ESTADÍSTICA PARA QUÍMICA

Resumen del programa:

1) Estadística descriptiva o análisis exploratorio de datos.
2) Probabilidades.
3) Inferencia Estadística (estimación, intervalos de confianza, test de hipótesis).
4) Algunos modelos para aplicar inferencia estadística: comparación de dos muestras, modelos de regresión, modelos de analisis de la varianza, etc.

Estadística descriptiva o Análisis exploratorio de datos: técnicas para presentar y resumir un conjunto de datos para facilitar su comprensión y/o para generar hipótesis.

Inferencia estadística: técnicas que permiten, a partir de los datos de una muestra, obtener alguna información sobre la población (de la que se extrajo la muestra).

MUESTRA $\Rightarrow$ POBLACION

Teoría de Probabilidades: rama de las matemáticas que nació para explicar los juegos de azar. Luego empezó a tener aplicaciones "más útiles". Constituye la base matemática de la teoría de inferencia estadística.

Antes de comenzar con estadística descriptiva, vamos a ver una clasificación de los datos.

Tipo de datos

1. Categóricos

   a) Con dos categorías: sexo, diabetes (sí o no)
   b) Con más de dos categorías
      i) las categorías tienen un orden: clase social (registrada como alta, media o baja), nivel de gravedad de una enfermedad (leve, moderada, severa). Suelen llamarse simplemente "datos ordinales".
      ii) las categorías no tienen orden: nacionalidad, estado civil, etc. (datos cualitativos o nominales).

2. Numéricos o cuantitativos

   a) Discretos: número de hijos de una mujer, número de parásitos en 1 cm^3 de sangre.
   b) Continuos: estatura, edad, glucemia.

La razón por la que distinguimos entre datos cualitativos y cuantitativos es porque se usan métodos estadísticos diferentes para cada tipo de datos.

ESTADISTICA DESCRIPTIVA

Tablas Gráficos Medidas de resumen.

1. Tablas y Gráficos para describir los datos de una variable

Tabla de distribución de frecuencias para describir una variable numérica.

Ejemplo: En una planta industrial se elabora un producto químico. Se registró el rendimiento obtenido en 210 lotes producidos consecutivamente (el ejemplo está en Box, Hunter y Hunter, Estadística para Experimentadores).

Queremos visualizar los datos registrados. Los datos (en %) para los primeros lotes son:

85,5 81,7 80,6 84,7 88,2 etc.

(los datos están en el archivo EXCEL ProduccionIndustrial.xls)

¿De qué forma se pueden presentar para poder "ver algo"?

Una posibilidad es hacer una tabla de "distribución de frecuencias".

Antes de hacerla necesito conocer el mínimo y el máximo de los valores observados:

```
STATISTIX 7.1                                                16/08/03

DESCRIPTIVE STATISTICS

VARIABLE          N        MEAN          SD     MINIMUM     MAXIMUM
RENDIMIEN       210      84.121      2.8809      76.500      91.700
```

Para hacer una tabla de frecuencia hay que elegir los "intervalos". Si por ejemplo elijo los intervalos de 2 unidades de longitud, puedo obtener la siguiente salida con el programa Statistix:

```
FREQUENCY DISTRIBUTION OF RENDIMIEN

                                        CUMULATIVE
    LOW    HIGH     FREQ  PERCENT     FREQ  PERCENT
     76      78        3      1.4        3      1.4
     78      80       12      5.7       15      7.1
     80      82       33     15.7       48     22.9
     82      84       53     25.2      101     48.1
     84      86       54     25.7      155     73.8
     86      88       34     16.2      189     90.0
     88      90       15      7.1      204     97.1
     90      92        6      2.9      210    100.0
TOTAL              210    100.0
```

Nota: Este programa no incluye al valor superior de cada intervalo, por ejemplo cuando dice de 76 a 78, no incluye al valor 78 (en otros programas la regla puede ser diferente) o sea que otra forma más clara de presentar la tabla es:

Distribución de frecuencias de los rendimientos obtenidos en 210 lotes producidos consecutivamente.

Rendimiento (en porcentajes)	Nro de lotes (%)
76 a 77,9	3 (1,4%)
78 a 79,9	12 (5,7%)
80 a 81,9	33 (15,7%)
82 a 83,9	53 (25,2%)
84 a 85,9	54 (25,7%)
86 a 87,9	34 (16,2%)
88 a 89,9	15 (7,1%)
90 a 91,9	6 (2.9%)
Total	210 (100%)

Si quiero más información sobre la distribución de la variable "rendimiento" puedo usar más intervalos, o lo que es lo mismo tomarlos de menor longitud. Por ejemplo si elijo intervalos de 1 unidad obtengo:

FREQUENCY DISTRIBUTION OF RENDIMIEN

LOW	HIGH	FREQ	PERCENT	CUMULATIVE FREQ	CUMULATIVE PERCENT
76	77	1	0.5	1	0.5
77	78	2	1.0	3	1.4
78	79	2	1.0	5	2.4
79	80	10	4.8	15	7.1
80	81	16	7.6	31	14.8
81	82	17	8.1	48	22.9
82	83	23	11.0	71	33.8
83	84	30	14.3	101	48.1
84	85	29	13.8	130	61.9
85	86	25	11.9	155	73.8
86	87	23	11.0	178	84.8
87	88	11	5.2	189	90.0
88	89	9	4.3	198	94.3
89	90	6	2.9	204	97.1
90	91	5	2.4	209	99.5
91	92	1	0.5	210	100.0
TOTAL		210	100.0		

