Cap. 1 : Introducción y estadística descriptiva

Alexandre Blondin Massé

Departamento de Informática y Matematica
Université du Québec à Chicoutimi

13 de junio del 2015
Modelado de sistemas aleatorios
Ingeniería de sistemas, producción y ambiental

Tabla de contenidos

1. Presentación del curso

2. Introducción

3. Métodos gráfico y tabular

4. Medidas de localización

5. Medidas de variabilidad

Profesor

- **Nombre**: Alexandre Blondin Massé;
- Conseguí mi **Ph.D.** en 2011;
- **Especialidad**: informática teórica, algorítmico, estructuras de datos, teoría de grafos, geometría digital.
- **Profesor** al departamento de **informatica** de UQAM (Université du Québec à Montréal);
- Fue **profesor** al departamento de **informatica y matematica** de UQAC (Université du Québec à Chicoutimi) de 2011-2014;
- **Sitio web**: `http://thales.math.uqam.ca/~blondin/es/8mqg210`

Capítulos

- Estadística **descriptiva**;
- **Probabilidad**;
- Distribuciones de variables **discretas**;
- Distribuciones de variables **continuas**;
- Distribución **conjunta** y estimación **puntual**;
- Intervalos de **confianza** y pruebas de **hipótesis**;
- Inferencia basada en **dos pruebas**;
- **Regresión** lineal simple y **correlación**.

Curso intensivo

- 8 cursos de 4 horas;
- 7 días (de sábado a sábado);
- Para retener el **máximo de información**, propongo los siguientes elementos de evaluación:
  - 8 series de ejercicios (5 % cada una);
  - 5 quizz (5 % cada una);
  - 1 examen (40 %).
  - Se conserva los 12 **mejores** resultados de los $8 + 5 = \mathbf{13}$ primeras evaluaciones;
  - Si el resultado del examen es **mejor** que los otros resultados, cuenta para 100 %.

Tabla de contenidos

1. Presentación del curso

2. Introducción

3. Métodos gráfico y tabular

4. Medidas de localización

5. Medidas de variabilidad

Caso de estudio: algoritmos de ordenación

- A veces, en informatica, debemos **ordenar datos**;
- Existen varios algoritmos:
  - **de burbuja** (**bubble sort**);
  - **por selección** (**selection sort**);
  - **rápido** (**quick sort**), etc.
- ¿Cual algoritmo es el **mejor**? ¿El **peor**?
- ¿Como podemos **decidir**lo o **demostrar**lo?

Implementación en Python (1/6)

```python
# Bubble sort
# -----------

def bubble_sort(L):
    n = len(L)
    while True:
        swapped = False
        for i in range(1, n):
            if L[i - 1] > L[i]:
                swap(L, i-1, i)
                swapped = True
        if not swapped: break

# Selection sort
# --------------

def selection_sort(L):
    n = len(L)
    for i in range(n - 1):
        k = i
        for j in range(i + 1, n):
            if L[j] < L[k]: k = j
        swap(L, i, k)
```

Implementación en Python (2/6)

```
1  # Quicksort
2  # ---------
3
4  def quick_sort(L):
5      quick_sort_recursive(L, 0, len(L) - 1)
6
7  def quick_sort_recursive(L, i, j):
8      if i < j:
9          p = partition(L, i, j)
10         quick_sort_recursive(L, i, p - 1)
11         quick_sort_recursive(L, p + 1, j)
12
13 def partition(L, i, j):
14     pi = randint(i, j)
15     pv = L[pi]
16     swap(L, pi, j)
17     p = i
18     for k in range(i, j):
19         if L[k] <= pv:
20             swap(L, k, p)
21             p += 1
22     swap(L, p, j)
23     return p
```

Implementación en Python (3/6)

```
 1  from random import randint
 2  from time import time
 3  from sort import bubble_sort, selection_sort, quick_sort
 4
 5  lengths = [5, 10, 20, 50, 100, 500, 1000, 2000, 3000, 4000, 5000]
 6
 7  s = "Number\tBubble sort\tSelection sort\tQuicksort\n"
 8
 9  for length in lengths:
10      s += "%s\t" % length
11      L = [randint(1, length) for _ in range(length)]
12      M = L[:]
13      # Timing bubble sort
14      before = time()
15      bubble_sort(L)
16      after = time()
17      s += "%.6f\t" % (after - before)
18      # Timing selection sort
19      before = time()
20      selection_sort(L)
21      after = time()
22      s += "%.6f\t" % (after - before)
23      # Timing quick sort
24      before = time()
25      quick_sort(M)
26      after = time()
27      s += "%.6f\n" % (after - before)
28  print s
```

Implementación en Python (4/6)

Number	Bubble sort	Selection sort	Quicksort
5	0.000005	0.000005	0.000012
10	0.000008	0.000008	0.000020
20	0.000041	0.000019	0.000041
50	0.000239	0.000072	0.000099
100	0.000946	0.000259	0.000213
500	0.024843	0.005279	0.001264
1000	0.097104	0.027886	0.002650
2000	0.406551	0.082428	0.005686
3000	0.875737	0.181358	0.008682
4000	1.557427	0.342378	0.012246
5000	2.508060	0.519824	0.015225

Observaciones:

- Las muestras son **aleatorias**;
- Parece que **de burbuja** es más lento que **por selección**;
- **Por selección** es más rápido que **quicksort** para $n \leq 80$;
- A partir de ≈ 80, quicksort es **más rápido**.

Implementación en Python (5/6)

```
1  from random import randint
2  from time import time
3  from sort import bubble_sort, selection_sort, quick_sort
4
5  NUM_TRIALS = 20
6  LENGTH = 2000
7
8  s = "Bubble sort\tSelection sort\tQuicksort\n"
9
10 for trial in range(NUM_TRIALS):
11     L = [randint(1, LENGTH) for _ in range(LENGTH)]
12     M = L[:]
13     # Timing bubble sort
14     before = time()
15     bubble_sort(L)
16     after = time()
17     s += "%s\t" % (after - before)
18     # Timing selection sort
19     before = time()
20     selection_sort(L)
21     after = time()
22     s += "%s\t" % (after - before)
23     # Timing quick sort
24     before = time()
25     quick_sort(M)
26     after = time()
27     s += "%s\n" % (after - before)
28 print s
```

