CENTRE DE RECHERCHE PUBLIC

Microarray Center

APPLIED STATISTICS

Lecture 1

Data Presentation and Numerical Measures

Petr Nazarov
petr.nazarov@crp-sante.lu

10-11-2009

The course:

- Enhance understanding of statistical basics
- Gives the methodological tools for the research
- Provides practical skill for fast data analysis

Organization

- 12 lectures organized in 6 sections (18 hours)
- 5 tests for self-control
- Lectures are integrated with practical work

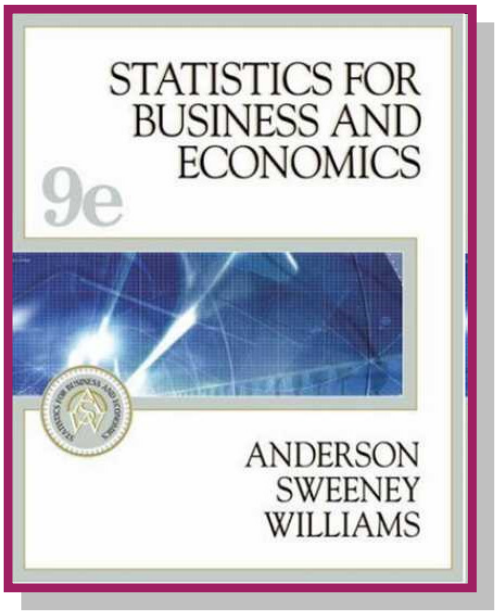
Software

- Microsoft® Excel
  - with Data Analysis Add-In installed

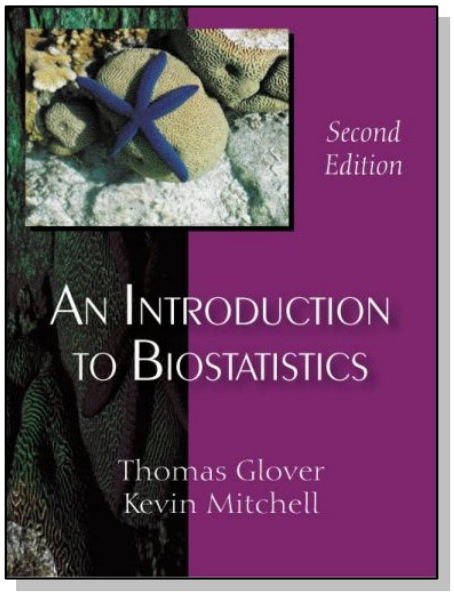
Web page

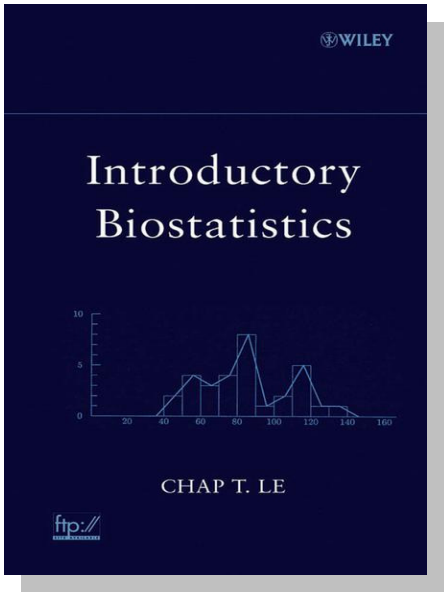
http://edu.sablab.net/stat

Recommended Literature

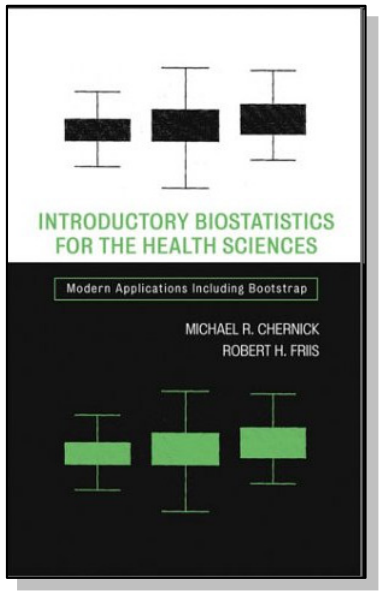


presentation
methodology





examples,
advanced topics



- **Data and statistic**
  - elements, variables and observation
  - types of data (qualitative and quantitative) and scales (nominal, ordinal, interval, ratio)
- **Descriptive statistics: tabular and graphical presentation**
  - frequency distribution
  - pie, bar chart and histogram representation
  - cumulative distributions
  - exploratory analysis: dot plot, stem-end-leaf
  - crosstabulation and scatter diagram
- **Descriptive statistics: numerical measures**
  - measures of location: mean, mode, median, quantiles/quartiles/percentiles
  - measure of variability: variance, standard deviation, MAD, coefficient of variation
  - other measures: skewness of distribution
  - z-score. Chebyshev's theorem. Detection of outliers.
- **Exploratory analysis.**
  - 5 number summary
  - box plot
- **Measure of association between two variables**
  - covariance and correlation coefficient
  - interpretation of correlation coefficient

Data
The facts and figures collected, analyzed, and summarized for presentation and interpretation.

elements

variables

observation

Person	Place	Gender	Net Worth ($BIL)	Age	Source	Internet Fame Score
William Gates III	1	M	40	53	Microsoft	9.5
Warren Buffett	2	M	37	79	Berkshire Hathaway	6.6
Carlos Slim Helu	3	M	35	69	telecom	2.1
Lawrence Ellison	4	M	22.5	64	Oracle	2.8
Ingvar Kamprad	5	M	22	83	IKEA	2.4
Karl Albrecht	6	M	21.5	89	Aldi	3.6
Mukesh Ambani	7	M	19.5	51	petrochemicals	4.4
Lakshmi Mittal	8	M	19.3	58	steel	5.4
Theo Albrecht	9	M	18.8	87	Aldi	1.5
Amancio Ortega	10	M	18.3	73	Zara	1.9
Jim Walton	11	M	17.8	61	Wal-Mart	3.9
Alice Walton	12	F	17.6	59	Wal-Mart	2.9

Can we consider the "Place" as element?

$$IFS = 3(\log_{10} N - 4.5)$$

Data *scales:*

Qualitative

Nominal scale
data use labels or names to identify an attribute of an element.

Ex.1: Male, Female

Ex.2: Rooms #: 101, 102, 103, …

Ordinal scale
data exhibit the properties of nominal data and the order or rank of the data is meaningful.

Ex.1: Winners: The 1st, 2nd, 3rd places

Ex.2: Marks: A, B, C, …

Quantitative

Interval scale
data demonstrate the properties of ordinal data and the interval between values is expressed in terms of a fixed unit of measure

Ex.1: Examination score 0 -100

Ex.2: Internet fame score ☺

Ratio scale
data demonstrate all the properties of interval data and the ratio of two values is meaningful.

