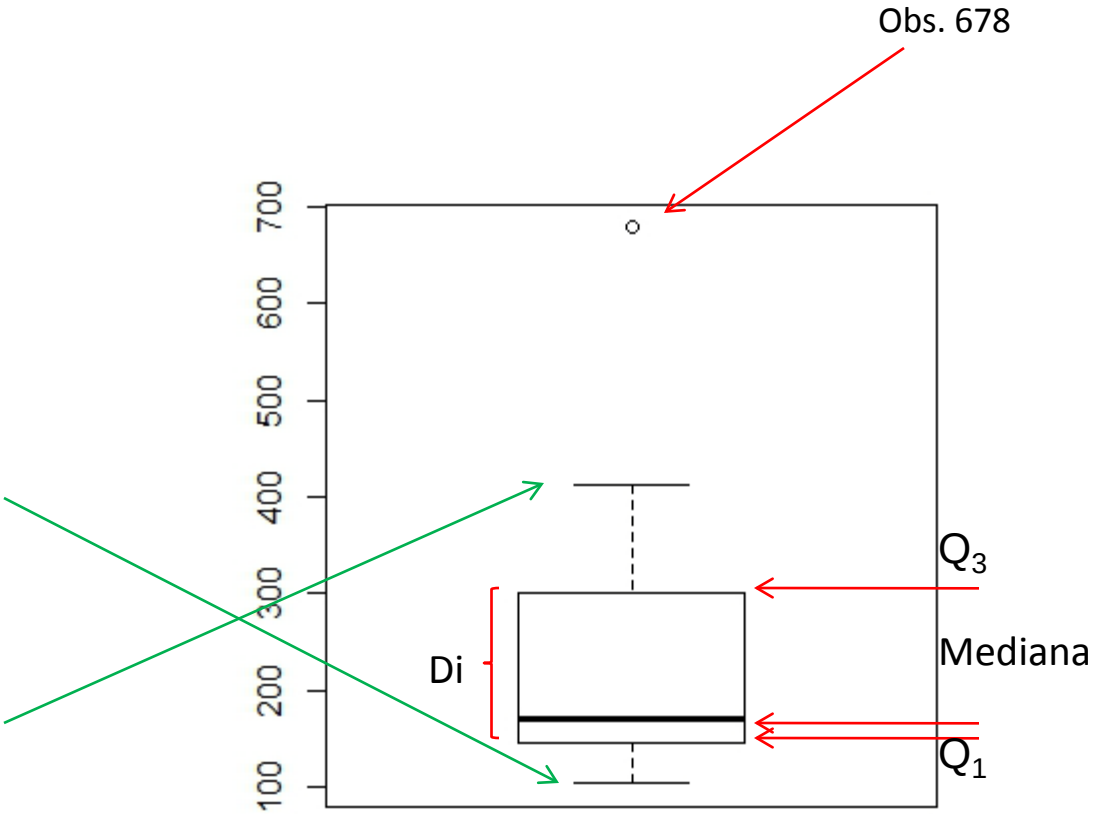
Q_1: calculo 0.25*(13+1)=3.5
Entonces Q_1 =146

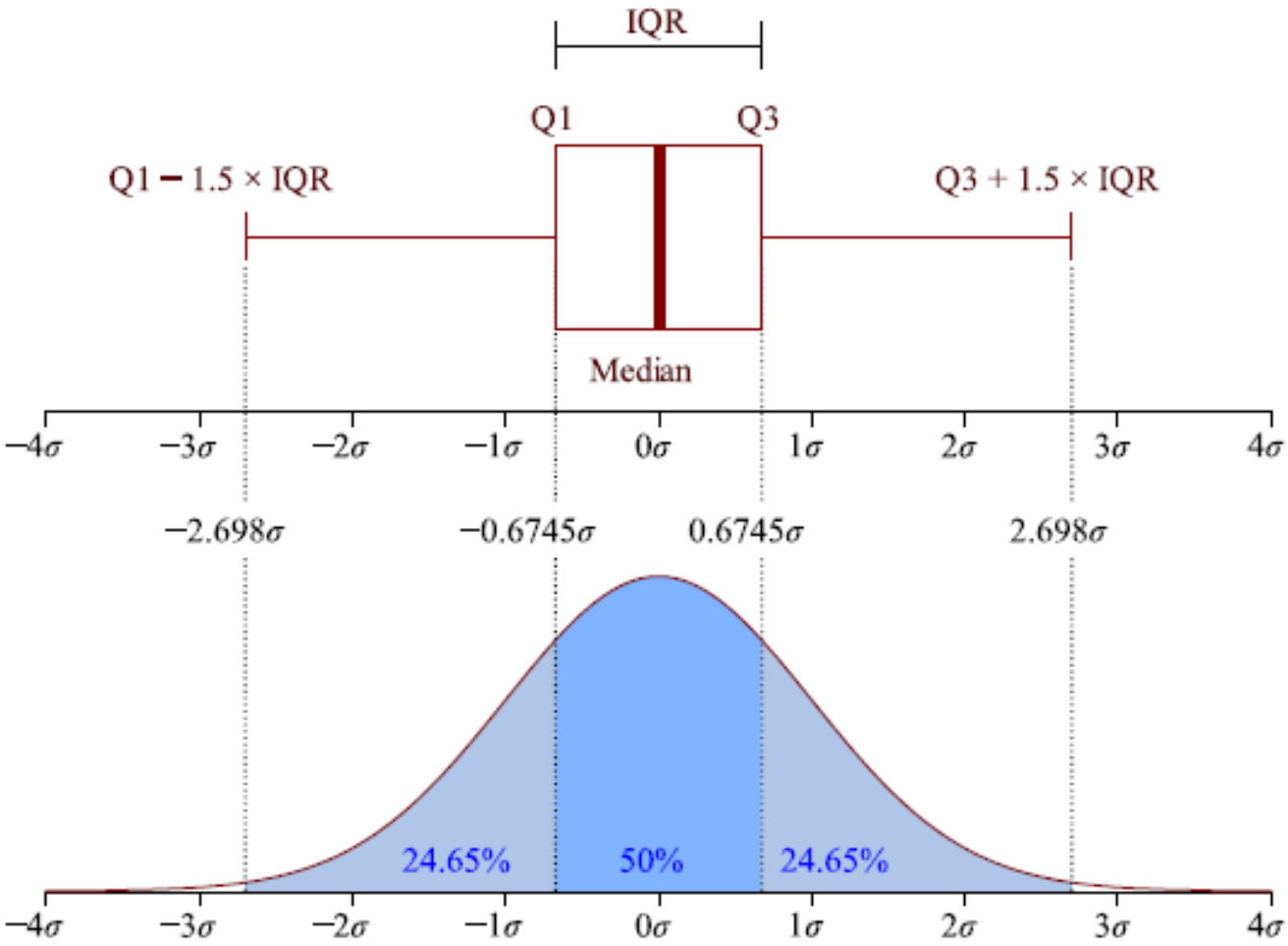
Q_3 : calculo 0.75*(13+1)=10.5
Entonces Q_3 =302

d_I=302-146=156

Calculamos
Li= primera cota inferior
= Q_1 -1.5*d_I=146-1.5*156=-88
Llego hasta la obs. 104

Ls= primera cota superior
= Q_3 +1.5*d_I=320+1.5*156=536
Llego hasta la obs. 412





Gracias Wikipedia!