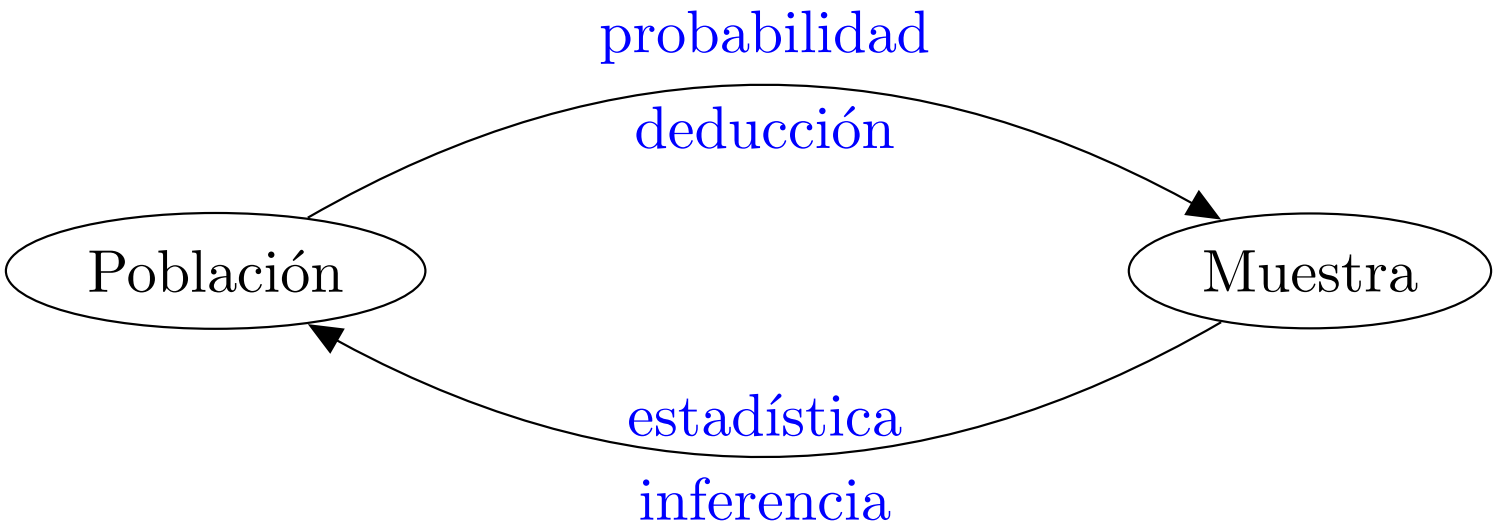
Implementación en Python (6/6)

```
Bubble sort     Selection sort  Quicksort
0.376595020294  0.0802211761475 0.00616908073425
0.382798910141  0.0806231498718 0.0060179233551
0.38033914566   0.0823440551758 0.00569605827332
0.373054027557  0.0820891857147 0.00556492805481
0.387526035309  0.0821340084076 0.0059380531311
0.386576890945  0.0819408893585 0.00588917732239
0.383502006531  0.0827250480652 0.00591397285461
0.383842945099  0.0832149982452 0.00544500350952
0.37625002861   0.0814909934998 0.00559282302856
0.380059957504  0.0818281173706 0.00566816329956
0.376256942749  0.0825319290161 0.00544714927673
0.383063077927  0.0813138484955 0.00535702705383
0.377019882202  0.0839569568634 0.00557494163513
0.394201993942  0.0842108726501 0.00587892532349
0.393414020538  0.0815269947052 0.00568079948425
0.374377965927  0.0817499160767 0.00585389137268
0.386357784271  0.0814430713654 0.00562405586243
0.375653982162  0.0817761421204 0.00550889968872
0.379374027252  0.0824038982391 0.00609111785889
0.384421110153  0.0818710327148 0.00570011138916
```

Observaciones:

- Los tiempos son **muy similares**;
- En otras palabras, la **varianza** es **pequeña**.

Relación entre probabilidad y estadística



- **Probablidad** describe la **distribución** de las muestras;
- **Estadística** permite **inferir característicos** de la población.

Definiciones básicas

- **Población**: conjunto de objetos estudiados;
- **Muestra**: subconjunto de una población;
- **Datos**: información (interesante o no) que se puede convertir en muestras. Pueden ser
  - **univariados** o
  - **multivariados**.
- **Variable**: característico de datos que son diferentes en función del objeto.

Tabla de contenidos

1. Presentación del curso

2. Introducción

3. Métodos gráfico y tabular

4. Medidas de localización

5. Medidas de variabilidad

Notación

- Una muestra de n **observaciones** x_1, x_2, x_3, ..., x_n es un **vector**

$$x = (x_1, x_2, x_3, \ldots, x_n);$$

- El **tamaño de muestra** est n;
- x_i es la i**-ésima** observación del conjunto de datos;
- También utiliza el símbolo $\sum$ para denotar una **suma** :

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \cdots + x_n$$

$$\sum_{i=1}^{n} x_i^2 = x_1^2 + x_2^2 + x_3^2 + \cdots + x_n^2$$

Diagrama de tallo y hojas

- Une forma rápida de obtener una representación visual es construir un **diagrama de tallo y hojas** :
- **Ejemplo :** La temperatura (en grados **Farenheit**) media anual de 35 ciudades de Estados Unidos :

2	9.0, 9.8
3	3.1, 4.1, 5.3, 5.8, 6.2, 9.0, 9.5, 9.5
4	0.0, 1.0, 2.4, 3.6, 3.7, 4.8, 5.0, 5.2, 6.0, 6.7, 8.1, 9.0, 9.2
5	1.0, 1.3, 2.0, 5.5, 7.1, 7.4, 7.6, 8.5, 9.3
6	9.0
7	0.0

- Columna de **izquierda** : **primer dígito**;
  Columna de **derecha** : **segundó dígito** y **decimal**.
- Por ejemplo, se obtiene **57.1** por la lectura de **5** y de **7.1**.

Diagrama de tallo y hojas

Esta representación tiene varias **ventajas**:

- Identificación de un valor **característico** o **representativo**;
- Grado de **dispersión** respecto al valor característico;
- **Simetría** en la distribución;
- **Picos** y **huecos**;
- Valores **atípicos**.

Tabla de distribución de frecuencias (1/2)

- Se recogieron los **salarios** de 42 estudiantes que tienen un títutol de ingeniería eléctrica que acaba de empezar a trabajar.
- Se obtuvieron los valores siguientes en **K$**:

```
60 56 52 50 47 47 52 47 52 49 50 54 52
51 52 57 50 52 52 57 54 51 56 51 50 51
52 49 57 51 54 50 52 52 47 51 54 54 51
49 48 51
```

- ¿Como presentar estos **datos**?

Tabla de distribución de frecuencias (2/2)

- Datos:

```
60 56 52 50 47 47 52 47 52 49 50 54 52
51 52 57 50 52 52 57 54 51 56 51 50 51
52 49 57 51 54 50 52 52 47 51 54 54 51
49 48 51
```

- Cuenta el **número de ocurrencias** de cada salario;
- Escribe el resultado en una **tabla de distribución de frecuencias**.
- Se obtiene:

Salario	47	48	49	50	51	52	54	56	57	60
Número	4	1	3	5	8	10	5	2	3	1

Representaciones gráficas

- En general, es más fácil darse una visión general de los datos utilizando un **gráfico** en lugar de una tabla.
- Las representaciones **más comunes** son:
  - Los **gráficos de líneas** (**line graphs**);
  - Los **gráficos de rectángulos** (**bar graphs**);
  - Los **polígonos de frecuencias** (**frequency polygons**);
  - Los **pictogramas**;
  - etc.