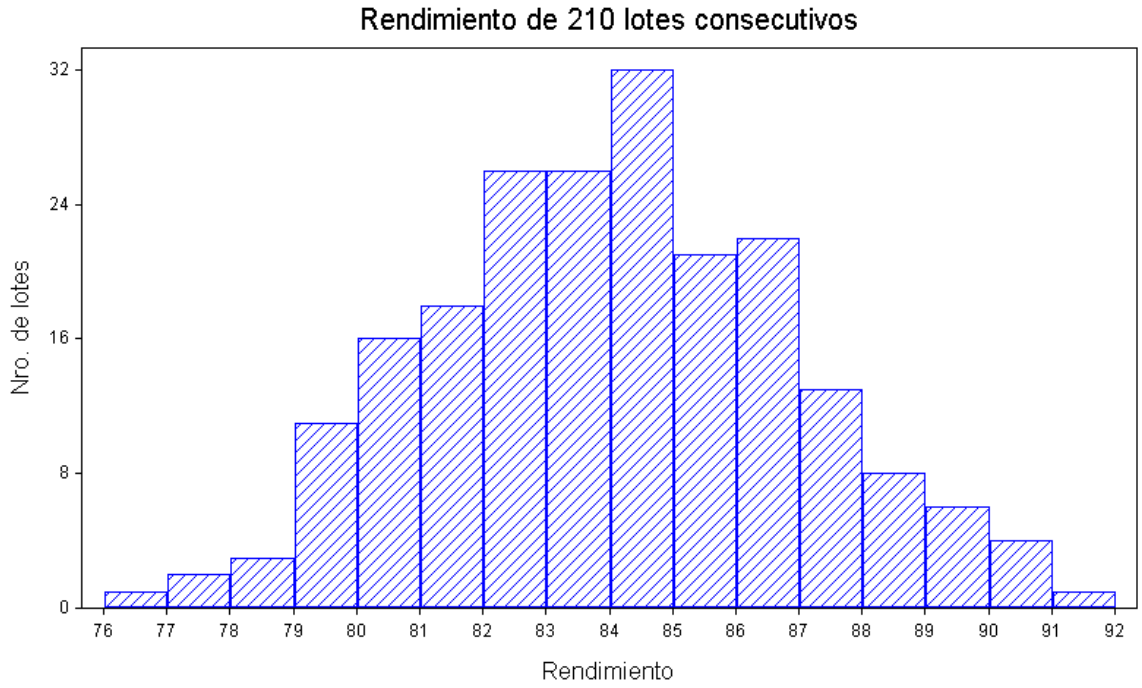
- Cuántos intervalos se toman? Una regla práctica es entre 5 y 20. Pero no es obligatoria: depende de la información que queremos. También depende del número de datos, por qué?
- Para facilitar la lectura los extremos de los intervalos conviene que sean números "fáciles" (no de 74,3 a 74,8 años por ejemplo).
- Los intervalos pueden ser de diferente longitud, aunque se prefiere, otra vez para facilitar la lectura, que sean de igual longitud. Salvo casos especiales en que interese distinguir más la distribución ciertos valores que para otros. Ejemplos?

Dos gráficos para mostrar la distribucion de una variable numérica:

a) Histogramas
b) Gráficos de frecuencias acumuladas

a) **Histograma:** es una representación gráfica de una tabla de distribución de frecuencias.

Ejemplo: el histograma que representa la distribución de los rendimientos obtenidos en 210 lotes consecutivos (tomando intervalos de 1 unidad) es el siguiente:



- El SX pone en el eje vertical del histograma el número de casos. En vez de esto se puede indicar en este eje el porcentaje de casos. Cambia la forma del histograma?
- En el histograma (por definición de histograma) el **área** de cada rectángulo representa el número (o el %) de datos en cada intervalo. Por ello si los intervalos son de diferente longitud, no es la altura sino el área la que tiene que representar el nro (o %) de casos. Si se hace que la altura represente el número de casos, el gráfico daría una impresión engañosa.

¿Qué se ve en el histograma?

Mínimo – Máximo

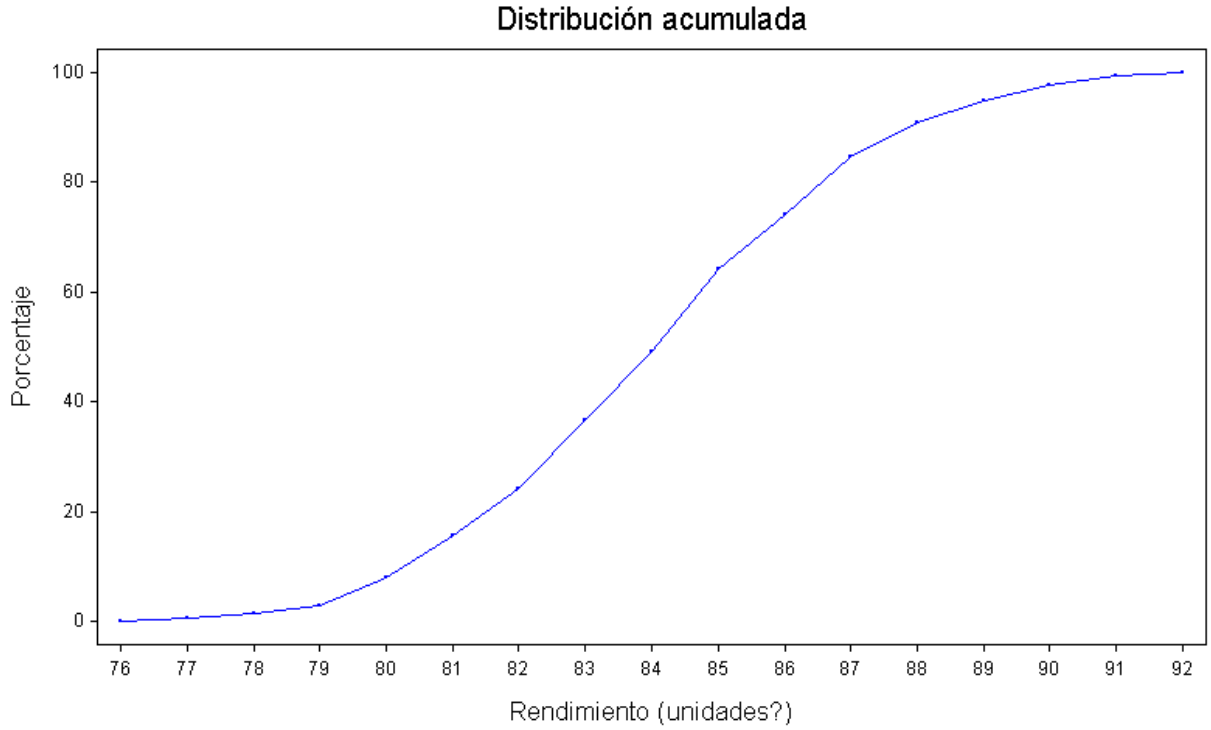
Intervalo o intervalos más frecuentes.

