Estadística Descriptiva o Análisis Exploratorio de Datos

- Nos ayudan a organizar la información que nos dan los datos de manera de detectar algún patrón de comportamiento así como también apartamientos importantes al modelo subyacente

- Nos presentan los datos de modo tal que sobresalga su estructura.

Explorar los datos, debe ser la primera etapa de todo análisis de datos.

Apuntes: Notas de Liliana Orellana – Clases de Ana Bianco

Estadística descriptiva

Hay varias formas de organizar los datos:

- Métodos gráficos: permiten detectar tanto las características sobresalientes como las características inesperadas.

- Medidas resumen: resumirlos en uno o dos números que pretenden caracterizar el conjunto con la menor distorsión o perdida de información posible.

Estadística descriptiva

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

(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

ESTADÍSTICO : una medida resumen calculada sobre la muestra

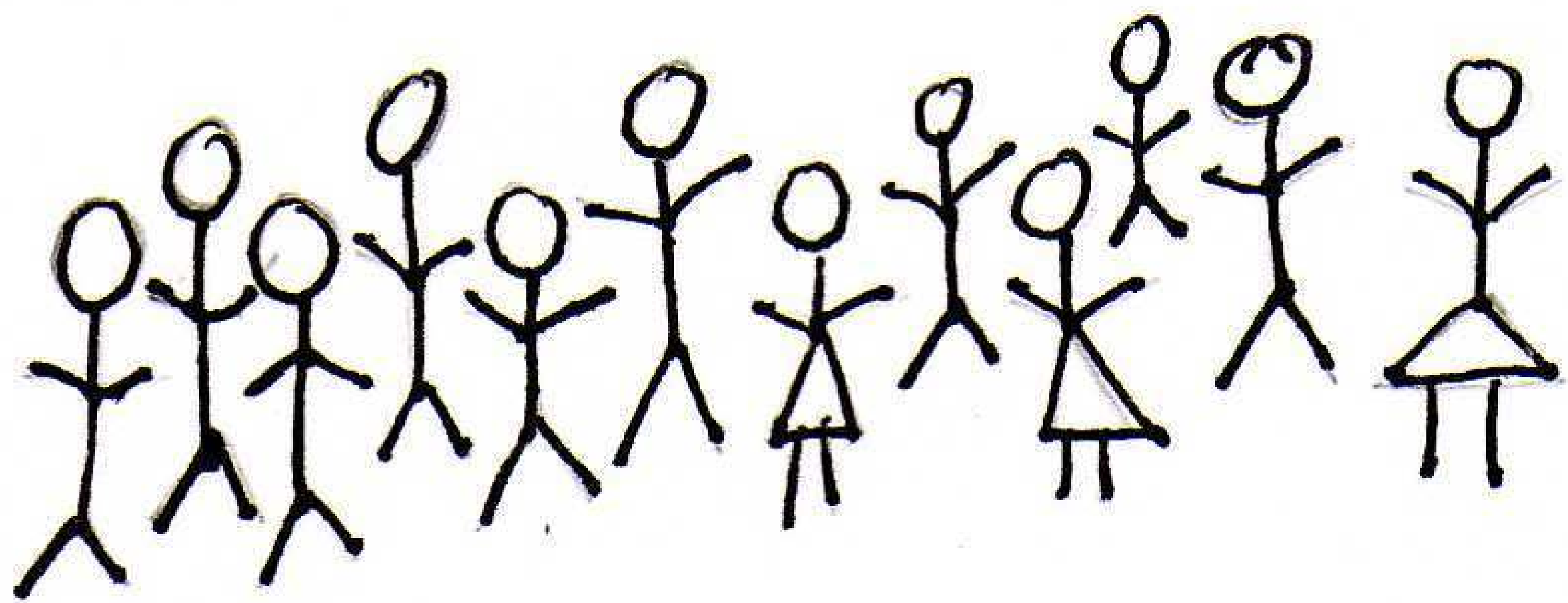
Estadística descriptiva

Cuando existen datos para toda la población (CENSO) no hay necesidad de usar métodos estadísticos, ya que es 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