Boxplot

¿Qué vemos en un box-plot?

- Posición
- Dispersión
- Asimetría
- Longitud de las colas
- Puntos anómalos o outliers.

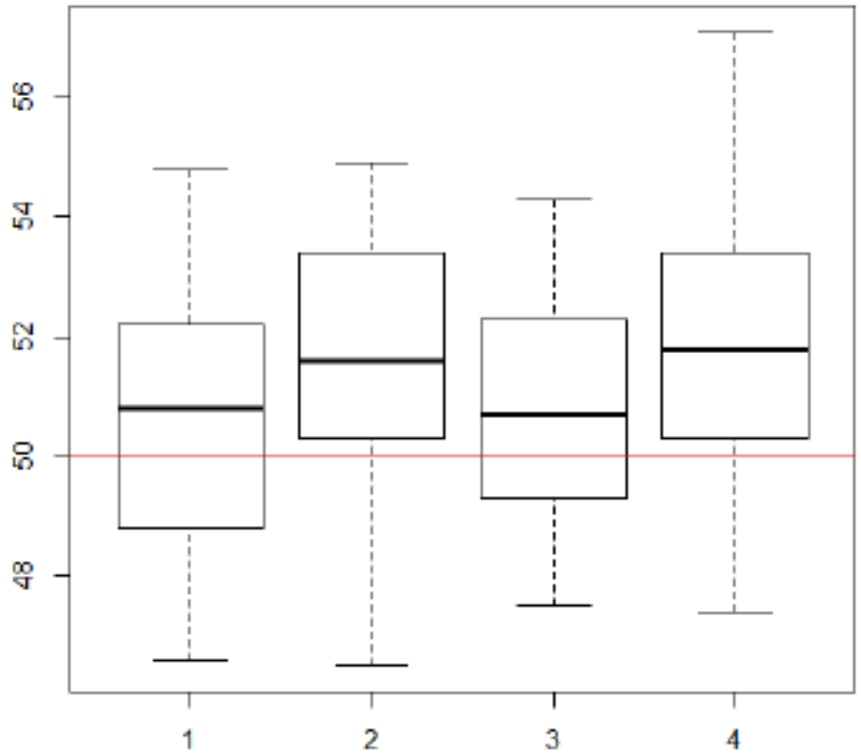
Los boxplots son muy útiles para comparar varios conjuntos de datos, pues nos dan una rápida impresión visual de sus características.

Boxplot

Ejemplo: Con el fin de estudiar las diferencias entre 4 laboratorios se miden 25 muestras con una concentración de analito de 50mg kg^{-1}.

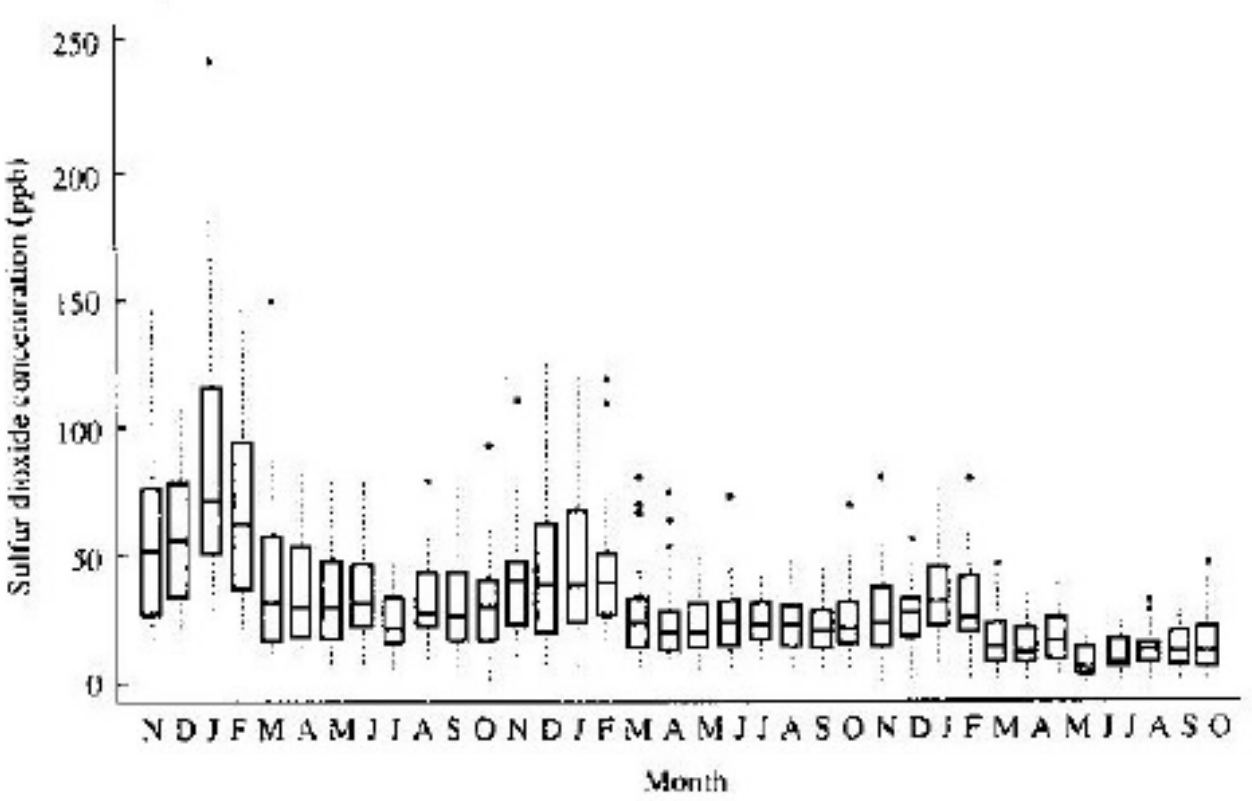
Se analizan los datos correspondientes a 25 mediciones realizadas en 4 laboratorios. Veamos que da este análisis.

```
boxplot(LAB1,LAB2,LAB3,LAB4)
abline(h=50,col="red")
```