Simetría o asimetría.

Tiene el rendimiento una distribución (aproximadamente) simétrica?

Pensar ejemplos de variables que tienen distribución simétrica y asimétrica.

b) Gráfico de frecuencias acumuladas



Esta función representa las frecuencias relativas acumuladas. Se usa mucho menos que el histograma. Este gráfico también está relacionado con la tabla presentada antes, verdad? Observar que la ordenada de este grafico sobre la abscisa 80 es 7%. Que significa este valor? Significa que el 7% de los lotes tienen rendimientos menores de 80 unidades.

Diagrama "intermedio" entre los datos ordenados y el histograma: diagrama de tallo y hoja (Tukey, 1977).

Para el ejemplo del rendimiento de 210 lotes, el diagrama de tallo y hoja puede ser:

```
STEM AND LEAF PLOT OF RENDIMIEN

   LEAF DIGIT UNIT = 0.1                MINIMUM  76.500
   76  5  REPRESENTS 76.5               MEDIAN   84.150
                                        MAXIMUM  91.700
         STEM  LEAVES
       1   76  5
       3   77  59
       5   78  14
      15   79  0355566778
      31   80  0012233455667799
      48   81  00144566678888999
      71   82  000122223333445667788 99
     101   83  000000111233333455555556667789
     (29)  84  00112223344445566777778888899
      80   85  000001122234444566678899
      55   86  012222345555666667777788
      32   87  12222233478
      21   88  002246899
      12   89  034779
       6   90  04558
       1   91  7
210 CASES INCLUDED   0 MISSING CASES
```

"Girando el diagrama" esta representación tiene la misma forma que el histograma, pero además están anotados todos los datos originales. Se observa que los datos, ordenados de menor a mayor son:

76,5 77,5 77,9 78,1 78,4 79,0 79,3 79,5 79,5 79,5 79,6 etc.

Como se construye este diagrama? Se seleccionan los primeros (uno, dos o más) dígitos y esos dígitos se llaman "tallos". Los dígitos de la derecha se llaman "hojas". En cada fila del diagrama se pone un tallo y a su lado todas las "hojas". Para decidir si tomar como tallos uno, dos o más dígitos, se puede tener en cuenta que, al igual que para la tabla de distribución de frecuencias y para el histograma, en general se recomienda tomar entre 5 y 20 "tallos". Este diagrama es un competidor del histograma. Para muestras grandes es un poco confuso, el histograma es más simple.

Los números que están a izquierda del gráfico se llaman "profundidad" y van contando el número de datos desde el principio y desde el final y se anota el menor de esos dos números. En la linea del "centro" se anota la cantidad de datos de esa linea entre parentesis. Esto es útil para calcular a mano la mediana y los cuartiles (que definiremos luego).

Tablas y Gráficos para describir los datos de una variable categórica

La tabla de distribución de frecuencias puede usarse no solo para variables numéricas, sino para cualquier tipo de variables. Si la variable es categórica es aún más fácil que para variables numéricas: no hay que preocuparse por elegir primero los intervalos. Para variables categóricas, el gráfico similar al histograma se lo suele llamar también gráfico de barras. Por el contrario el gráfico de la distribución acumulada tendría sentido para variables ordinales, pero no para cualitativas. El diagrama de tallo y hoja no tiene significado para variables no numéricas, que quiere decir "los primeros dígitos"?

2. Medidas de resumen

a) **Medidas de posición o de tendencia central.**
b) **Cuartiles y percentiles.**
c) **Medidas de dispersión.**

a) Medidas de posición o de tendencia central: Media, mediana, medias podadas.

Media (sólo para variables numéricas):

Si llamamos $x_1, x_2, \ldots, x_n$ a los datos, la media es:

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

Mediana (para datos numéricos u ordinales)**:**

Es un número tal que está “en el medio de los datos” en el sentido de que hay tantos datos “menores que ese número” como datos “mayores que ese número”.

Ejemplo 1: valores de proteína C reactiva (PCR) en mg/l para 5 pacientes cardíacos

31,4 17,2 28,9 47,1 5,2

Cálculo de la mediana:

1er paso: ordeno los datos de menor a mayor:

5,2 17,2 28,9 31,4 47,1

2do paso: tomo "el del centro" en este caso el 3º.