Ex.1: Weight

Ex.2: Price

Example: Pancreatitis Study

The role of smoking in the etiology of pancreatitis has been recognized for many years. To provide estimates of the quantitative significance of these factors, a hospital-based study was carried out in eastern Massachusetts and Rhode Island between 1975 and 1979. **53 patients** who had a hospital discharge diagnosis of **pancreatitis** were included in this unmatched case-control study. The **control group** consisted of 217 patients admitted for **diseases other** than those of the pancreas and biliary tract. Risk factor information was obtained from a standardized interview with each subject, conducted by a trained interviewer.

adapted from Chap T. Le, Introductory Biostatistics

pancreatitis.xls

Pancreatitis patients:

Smokers	Ex-smokers	Ex-smokers	Smokers	Smokers	Smokers
Ex-smokers	Smokers	Smokers	Smokers	Smokers	Smokers
Ex-smokers	Smokers	Smokers	Ex-smokers	Smokers	Smokers
Ex-smokers	Ex-smokers	Smokers	Ex-smokers	Smokers	
Smokers	Never	Smokers	Ex-smokers	Ex-smokers	
Smokers	Ex-smokers	Smokers	Smokers	Ex-smokers	
Smokers	Smokers	Smokers	Smokers	Smokers	
Ex-smokers	Smokers	Smokers	Smokers	Smokers	
Smokers	Smokers	Smokers	Smokers	Smokers	
Smokers	Never	Smokers	Smokers	Smokers	

Frequency Distribution

Frequency distribution
A tabular summary of data showing the number (frequency) of items in each of several nonoverlapping classes.

Marks
A
B
C
B
A
B
B
A
B
C

Frequency distribution:

Mark	Frequency
A	3
B	5
C	2
Total	**10**

Relative frequency distribution:

Mark	Frequency
A	0.3
B	0.5
C	0.2
Total	**1**

Percent frequency distribution:

Mark	Frequency
A	30%
B	50%
C	20%
Total	**100%**

`pancreatitis.xls`

Frequency distribution:

Smoking	Cases	Controls
Never	2	56
Ex-smokers	13	80
Smokers	38	81
Total	53	217

Relative frequency distribution:

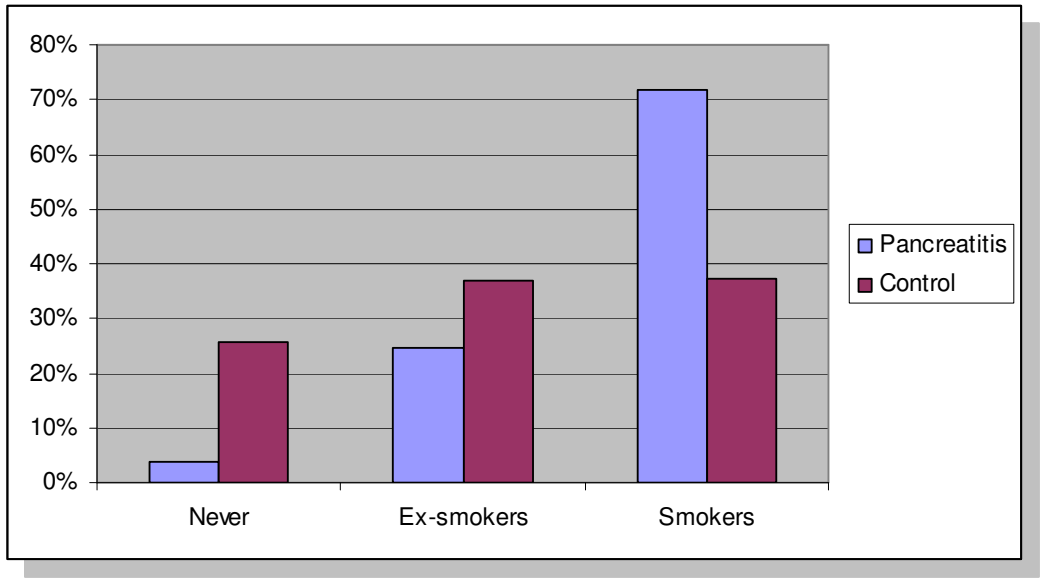
Smoking	Cases	Controls
Never	0.038	0.258
Ex-smokers	0.245	0.369
Smokers	0.717	0.373
Total	1	1

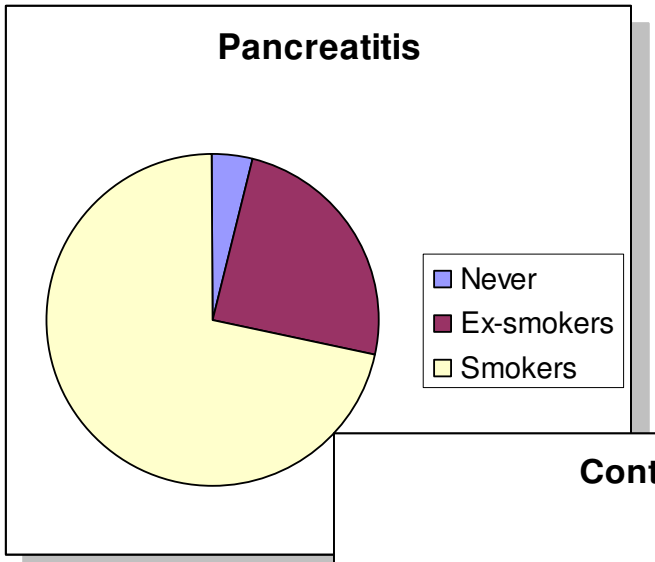
In MS Excel use the following functions:

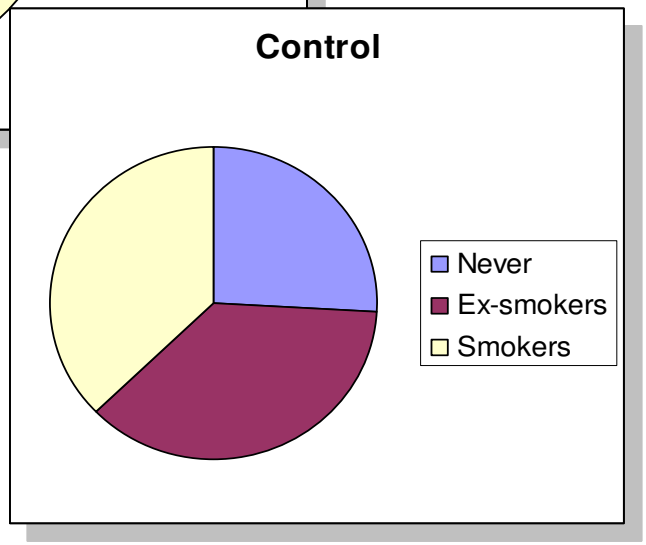
- `=COUNTIF(data,element)` to get number of “elements” found in the “data” area
- `=SUM(data)` to get the sum of the values in the “data” area

TABULAR AND GRAPHICAL PRESENTATION

Bar and Pie Charts







? Based on these data, can we say that smoking is one of the major causes of pancreatitis ?

In MS Excel use the following steps:

- Chart Wizard → Columns → Set data range (both columns of Percent freq. distribution)
- Chart Wizard → Pie → Set data range (one columns of Percent freq. distribution)

Histogram

The following are weights in pounds of 57 children at a day-care center:

children.xls

68	63	42	27	30	36	28	32	79	27
22	23	24	25	44	65	43	25	74	51
36	42	28	31	28	25	45	12	57	51
12	32	49	38	42	27	31	50	38	21
16	24	69	47	23	22	43	27	49	28
23	19	46	30	43	49	12			

Since the smallest number is 12, we may begin our first interval at 10.

bins

Weight Interval (lb)	Tally	Frequency	Relative Frequency (%)
10–19	𝍸	5	8.8
20–29	𝍸 𝍸 𝍸 ||||	19	33.3
30–39	𝍸 𝍸	10	17.5
40–49	𝍸 𝍸 |||	13	22.8
50–59	||||	4	7.0
60–69	||||	4	7.0
70–79	||	2	3.5
Total		57	100.0