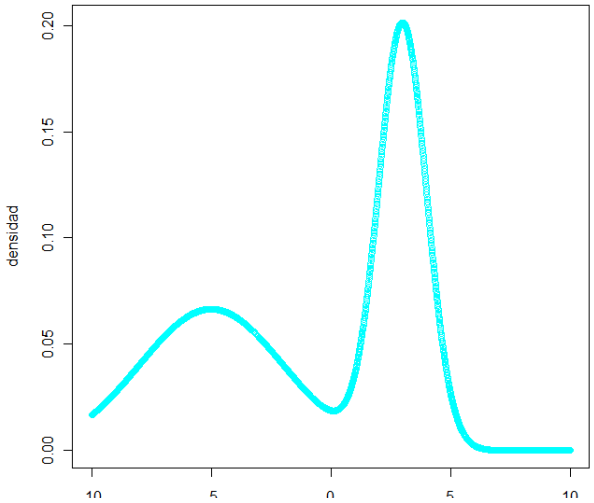


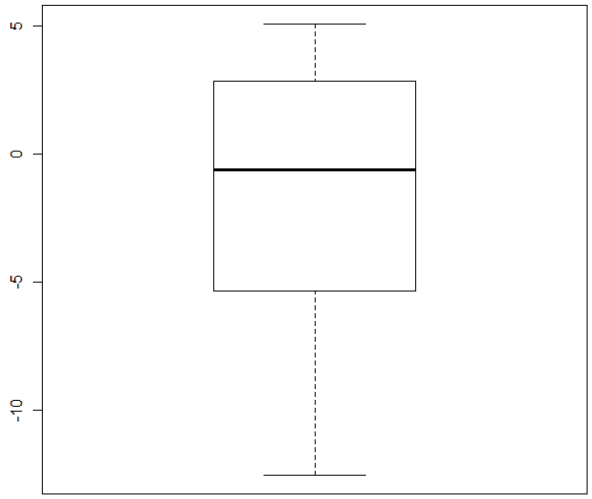
Boxplot

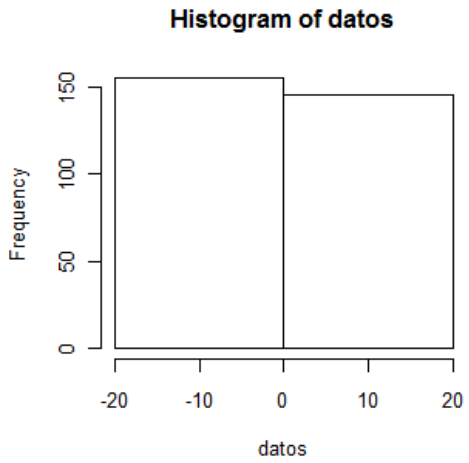
Ejemplo: Los siguientes boxplots corresponden a datos de concentración máxima diaria en partes por mil millones de dióxido de azufre en Bayonne, en el estado de Nueva Jersey, desde noviembre de 1969 hasta octubre de 1972 agrupados por meses. Hay 36 grupos de datos, cada uno de tamaño aproximadamente 30.

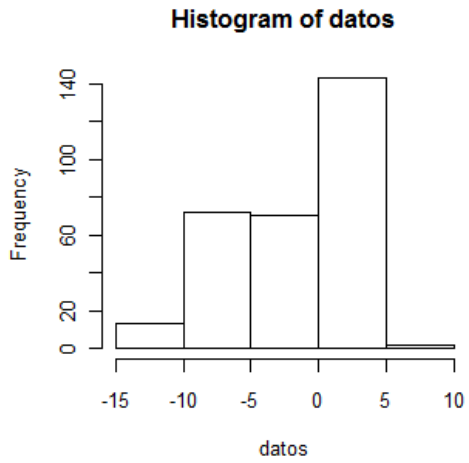


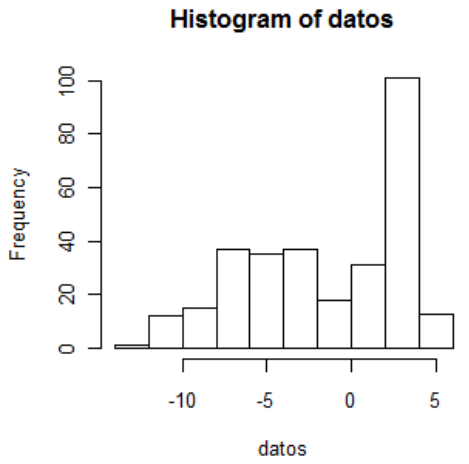
Boxplot vs. Histograma: información complementaria

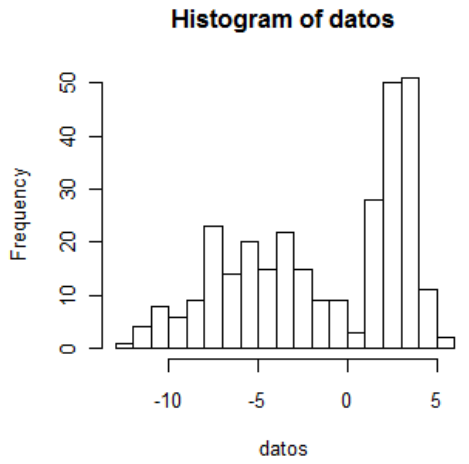












QQ-Plot o Grafico cuantil-cuantil

QQ-plot

El qq-plot es un gráfico que nos sirve para evaluar la cercanía a la distribución normal.

Para realizarlo se consideran los estadísticos de orden

$$X_{(1)} \leq X_{(2)} \leq \ldots\ldots \leq X_{(n)}$$

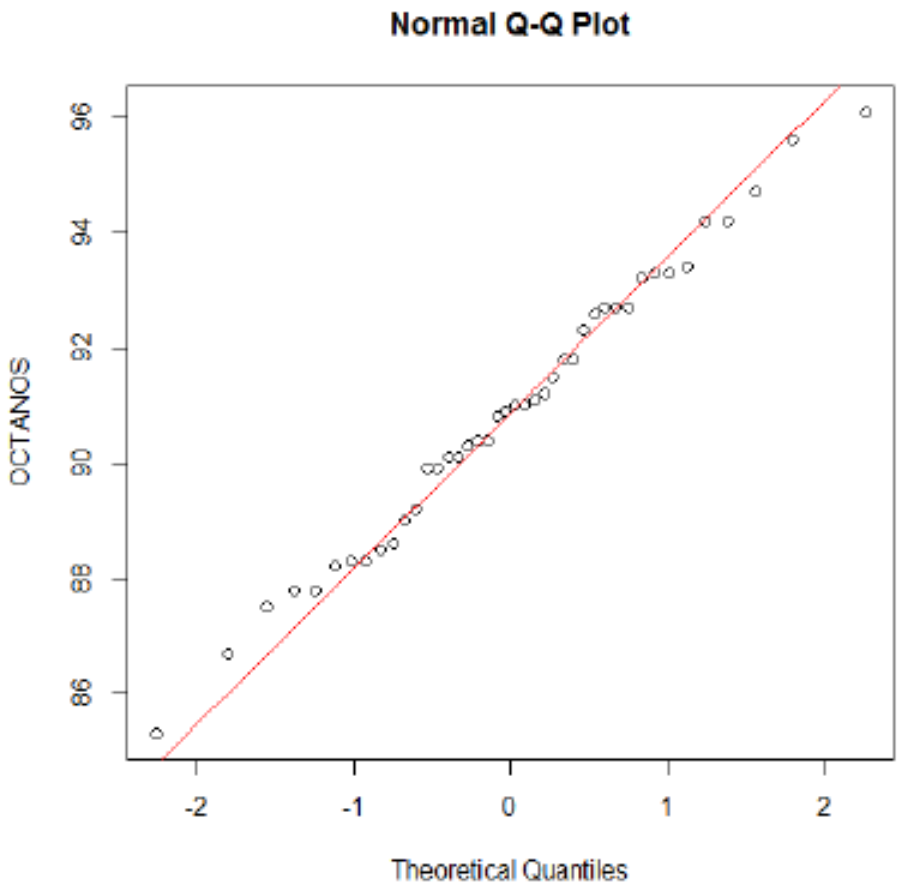
que se grafican versus el percentil $\frac{i-1/3}{n+1/3}$ de la normal, es decir $\phi^{-1}(\frac{i-1/3}{n+1/3})$ (algunos programas toman variaciones de estos valores)

Si los datos provienen de una distribución normal esperamos que el gráfico sea parecido a una recta.