Por lo tanto: mediana = 28,9 mg/l

Ejemplo 2: valores de PCR (mg/l) para 6 pacientes cardíacos

31,4 17,2 28,9 47,1 5,2 23,0

Cálculo de la mediana:

1er paso: ordeno los datos de menor a mayor:

5,2 17,2 23,0 28,9 31,4 47,1

2do paso: aquí no hay uno "en el centro": el centro está entre el 3º y el 4º. En este caso se toma como mediana el promedio de los dos valores centrales:

$$\text{Mediana} = \frac{23,0 + 28,9}{2} = 25,95 \text{ mg/l}.$$

Comparación de Media y Mediana

Se parecen para distribuciones **simétricas**?

Se puede observar que el histograma de rendimiento de 210 lotes es simétrico.

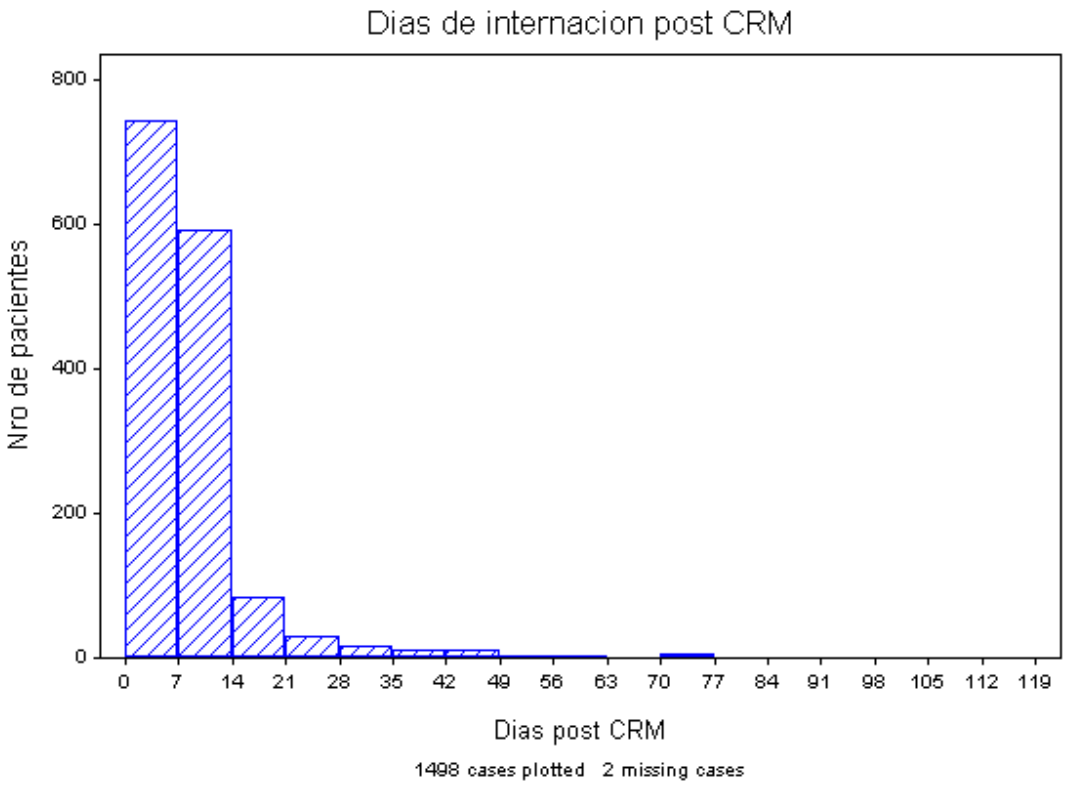
DESCRIPTIVE STATISTICS

VARIABLE	N	MEAN	MINIMUM	MEDIAN	MAXIMUM
RENDIMIEN	210	84.121	76.500	84.150	91.700

La media y la mediana son muy parecidas para estos datos la media es 84,12 y la mediana es 84,15.

Se parecen para distribuciones **asimétricas**?

Un ejemplo de variable que en casi todas las poblaciones de pacientes internados con cierta patología tiene distribución asimétrica es la duración de la internación. El siguiente histograma muestra la distribución de la duración de la internación post cirugía para 1500 pacientes sometidos a una operación cardíaca.



Para distribuciones con asimetría positiva la media es mayor que la mediana. La mediana de la duración de la internación para estos pacientes es 7 días, la media es 8,7 días.

¿Cuál de estas dos medidas (media o mediana) es más **resistente a valores atípicos**?

Veámoslo en un ejemplo: En el ejemplo 2 la mediana es 25,95, la media se calcula:

$$\bar{x} = \frac{31,4 + 17,2 + 28,9 + 47,1 + 5,2 + 23,0}{6} = 25,467 \cong 25,5$$

Ejemplo 3: A los 6 pacientes del ej. 2, se le agrega un nuevo pacientes que tiene un valor de PCR de 150,0

Cálculo de la mediana:

1er paso: ordeno los datos de < a >:

5,2 17,2 23,0 28,9 31,4 47,1 150,0

2do paso: cuál es el del medio en este ejemplo en que hay un número impar de datos, la mediana es el valor que ocupa el lugar (n+1)/2 = (7+1)/2 = 4. Luego

Mediana = 28,9 g/l

La media en este ejemplo es 43,257≅ 43,3

Se puede apreciar que un único dato muy alejado de los demás tuvo poca influencia sobre la mediana pero mucha sobre la media. La media es muy sensible a valores atípicos, la mediana no.

Una medida de posición intermedia entre mediana y media: medias podadas.

