CENTRE DE RECHERCHE PUBLIC

Microarray Center

UNIVERSITÉ DU LUXEMBOURG

STATISTICAL DATA ANALYSIS IN EXCEL

Part 1

Introduction to Statistics
Descriptive Statistics

dr. Petr Nazarov

petr.nazarov@crp-sante.lu

31-10-2011

The course

- Reminds statistical basics
- Gives the methodological tools for the research
- Provides practical skill for fast data analysis

Organization

- 5 topics, 8-9 hours in total = 1 days

PLEASE: ask questions. Understanding is extremely important for later parts

http://edu.sablab.net/sdae2011

Look for the data: **http://edu.sablab.net/data/xls**

COURSE OUTLINE

1. Introduction
- Descriptive statistics
- Exploratory analysis
- Discrete probability distribution
- Continues probability distribution

2. Interval Estimations
- Sampling distribution
- Interval estimation for mean
- Interval estimation for proportion
- Sample size selection

3. Testing Hypotheses about Means
- Hypotheses
- Comparing of a mean and a constant
- Unpaired t-test
- Paired t-test

4. ANOVA
- 1-way ANOVA
- 2-way ANOVA

5. Linear Regression
- Simple linear regression
- Multiple linear regression

Look for the data: `http://edu.sablab.net/data/xls`

- **Introduction**
  - descriptive statistics
  - numerical measures

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%**

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

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.

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

`pancreatitis.xls`

Frequency distribution:

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

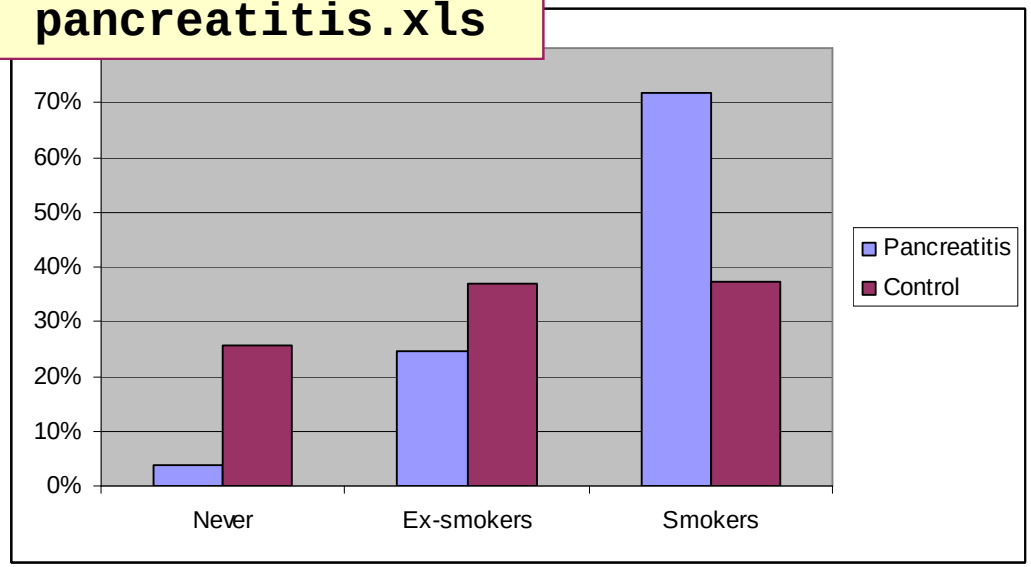
Relative frequency distribution:

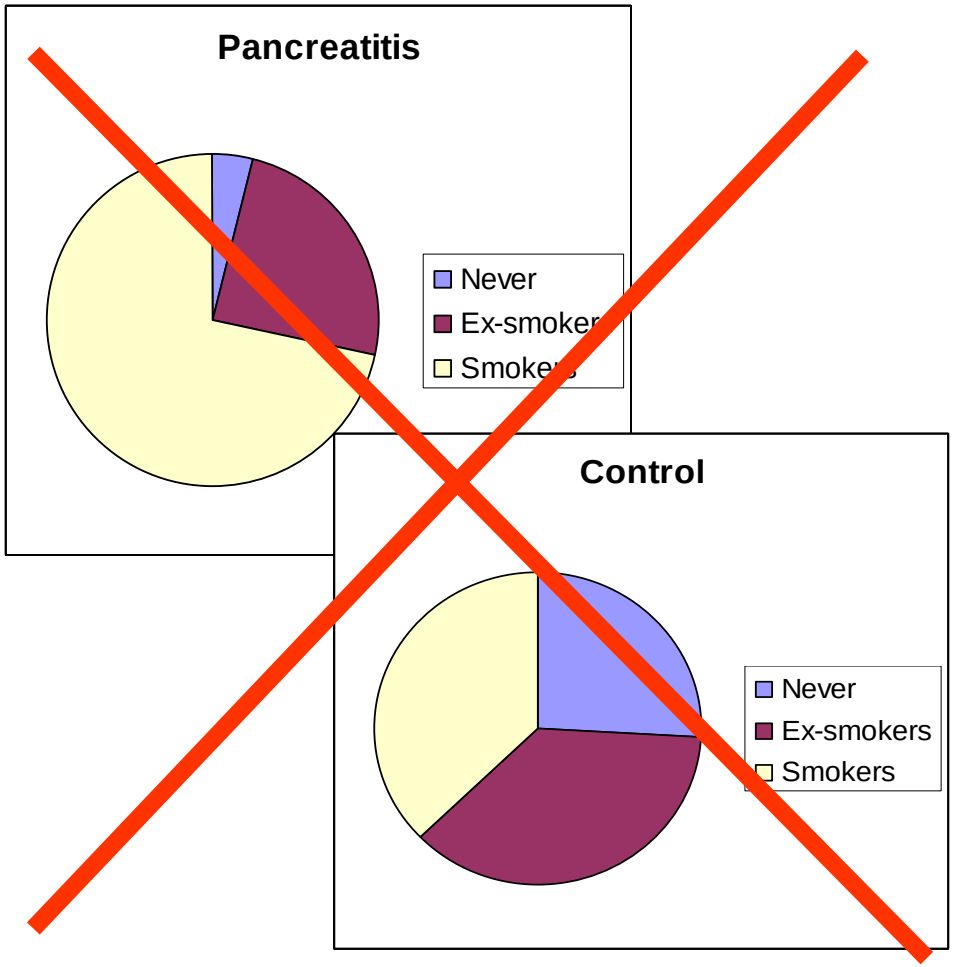
Smoking	Cases	Controls
Never	0.038	0.258
Ex-smokers	0.245	0.369
Smokers	0.717	0.373
Total	1	1