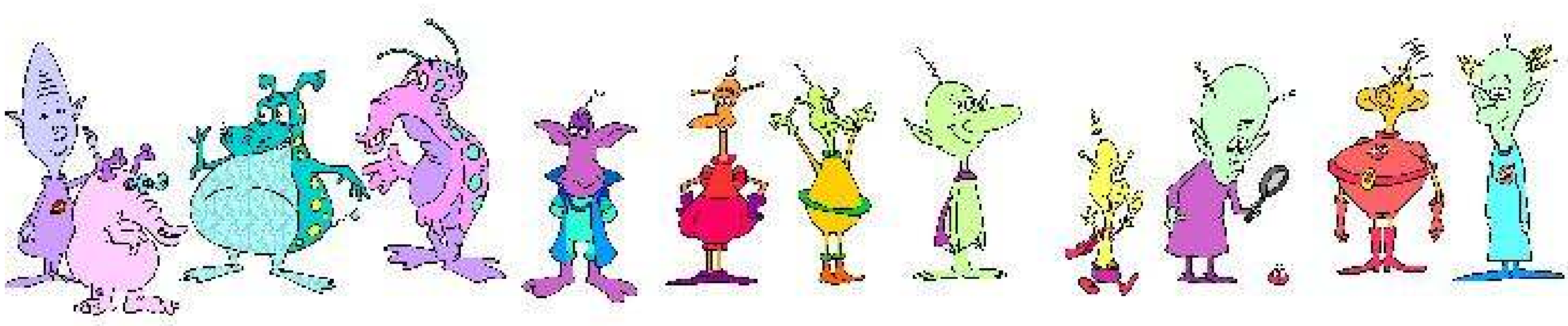
CENSO

~~CENSO~~

Limitaciones

Imposibilidad

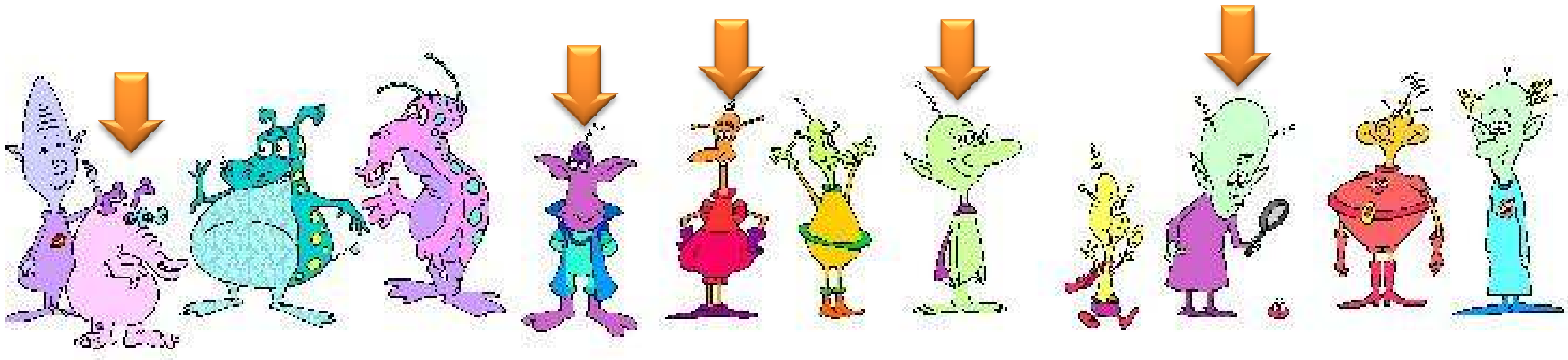
Población



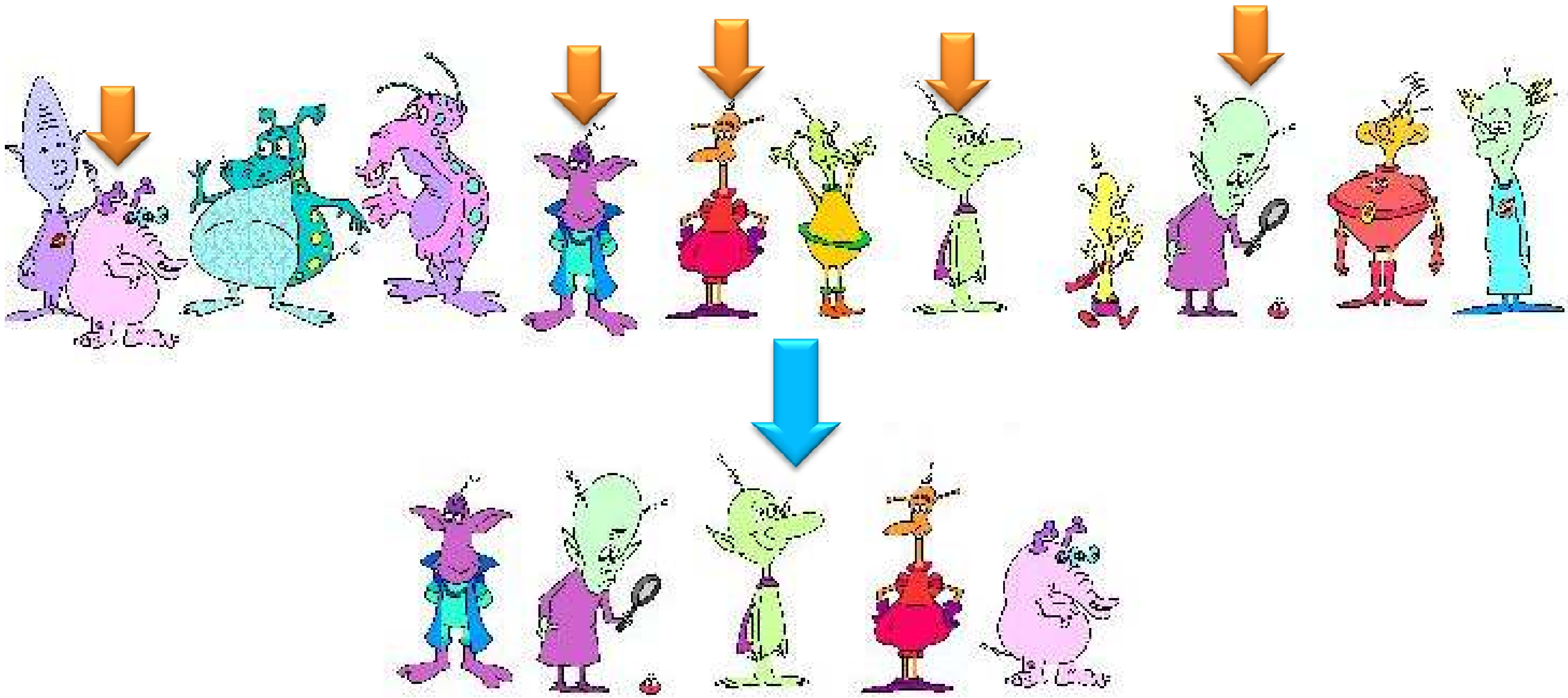
Población

Población

Población



Población



MUESTRA

Estadística descriptiva

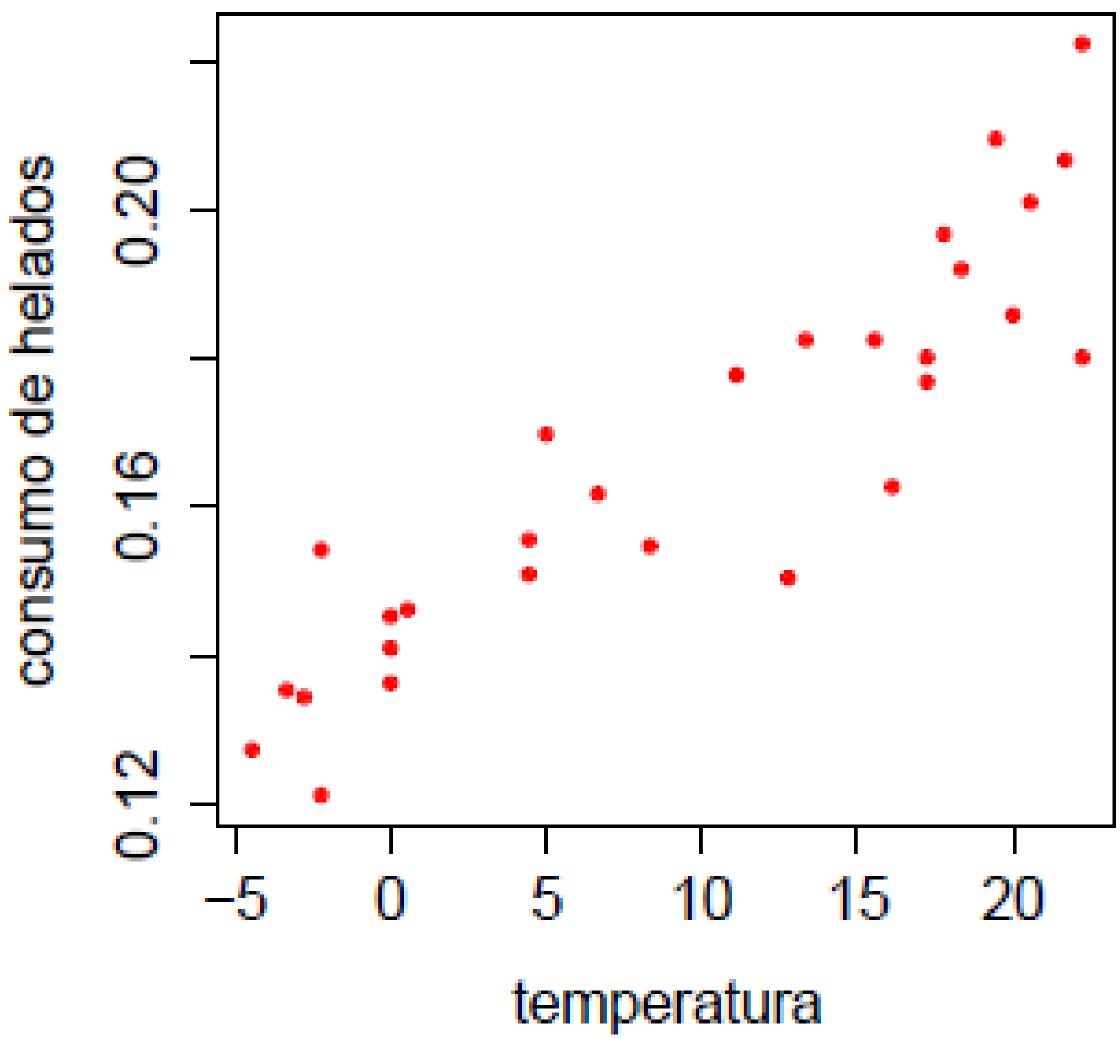
VARIABLE: Una variable es una característica que varía de individuo en individuo.

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

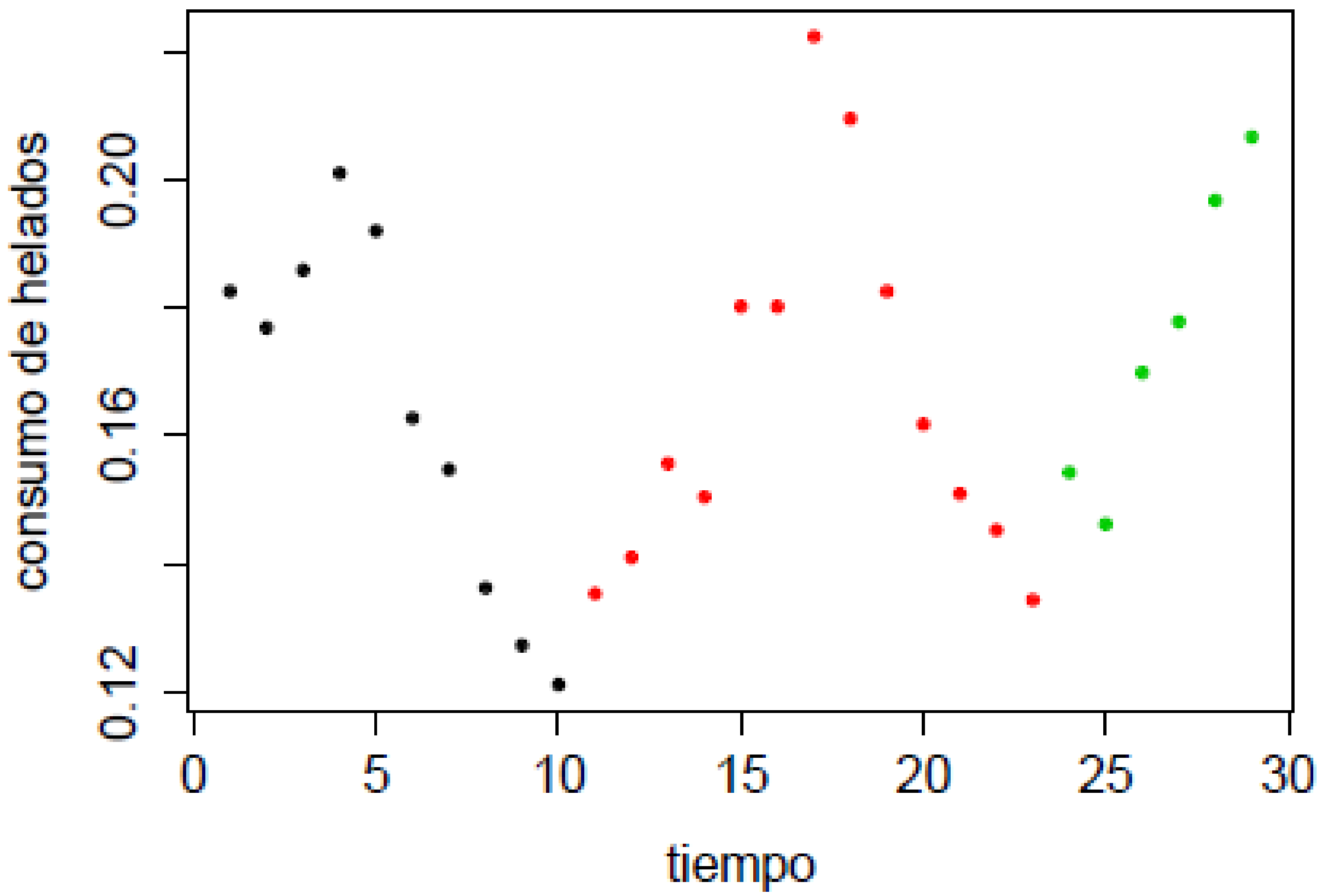
DATOS: son los valores de la variable en estudio

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.

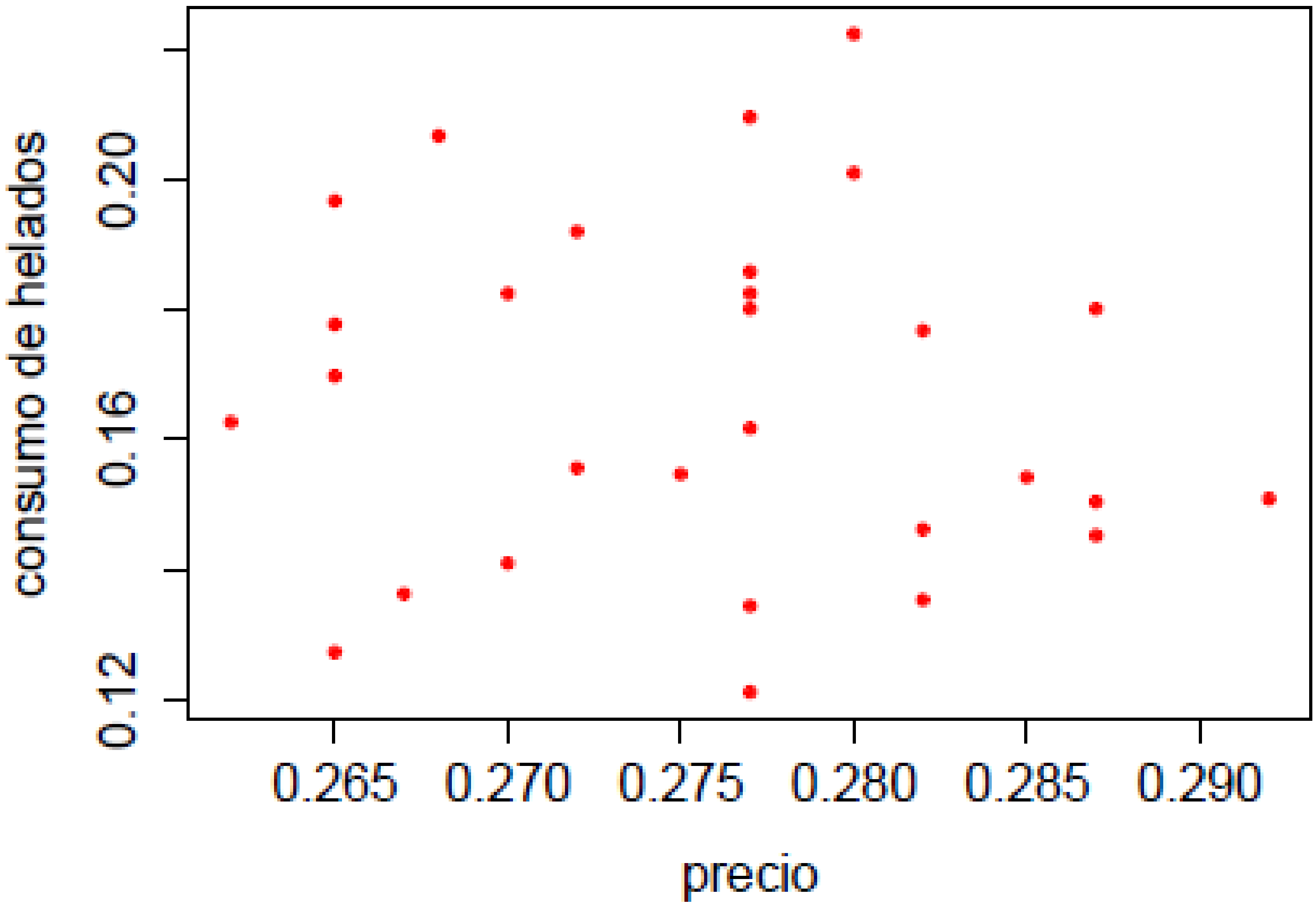
Consumo mensual de helado entre marzo de 1951 y julio de 1953