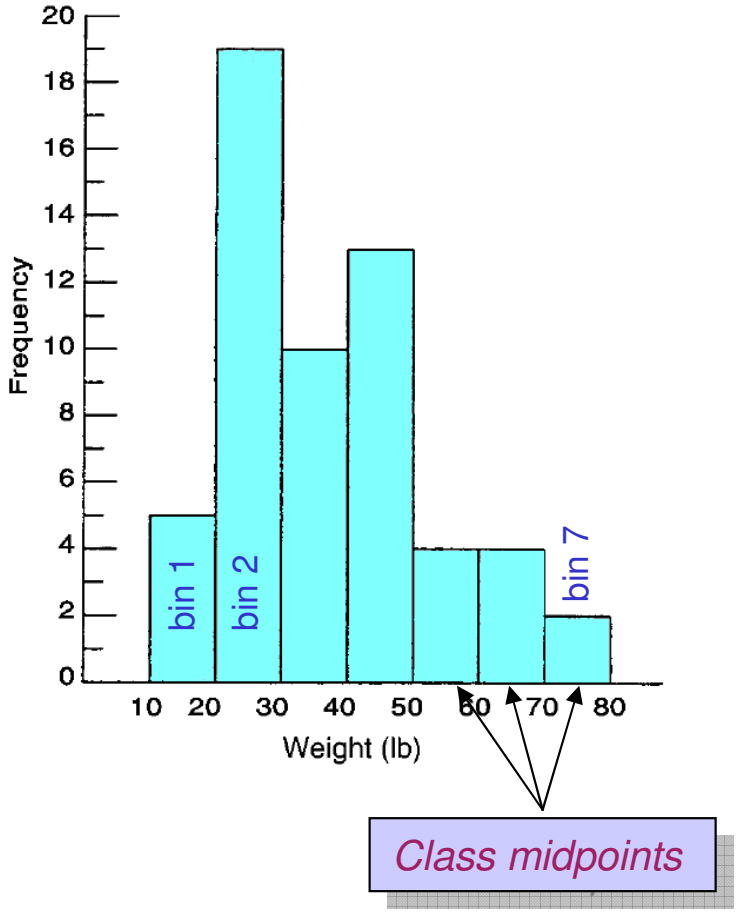
Chap T. Le, Introductory Biostatistics

TABULAR AND GRAPHICAL PRESENTATION

Histogram

68	63	42	27	30	36	28	32	79	27
22	23	24	25	44	65	43	25	74	51
36	42	28	31	28	25	45	12	57	51
12	32	49	38	42	27	31	50	38	21
16	24	69	47	23	22	43	27	49	28
23	19	46	30	43	49	12			

Bin	*Frequency*
20	5
30	21
40	8
50	14
60	3
70	4
80	2
More	0



In Excel use the following steps:

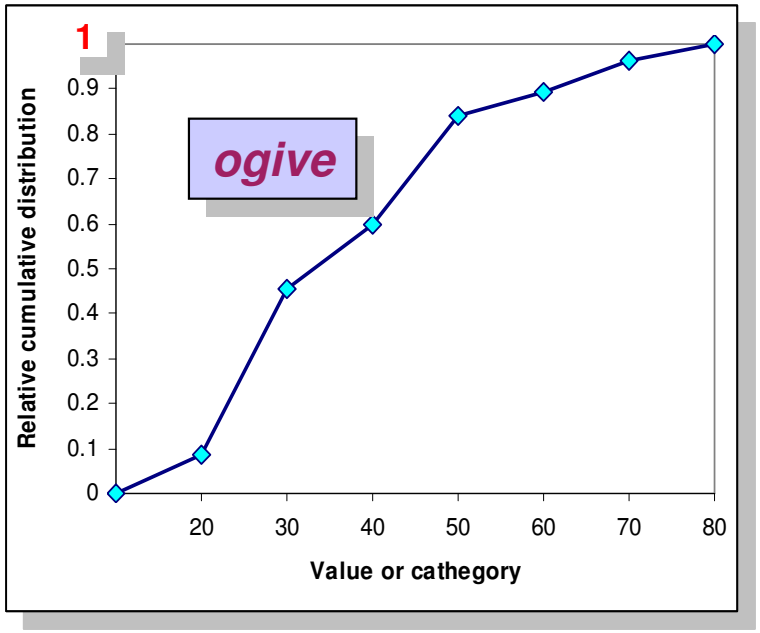
- Specify the column of bins (interval) upper-limits
- Tools → Data Analysis → Histrogram → select the input data, bins, and output (Analysis ToolPak should be installed)
- use Chart Wizard → Columns to visualize the results

Cumulative Frequency Distribution

Cumulative frequency distribution
A tabular summary of quantitative data showing the number of items with values less than or equal to the upper class limit of each class.

68	63	42	27	30	36	28	32	79	27
22	23	24	25	44	65	43	25	74	51
36	42	28	31	28	25	45	12	57	51
12	32	49	38	42	27	31	50	38	21
16	24	69	47	23	22	43	27	49	28
23	19	46	30	43	49	12			

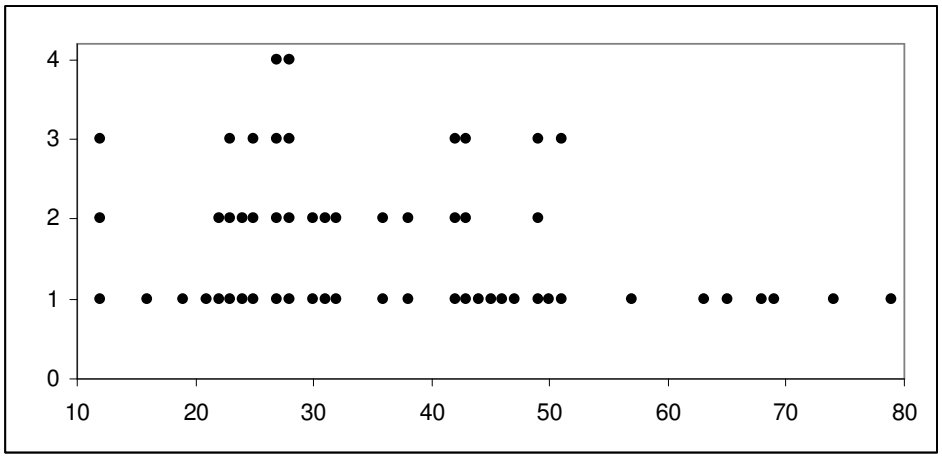
Bins	Frequency	Cumulative FD	Relative CFD	Percent CFD
10-20	5	5	0.09	9%
20-30	21	26	0.46	46%
30-40	8	34	0.60	60%
40-50	14	48	0.84	84%
50-60	3	51	0.89	89%
60-70	4	55	0.96	96%
70-80	2	57	1	100%
Total	57			



Exploratory data analysis

Methods that use simple arithmetic and easy-to-draw graphs to summarize data quickly.