TABULAR AND GRAPHICAL PRESENTATION

Bar and Pie Charts

pancreatitis.xls



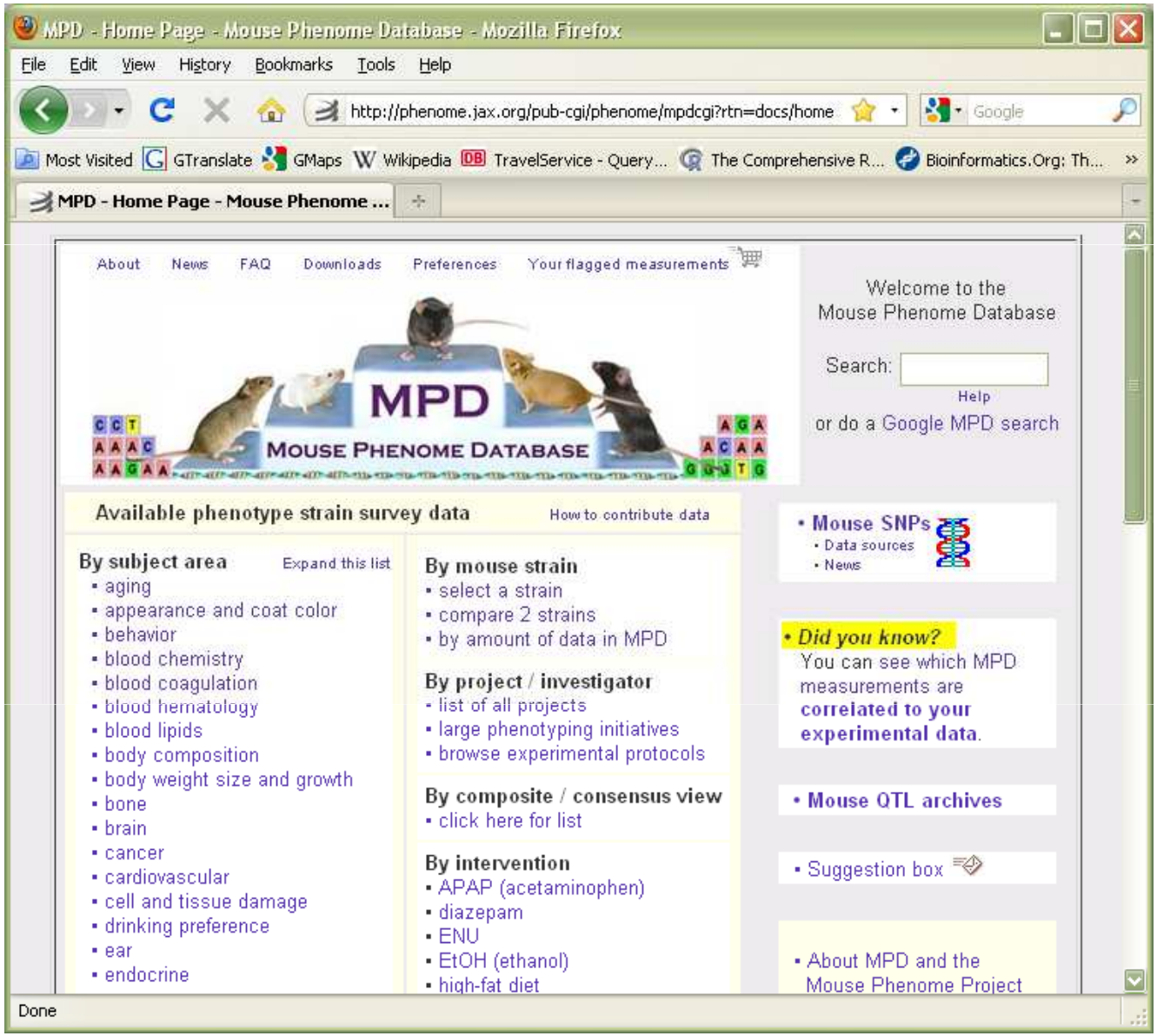


In MS Excel use the following steps:

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

TABULAR AND GRAPHICAL PRESENTATION

Mice Data Series



Tordoff MG, Bachmanov AA

Survey of calcium & sodium intake and metabolism with bone and body composition data

Project symbol: **Tordoff3**

Accession number: **MPD:103**

`mice.xls`

790 mice from different strains

http://phenome.jax.org

parameter
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

Histogram

The following are weights in grams for 970 mice:

mice.xls

20.5	23.2	24.6	23.5	26	25.9	23.9	22.8	19.9	…
20.8	22.4	26	23.8	26.5	26	22.8	22.9	20.9	…
19.8	22.7	31	22.7	26.3	27.1	18.4	21	18.8	…
21	21.4	25.7	19.7	27	26.2	21.8	22.2	19.2	…
21.9	22.6	23.7	26.2	26	27.5	25	20.9	20.6	…
22.1	20	21.1	24.1	28.8	30.2	20.1	24.2	25.8	…
21.3	21.8	23.7	23.5	28	27.6	21.6	21	21.3	…
20.1	20.8	24.5	23.8	29.5	21.4	21.5	24	21.1	…
18.9	19.5	32.3	28	27.1	28.2	22.9	19.9	20.4	…
21.3	20.6	22.8	25.8	24.1	23.5	24.2	22	20.3	…