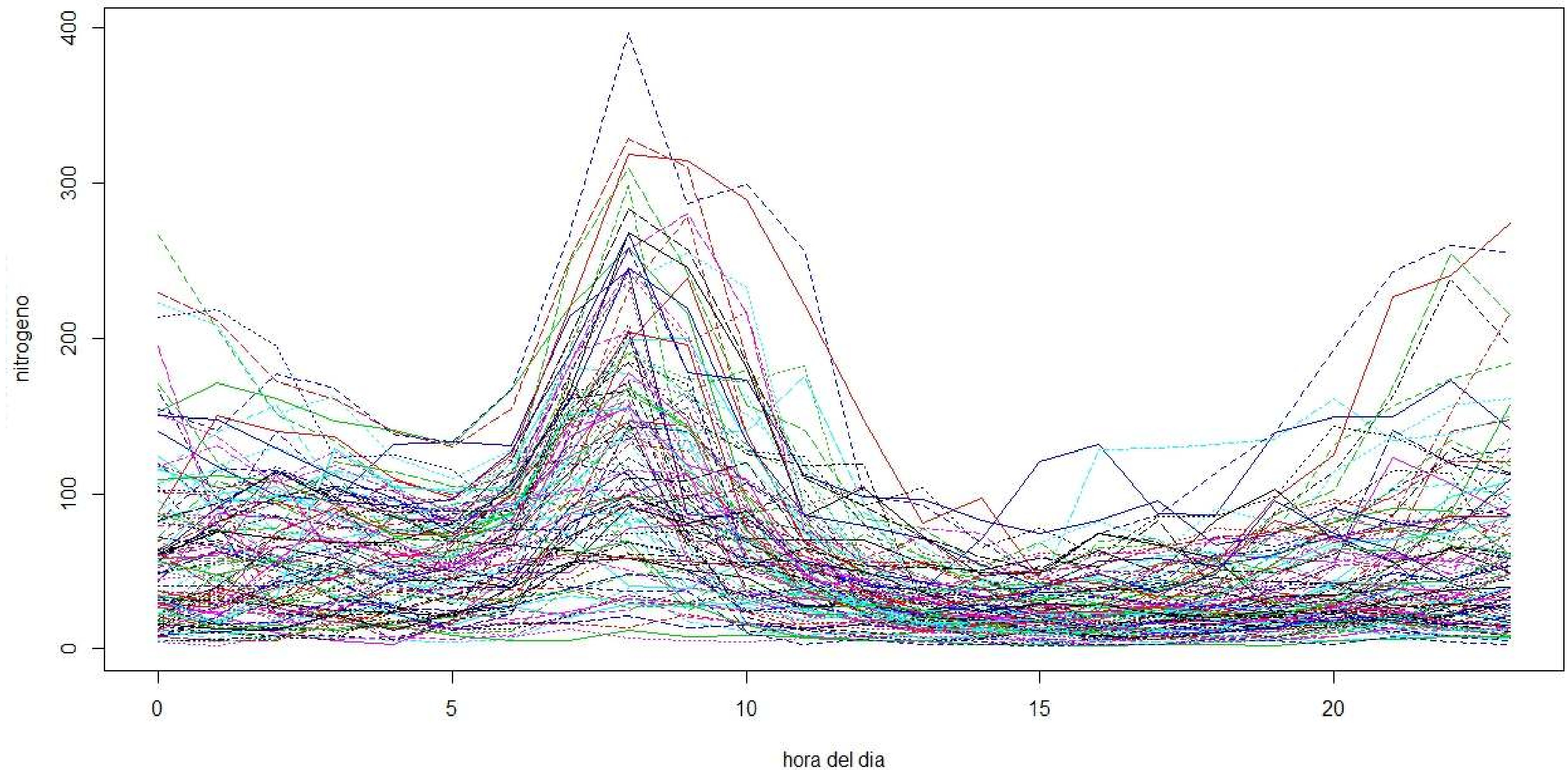
Consumo mensual de helado entre marzo de 1951 y julio de 1953



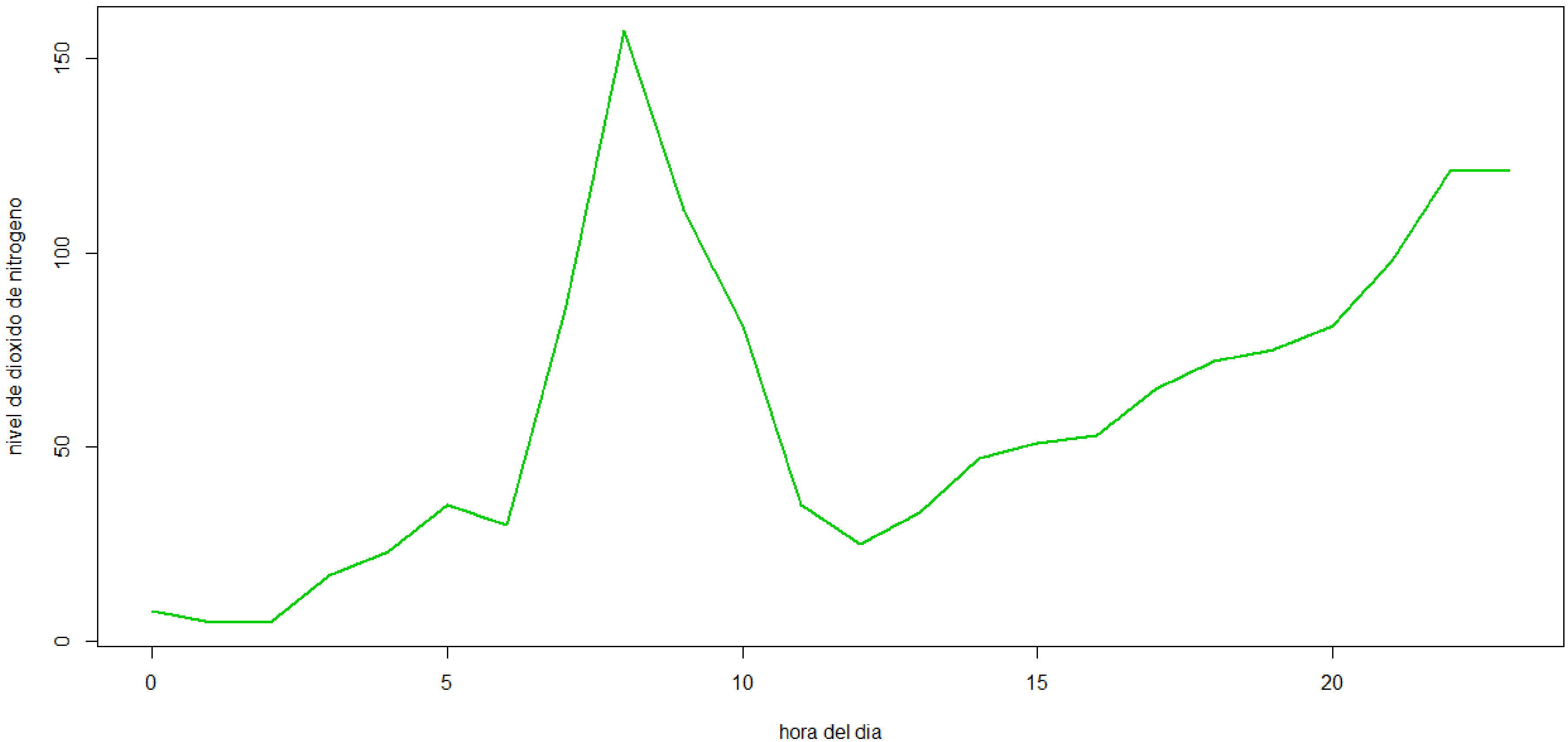
Consumo mensual de helado entre marzo de 1951 y julio de 1953



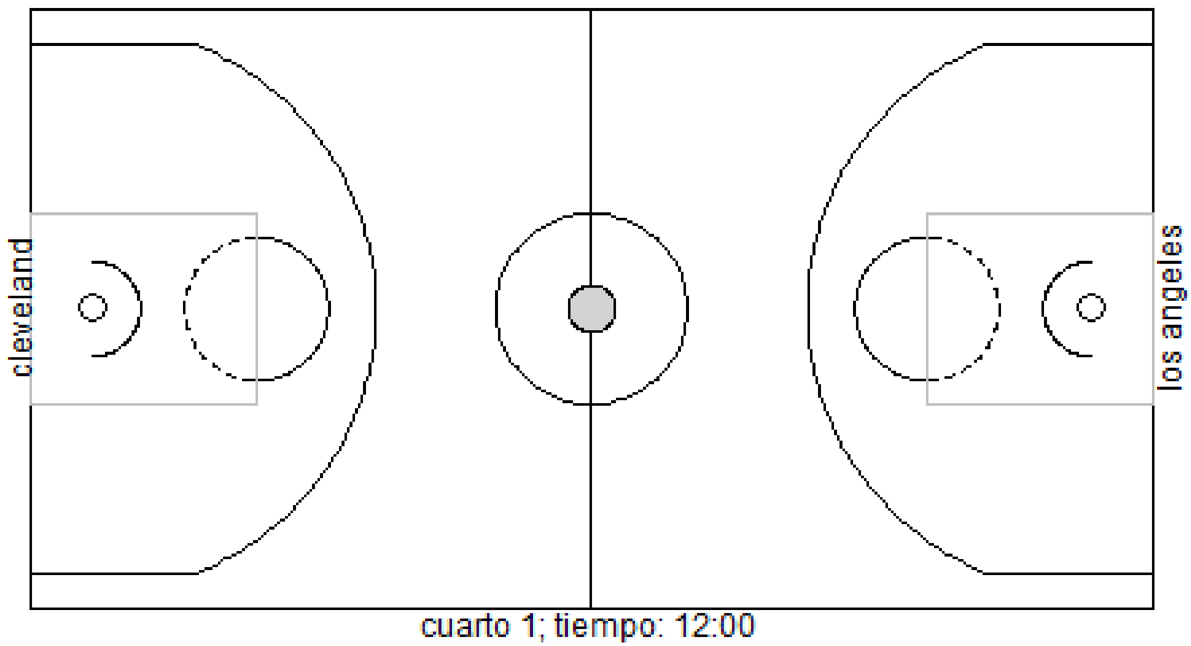
115 Curvas del nivel de dióxido de nitrógeno en Barcelona a lo largo del día durante el 2005



115 Curvas del nivel de dióxido de nitrógeno en Barcelona a lo largo del día durante el 2005



Partido de la NBA diciembre de 2009
Cleveland vs Los Angeles Lakers



- Los datos guardan información pero será necesario analizarlos o procesarlos para obtener respuestas y sacar 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 de números?

¿Difieren algunos pocos números notablemente del resto?

Estadística descriptiva

TIPOS DE DATOS:

Categóricos:

- dicotómicos: (dos categorías) (sexo, genero, fuma o no fuma)
- mas categorías:
  - nominales: No existe orden obvio entre las categorías.

    (país de origen, estado civil, diagnóstico.)
  - ordinales: Existe un orden natural entre las categorías.

    (Tabaquismo: No fuma / ex-fumador / fuma ≤ 10 cigarrillos diarios / fuma > 10 cigarrillos diarios)

    (Severidad de la patología: Ausente/leve/moderado/severo)

Estadística descriptiva

TIPOS DE DATOS:

Numéricos: el resultado de la observación o medición es un número

- Discretos: La variable sólo puede tomar un cierto conjunto de valores posibles. En general, aparecen por conteo.

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

- Continuos: Generalmente son el resultado de una medición que se expresa en unidades. Las mediciones pueden tomar teóricamente un conjunto infinito de valores posibles dentro de un rango. En la práctica los valores posibles de la variable están limitados por la precisión del método de medición o por el modo de registro.

  (altura, peso, pH, nivel de colesterol en sangre.)

Estadística descriptiva

El tipo de dato nos permite decidir que 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.

Estadística descriptiva

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.

Métodos Gráficos

REPRESENTACIÓN DE DATOS NUMERICOS

HISTOGRAMAS

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

Una virtud del gráfico de tallo-hojas es que retiene los valores de las observaciones, sin embargo, esta característica puede ser una desventaja para gran cantidad de datos.

Construir manualmente un histograma es más laborioso que construir un gráfico de tallo-hojas, pero la mayoría de los paquetes estadísticos producen histogramas.

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 de dicho intervalo.

Métodos Gráficos

REPRESENTACIÓN DE DATOS NUMERICOS

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**

Métodos Gráficos

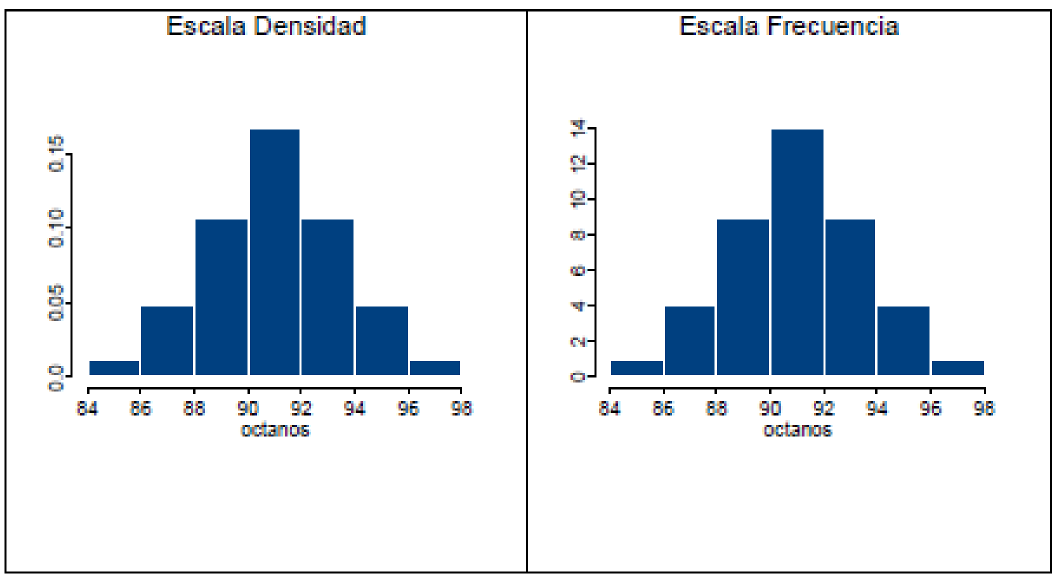
REPRESENTACIÓN DE DATOS NUMERICOS

HISTOGRAMAS

Los comandos son

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

Histogramas para datos de OCTANOS



Métodos Gráficos

REPRESENTACIÓN DE DATOS NUMERICOS

HISTOGRAMAS

No es necesario que todos los intervalos tengan la misma longitud, pero es recomendable que así sea. Esto facilita la lectura.