Gráfico de líneas

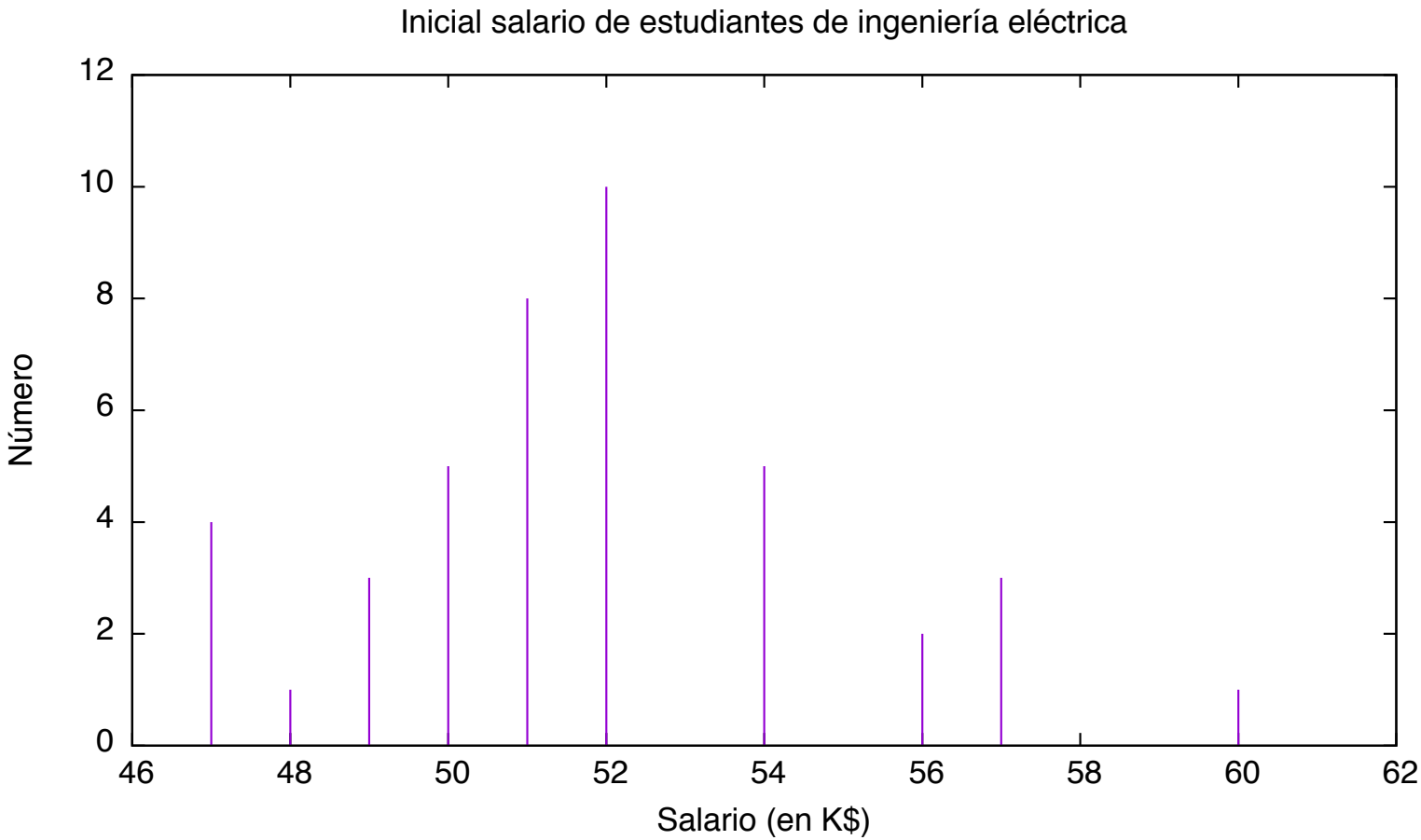
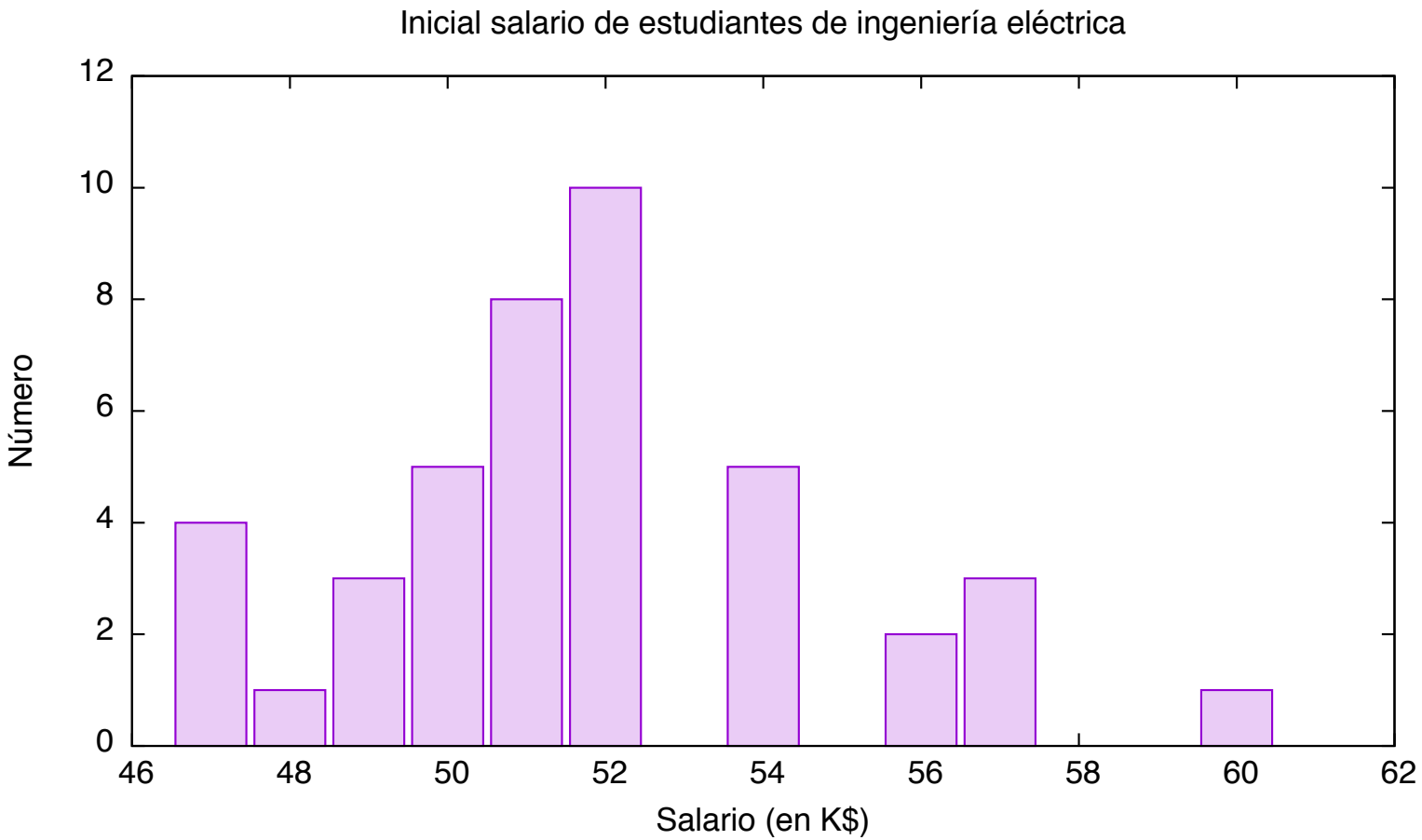
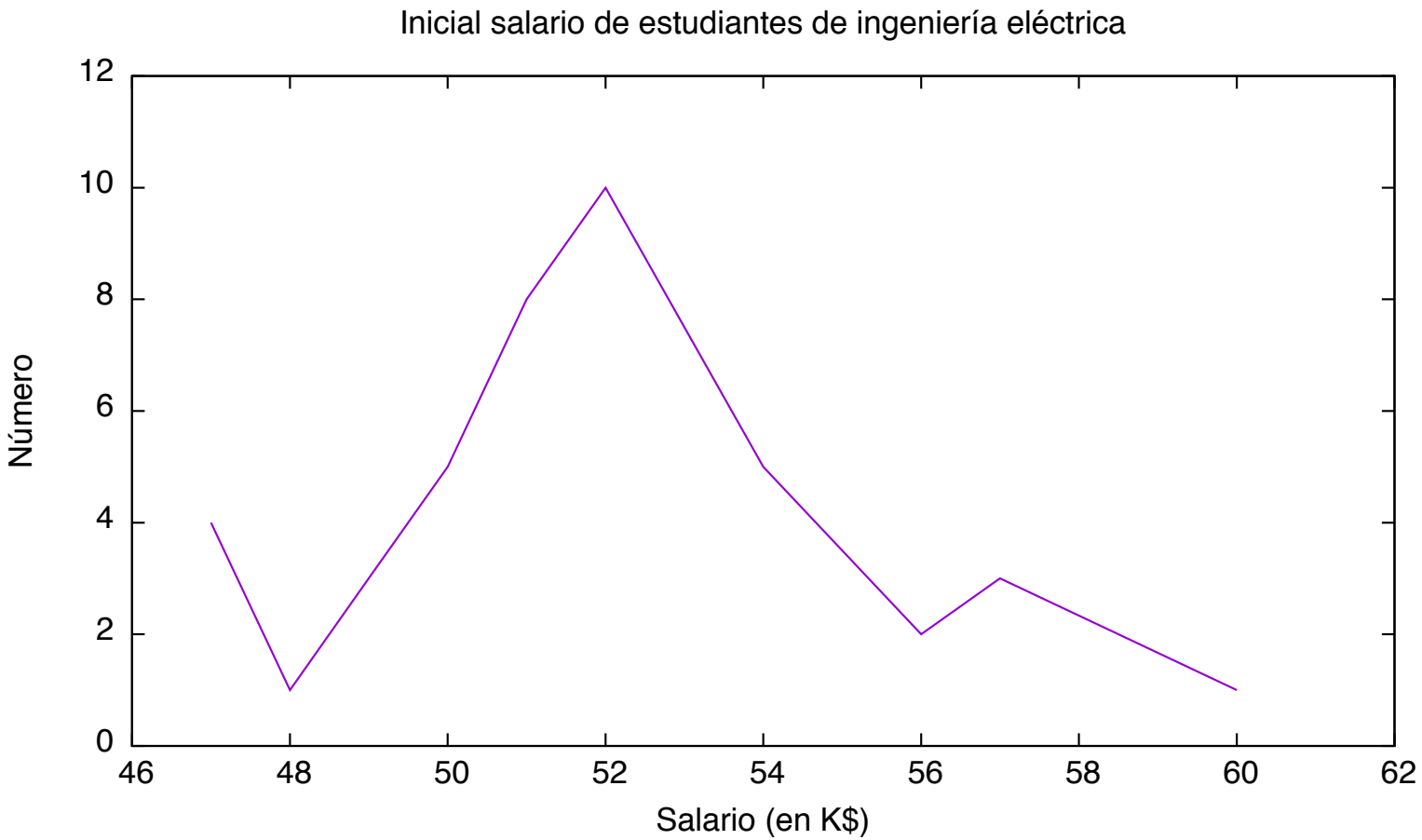