Dot-plot



In Excel use the following steps:

- Sort the values
- Assume values are in column A, starting from 2nd row. In cell B2 use function `=IF(A2=A1, B1+1, 1)`
- drag the function for all values
- use Chart Wizard → Scatter plot

Stem-and-Leaf

1	2	2	2	6	9													
2	1	2	2	3	3	3	4	4	5	5	5	7	7	7	7	8	8	8
3	0	0	1	1	2	2	3	3	8	8								
4	2	2	2	3	3	3	4	5	6	7	9	9	9					
5	0	1	1	7														
6	3	5	8	9														
7	4	9																

68	63	42	27	30	36	28	32	79	27
22	23	24	25	44	65	43	25	74	51
36	42	28	31	28	25	45	12	57	51
12	32	49	38	42	27	31	50	38	21
16	24	69	47	23	22	43	27	49	28
23	19	46	30	43	49	12			

Crosstabulation

`pancreatitis.xls`

Smoking	*Disease* other	pancreatitis	Total
Ex-smokers	80	13	93
Never	56	2	58
Smokers	81	38	119
Total	217	53	**270**

In Excel use the following steps:

- Data → Pivot Table and PivotChart→ MS Office list + Pivot Table
- Set the range, including the headers of the data
- Select output and set layout by drag-and-dropping the names into the table

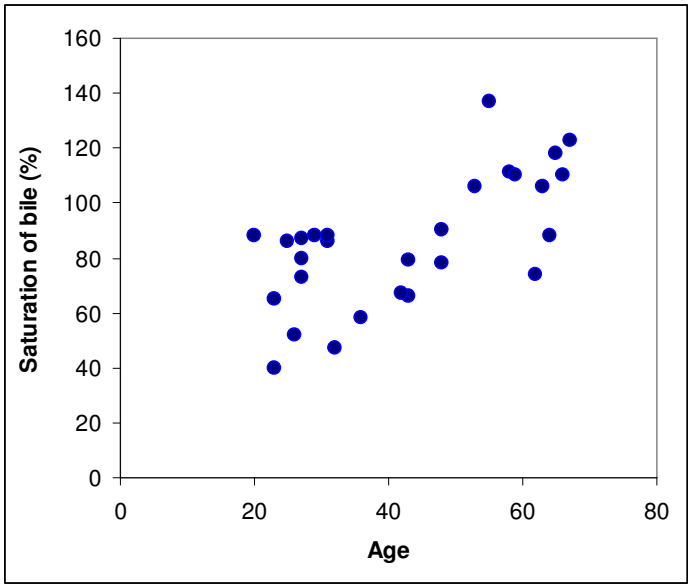
Scatter Plot

bile.xls

Date to the left, shows age and percentage saturation of bile for 28 male patients.

adapted from Chap T. Le, Introductory Biostatistics

Age	Saturation
23	40
31	86
58	111
25	86
63	106
43	66
67	123
48	90
26	52
64	88
55	137
31	88
20	88
23	65
43	79
27	87
59	110
53	106
66	110
48	78
27	80
32	47
62	74
36	58
29	88
27	73
65	118
42	67



In Excel use the following steps:

- Select the data region
- Use Chart Wizard → XY (Scatter)

NUMERICAL MEASURES

Population and Sample

Population parameter
A numerical value used as a summary measure for a population (e.g., the population mean μ, variance σ^2, standard deviation σ)

POPULATION

μ – mean
σ^2 – variance
N – number of elements (usually $N=\infty$)

SAMPLE

$m, \bar{x}$ – mean
s^2 – variance
n – number of elements

Sample statistic
A numerical value used as a summary measure for a sample (e.g., the sample mean x, the sample variance s^2, and the sample standard deviation s)

All existing laboratory
Mus musculus

mice.xls

790 mice from different strains
http://phenome.jax.org

ID	Strain	Sex	Starting age	Ending age	Starting weight	Ending weight	Weight change	Bleeding time	Ionized Ca in blood	Blood pH	Bone mineral density	Lean tissues weight	Fat weight
1	129S1/SvImJ	f	66	116	19.3	20.5	1.062	64	1.2	7.24	0.0605	14.5	4.4
2	129S1/SvImJ	f	66	116	19.1	20.8	1.089	78	1.15	7.27	0.0553	13.9	4.4
3	129S1/SvImJ	f	66	108	17.9	19.8	1.106	90	1.16	7.26	0.0546	13.8	2.9
368	129S1/SvImJ	f	72	114	18.3	21	1.148	65	1.26	7.22	0.0599	15.4	4.2
369	129S1/SvImJ	f	72	115	20.2	21.9	1.084	55	1.23	7.3	0.0623	15.6	4.3
370	129S1/SvImJ	f	72	116	18.8	22.1	1.176		1.21	7.28	0.0626	16.4	4.3
371	129S1/SvImJ	f	72	119	19.4	21.3	1.098	49	1.24	7.24	0.0632	16.6	5.4
372	129S1/SvImJ	f	72	122	18.3	20.1	1.098	73	1.17	7.19	0.0592	16	4.1
4	129S1/SvImJ	f	66	109	17.2	18.9	1.099	41	1.25	7.29	0.0513	14	3.2
5	129S1/SvImJ	f	66	112	19.7	21.3	1.081	129	1.14	7.22	0.0501	16.3	5.2
10	129S1/SvImJ	m	66	112	24.3	24.7	1.016	119	1.13	7.24	0.0533	17.6	6.8
364	129S1/SvImJ	m	72	114	25.3	27.2	1.075	64	1.25	7.27	0.0596	19.3	5.8
365	129S1/SvImJ	m	72	115	21.4	23.9	1.117	48	1.25	7.28	0.0563	17.4	5.7
366	129S1/SvImJ	m	72	118	24.5	26.3	1.073	59	1.25	7.26	0.0609	17.8	7.1
367	129S1/SvImJ	m	72	122	24	26	1.083	69	1.29	7.26	0.0584	19.2	4.6
6	129S1/SvImJ	m	66	116	21.6	23.3	1.079	78	1.15	7.27	0.0497	17.2	5.7
7	129S1/SvImJ	m	66	107	22.7	26.5	1.167	90	1.18	7.28	0.0493	18.7	7
8	129S1/SvImJ	m	66	108	25.4	27.4	1.079	35	1.24	7.26	0.0538	18.9	7.1
9	129S1/SvImJ	m	66	109	24.4	27.5	1.127	43	1.29	7.29	0.0539	19.5	7.1

Mean
A measure of central location computed by summing the data values and dividing by the number of observations.

Median
A measure of central location provided by the value in the middle when the data are arranged in ascending order.

Mode
A measure of location, defined as the value that occurs with greatest frequency.