El histograma representa la frecuencia o la frecuencia relativa a través del **área y no a** través de la altura.

Es recomendable tomar

$$\textbf{Altura del rectángulo} = \frac{\textbf{frecuencia relativa}}{\textbf{Long. del intervalo}}$$

De esta manera el área es 1 y dos histogramas son fácilmente comparables independientemente de la cantidad de observaciones en las que se basa cada uno.

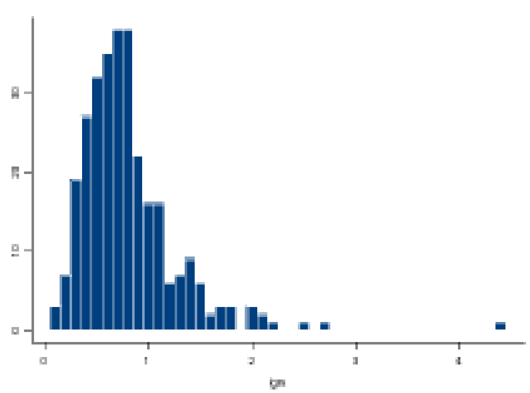
Métodos Gráficos

REPRESENTACIÓN DE DATOS NUMERICOS

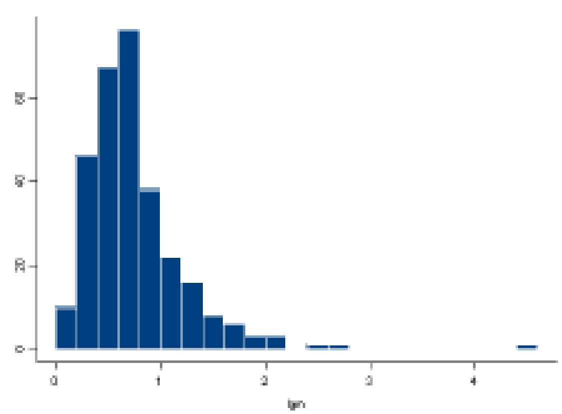
HISTOGRAMAS

Ejemplo: Concentración de Inmunoglobulina

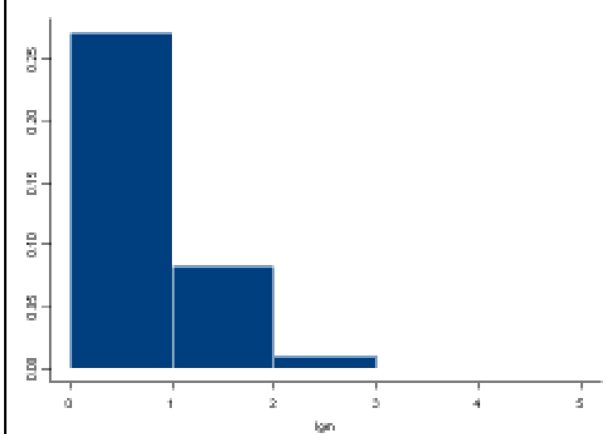
En este ejemplo vemos cómo la elección del ancho de las clases afecta el gráfico.



Longitud de Clase= 0.1 g/l



Longitud de Clase= 0.2 g/l



Longitud de Clase=1g/l

Métodos Gráficos

REPRESENTACIÓN DE DATOS NUMERICOS

HISTOGRAMAS

No existen criterios óptimos para elegir la cantidad de intervalos. En general, entre 8 y 15 intervalos deberían ser suficientes. Muchos o muy pocos intervalos puede ser poco informativo.

Se busca un equilibrio entre un histograma muy irregular y uno demasiado suavizado.

Métodos Gráficos

REPRESENTACIÓN DE DATOS NUMERICOS

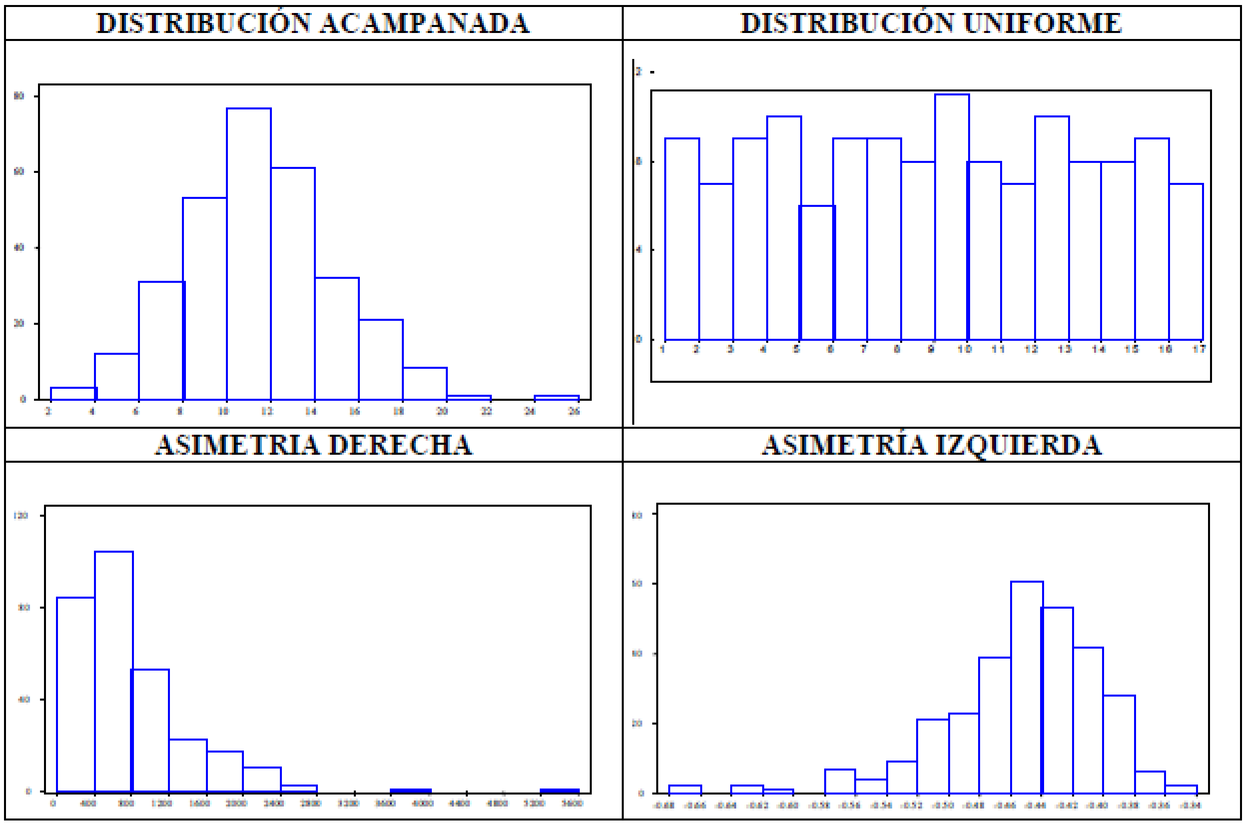
HISTOGRAMAS

¿Qué podemos ver en un histograma?

Rango de variación de los datos (Mínimo – Máximo)

Intervalos más frecuentes

Simetría o Asimetría



Métodos Gráficos

REPRESENTACIÓN DE DATOS NUMERICOS

HISTOGRAMAS

En que difieren un gráfico de barras y un histograma?

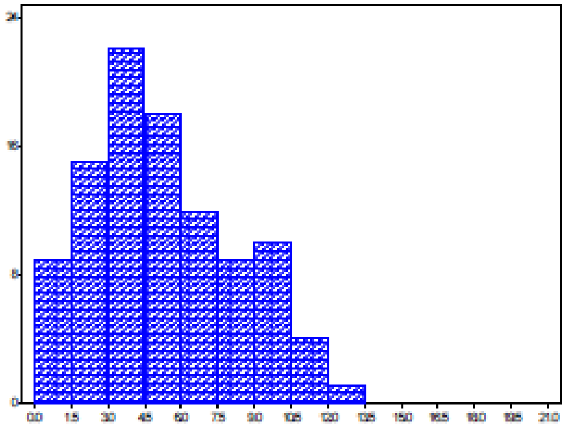
-El gráfico de barras representa el porcentaje en la altura de la barra. Mientras que en un histograma el porcentaje se representa en el área de la barra.

- 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.

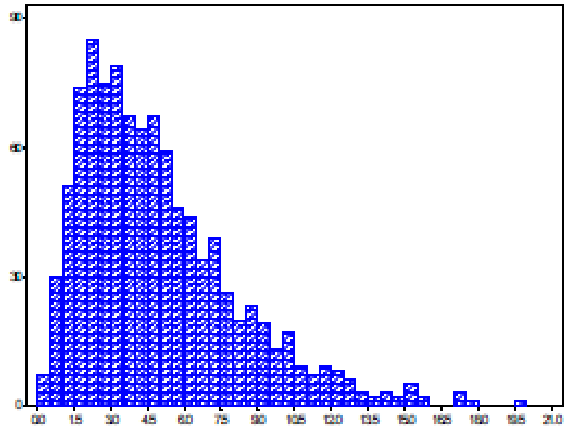
Métodos Gráficos

REPRESENTACIÓN DE DATOS NUMERICOS

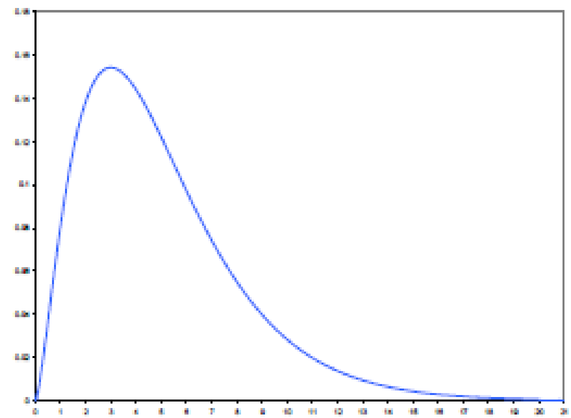
HISTOGRAMAS



Muestra n = 100



Muestra n = 1000



Población

Medidas resumen

Resumiremos la información de los datos mediante medidas de fácil interpretación que reflejen sus características más relevantes. Las medidas resúmenes 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 cuan variable es un conjunto
de datos.