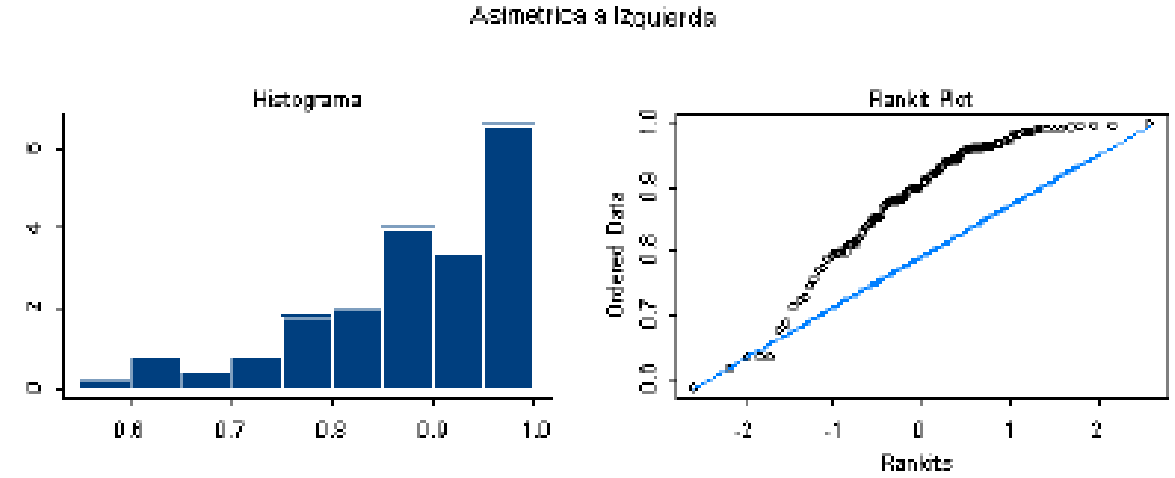
El alejamiento de la normalidad se ve reflejado por la forma del gráfico.

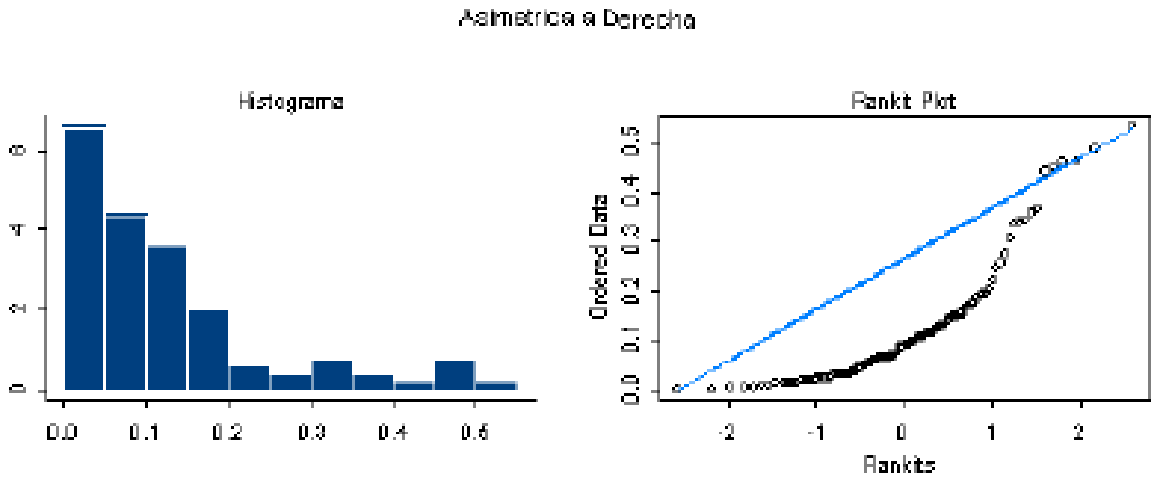
QQ-Plot o Grafico cuantil-cuantil

```
qqnorm(octanos.per,ylab="OCTANOS" )
qqline(octanos.per, col = 2)
```