$$\bar{x} = \frac{\sum x_i}{n}$$

$$\mu = \frac{\sum x_i}{N}$$

$$p = \frac{\sum (x_i = true)}{n}$$

Weight
12
16
19
22
23
23
24
32
36
42
63
68

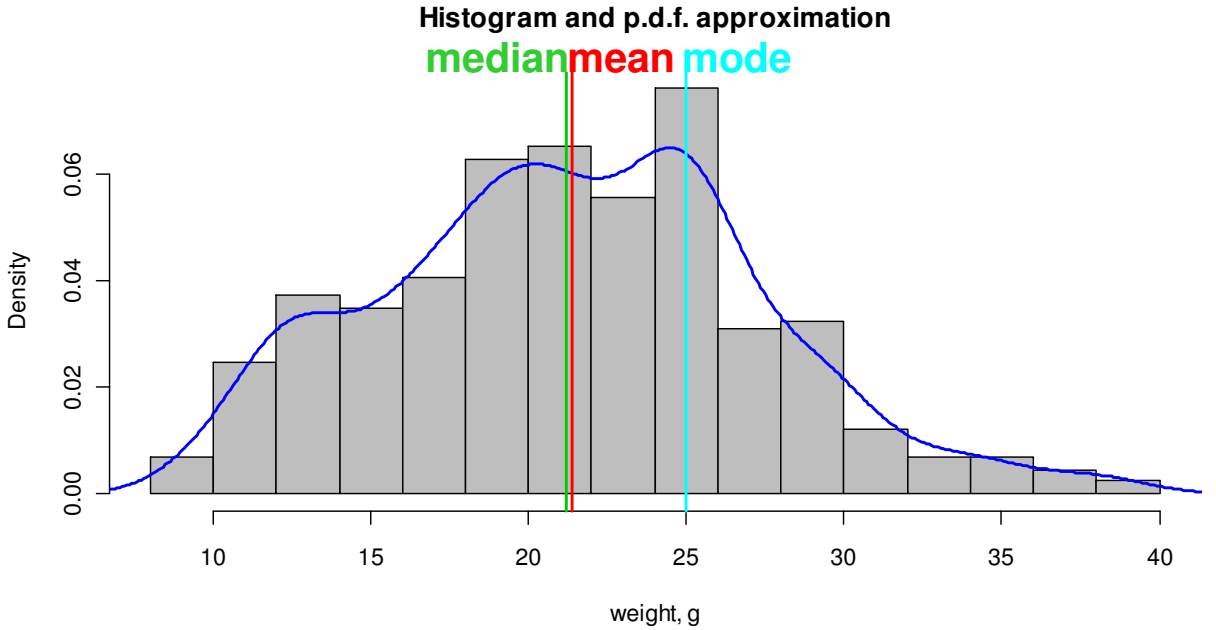
Mode = 23

Median = 23.5

Mean = 31.7

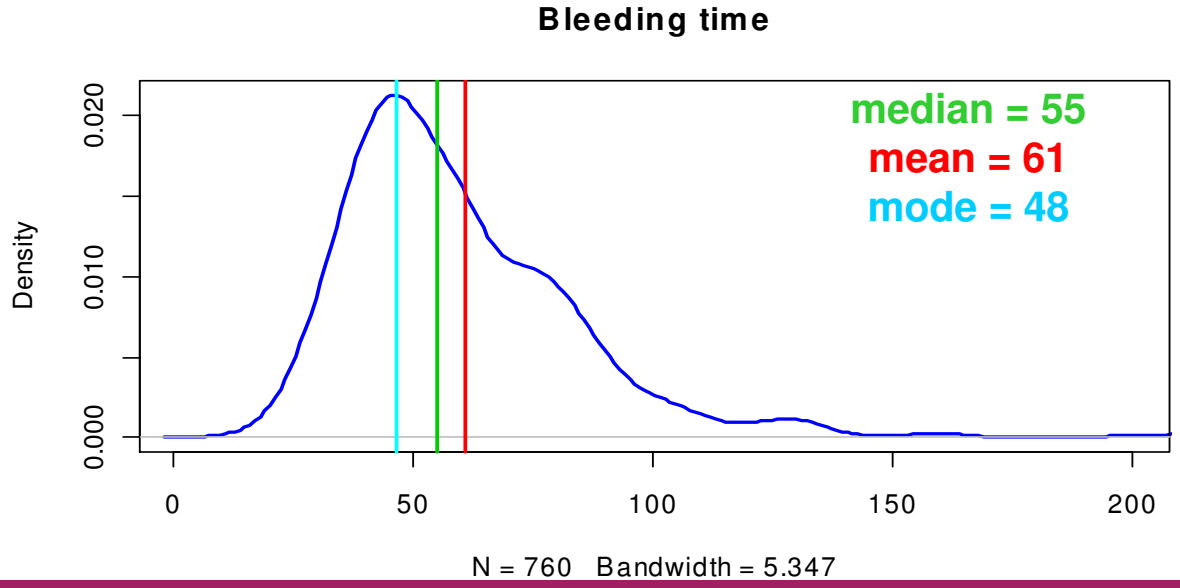
mice.xls

Female proportion
$p_f = 0.501$



In Excel use the following functions:

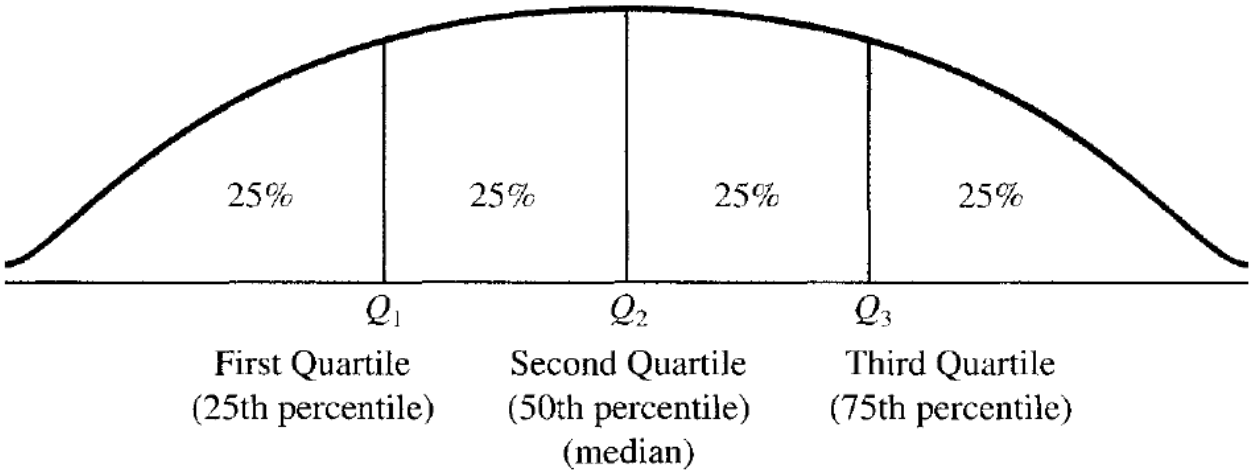
- = **AVERAGE(data)**
- = **MEDIAN(data)**
- = **MODE(data)**



Quantiles, Quartiles and Percentiles

Percentile
A value such that at least p% of the observations are less than or equal to this value, and at least (100-p)% of the observations are greater than or equal to this value. The 50-th percentile is the ***median***.

Quartiles
The 25th, 50th, and 75th percentiles, referred to as the first quartile, the second quartile (median), and third quartile, respectively.



In Excel use the following functions:

- `=PERCENTILE(data,p)`

Weight	12	16	19	22	23	23	24	32	36	42	63	68

$Q_1 = 21$ $Q_2 = 23.5$ $Q_3 = 39$

Measures of Variability

Interquartile range (IQR)
A measure of variability, defined to be the difference between the third and first quartiles.

$$IQR = Q_3 - Q_1$$

Variance
A measure of variability based on the squared deviations of the data values about the mean.

population

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

sample

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

Standard deviation
A measure of variability computed by taking the positive square root of the variance.

$$Sample\ standard\ deviation = s = \sqrt{s^2}$$

$$Population\ standard\ deviation = \sigma = \sqrt{\sigma^2}$$

Weight	12	16	19	22	23	23	24	32	36	42	63	68

IQR = 18 *Variance* = 320.2 *St. dev.* = 17.9

In Excel use the following functions:

- **`=VAR(data), =STDEV(data)`**

Coefficient of variation
A measure of relative variability computed by dividing the standard deviation by the mean.

Weight	12	16	19	22	23	23	24	32	36	42	63	68

$$\left(\frac{Standard\ deviation}{Mean}\times 100\right)\%$$

$CV = 57\%$

Median absolute deviation (MAD)
MAD is a robust measure of the variability of a univariate sample of quantitative data.