Gráfico de rectángulos



Polígono de frecuencias



El formato CSV

- Es un formato de archivo de **texto** que contiene **datos**;
- Se reconoce por las **hojas de cálculo**, como **Excel**, **LibreOffice Calc**, **R**, **Gnuplot**, etc.
- Ejemplo :

```
Genero,Nombre,Nacimiento
M,Daniel,1979
F,Lucia,1988
F,Clara,1992
M,Hugo,1989
M,Javier,1991
```

- En ese **curso**, deben utilizar el **formato CSV** para representar sus datos.

Gnuplot (1/2)

- Es un **software** que produce **gráficos** en 2D o 3D a partir de datos.
- Sitio web : `http://www.gnuplot.info/`.
- Ejemplo de **script** :

```
 1  set terminal pdf
 2  unset key
 3  set output 'rect-graph.pdf'
 4  set datafile separator ","
 5  set title "Inicial salario ... ingenieria electrica"
 6  set xlabel "Salario (en K\$)"
 7  set ylabel "Numero"
 8  set xrange [46:62]
 9  set yrange [0:12]
10  set boxwidth 0.9
11  plot 'salaire-freq.dat' with boxes fs solid 0.2
```

Frecuencia relativa (1/2)

- La mayoría del tiempo, es más interesante la **frecuencia relativa** que la **frecuencia absoluta**.
- Ejemplo de los salarios :