Medidas resumen

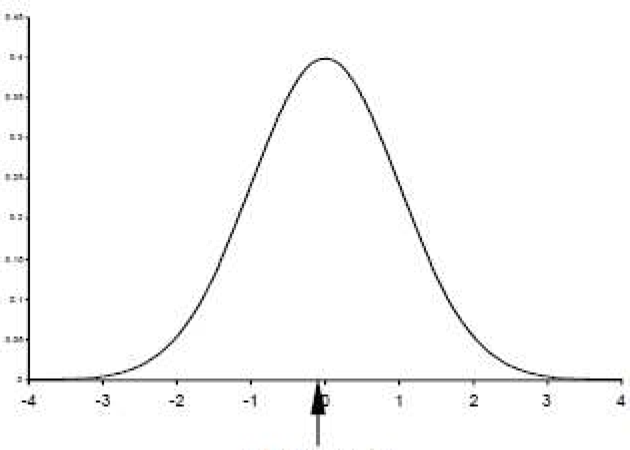
Medidas de Posición o Centrado

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

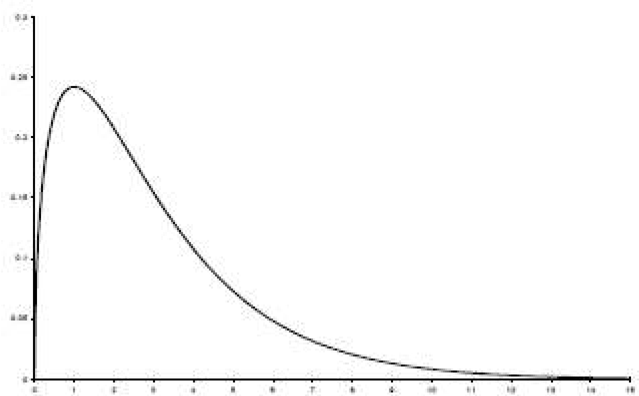
Buscamos un valor típico que represente a los datos.

Si la distribución es simétrica diferentes medidas darán resultados similares y hay un claro valor de centrado.

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



CENTRO



¿CENTRO?

Medidas resumen

Medidas de Posición o Centrado

Promedio o Media Muestral

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

$$\bar{x} = \frac{x_1 + x_2 + \ldots\ldots + x_n}{n}$$

Medidas resumen

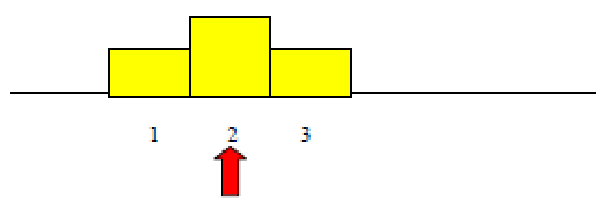
Medidas de Posición o Centrado

Ejemplo: Fuerza de compresión de muestras de Aleación de Aluminio-Litio

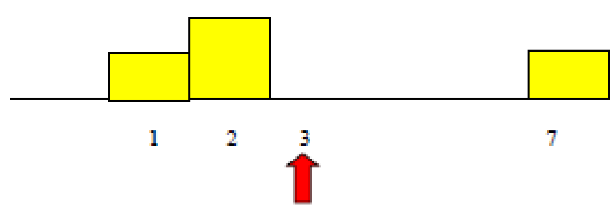
$$\bar{x} = \frac{\sum_{i=1}^{45} x_i}{45} = \frac{5350}{45} = 118.89$$

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 resumen

Medidas de Posición o Centrado

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 (hombres/ mujeres) $(\bar{x}_1, n_1, \bar{x}_2, n_2)$

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

Medidas resumen

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.

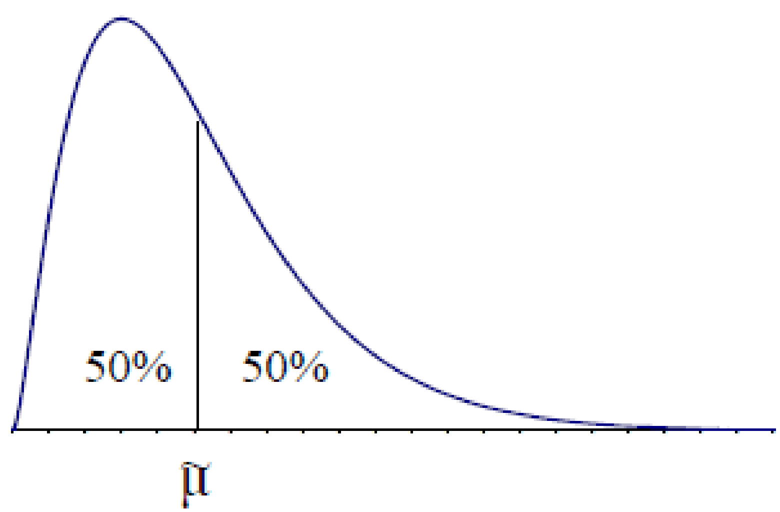
La mediana es resistente a la presencia de datos atípicos.

Medidas resumen

Medidas de Posición o Centrado

Mediana poblacional

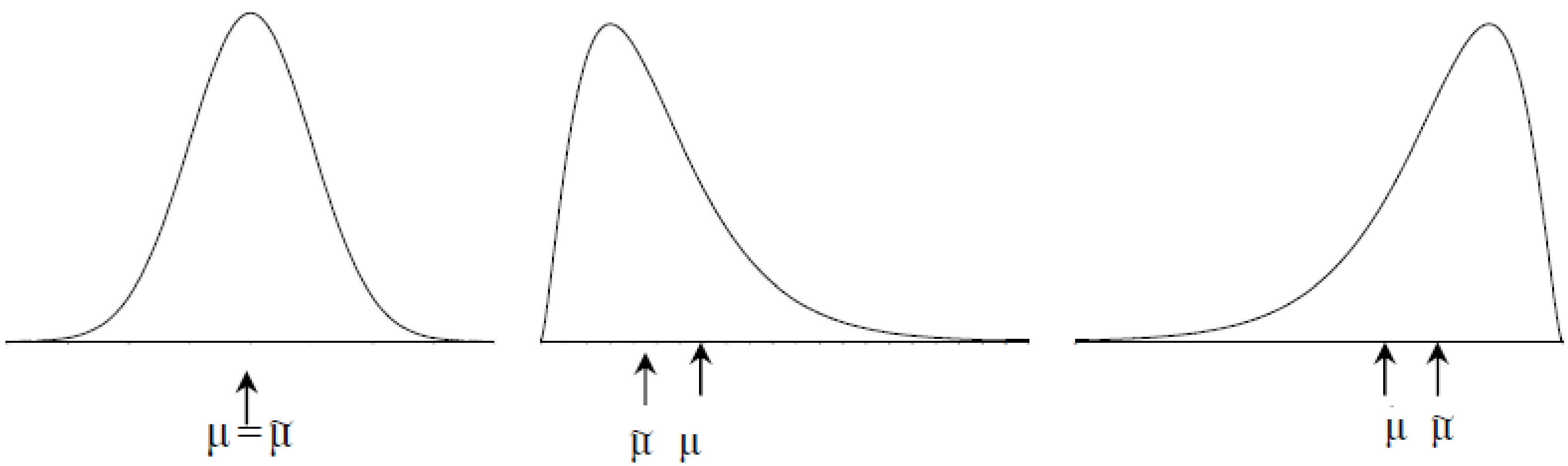
La *mediana poblacional se define de modo equivalente a la mediana muestral y es el valor* de la variable por debajo del cual se encuentra a lo sumo el 50% de la población y por encima del cual se encuentra a lo sumo el 50% de la población.



Medidas resumen

Medidas de Posición o Centrado

Mediana poblacional



Medidas resumen

Medidas de Posición o Centrado

Dra. Ana M. Bianco

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 resumen

Medidas de Posición o Centrado

Medias α-Podadas

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

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

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 resumen

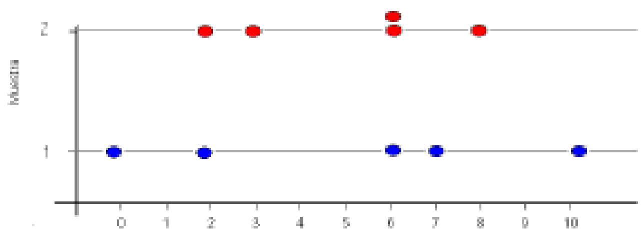
Medidas de Dispersión

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



$$\bar{X} = \bar{Y} = 5$$
$$\tilde{X} = \tilde{Y} = 6$$

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

Medidas resumen

Medidas de Dispersión

Rango Muestral

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

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

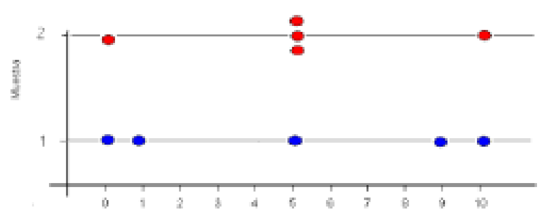
Ejemplo: en nuestros conjuntos de datos:

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

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

Veamos otro ejemplo:

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



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

Medidas resumen

Medidas de Dispersión

Estadística (Química) 2012
FCEyN UBA

Dra. Ana M. Bianco

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: en los dos ejemplos anteriores obtenemos

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

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

Medidas resumen

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 resumen

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. Llamaremos cuartil inferior (o primer cuartil) al percentil 25% y cuartil superior (o tercer cuartil) al percentil 75%.

Los cuartiles y la mediana dividen a la muestra ordenada en cuatro partes igualmente pobladas (aproximadamente un 25 % de los datos en cada una de ellas). Entre los cuartiles se hallan aproximadamente el 50% central de los datos y el rango de éstos es:

$$\boxed{\mathbf{d_I = distancia\ intercuartil = cuartil\ superior - cuartil\ inferior}}$$

Observación: Si en ejemplo cambiáramos el último dato por 110, la distancia intercuartil no cambiaría, mientras que el desvío pasaría de 3.2 a 24.13!!!!

Medidas resumen

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.

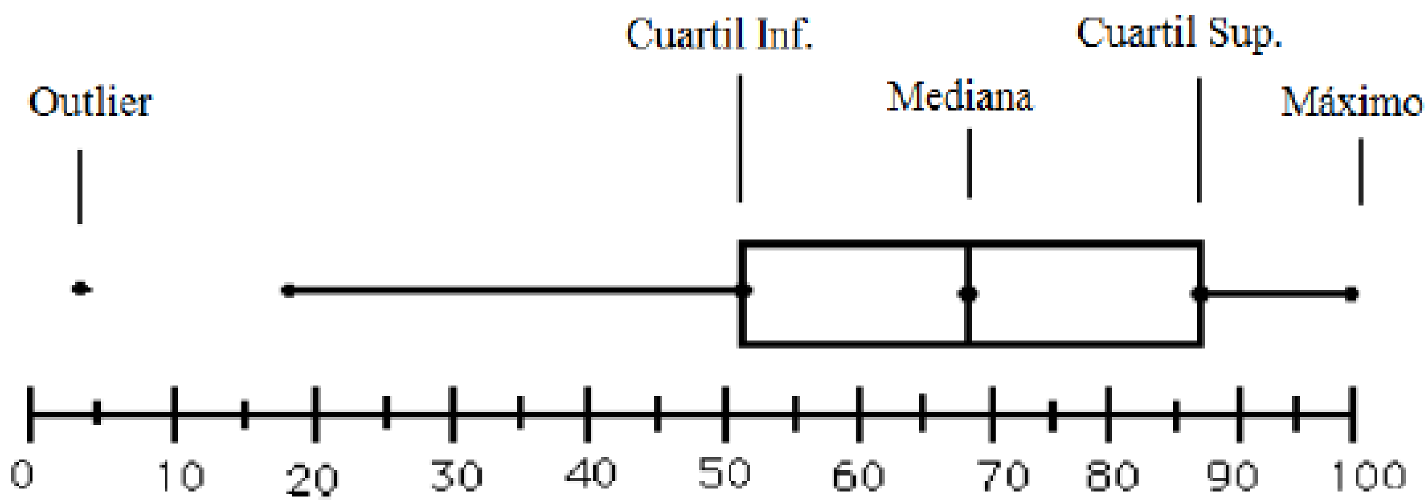
Observación: Si deseamos comparar la distancia intercuartil y la MAD con el desvío standard es conveniente dividirlas por constantes adecuadas. En ese caso se compara a S mediante

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

Métodos Gráficos

REPRESENTACIÓN DE DATOS NUMERICOS

Boxplot



Métodos Gráficos

REPRESENTACIÓN DE DATOS NUMERICOS

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 entre 1.5 d_I y 3 d_I de cada extremo y con o a aquellos que están a más