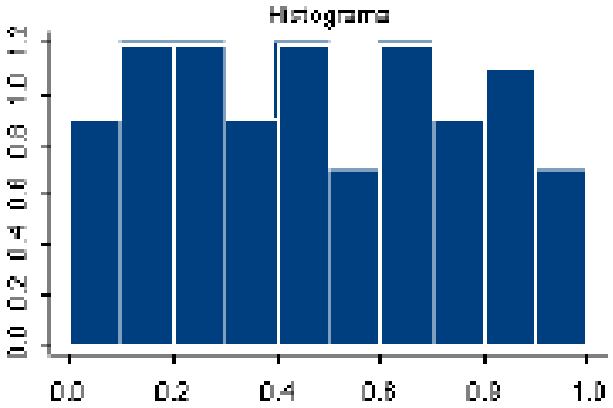
QQ-Plot o Grafico cuantil-cuantil

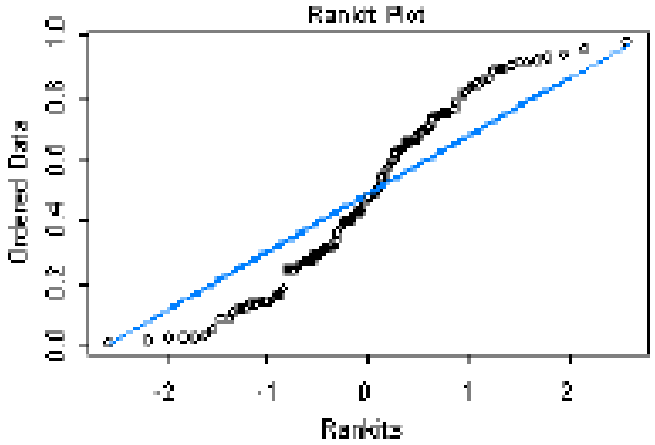




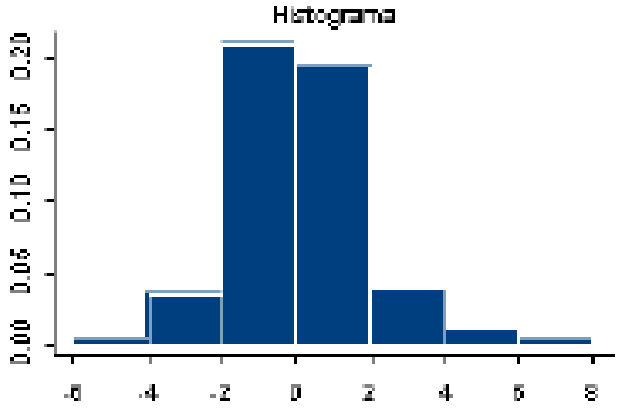
QQ-Plot o Grafico cuantil-cuantil

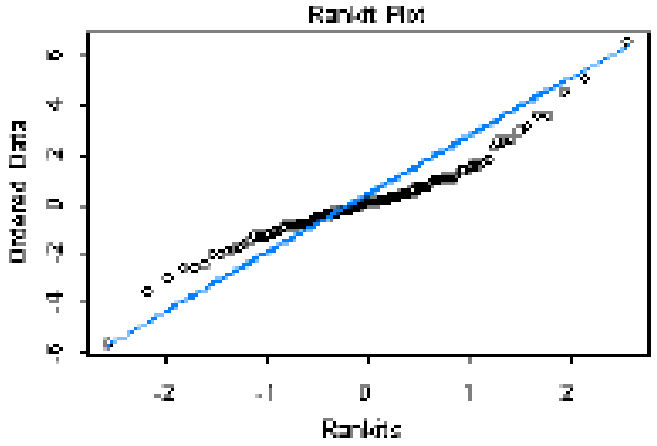
Simetría con colas Livianas

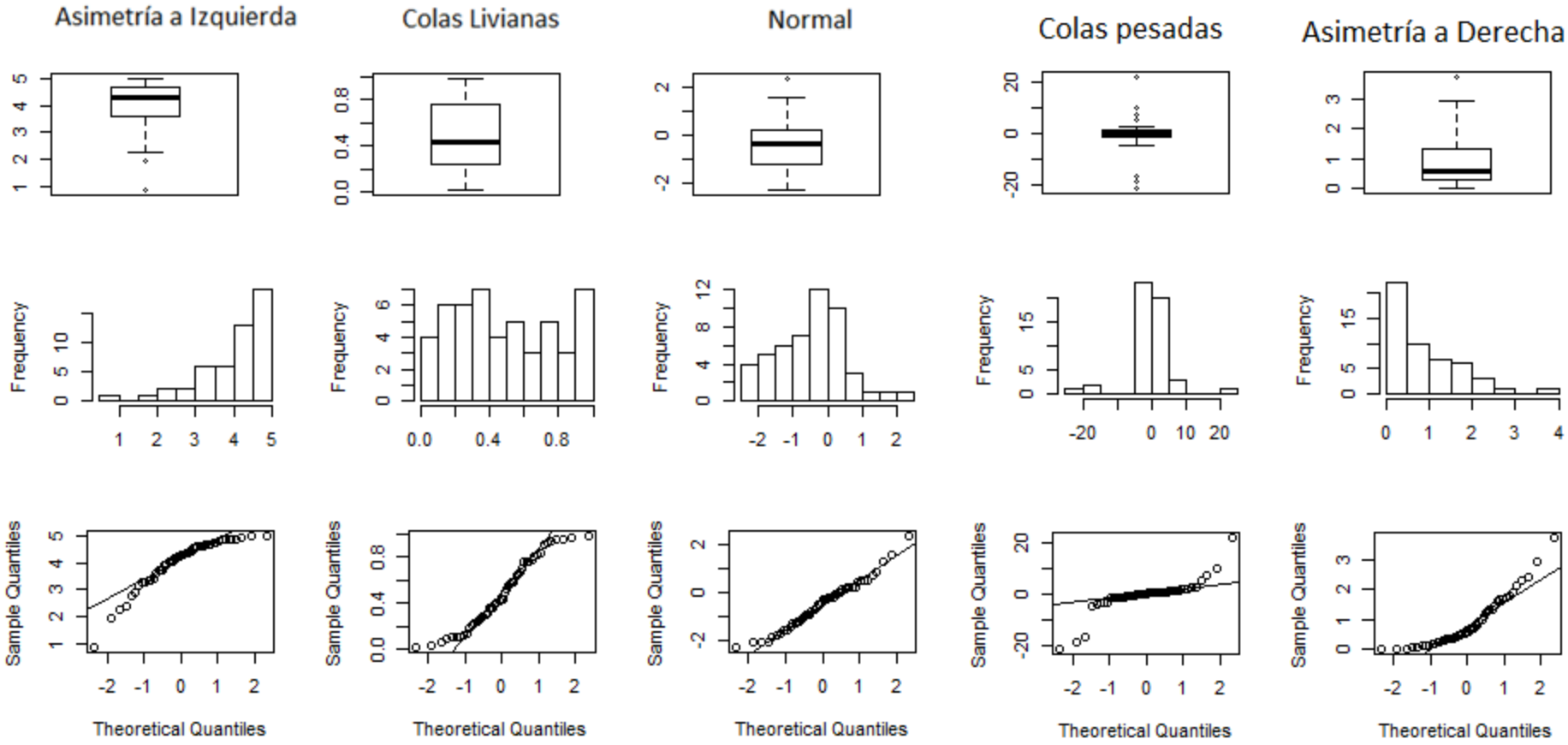




Simetría con colas Pesadas






Asimetría a Izquierda
Colas Livianas
Normal
Colas pesadas
Asimetría a Derecha
Frequency
Sample Quantiles
Theoretical Quantiles

ERRORES

Errores en el Proceso de Medición

En todo proceso de medición existen limitaciones dadas por

- los instrumentos usados
- el método de medición
- el observador

El mismo proceso de medición introduce errores o incertezas.

Ejemplo: Si usamos un termómetro para medir una temperatura, parte del calor del objeto fluye al termómetro, de modo que el resultado de la medición es un valor modificado del original debido a la interacción. Esta interacción podrá o no ser significativa, de acuerdo a si medimos la temperatura de un metro cúbico de agua si el volumen en cuestión es una fracción del mililitro.

ERRORES

Errores en el Proceso de Medición

Los instrumentos que usamos para medir como las magnitudes mismas son fuente de incertezas al momento de medir.

Los instrumentos tienen una *precisión finita, por lo tanto* siempre existe una variación mínima de la magnitud que puede detectar.

Ejemplo: con una regla graduada en milímetros, no podemos detectar variaciones menores que una fracción del milímetro.

Las magnitudes a medir no están definidas con infinita precisión.

Ejemplo: Si queremos medir el largo de una mesa, si usamos instrumentos cada vez más precisos empezamos a notar las irregularidades

ERRORES

Errores en el Proceso de Medición:

Tipos de Errores:

Errores sistemáticos: (sesgo) surgen por falla del equipo o del diseño. No se pueden evaluar realizando medidas repetidas.