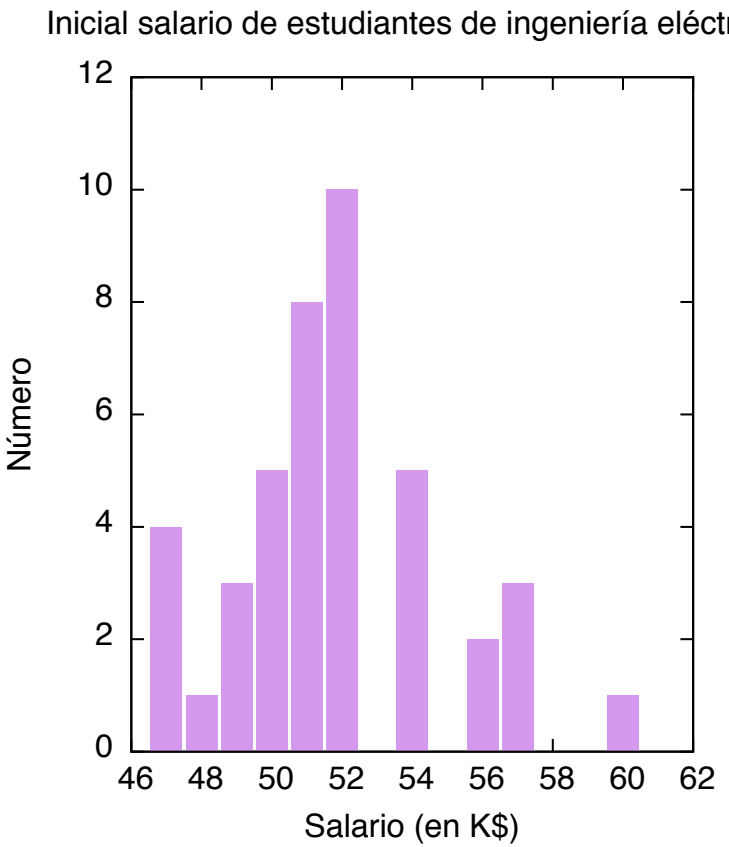
Salario	47	48	49	50	51	52	54	56	57	60
Número	4	1	3	5	8	10	5	2	3	1

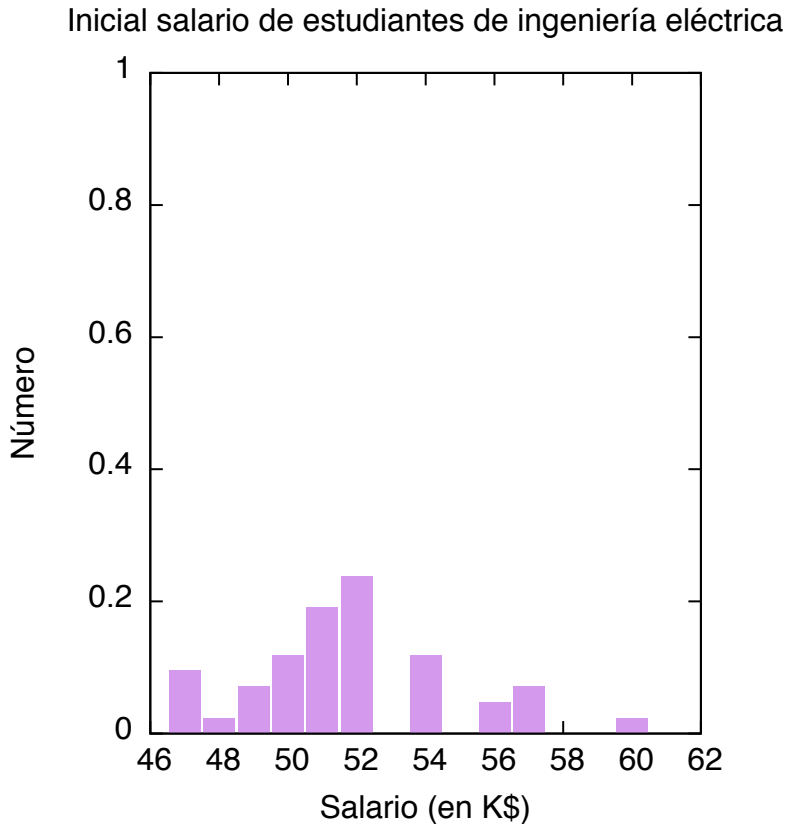
- En total, hay **42** valores. Dividiendo cada valor por 42 se obtiene la frecuencia **relativa**:

Salario	47	48	49	50	51
Frecuencia	0.095	0.024	0.071	0.119	0.190

Salario	52	54	56	57	60
Frequencia	0.238	0.119	0.048	0.071	0.024

Frecuencia relativa (2/2)





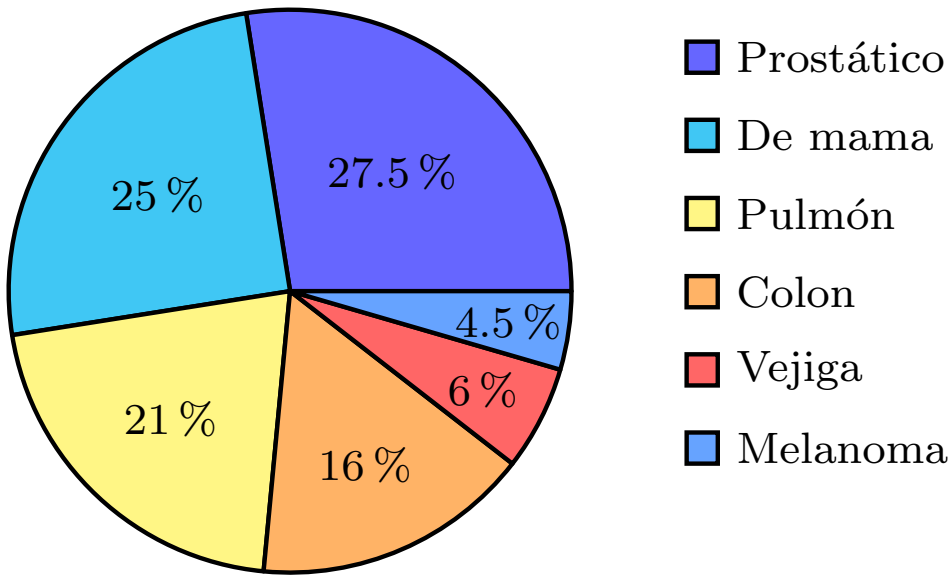
Datos cualitativos

- De vez en cuando, los datos no son **numéricos**.
- **Ejemplo :** En una clínica de **oncología**, hay diferentes tipos de cáncer que aflige los 200 últimos pacientes que se han encontrado :

Tipo de cáncer	Número	Frecuencia
Pulmón	42	0.210
De mama	50	0.250
Colon	32	0.160
Prostático	55	0.275
Melanoma	9	0.045
Vejiga	12	0.060

Gráfico circular

- **Cualitativos** datos se pueden representar con gráficos de líneas, rectángulos, etc.
- También es posible utilizar un **gráfico circular** (**pie chart**) :



- Ángulo de un sector circular :

$$\textbf{ángulo} = \textbf{frecuencia relativa} \times 360^\circ.$$

Datos agrupados (1/3)

- A veces, el **tamaño** de los datos es **muy grande**.
- **Ejemplo :** Vida de 200 incandescantes bulbos en **horas**:

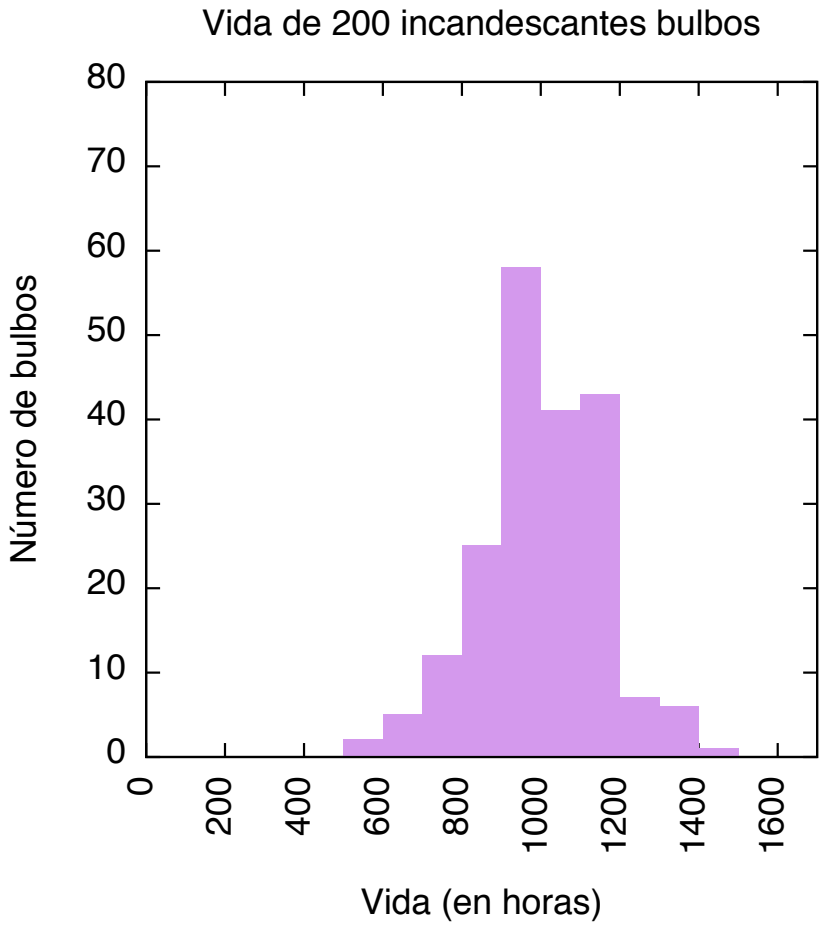
1067	855	1157	1022	923	521	930	999	901	996
1187	824	844	1037	1026	1039	1023	1134	998	610
919	1196	1092	1162	1195	1195	978	832	1333	811
933	928	807	954	932	1035	1324	818	780	900
1067	1118	653	980	814	1103	1151	863	1147	883
1083	1040	984	856	932	938	996	1133	916	1001
785	1170	1340	1009	1217	1153	1063	944	1250	1106
1037	935	1000	990	867	1289	924	1078	765	895
1126	936	918	929	950	905	1122	938	970	1157
1151	1009	1085	896	958	946	858	1071	1002	909
1077	1049	940	1122	1203	1078	890	704	621	854
958	760	1101	878	934	910	788	1143	935	1035
1112	931	990	1258	1192	699	1083	880	801	1122
1292	1180	1106	1184	775	1105	1081	709	860	1110
1156	972	1237	765	1311	1069	1021	1115	1303	1178
949	1058	1069	970	922	1029	1116	954	1171	1149
920	948	1035	1045	956	1102	958	902	1037	702
830	1063	1062	1157	833	1320	1011	1102	1138	951
992	966	730	980	1170	1067	932	904	1150	1091
658	912	880	1173	824	529	705	1425	972	1002

Datos agrupados (2/3)

- **Solución :** los datos son agrupados en **clases**.
- ¿Cuántas clases? Típicamente entre 5 **y** 10.
- Hay que hacer un **compromiso** entre
  - **demasiadas clases** : difícil de detectar un **patrón**;
  - **demasiado pocas clases** : pérdida de información en cada clase.
- Es **ideal** (pero no es obligatorio) elegir intervalos **regulares**.
- Por convención,
  - **extremo izquierdo** : valor **incluido** en el intervalo;
  - **extremo derecho** : valor **excluido**.

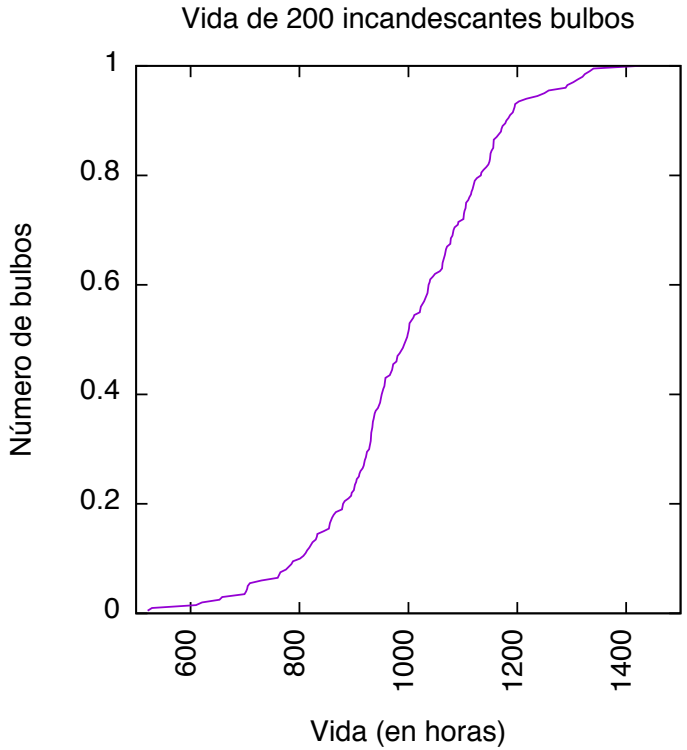
Datos agrupados (3/3)

Intervalo	Frecuencia
500 – 600	2
600 – 700	5
700 – 800	12
800 – 900	25
900 – 1000	58
1000 – 1100	41
1100 – 1200	43
1200 – 1300	7
1300 – 1400	6
1400 – 1500	1



Se llama **histograma de frecuencias**.

Frecuencias acumuladas



- Este gráfico se llama **ojiva**.
- ¿Qué proporción de bulbos tienen una vida de 1000 horas o menos ?
- Respuesta : Cerca de 52 %.

Tabla de contenidos

1. Presentación del curso

2. Introducción

3. Métodos gráfico y tabular

4. Medidas de localización

5. Medidas de variabilidad

Indicador estadístico

Has cuatro tipos de **indicadores**:

- de **tendencia central** : media, mediana, moda;
- de **dispersión** : alcance, varianza, desviación estándar;
- de **posición** : cuartil, percentil.
- **qualitativos** : simétrico, simétrica a la izquierda, a la derecha.