Vimos que la media es muy sensible a valores atípicos y la mediana no. Esta es una ventaja de la mediana. Sin embargo la mediana tiene intuitivamente el problema de que es el "valor central" sin tener en cuenta en su calculo los otros valores. A veces se desea algo intermedio. La media con poda α% se basa en la siguiente idea: se ordenan los datos de menor a mayor, se "desprecian" el α% de los valores más pequeños y el α% de los valores más grandes y se promedian los valores centrales "que quedan". Por ejemplo la media con poda 10% es el promedio del 80% de valores centrales. Esta es la idea, después solo queda aclarar que se hace cuando el α% del número de datos no es un número entero. Por ejemplo para calcular la media con poda 10% para las PCR's de 29 pacientes, que se hace? Nota: Statistix no calcula la media podada. Splus 6.1 trunca el numero de casos que desprecia.

b) Cuartiles y percentiles.

¿Qué es la mediana? El número que deja aproximadamente la mitad de los datos "a izquierda" (si los representamos en la recta) y la mitad "a derecha".

¿Qué es el primer cuartil? La idea es: el número que deja aproximadamente la cuarta parte de los datos "a izquierda" (o sea la cuarta parte de los datos son menores que ese número) y las ¾ partes "a derecha". Se lo denota Q_1 . Como se lo calcula?

¿Qué es el tercer cuartil o Q_3 ? ¿Qué es el segundo cuartil?

¿Qué es el percentil 10? El número que deja aproximadamente el 10% de los datos "a izquierda". O lo que es lo mismo el 10% de los datos son menores que el Percentil 10. Se lo denota P_{10}.

¿Qué es P_{90}? ¿El P_{25} tiene otro nombre? ¿Y el P_{50}?

c) Medidas de dispersión

i) Rango ii) Desviación Standard iii) Rango intercuartil.

i) Rango: es la diferencia entre el máximo y el mínimo. En el ej. 2 (TnT para 6 pacientes) es 0,07 - 0,01 = 0,06 (es más común presentar el mínimo y el máximo, sin hacer la resta, **rango: 0,01 a 0,07**).

ii) Varianza y Desviación Standard. Se llama **varianza** a

$$\text{Varianza} = s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

y **Desviación Standard** (se abrevia DS) a su raíz cuadrada.

$$\text{D.S.} = s = \sqrt{s^2}$$

En el ejemplo 2 la DS es 14,14, en el ejemplo 3 (agregando un valor muy grande) sube a 48,8.

iii) Rango intercuartil.

Es la diferencia entre el cuartil 3 y el cuartil 1, o sea Q3 - Q1.
Igual que con el rango, en general se muestran los cuartiles y no la resta. Cuando se calculan los cuartiles, también se suele calcular la mediana y presentar "los tres cuartiles": la mediana y los cuartiles inferior y superior.

Comparación de las medidas de dispersión: definimos tres medidas de dispersión.

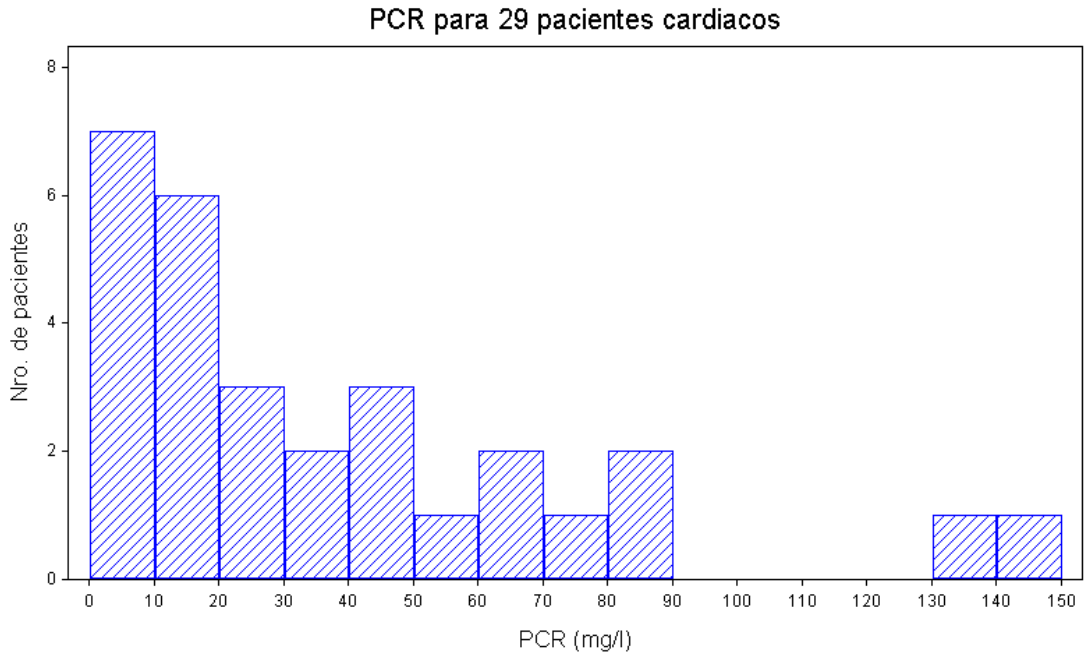
¿Cuál es la más resistente a valores atípicos? ¿Cuál la menos resistente?

¿Cuales medidas de posición y de dispersión son las más usadas?

Para variables cuya distribución se parece a la curva normal de Gauss conviene usar media y DS como medidas de resumen. Pero tampoco es incorrecto usar mediana y cuartiles. Para variables con distribución asimétrica o con valores atípicos, es más frecuente usar mediana y cuartiles. Esto es particularmente cierto en el caso de la presencia de valores atípicos que hacen muy inestables la media y la DS.