Errores aleatorios: surgen por efectos de variables no controladas. Siempre está presente, nunca se pueden eliminar. Podemos minimizarlos y realizando medidas repetidas independientes se pueden evaluar, usando procedimientos estadísticos .

ERRORES

Errores en el Proceso de Medición

Precisión: la precisión de un instrumento o un método de medición está asociada a la sensibilidad o menor variación de la magnitud que se pueda detectar con dicho instrumento o método.

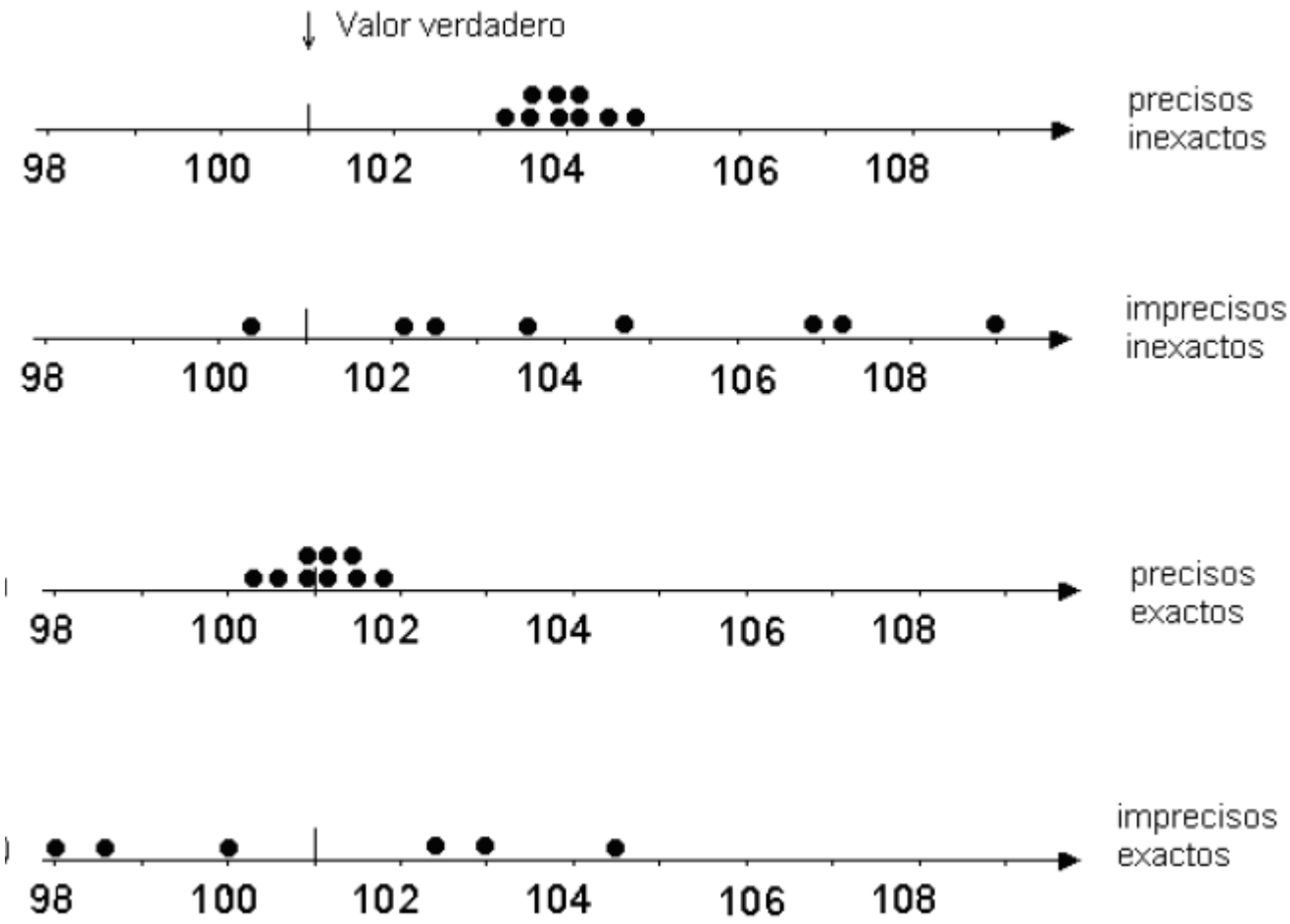
Ejemplo: un cronómetro es más preciso que un reloj común

Exactitud: La exactitud de un instrumento o método de medición está asociada a la calidad de la calibración del mismo, a la proximidad del valor verdadero.

Ejemplo: Imaginemos que el cronómetro que usamos es capaz de determinar la centésima de segundo pero adelanta dos minutos por hora, mientras que un reloj de pulsera común no lo hace. En este caso decimos que el cronómetro es todavía más preciso que el reloj común, pero menos exacto.

ERRORES

Errores en el Proceso de Medición: Precisión y Exactitud



ERRORES

Errores en el Proceso de Medición

Tenemos errores por diversos origenes

Error de apreciación (minima division de escala)
Error de definición (falta de definición del objeto)
Error de interacción (interacción en el método de medición)

ERRORES

Errores en el Proceso de Medición

$$\text{MEDICION} = \mu + \varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \ldots + \varepsilon_n$$

errores

Cantidad a medir: (desconocido pero no aleatorio)

$$\text{VAR}(\mu + \varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \ldots + \varepsilon_n) = \sigma^2_1 + \sigma^2_2 + \sigma^2_3 + \ldots + \sigma^2_n = \sigma^2$$

Si llamamos $\varepsilon = \varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \ldots + \varepsilon_n$, bajo ciertas condiciones:

$$X = \mu + \varepsilon \text{ donde } \varepsilon \sim N(0, \sigma^2) \text{ esto es equivalente a } X \sim N(\mu, \sigma^2)$$

MODELO

PROPAGACION DE INCERTIDUMBRE

$$X = \mu + \xi$$

si $Z = f(X)$ el método de propagación de la incertidumbre dice que es una buena aproximación tomar

$$V(Z) = [f'(\mu)]^2 V(\xi)$$

Para justificar el método basta tomar el polinomio de Taylor de orden 1 centrado en X, evaluarlo en μ para obtener que

$$f(X) = f(\mu) + f'(\mu)\xi + R$$

Si se desprecia el termino de error R y se utiliza las propiedades de la varianza.

$$V(f(X)) = V(f(\mu) + f'(\mu)\xi + R) = V(f'(\mu)\xi) = [f'(\mu)]^2 V(\xi)$$

Si tenemos más de una medición

$$(X_1, \ldots, X_n) = (\mu_1, \ldots, \mu_n) + (\xi_1, \ldots, \xi_n).$$

Si $Z = f(X_1, \ldots, X_n)$

$$V(Z) = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial f}{\partial \mu_i} \frac{\partial f}{\partial \mu_j} Cov(\xi_i, \xi_j)$$

donde las derivadas parciales están evaluadas en $(X_1, \ldots, X_n)$. Para el caso particular donde los errores no están correlacionados la ecuación anterior queda

$$V(Z) = \sum_{i=1}^{n} \frac{\partial f}{\partial \mu_i}^2 V(\xi_i)$$

INTERVALOS DE CONFIANZA

Cuando se obtiene una estimación puntual de un parámetro, es conveniente acompañar dicha estimación por una **medida** de la precisión de la estimación.