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

Ci=0.25-percentil calculo 0.25*(13+1)=3.5
Entonces Ci=146

CS=0.75-percentil calculo 0.75*(13+1)=10.5
Entonces Cs=302

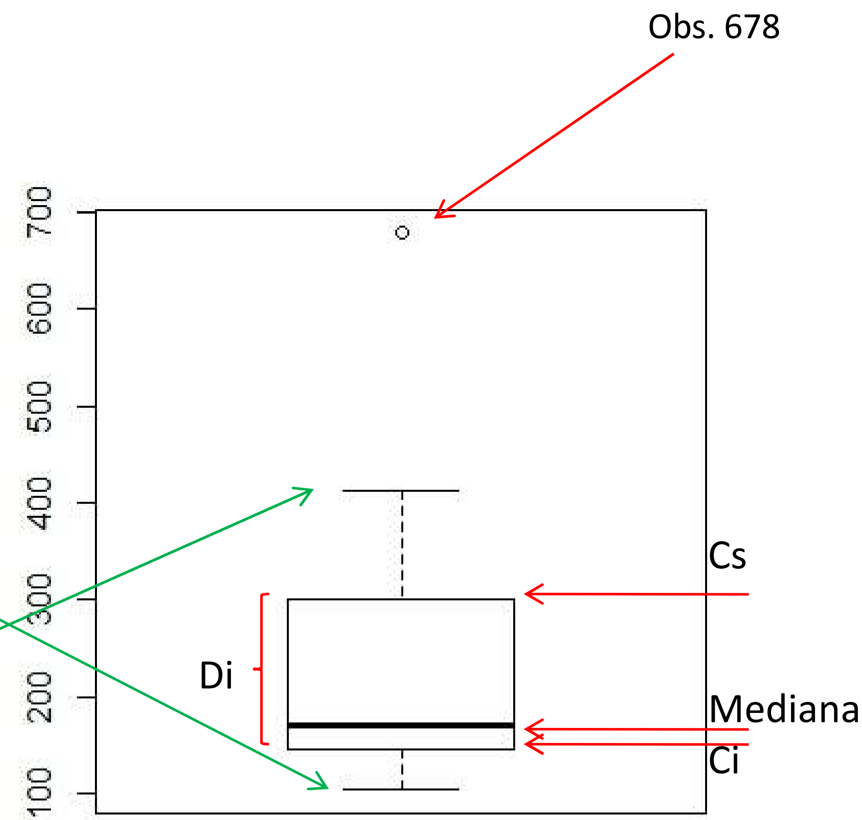
Di=302-146=156

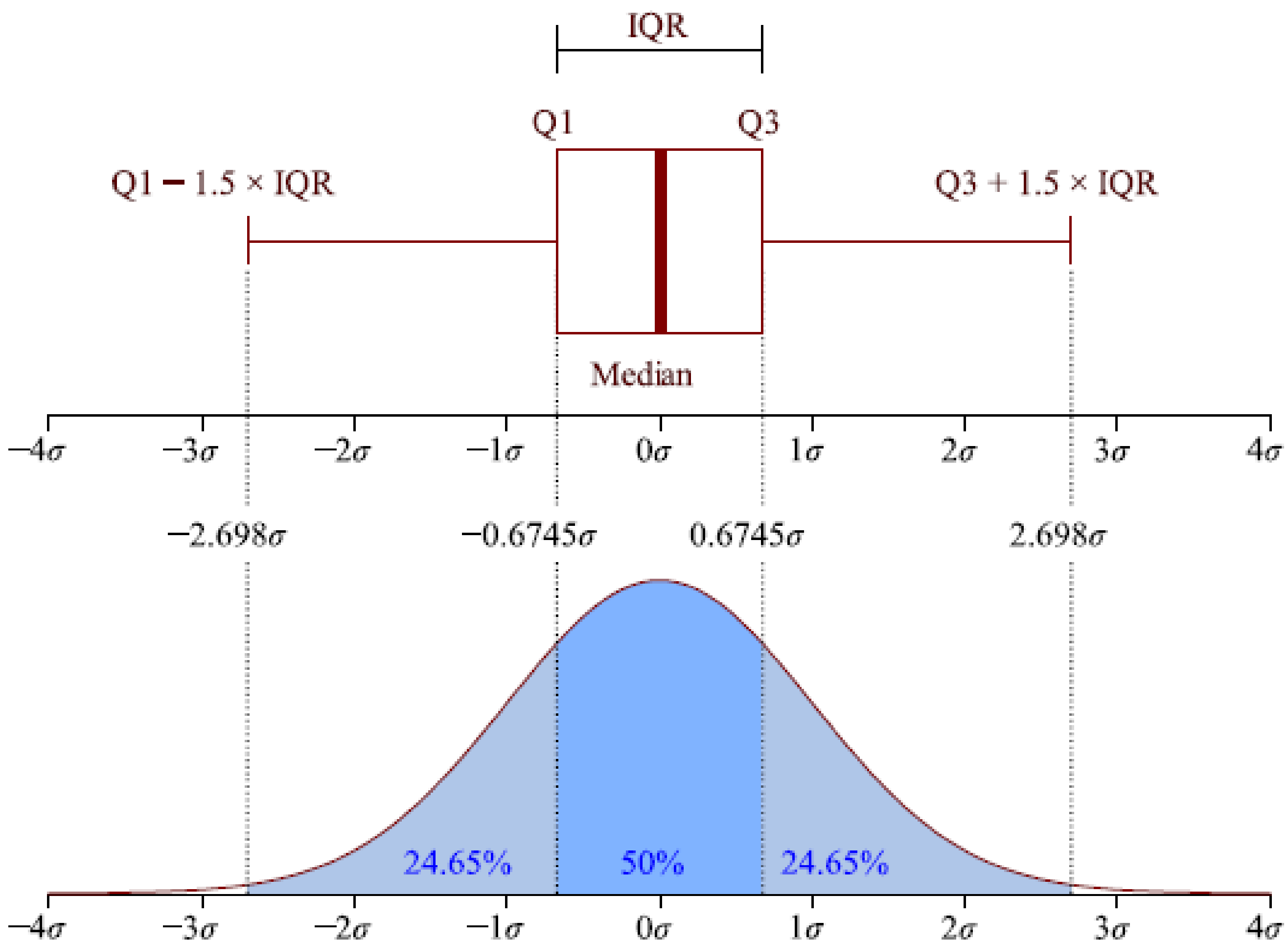
Calculamos
Li= primera cota inferior
= Ci-1.5*Di=146-1.5*156=-88
Llego hasta la obs. 104

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

Si= segunda cota inferior
= Ci-3*Di=146-3*156=-322 No tengo outliers

Ss= segunda cota superior
= Cs+3*Di=302+3*156=770 marco la obs. 678





Gracias Wikipedia!

Métodos Gráficos

REPRESENTACIÓN DE DATOS NUMERICOS

Boxplot

Que vemos en un box-plot?

- posición
- Dispersión
- Asimetría
- longitud de las colas
- puntos anómalos o outliers.