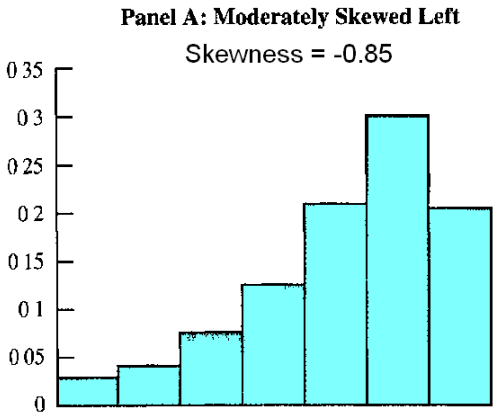
$$MAD = median\left(\left|x_i - median(x)\right|\right)$$

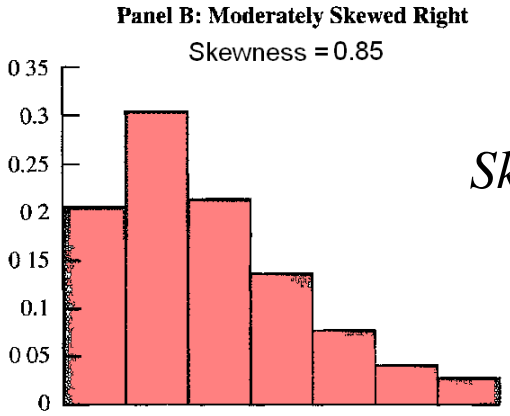
Set 1	Set 2
23	23
12	12
22	22
12	12
21	21
18	81
22	22
20	20
12	12
19	19
14	14
13	13
17	17

	Set 1	Set 2
Mean	17.3	22.2
Median	18	19
St.dev.	4.23	18.18
MAD	5.93	5.93

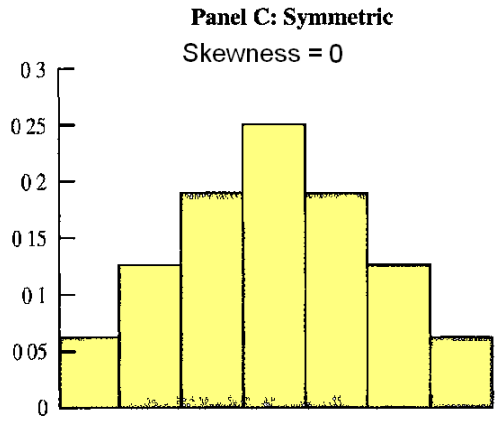
Skewness

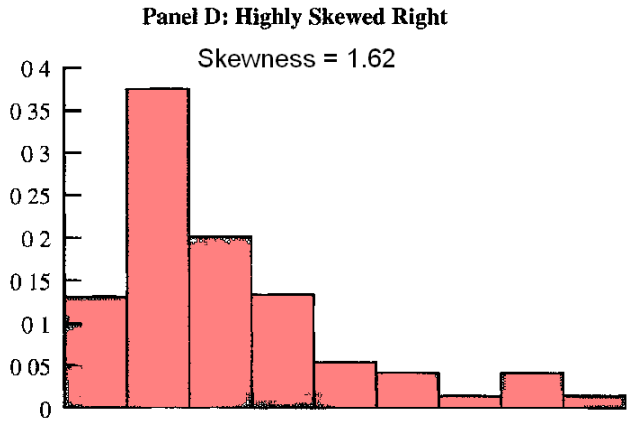
A measure of the shape of a data distribution. Data skewed to the left result in negative skewness; a symmetric data distribution results in zero skewness; and data skewed to the right result in positive skewness.





$$Skewness = \frac{n}{(n-1)(n-2)} \sum_i \left(\frac{x_i - \bar{x}}{s} \right)^3$$





adapted from Anderson et al Statistics for Business and Economics

z-score

z-score
A value computed by dividing the deviation about the mean (x_i - $\bar{x}$) by the standard deviation s. A ***z-score*** is referred to as a standardized value and denotes the number of standard deviations x_i is from the mean.

$$z_i = \frac{x_i - \bar{x}}{s}$$

Weight	z-score
12	-1.10
16	-0.88
19	-0.71
22	-0.54
23	-0.48
23	-0.48
24	-0.43
32	0.02
36	0.24
42	0.58
63	1.75
68	2.03

Chebyshev's theorem
For **any data set**, at least **$(1 - 1/z^2)$** of the data values must be within ***z*** standard deviations from the mean, where z – any value > 1.

For ANY distribution:

- At least **75 %** of the values are within **z = 2** standard deviations from the mean
- At least **89 %** of the values are within **z = 3** standard deviations from the mean
- At least **94 %** of the values are within **z = 4** standard deviations from the mean
- At least **96%** of the values are within **z = 5** standard deviations from the mean

Detection of Outliers

For bell-shaped distributions:

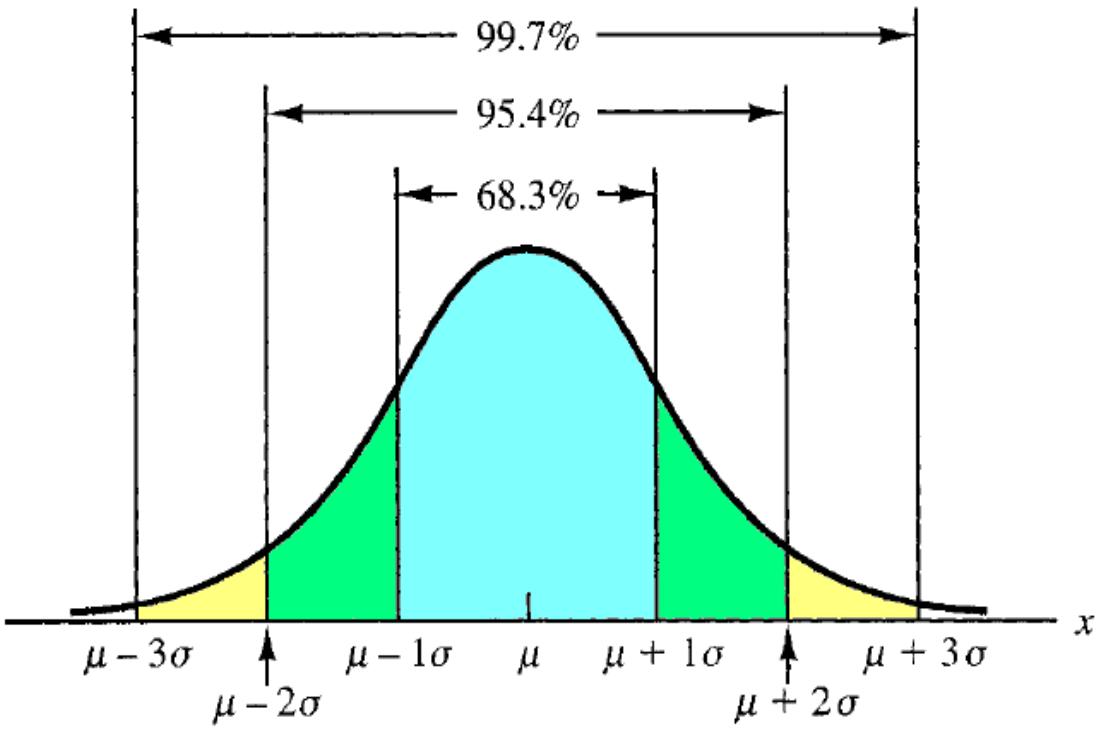
- Approximately 68 % of the values are within 1 st.dev. from mean
- Approximately 95 % of the values are within 2 st.dev. from mean
- Almost all data points are inside 3 st.dev. from mean

Outlier
An unusually small or unusually large data value.

For bell-shaped distributions data points with |z|>3 can be considered as outliers.