Un modo de hacerlo es informar el estimador y su error standard.

Otro modo es reemplazar la estimación puntual por un intervalo de valores posibles para el parámetro.

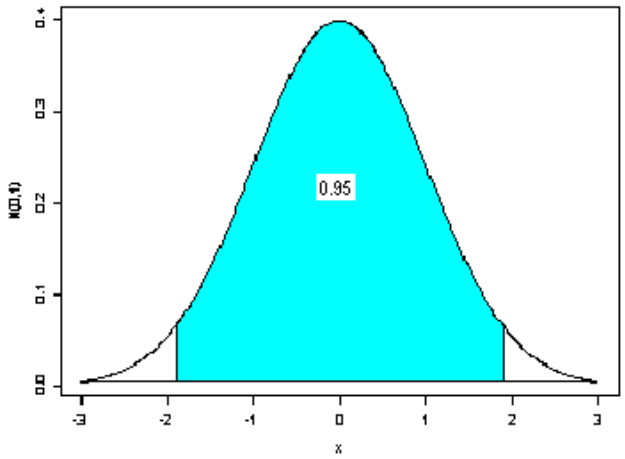
Ejemplo: Supongamos que tenemos una m.a. $X_1, X_2, \ldots, X_n$ de una distribución $N(\mu, \sigma_o^2)$ con varianza σ_o^2 conocida.

Por ser los datos normales, sabemos que

$$\overline{X} \sim N\left(\mu, \frac{\sigma_o^2}{n}\right) \quad \Leftrightarrow \quad \frac{\overline{X} - \mu}{\frac{\sigma_o}{\sqrt{n}}} \sim N(0,1)$$

y, por lo tanto,

$$P\left(-1.96 \leq \sqrt{n}\,\frac{\overline{X} - \mu}{\sigma_o} \leq 1.96\right) = 0.95$$



A partir de esta expresión obtenemos

$$P\left(-1.96\frac{\sigma_o}{\sqrt{n}} \le \overline{X} - \mu \le 1.96\frac{\sigma_o}{\sqrt{n}}\right) = 0.95 \quad \Leftrightarrow \quad P\left(\overline{X} - 1.96\frac{\sigma_o}{\sqrt{n}} \le \mu \le \overline{X} + 1.96\frac{\sigma_o}{\sqrt{n}}\right) = 0.95$$

Es decir, que la probabilidad de que el intervalo

$$\left[\overline{X} - 1.96\frac{\sigma_o}{\sqrt{n}}, \overline{X} + 1.96\frac{\sigma_o}{\sqrt{n}}\right]$$

contenga al verdadero valor del parámetro μ es 0.95. Este intervalo se denomina **intervalo de confianza para μ de nivel de confianza 0.95**.

A partir de esta expresión obtenemos

$$P\left(-1.96\frac{\sigma_o}{\sqrt{n}} \le \overline{X} - \mu \le 1.96\frac{\sigma_o}{\sqrt{n}}\right) = 0.95 \quad \Leftrightarrow \quad P\left(\overline{X} - 1.96\frac{\sigma_o}{\sqrt{n}} \le \mu \le \overline{X} + 1.96\frac{\sigma_o}{\sqrt{n}}\right) = 0.95$$

Es decir, que la probabilidad de que el intervalo

$$\left[\overline{X} - 1.96\frac{\sigma_o}{\sqrt{n}}, \overline{X} + 1.96\frac{\sigma_o}{\sqrt{n}}\right]$$

contenga al verdadero valor del parámetro μ es 0.95. Este intervalo se denomina **intervalo de confianza para μ de nivel de confianza 0.95**.

En general, tendremos

$$P\left(-z_{\alpha/2} \le \sqrt{n}\,\frac{\overline{X} - \mu}{\sigma_o} \le z_{\alpha/2}\right) = 1 - \alpha$$

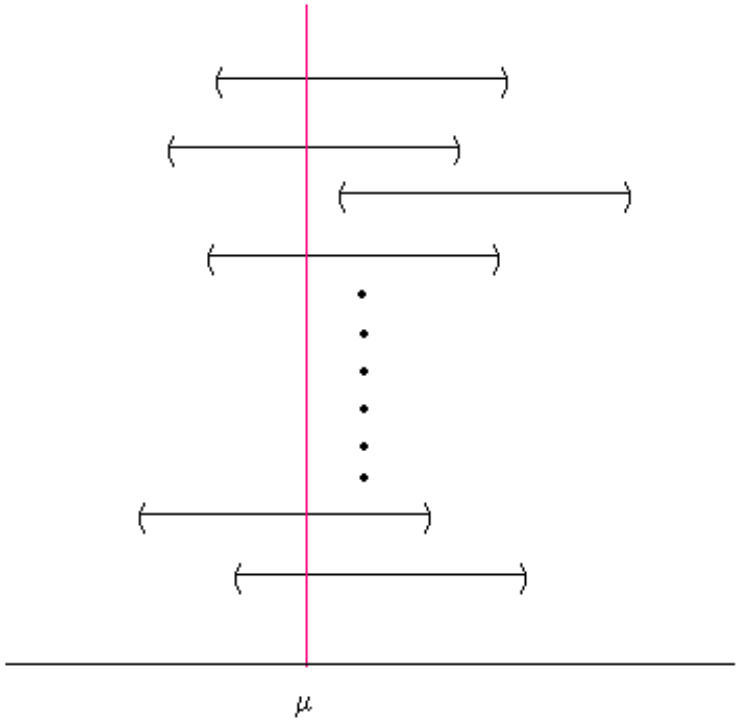
luego el siguiente **intervalo de confianza es de nivel 1 - α para μ**

$$\left[\overline{X} - z_{\varepsilon/2}\,\frac{\sigma_o}{\sqrt{n}}, \overline{X} + z_{\varepsilon/2}\,\frac{\sigma_o}{\sqrt{n}}\right]$$

Interpretación:

Supongamos que, en base a diferentes muestras calculamos los correspondientes intervalos de confianza para μ.

Entonces el (1 - α) 100% de ellos contendrán al verdadero valor μ.



,

Ejemplo:

Supongamos que tenemos una muestra normal con n=49 con verdadero valor del desvío standard es $\sigma_o = 35$ y que se observa $\bar{x} = 160$ y construimos un intervalo de confianza para la media de nivel 0.95.