La media (1/4)

Definición

Sea x una muestra. La ***media*** *de x,* ***denotada por*** $\bar{x}$*, es*

$$\bar{x} = \left(\sum_{i=1}^{n} x_i \right) / n.$$

Proposición

Sean $x = (x_1, x_2, \ldots, x_n)$, $y = (y_1, y_2, \ldots, y_n)$ dos muestras tales que

$$y_i = ax_i + b, \quad i = 1, 2, \ldots, n,$$

donde a ***y*** *b* ***son constantes****. Entonces*

$$\bar{y} = a\bar{x} + b.$$

La media (2/4)

- Las puntuaciones de los campeones de los **Masters Golf Tournaments** en Estados-Unidos entre 1982 y 1991 son:

$$284, 280, 277, 282, 279, 285, 281, 283, 278, 277.$$

- Entonces, la **media** es

$$\begin{aligned} \bar{x} &= \frac{284 + 280 + 277 + 282 + 279 + 285 + 281 + 283 + 278 + 277}{10} \\ &= \frac{2806}{10} \\ &= 280{,}6. \end{aligned}$$

La media (3/4)

- Sea una **tabla de frecuencias** con
  - **valores** $v_1, v_2, \ldots, v_k$
  - **frecuencias** $f_1, f_2, \ldots, f_k$.
- ¿Como podemos calcular la **media**?

La media (4/4)

- En un grupo, hay 54 personas cuya edad es

Edad	Frecuencia
15	2
16	5
17	11
18	9
19	14
20	13

- Entonces, la **edad promedio** es

$$\begin{aligned} \bar{x} &= \frac{15 \cdot 2 + 16 \cdot 5 + 17 \cdot 11 + 18 \cdot 9 + 19 \cdot 14 + 20 \cdot 13}{54} \\ &= \frac{985}{54} \\ &\approx 18{,}24. \end{aligned}$$

La mediana (1/2)

Definición

Sea $x = (x_1, x_2, \ldots, x_n)$ *una muestra* ***ordenada*** ($x_i \leq x_{i+1}$ *para* $i = 1, 2, \ldots, n-1$). *La* ***mediana*** *de* x *es*

$$\begin{cases} x_{(n+1)/2}, & \text{si } n \text{ es } \textbf{impar}; \\ \frac{x_{n/2} + x_{n/2+1}}{2}, & \text{si } n \text{ est } \textbf{par}. \end{cases}$$

En otras palabras, es el dato ***central*** *de* x.

La mediana (2/2)

- El ejemplo anterior

Edad	Frecuencia
15	2
16	5
17	11
18	9
19	14
20	13

- Hay 54 datos, tenemos que calcular la media de los datos números 27 y 28:

Diferencia entre media y mediana (1/3)

- Cuando los datos son «**normales**», la media y la mediana están muy **cerca**.
- En contraste a la **mediana**, la media es **sensible** a los valores **extremos**. Por ejemplo, si

$$x = (10, 15, 20, 25, 30, 1000),$$

entonces $\bar{x} \approx$ **183,33**, mientras que la mediana es $(20 + 25)/2 =$ **22,5**.

- Estos dos indicadores dan información **diferente**.

Diferencia entre media y mediana (2/3)

- Se lleva a cabo un experimento en ratones, que se dividen en **dos grupos**: (1) un entorno **desinfectado** y un entorno **convencional**.
- Se cuenta el tiempo que viven los ratones en **días**:

Entorno desinfectado

1	58, 92, 93, 94, 95
2	02, 12, 15, 29, 30, 37, 40, 44, 47, 59
3	01, 01, 21, 37
4	15, 34, 44, 85, 96
5	29, 37
6	24
7	07
8	00

Entorno convencional

1	59, 89, 91, 98
2	35, 45, 50, 56, 61, 65, 66, 80
3	43, 56, 83
4	03, 14, 28, 32

Diferencia entre media y mediana (3/3)

- Las **medias** son 344,07 y 292,32;
- Las **medianas** son 259 y 265;
- Las medias son claramente **diferentes**, pero las medianas son **similares**. ¿Por qué?
- Parece que los ratones que **viven más tiempo** tienen una vida más larga cuando son en un **entorno desinfectado**;
- En contraste, no hay **diferencia** en el caso donde viven menos tiempo.

La moda

Definición

*Sea x una muestra. Un número m es una **moda** de x si no existe un valo en x que aparece con más frecuencia que m.*
***Nota :** La moda m no es necesariamente **única**.*

Lanza un dado 40 veces et se obtienen los resultados siguientes:

Valeur	1	2	3	4	5	6
Fréquence	9	8	5	5	6	7

Calculen la **media**, la **mediana** y la **moda**.

Tabla de contenidos

1. Presentación del curso

2. Introducción

3. Métodos gráfico y tabular

4. Medidas de localización

5. Medidas de variabilidad

La varianza (1/4)

Definición

Sea x una muestra. La ***varianza*** *de x es*

$$s_n^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 / n.$$

La ***varianza insesgada*** *de x es*

$$s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 / (n-1).$$

Por razones que se aclararán **más adelante**, se prefiere la **varianza insesgada**.

La varianza (2/4)

Calculamos las **varianzas insesgadas** de las muestras

$$\begin{aligned} x &= (3, 4, 6, 7, 10) \\ y &= (-20, 5, 15, 24)? \end{aligned}$$

Obtenemos

$$\begin{aligned} \bar{x} &= \frac{3+4+6+7+10}{5} = 6 \\ \bar{y} &= \frac{-20+5+15+24}{4} = 6 \\ s_x^2 &= \frac{(-3)^2+(-2)^2+0^2+1^2+4^2}{4} = 7{,}5 \\ s_y^2 &= \frac{(-26)^2+(-1)^2+9^2+18^2}{3} \approx 360{,}67 \end{aligned}$$

La varianza (3/4)

Una **fórmula** muy útil:

Proposición

Sea $x = (x_1, x_2, \ldots, x_n)$ *una muestra. Entonces*

$$s_x^2 = \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right) / (n-1).$$

Por otra parte, si $y = (y_1, y_2, \ldots, y_n)$ *es una otra muestra y* $y_i = ax_i + b$ *para* $i = 1, 2, \ldots, n$, *entonces*

$$s_y^2 = a^2 s_x^2.$$

La varianza (4/4)

Anno	Accidentes
1985	22
1986	22
1987	26
1988	28
1989	27
1990	25
1991	30
1992	29
1993	24

Se cuenta el **número de accidentes** de aviación por año entre 1985 y 1993.

Qué es la varianza (insesgada) ?

La varianza (4/4)

Anno	Accidentes
1985	22
1986	22
1987	26
1988	28
1989	27
1990	25
1991	30
1992	29
1993	24

Se cuenta el **número de accidentes** de aviación por año entre 1985 y 1993.

Qué es la varianza (insesgada) ?