Weight	z-score
23	0.04
12	-0.53
22	-0.01
12	-0.53
21	-0.06
81	**3.10**
22	-0.01
20	-0.11
12	-0.53
19	-0.17
14	-0.43
13	-0.48
17	-0.27

Example: Gaussian distribution



Five-number summary
An exploratory data analysis technique that uses five numbers to summarize the data: smallest value, first quartile, median, third quartile, and largest value

children.xls

Min. :	12
Q_1 :	25
Median:	32
Q_3 :	46
Max. :	79

In Excel use:

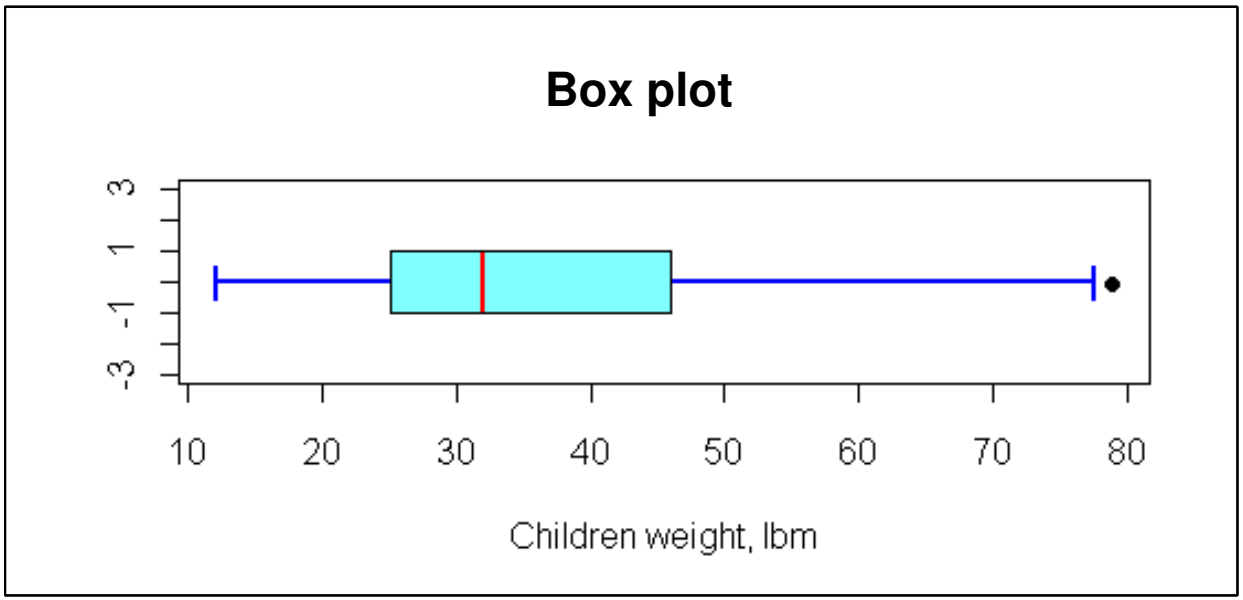
- Tool → Data Analysis → Descriptive Statistics

Box plot
A graphical summary of data based on a five-number summary

In Excel use (indirect):

- Chart Wizard → Stock → Open-high-low-close

open	Q3
high	Q3+1.5*IQR
low	Q1-1.5*IQR
close	Q1



Example: Mice Weight

Example
Build a box plot for weights of male and female mice

mice.xls

1. Build 5 number summaries for males and females

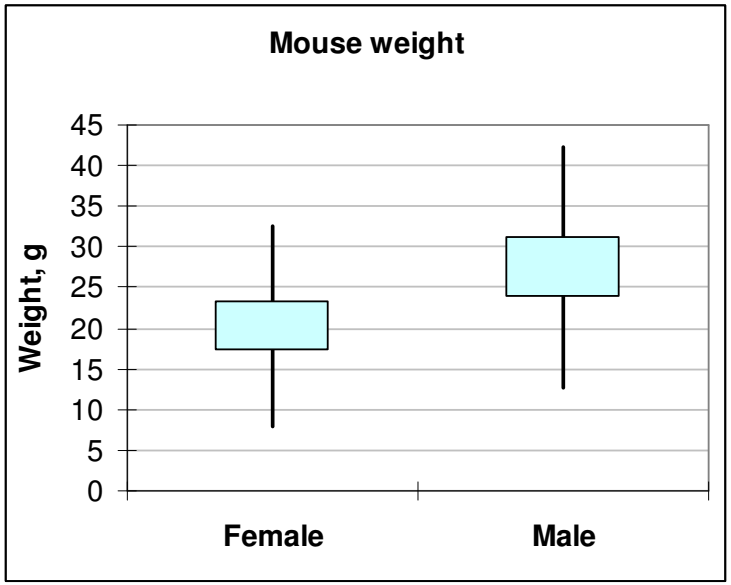
	Female	Male
Min	10.0	12.0
Q1	17.2	23.8
Q2	20.7	27.1
Q3	23.3	31.2
Max	41.5	49.6

2. Combine the numbers into the following order

open	Q3
high	Q3+min(1.5*(Q3-Q1),Max)
low	Q1-max(1.5*(Q3-Q1),Min)
close	Q1

In Excel use:

- Chart Wizard → Stock → Open-high-low-close
- Put “series-in-rows”
- Adjust colors, etc



Covariance
A measure of linear association between two variables. Positive values indicate a positive relationship; negative values indicate a negative relationship.

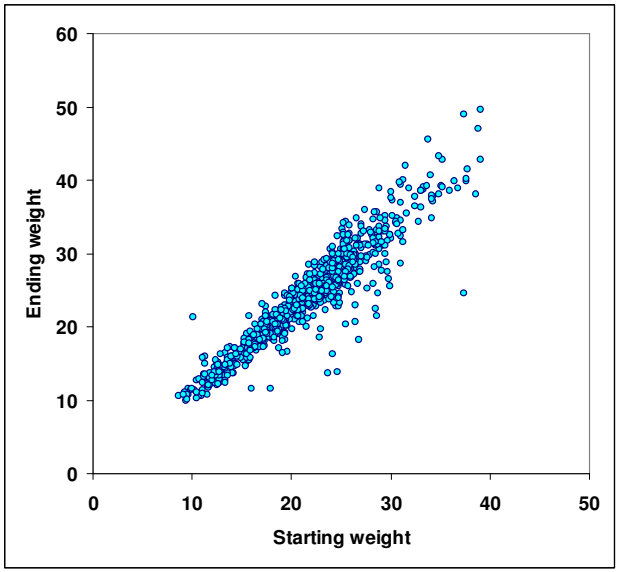
population

$$\sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}$$

sample

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

mice.xls



Ending weight vs. Starting weight

In Excel use function:

- `=COVAR(data)`

s_{xy} = 39.8

hard to interpret

Measure of Association between 2 Variables

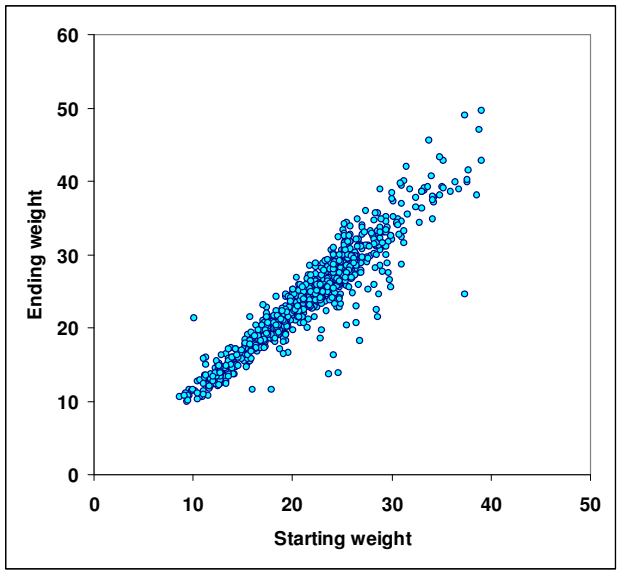
Correlation (Pearson product moment correlation coefficient)
A measure of linear association between two variables that takes on values between -1 and +1. Values near +1 indicate a strong positive linear relationship, values near -1 indicate a strong negative linear relationship; and values near zero indicate the lack of a linear relationship.