El archivo PCRCPKTn contiene valores de PCR para 29 pacientes cardíacos.

Observemos el histograma:



Supongamos ahora que queremos comparar los valores de PCR para pacientes varones y mujeres. Calculamos las medidas de resumen para cada sexo:

```
DESCRIPTIVE STATISTICS FOR SEXO = F

                     PCR
N                     11
MEAN              50.129
SD                45.674
MINIMUM           1.5000
1ST QUARTI        3.9000
MEDIAN            47.120
3RD QUARTI        87.860
MAXIMUM           149.57


DESCRIPTIVE STATISTICS FOR SEXO = M

                     PCR
N                     18
MEAN              31.659
SD                32.892
MINIMUM           5.2000
1ST QUARTI        9.8325
MEDIAN            17.465
3RD QUARTI        45.375
MAXIMUM           134.00
```

Teniendo en cuenta que la variable tiene, para estos pacientes, distribución asimétrica y con datos atípicos, elegimos como medidas de resumen la mediana y cuartiles. Informamos que
De los 29 pacientes estudiados, 18 eran hombres. La mediana (P25; P75) de PCR fue 17,5 (9,8; 45,4) para los hombres y 47,1 (3,9; 87,9) para las mujeres.

Otra medida de dispersión: mediana de las desviaciones absolutas.

Otra medida de dispersión que se ha propuesto, que es una medida muy poco sensible a datos atípicos es la mediana de las desviaciones absolutas (abreviada MAD)

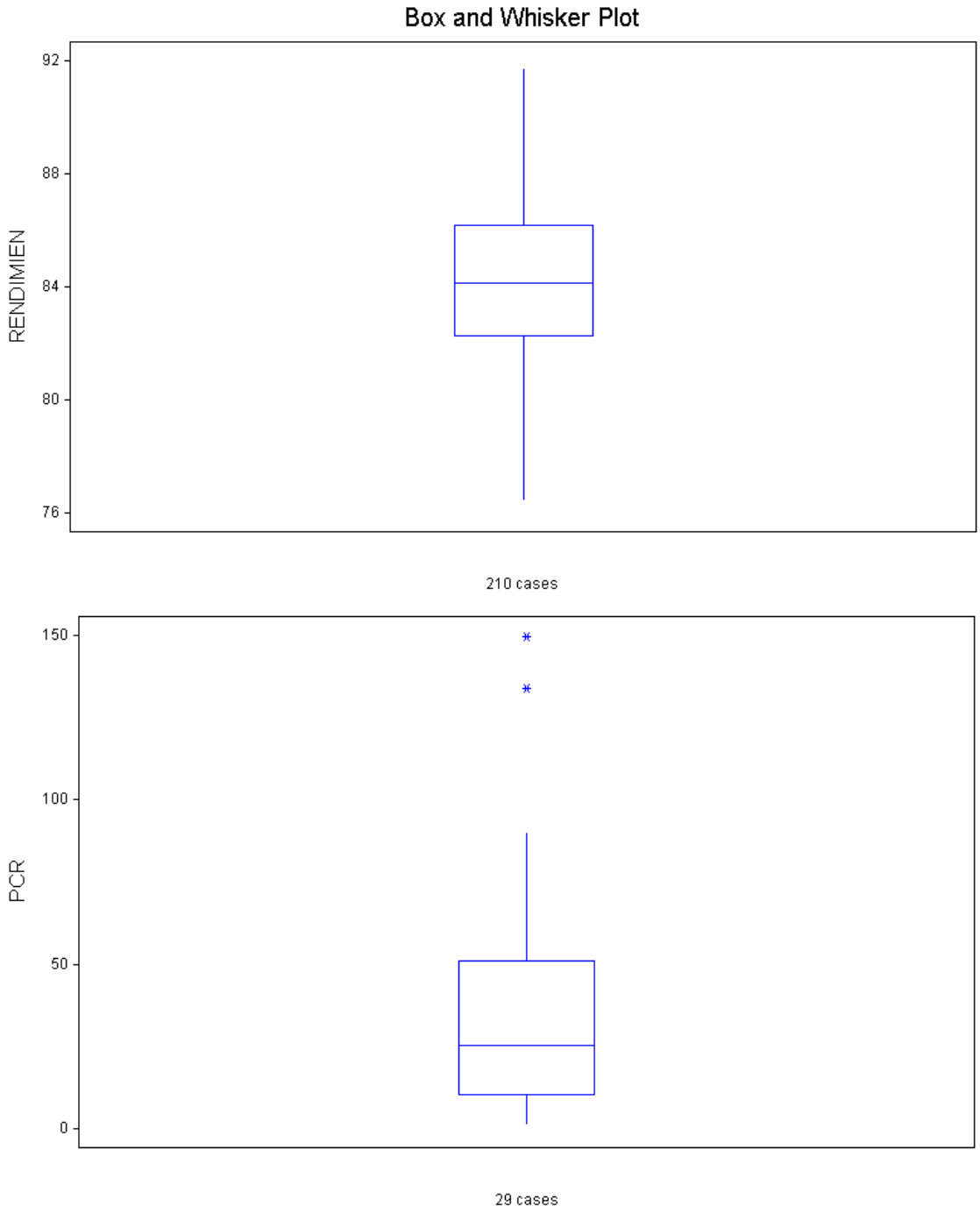
$$MAD = \text{mediana } |x_i - \tilde{x}|$$

donde indicamos con $\tilde{x}$ a la mediana de los valores $x_1, x_1, ..., x_n$. Esta medida es menos usada que la DS y el rango intercuartil.