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

Métodos Gráficos

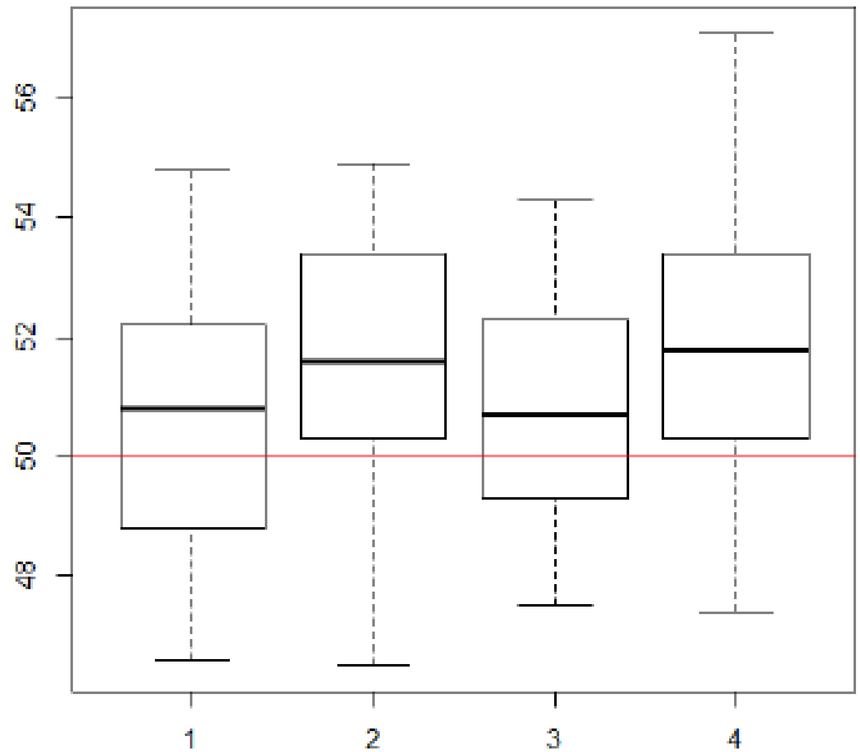
REPRESENTACIÓN DE DATOS NUMERICOS

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")
```

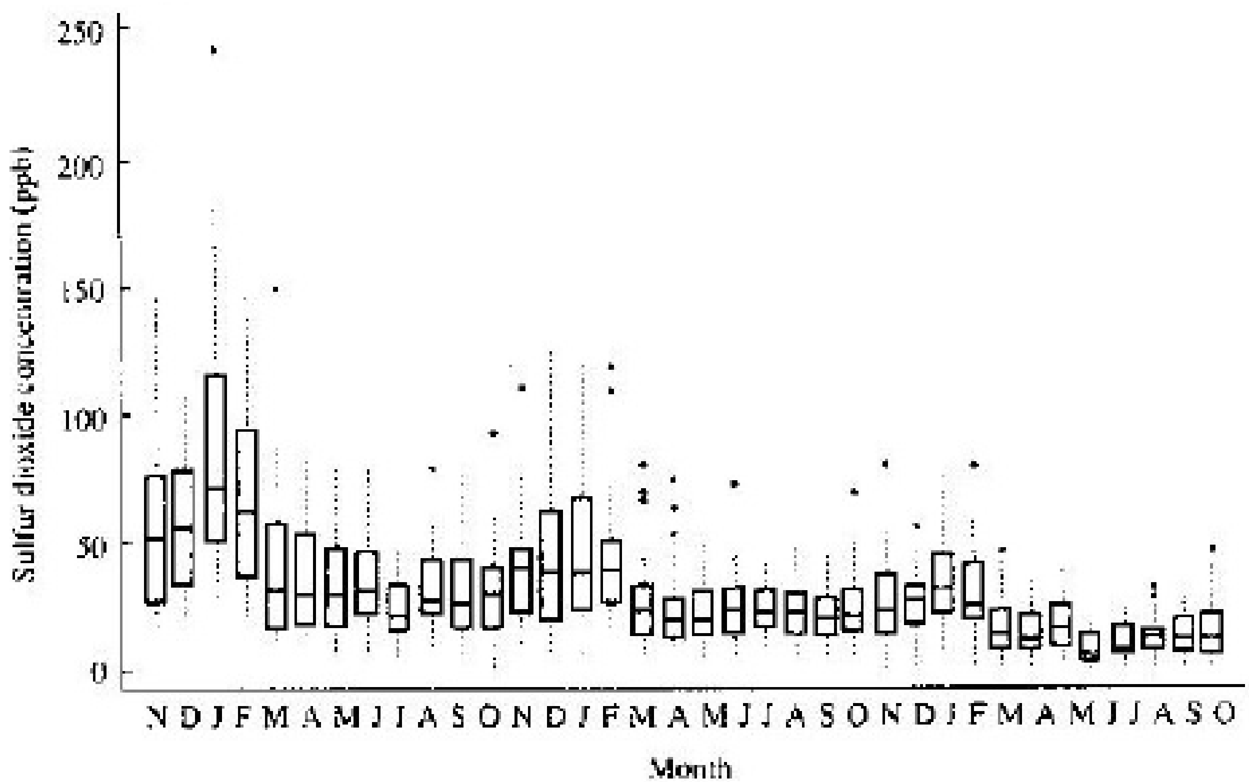


Métodos Gráficos

REPRESENTACIÓN DE DATOS NUMERICOS

Boxplot

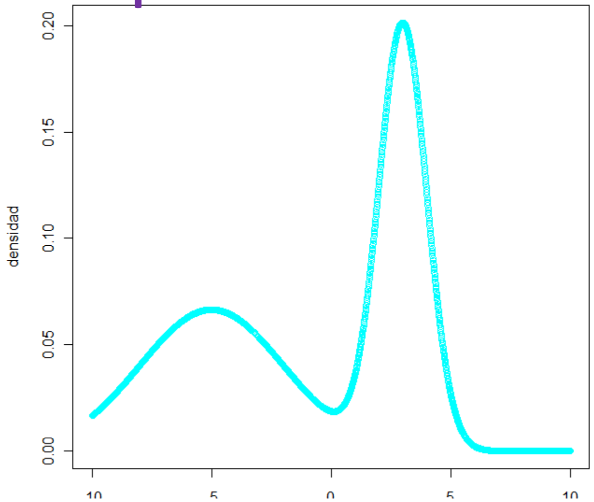
Ejemplo: Los siguientes boxplots corresponden a datos de concentración máxima diaria en partes por mil millones de dióxido de azufre en Bayonne, en el estado de Nueva Jersey, desde noviembre de 1969 hasta octubre de 1972 agrupados por meses. Hay 36 grupos de datos, cada uno de tamaño aproximadamente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

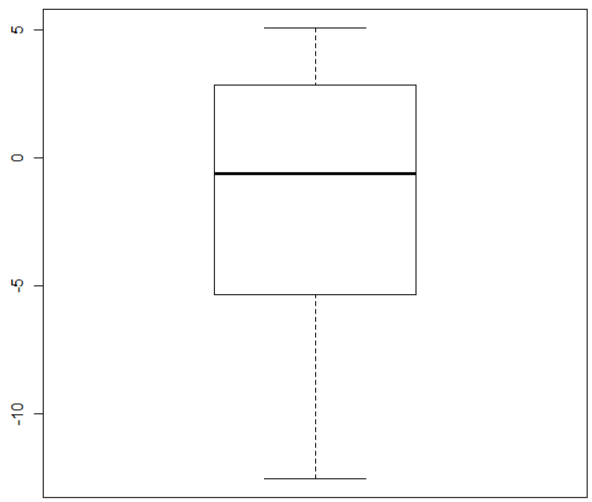


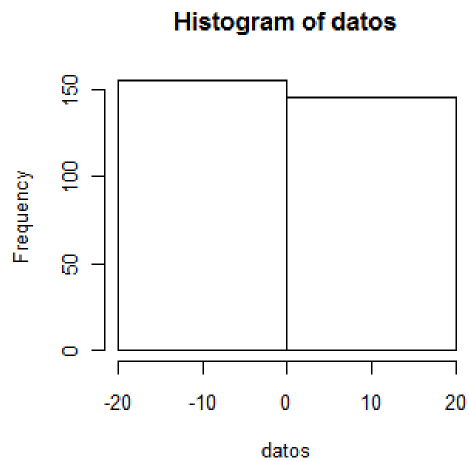
Métodos Gráficos

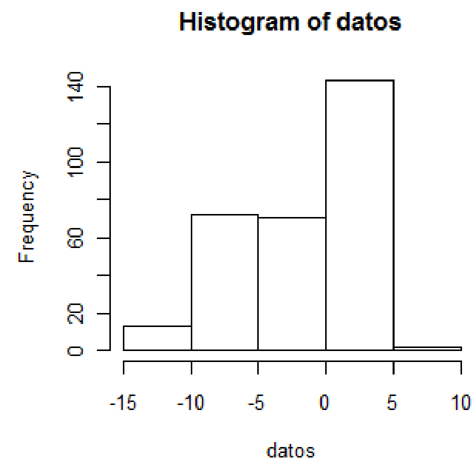
REPRESENTACIÓN DE DATOS NUMERICOS

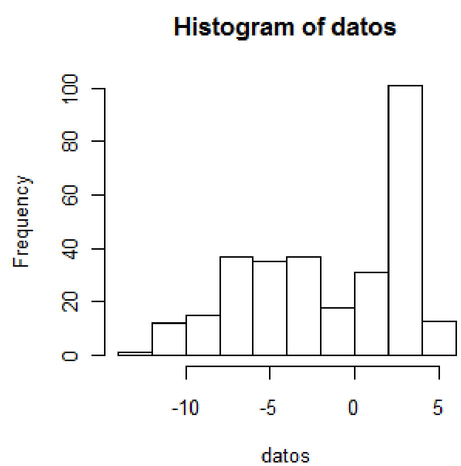
Boxplot

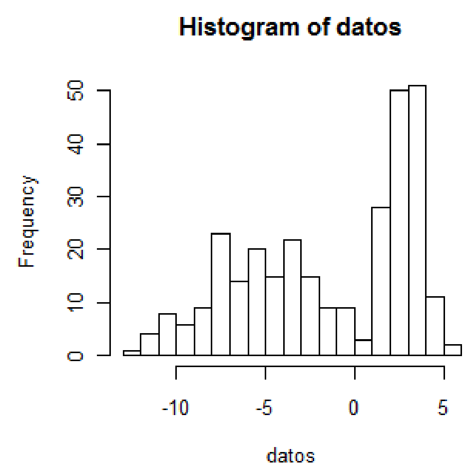












Métodos Gráficos

REPRESENTACIÓN DE DATOS NUMERICOS

QQ-Plot o Grafico cuantil-cuantil

Estadística (Química) 2012
FCEyN UBA

Dra. Ana M. Bianco

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

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

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.

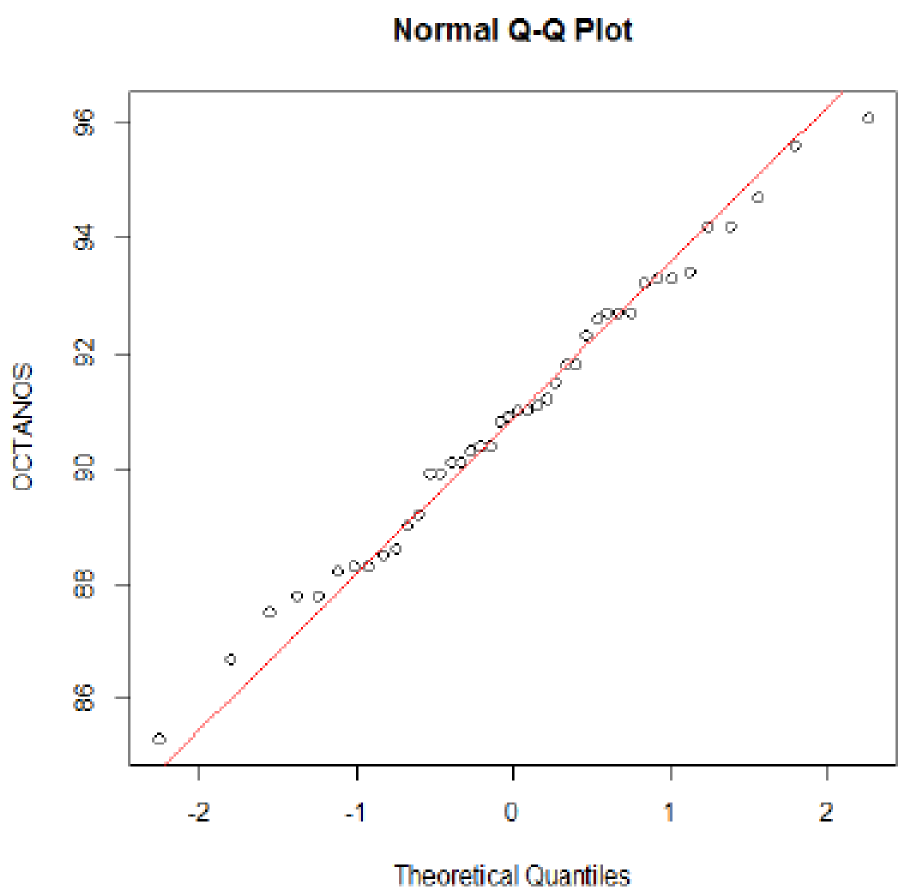
El alejamiento de la normalidad se ve reflejado por la forma del gráfico.

Métodos Gráficos

REPRESENTACIÓN DE DATOS NUMERICOS

QQ-Plot o Grafico cuantil-cuantil

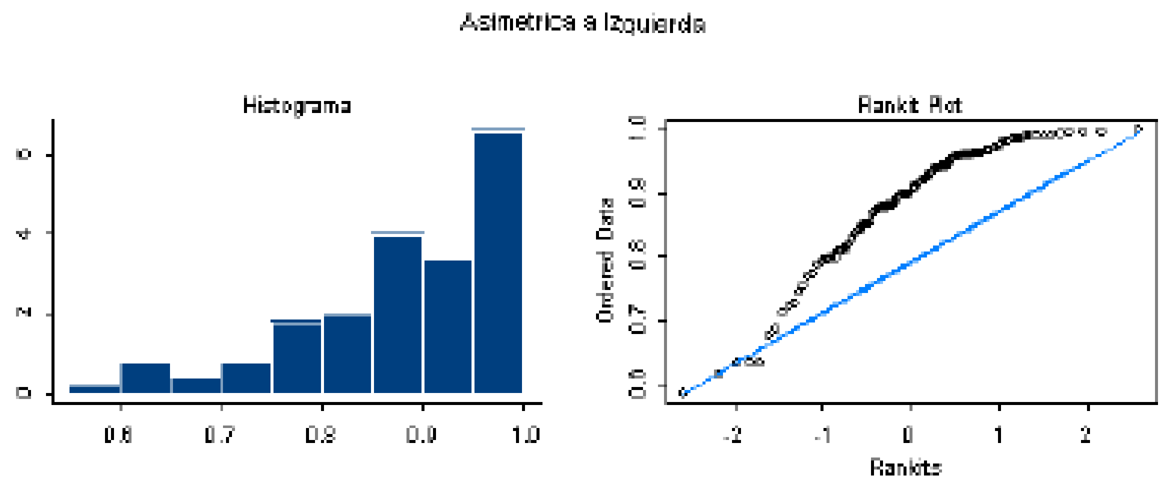
```
qqnorm(octanos.per,ylab="OCTANOS" )
qqline(octanos.per, col = 2)
```

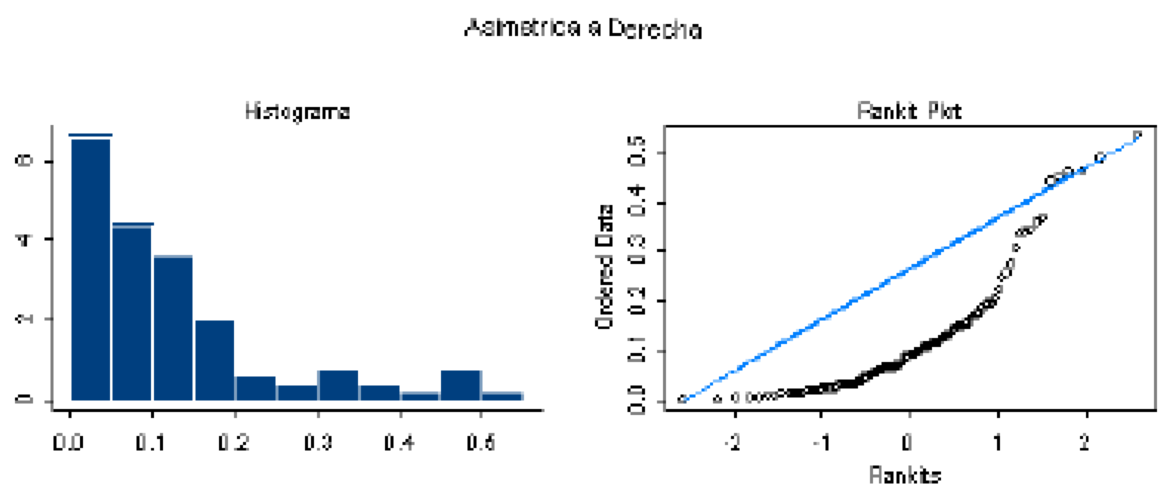