Respuesta:

$$s^2 = \frac{203 - 9(35/9)^2}{8} \approx 8{,}361.$$

La desviación estándar

Definición

*La **desviación estándar insesgada** s (**standard deviation**) de una muestra $x = (x_1, x_2, \ldots, x_n)$ es*

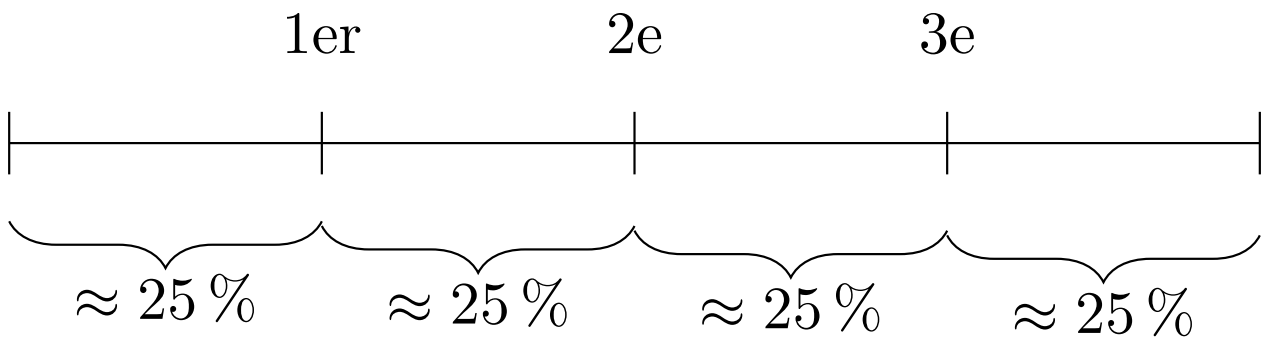
$$s_x = \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2/(n-1)} = \sqrt{s_x^2}.$$

Como para la **varianza**, se divide por $n-1$ en lugar de n.

Cuartiles (1/4)

Definición

Sea x una muestra e $i \in \{1, 2, 3\}$. El ***i-ésimo cuartil*** *de x es el número q en x de tal manera que $25i$ por ciento de los datos en x son* ***más pequeños o igual*** *a q y $100 - 25i$ por ciento de los datos son* ***más grandes o igual*** *a q. Si hay dos de tales valores q_1 y q_2, entonces es* ***la media*** *$q = (q_1 + q_2)/2$.*

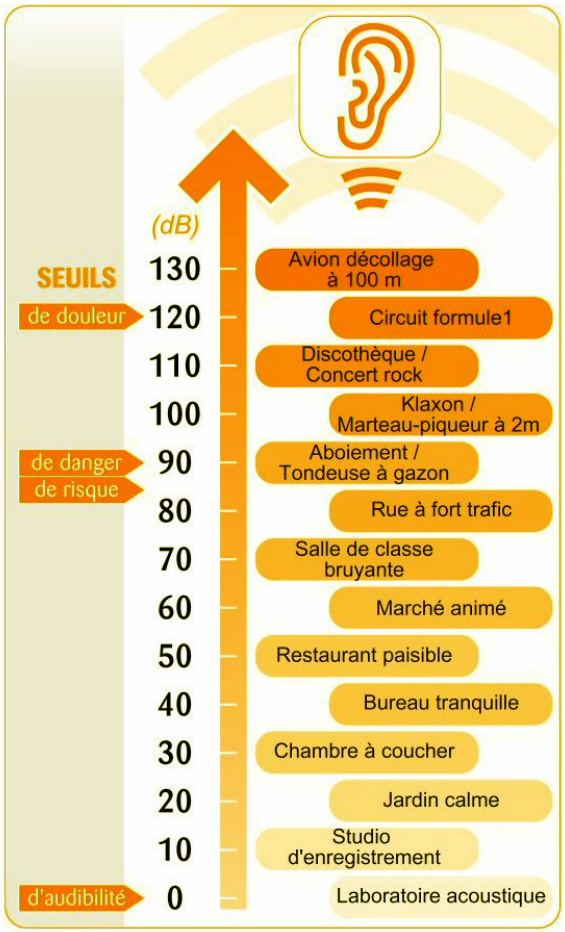


Cuartiles (2/4)

La unidad de medida del sonido est **el decibelio (dB)**. Se toma el nivel de ruido en 34 diferentes momentos en Grand Central Station, Manhattan :

```
 82   89   94   110 74   122 112
 95   100 78   65   60   90   83
 87   75   114 85   69   94   124
 115 107 88   97   74   72   68
 83   91   102 77   125 108
```

Calculen los cuartiles.



Cuartiles (3/4)

En primer lugar, hay que **ordenar los datos**. Un **diagrama de tallo y hojas** es adecuado :

6	0,5,8,9
7	2,4,4,5,7,8
8	2,3,3,5,7,8,9
9	0,1,4,4,5,7
10	0,2,7,8
11	0,2,4,5
12	2,4,5

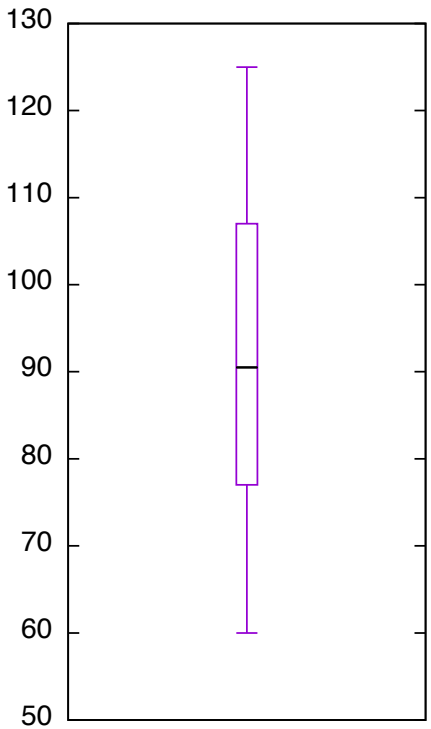
Calculamos los cuartiles:

$$\begin{aligned} 0{,}25 \cdot 34 &= 8{,}5 \to \mathbf{9} \\ 0{,}5 \cdot 34 &= 17 \to \mathbf{17—18} \\ 0{,}75 \cdot 34 &= 25{,}5 \to \mathbf{26} \end{aligned}$$

Los cuartiles son **77, 89,5 y 107**.

Cuartiles (4/4)

- En general, para calcular **el cuartil** i, se utiliza la estrategia siguiente;
- Se calcula $r = 0{,}25 \cdot i$;
- Si r no es **entero**, entonces se calcula el cuartil i por **redondeando a la unidad superior**.
- Si r es **entero**, entonces se calcula la media de los valores x_r y x_{r+1};
- Cuartiles son representados por un **diagrama de caja**;
- La **cuarta dispersión** es $q_3 - q_1$.



Percentiles

Definición

Sea x una muestra y $i \in \{1, 2, \ldots, 99\}$. El ***percentil*** i *de x es el valor c de x de tal manera que i por ciento de los datos de x son* ***más pequeños o igual*** *a c y $100 - i$ por ciento de los datos son* ***más grandes o igual*** *a c. Si hay dos de tales valores c_1 y c_2, entonces* ***la media*** *$c = (c_1 + c_2)/2$.*