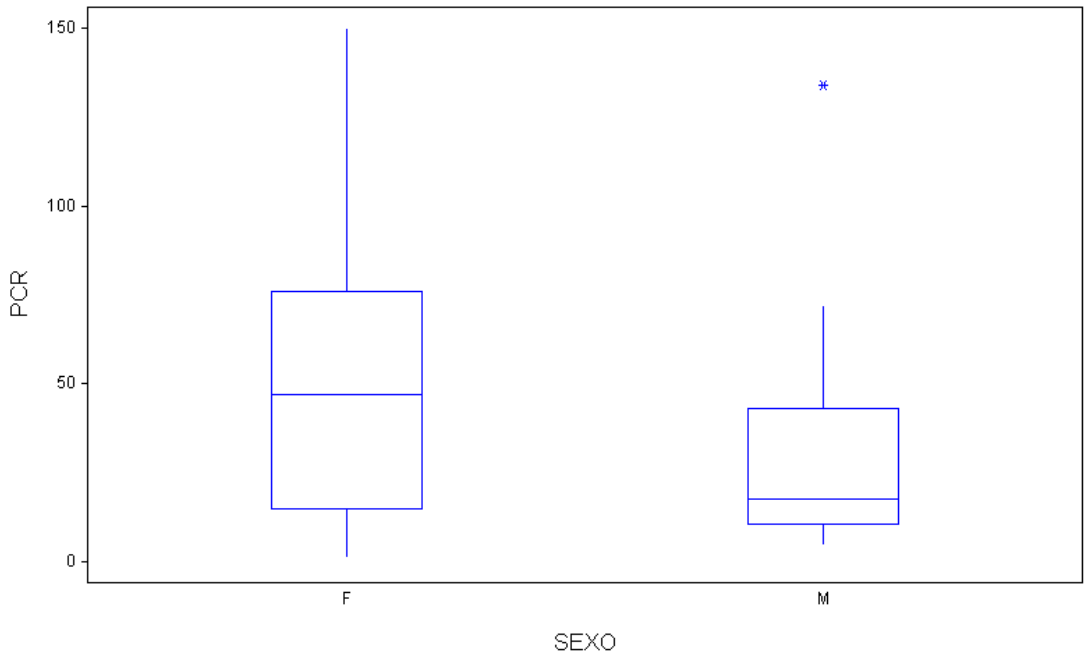
Otro gráfico para representar la distribución de una variable: Gráfico de caja (Box plot)

Es otro gráfico propuesto por un muy prestigioso estadístico: John Tukey (Exploratory Data Análisis, 1977). Es una representación gráfica de cinco números: mínimo, primer cuartil, mediana, tercer cuartil y máximo. Además se representan como puntos separados los datos atípicos (outliers).

Los diagramas de caja para los datos de rendimiento de 210 lotes y para los datos de las PCR para 29 pacientes son los siguientes.



Se ve que el primero no muestra valores atípicos y el segundo sí. Estos box plots se usan poco. El histograma da más información que este gráfico sobre la distribución de una variable numérica. Pero el Box plot es muy útil cuando se desea comparar la distribución de una variable numérica en dos o más poblaciones. No es fácil comparar visualmente dos histogramas, más difícil aún es comparar 3 o más. Por ejemplo si deseamos comparar los valores de PCR para pacientes varones y mujeres podemos graficar:



Más adelante en el curso veremos "tests estadísticos" que permiten decir si hay suficiente información como para afirmar (con baja probabilidad de error) que hay diferencias entre los valores de PCR para ambos sexos en las poblaciones de pacientes con las mismas características que los observados.

Datos transformados:

Hemos insistido en que si los datos tienen valores atípicos no conviene usar la media y la DS como medidas de resumen, sino otras medidas. Otra solución que a veces da resultados es aplicarle a los datos alguna transformación (la más usada es el logaritmo). Cuando una variable tiene distribución con asimetría positiva, el logaritmo de esa variable tiene una distribución más simétrica y los valores atípicos "del lado derecho" se acercan a los otros datos. A veces la transformación no dá el resultado deseado (los datos transformados siendo asimétricos o con datos atipicos) , pero otras veces se consigue que la variable transformada tenga distribución simétrica y sin datos atípicos. En estos casos, simplemente se utiliza en los análisis estadisticos (calculo de medidas de resumen, tests de hipotesis, regresiones, etc) la variable transformada. Se ven ejemplos de aplicación de transformación logarítmica en la práctica (por ejemplo concentraciones de IgM en sangre de 298 niños).

Utilización de las medidas de posición y de dispersión para evaluar una metodología de medición. Precisión y exactitud.

Para evaluar la calidad de un método de medición, se necesita realizar varias mediciones. Si realizamos varias mediciones del mismo material con la misma metodología, esperamos que las mediciones sean parecidas, pero generalmente no van a ser exactamente iguales. Cuando la dispersión de los distintos valores obtenidos es pequeña diremos que el método tiene una precisión alta. La dispersión se mide con las medidas de dispersión que hemos estudiado, la usada con mayor frecuencia es la desviación standard.