population

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y N}$$

sample

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y (n-1)}$$

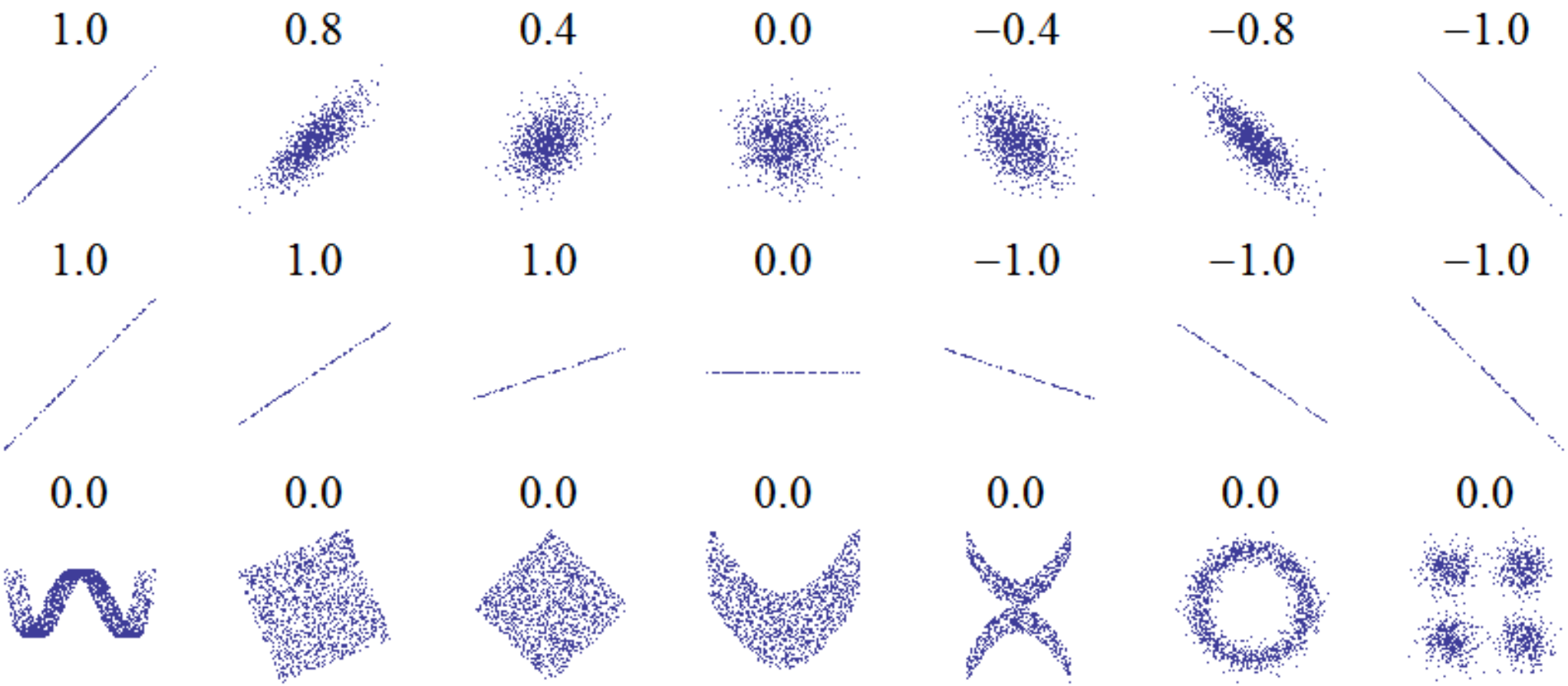


In Excel use function:

- `=CORREL(data)`

$r_{xy} = 0.94$

Correlation Coefficient



Wikipedia

? If we have only 2 data points in *x* and y datasets, which values would you expect for correlation b/w *x* and y ?

Weighted mean
The mean obtained by assigning each observation a weight that reflects its importance

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i}$$

As an example of the need of weighted mean, consider the following sample of five purchases of a raw material over several months

Purchase	Cost per Pound ($)	Number of Pounds
1	3.00	1200
2	3.40	500
3	2.80	2750
4	2.90	1000
5	3.25	800

Note that the cost per pound varies from $2.80 to $3.40, and quantity purchased has varied from 500 to 2750. Suppose that manager asked for information about the mean cost per pound of the raw material.

$$\bar{x} = \frac{1200(3.00) + 500(3.40) + 2750(2.80) + 1000(2.90) + 800(3.25)}{1200 + 500 + 2750 + 1000 + 800} = \frac{18{,}500}{6250} = 2.96$$

If we would use a simple mean of the cost p.p.:

$(3.00 + 3.40 + 2.80 + 2.90 + 3.25)/5 = 15.35/5 = \$3.07$

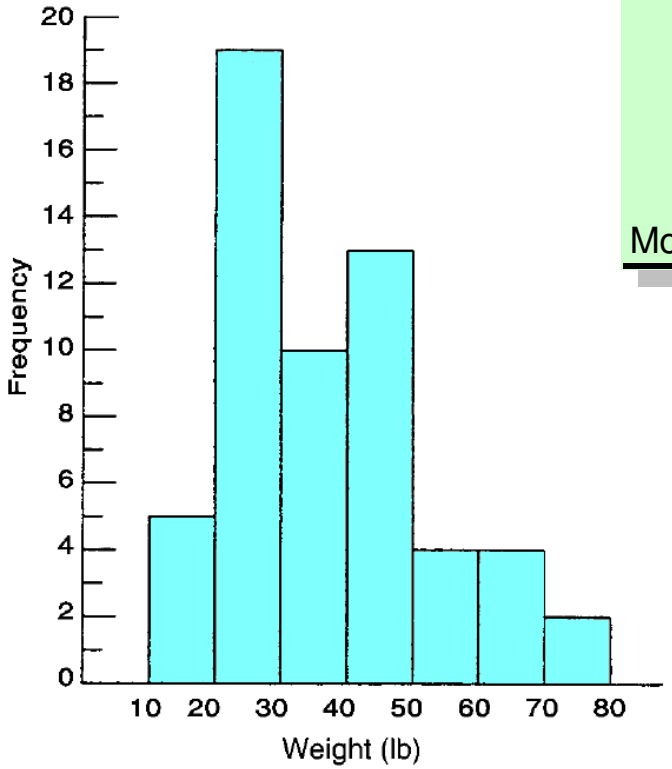
we overestimate the average cost!

Anderson et al Statistics for Business and Economics

Grouped Mean

Grouped data
Data available in class intervals as summarized by a frequency distribution. Individual values of the original data are not available.

~~children.xls~~ not available



Bin	Frequency
20	5
30	21
40	8
50	14
60	3
70	4
80	2
More	0

Mean for grouped data

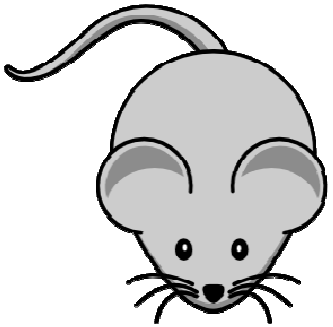
$$\bar{x} = \frac{\sum_{i}^{k} f_i M_i}{n}$$

M_i = the midpoint for class i
f_i = the frequency for class i
n = the sample size

Variance for grouped data

$$s^2 = \frac{\sum_{i}^{k} f_i (M_i - \bar{x})^2}{n-1}$$

Thank you for your attention



to be continued…