Como las v.a. son normales y la varianza es conocida, el intervalo para μ será de la forma

$$\left(\overline{X} - z_{\varepsilon/2} \frac{\sigma_o}{\sqrt{n}}, \overline{X} + z_{\varepsilon/2} \frac{\sigma_o}{\sqrt{n}} \right)$$

Como, $z_{\alpha/2} = z_{0.025} = 1.96$ y $\sigma_o = 35, n = 49$ obtenemos

$$\left(160 - 1.96 \frac{35}{\sqrt{49}}, 160 + 1.96 \frac{35}{\sqrt{49}} \right) = (160 - 9.8, 160 + 9.8) = (150.2, 169.8)$$

INTERVALOS DE CONFIANZA PARA LOS PARAMETROS DE LA DISTRIBUCION NORMAL

Distribución t:

Sean dos v.a. Z ~N(0,1) y U $\sim \chi^2_n = \Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$ independientes, entonces

$$T = \frac{Z}{\sqrt{U/n}} \sim t_n$$

Se dice que *T* tiene distribución ***t* de Student con *n* grados de libertad**.

Esta distribución está tabulada para diferentes valores de *n*. Su densidad es simétrica respecto al 0 y tiene forma de campana, pero tiene colas más pesadas que la distribución normal standard.

Cuando n tiende a infinito, la distribución de Student tiende a la distribución normal standard.

Proposición: Sea $X_1, X_2, \ldots, X_n$ una m.a. de una distribución N(μ, σ^2), entonces

a) $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \Leftrightarrow \quad \sqrt{n}\frac{\overline{X}-\mu}{\sigma} \sim N(0,1)$

b) $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ con $S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$

c) $\overline{X}$ y S^2 son independientes

d) $\sqrt{n}\frac{\overline{X}-\mu}{S} \sim t_{n-1}$

Intervalo de confianza para la media de la distribución normal con varianza desconocida:

Sea $X_1, X_2, \ldots, X_n$ una m.a. de una distribución N(μ, σ^2), entonces

$$\sqrt{n}\,\frac{\bar{X}-\mu}{S} \sim t_{n-1}$$

$$P\left(-t_{n-1,\alpha/2} \le \sqrt{n}\,\frac{\bar{X}-\mu}{S} \le t_{n-1,\alpha/2}\right) = 1-\alpha$$

de donde se deduce el siguiente **intervalo de confianza de nivel 1 - α para μ,**

$$\left[\bar{X} - t_{n-1,\alpha/2}\,\frac{S}{\sqrt{n}}, \bar{X} + t_{n-1,\alpha/2}\,\frac{S}{\sqrt{n}}\right]$$

Intervalo de confianza para la varianza de la distribución normal con media conocida

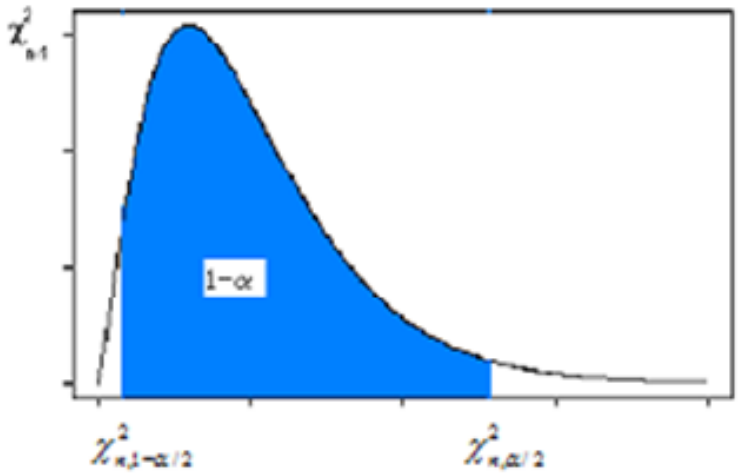
Sea $X_1, X_2, \ldots, X_n$ una m.a. de una distribución $N(\mu_o, \sigma^2)$, con media μ_o conocida, entonces

$$\frac{X_i - \mu_o}{\sigma} \sim N(0,1) \quad \forall 1 \le i \le n \quad \Rightarrow \quad \left(\frac{X_i - \mu_o}{\sigma}\right)^2 \sim \chi_1^2 = \Gamma\left(\frac{1}{2}, \frac{1}{2}\right) \quad \forall 1 \le i \le n$$

Como además las v.a. son independientes

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu_o}{\sigma}\right)^2 \sim \chi_n^2 = \Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$$

¿Cómo elegimos los percentiles de la distribución χ^2 que encierran un área igual a $1 - \alpha$?



Los elegimos de manera tal que quede un área igual a $\alpha/2$ en cada extremo. Entonces,

$$P\left(\chi^2_{n,1-\alpha/2} \le \frac{\sum_{i=1}^{n}(X_i-\mu_o)^2}{\sigma^2} \le \chi^2_{n,\alpha/2}\right) = 1-\alpha$$

Se obtiene el siguiente intervalo

$$\left[\frac{\sum_{i=1}^{n}(X_i-\mu_o)^2}{\chi^2_{n,\alpha/2}}, \frac{\sum_{i=1}^{n}(X_i-\mu_o)^2}{\chi^2_{n,1-\alpha/2}}\right]$$

Intervalo de confianza para la varianza de la distribución normal con media desconocida

Sea $X_1, X_2, \ldots, X_n$ una m.a. de una distribución N(μ, σ^2), entonces

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Por lo tanto,

$$P\left(\chi^2_{n-1,1-\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{n-1,\alpha/2} \right) = 1-\alpha$$

Se obtiene el siguiente intervalo

$$\left[\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}} \right]$$

Volviendo al ejemplo:

Supongamos ahora que la varianza es desconocida, pero que el valor observado de S es s=35.

El correspondiente intervalo de confianza para μ será de la forma

$$\left(\bar{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right)$$

con $t_{n-1,\alpha/2} = t_{48,0.025} = 2.01$. Obtenemos

$$\left(160 - 2.01\frac{35}{\sqrt{49}}, 160 + 2.01\frac{35}{\sqrt{49}}\right) = (160 - 10.05, 160 + 10.05) = (149.95, 170.05)$$

y como es lógico, resulta un intervalo más largo.

Ahora un intervalo para σ

Suponiendo como antes que observamos $\bar{x} = 160$ y $s = 35$, hallemos un intervalo de confianza para σ de nivel 0.95.

Por tratarse de una muestra normal con media desconocida, el intervalo para σ^2 será de la forma

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1,\varepsilon/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,1-\varepsilon/2}} \right)$$

con $\chi^2_{n-1,\alpha/2} = \chi^2_{48,0.025} = 69.02$ y $\chi_{n-1,1-\alpha/2} = \chi_{48,0.975} = 30.75$. Obtenemos

$$\left(\frac{48 \cdot 35^2}{69.02}, \frac{48 \cdot 35^2}{30.75} \right) = (851.93, 1912.20)$$

y tomando raíz cuadrada, un intervalo de confianza para σ de nivel 0.95 será

$$\left(\sqrt{\frac{48 \cdot 35^2}{69.02}}, \sqrt{\frac{48 \cdot 35^2}{30.75}} \right) = \left(\sqrt{851.93}, \sqrt{1912.20} \right) = (29.19, 43.73)$$