Sin embargo puede ocurrir que un método de medición tenga alta precision (poca dispersión) y que sin embargo las medidas no estén cerca del verdadero valor. Supongamos por ejemplo que

se sabe que una aleación "standard" tiene 4.44% de Niquel. Se envían muestras a cuatro laboratorios y en cada laboratorio se hacen ocho determinaciones obteniéndose:

Laboratorio A 4.61 4.61 4.72 4.60 4.60 4.64 4.54 4.49

Laboratorio B 4.45 4.30 4.18 4.49 4.36 4.66 4.51 4.50

En el primer laboratorio los datos observados son todos más altos que el valor verdadero. Por el contrario los del laboratorio B oscilan alrededor del valor verdadero. El problema del laboratorio A no es la precisión: los datos tienen menor dispersión que los de B. En efecto si calculamos la media y la desviación standard para cada laboratorio obtenemos:

Laboratorio A media=4.601 DS=0.067
Laboratorio B media=4.431 DS=0.148

Observamos que los valores del laboratorio A tienen menor desviación standard que los de B, pero el promedio está lejos del verdadero valor. Se dice que tienen buena precisión pero son **inexactos.** Las determinaciones de B son mas dispersas que las de A, pero la media esta cerca del verdadero valor (menor precisión pero mayor exactitud).

Agreguemos a nuestro hipotético experimento otros dos laboratorios:

Laboratorio C 4.42 3.83 4.34 4.28 4.03 4.28 4.29 4.54

Laboratorio D 4.43 4.49 4.42 4.40 4.41 4.47 4.44 4.42

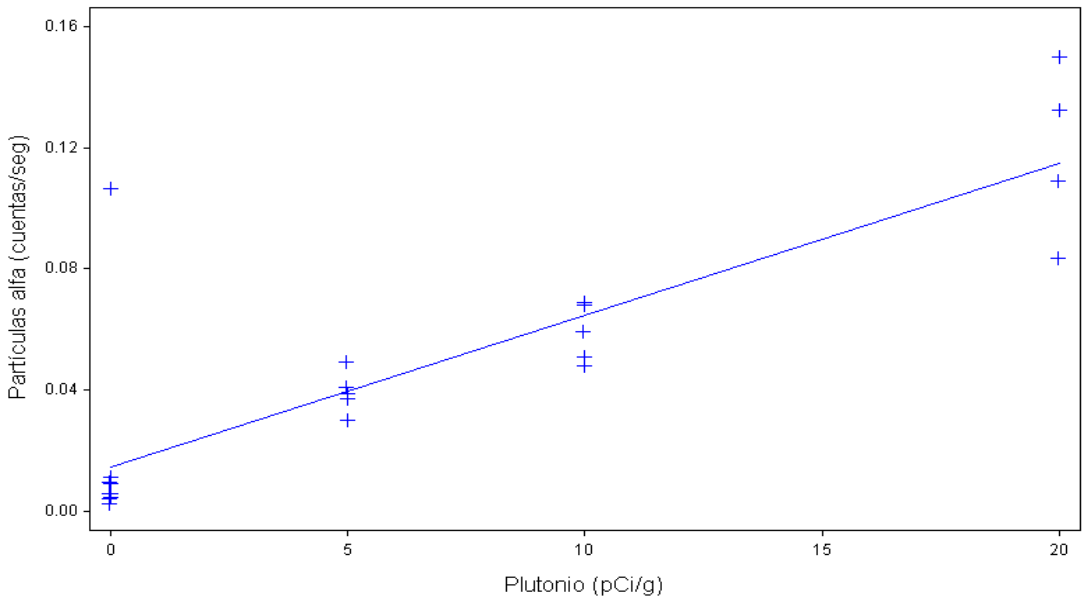
Laboratorio C media=4.251 DS=0.224
Laboratorio D media=4.435 DS=0.031

Cuál de los cuatro laboratorios está trabajando mejor? Evidentemente el laboratorio D tiene los resultados más exactos y más precisos.. El laboratorio C por el contrario es el que está trabajando peor: es el que tiene la menor exactitud y la menor precisión de los cuatro.
Es importante destacar que para estudiar la exactitud de un método de medición, se necesita hacer mediciones de muestras para las que se conocer el " valor verdadero". Pero este valor nunca se conoce exactamente. Lo que a veces se puede hacer es preparar las muestras a medir de modo tal que se conozca aproximadamente el " valor verdadero". Por ejemplo se pueden preparar las muestras mezclando proporciones conocidas (medidas lo mejor posible) de distintos materiales de alta pureza. Las mediciones se hacen entonces sobre estas muestras " artificiales".
Los errores sistemáticos en el proceso de medición afectan la exactitud. Los errores aleatorios provocan variabilidad de las mediciones y por lo tanto afectan la dispersión de los datos (la precisión).

Nota: antes de calcular las medias y las DS para cada laboratorio, convendría hacer algún gráfico (puede ser de puntos) para detectar si hay valores atípicos que puedan tener mucha influencia sobre estas dos medidas. En ese caso conviene calcular otras medidas de posición y dispersión o aplicar alguna transformación a los datos.

Gráfico de dispersión (scatter plot).

Para estudiar la relación entre dos variables numéricas se usa el "gráfico de dispersión" que es simplemente representar el valor de una de las dos variables en el eje de las abscisas y el de la otra en el eje de las ordenadas.
Ejemplo: cuando el plutonio está presente en pequeñas cantidades mezclado con otros materiales es difícil detectarlo. Una forma de detectarlo es medir las partículas alfa que emite. En una investigación para estudiar la relación entre la cantidad de plutonio y la emisión de partículas alfa, se midieron varias veces cuatro materiales standards para los que se sabe que la actividad de plutonio (0, 5, 10 y 20 picocuries por gramo (pCi/g). Los resultados de estas mediciones están en el archivo plutonio.xls y en el siguiente gráfico se puede apreciar la relación entre las dos variables



Este ejemplo es un problema de "calibración": queremos a partir del valor medido de partículas alfa, conocer aproximadamente la actividad del plutonio. Para ello se emplea un modelo de "regresión" (que veremos más adelante).