Métodos Gráficos

REPRESENTACIÓN DE DATOS NUMERICOS

QQ-Plot o Grafico cuantil-cuantil



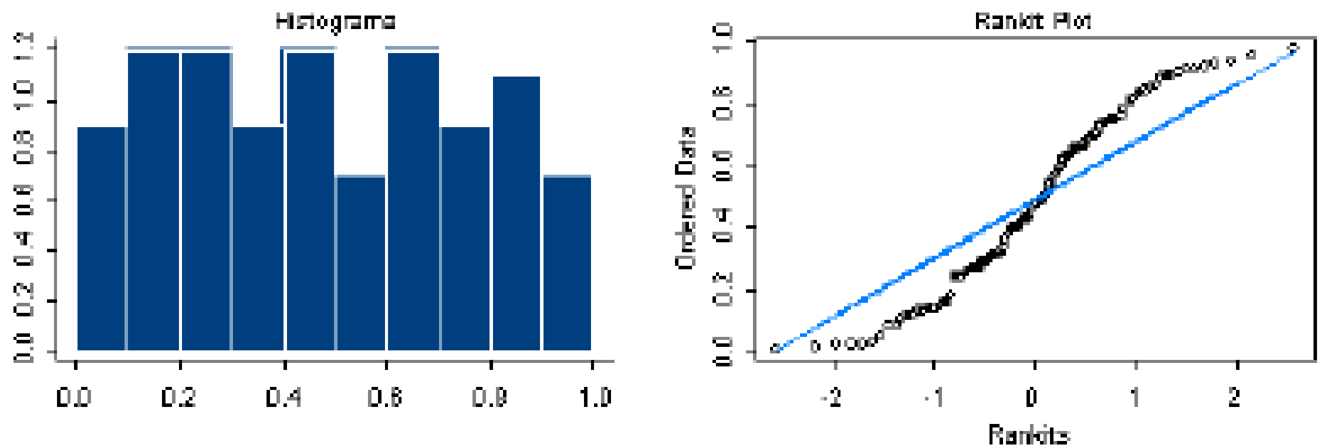


Métodos Gráficos

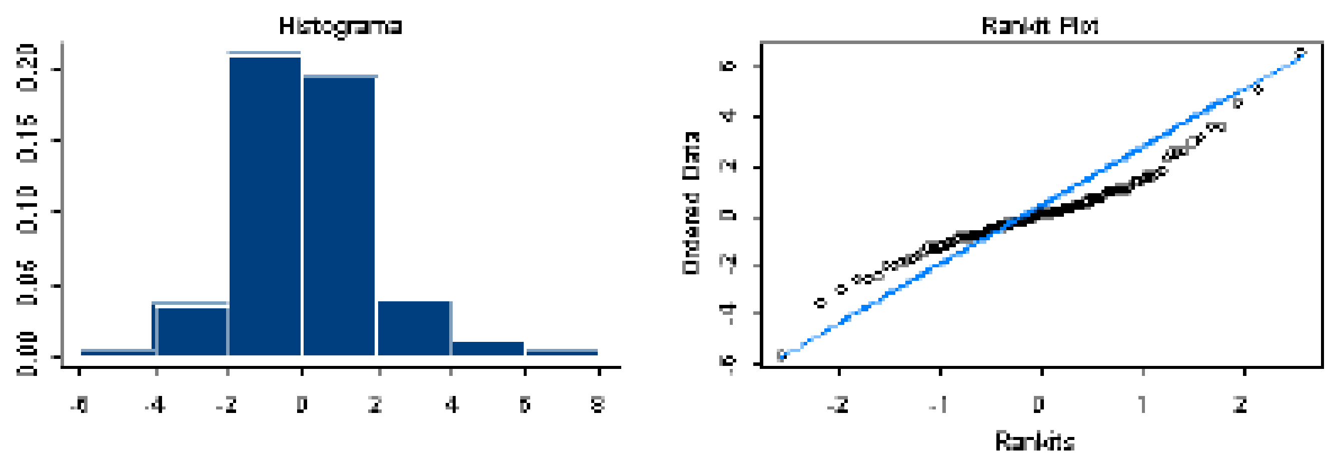
REPRESENTACIÓN DE DATOS NUMERICOS

QQ-Plot o Grafico cuantil-cuantil

Simetría con colas Pesadas



Simetría con colas Livianas



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 empecemos 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 esta 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.

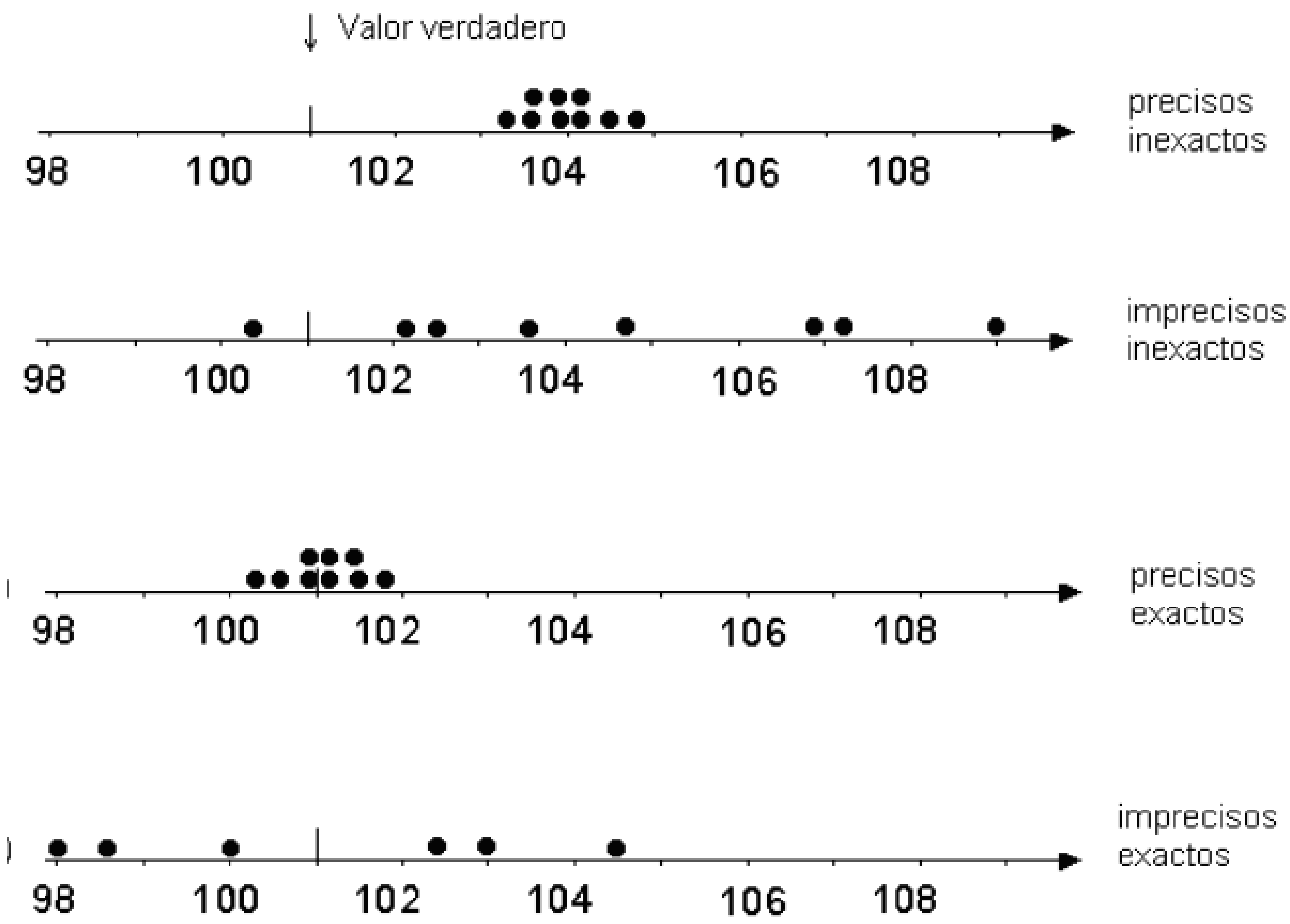
Ejemplo: un cronómetro es más preciso que un reloj común

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 metodo de medición)

ERRORES

Errores en el Proceso de Medición

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

Cantidad a medir: (desconocido pero no aleatorio)

$VAR(\mu+\varepsilon_1+\varepsilon_2+\varepsilon_3+\ldots+\varepsilon_n) = \sigma_1^2+\sigma_2^2+\sigma_3^2+\ldots.+\sigma_n^2 = \sigma^2$ si llamamos $\varepsilon = \varepsilon_1+\varepsilon_2+\varepsilon_3+\ldots+\varepsilon_n$

$X = \mu+\varepsilon$ donde $\varepsilon \sim N(0, \sigma^2)$ 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

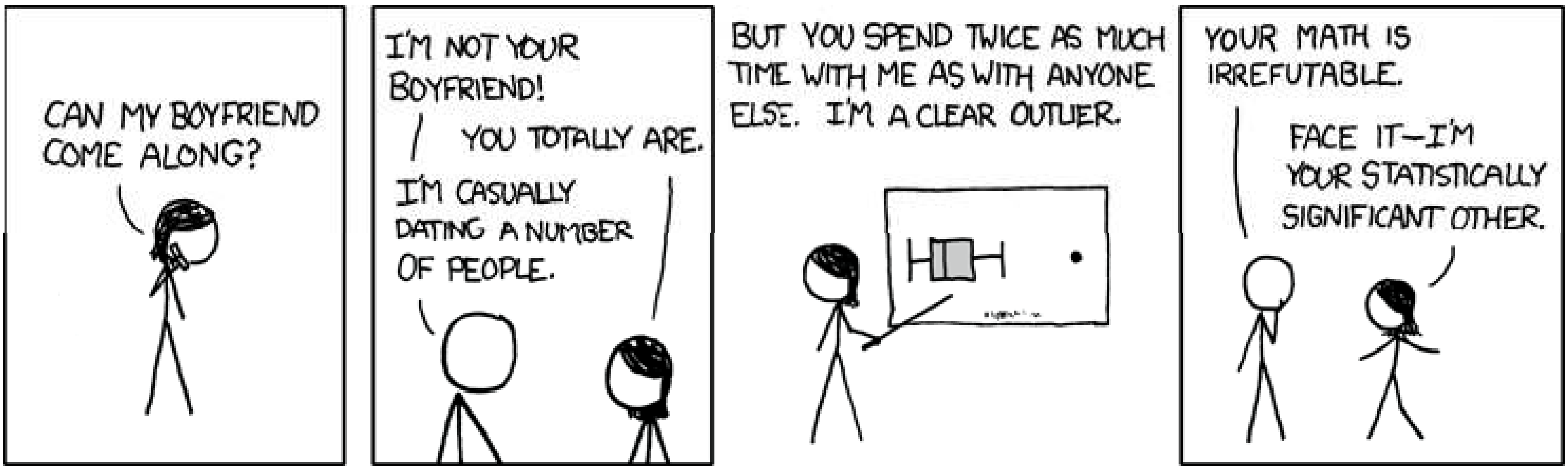
$$(X_1, \ldots, X_n) = (\mu_1, \ldots, \mu_n) + (\xi_1, \ldots, \xi_n).$$

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

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


CAN MY BOYFRIEND COME ALONG?
I'M NOT YOUR BOYFRIEND!
YOU TOTALLY ARE.
I'M CASUALLY DATING A NUMBER OF PEOPLE.
BUT YOU SPEND TWICE AS MUCH TIME WITH ME AS WITH ANYONE ELSE. I'M A CLEAR OUTLIER.
YOUR MATH IS IRREFUTABLE.
FACE IT—I'M YOUR STATISTICALLY SIGNIFICANT OTHER.