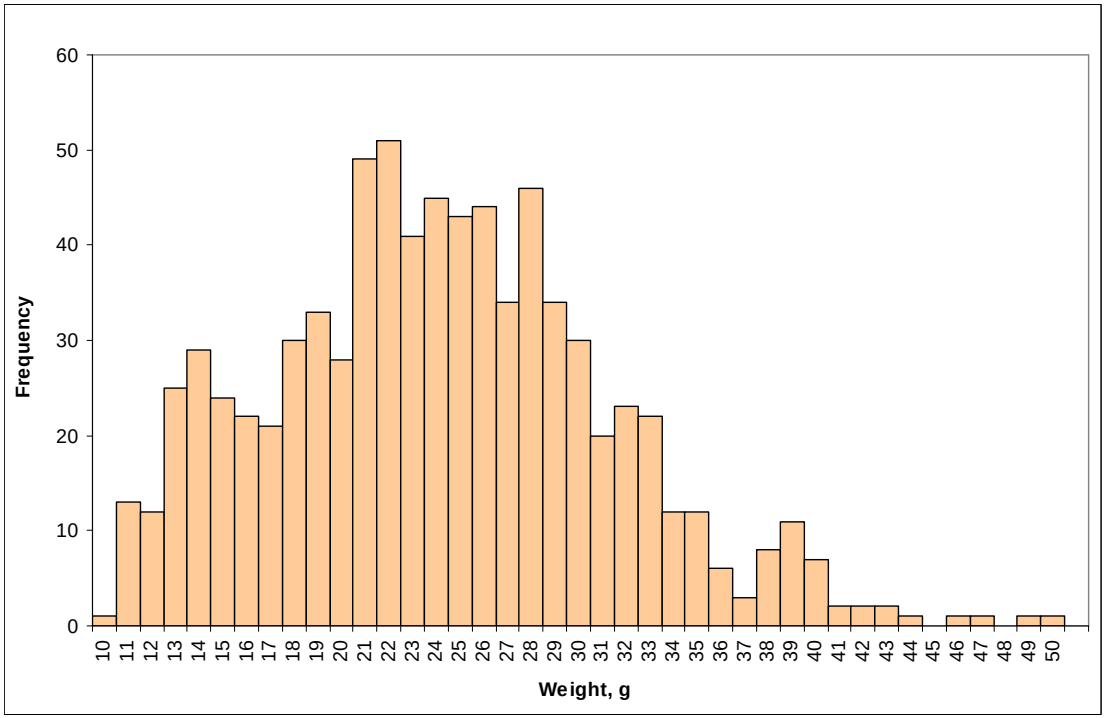
Sorted weights show that the values are in the 10 – 49.6 grams.
Let us divide the weight into the “bins”

bins

Weight,g	*Frequency*
>=10	1
10-20	237
20-30	417
30-40	124
40-50	11
More	0

Now, let us use bin-size = 1 gram

Bin	*Frequency*
10	1
11	13
12	12
13	25
14	29
...	...
46	1
47	1
48	0
49	1
50	1
More	0



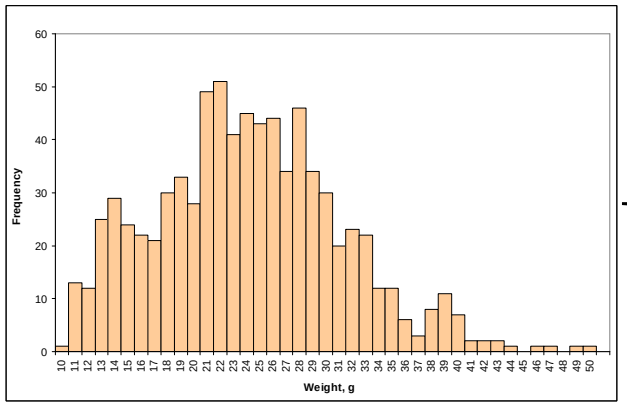
In Excel use the following steps:

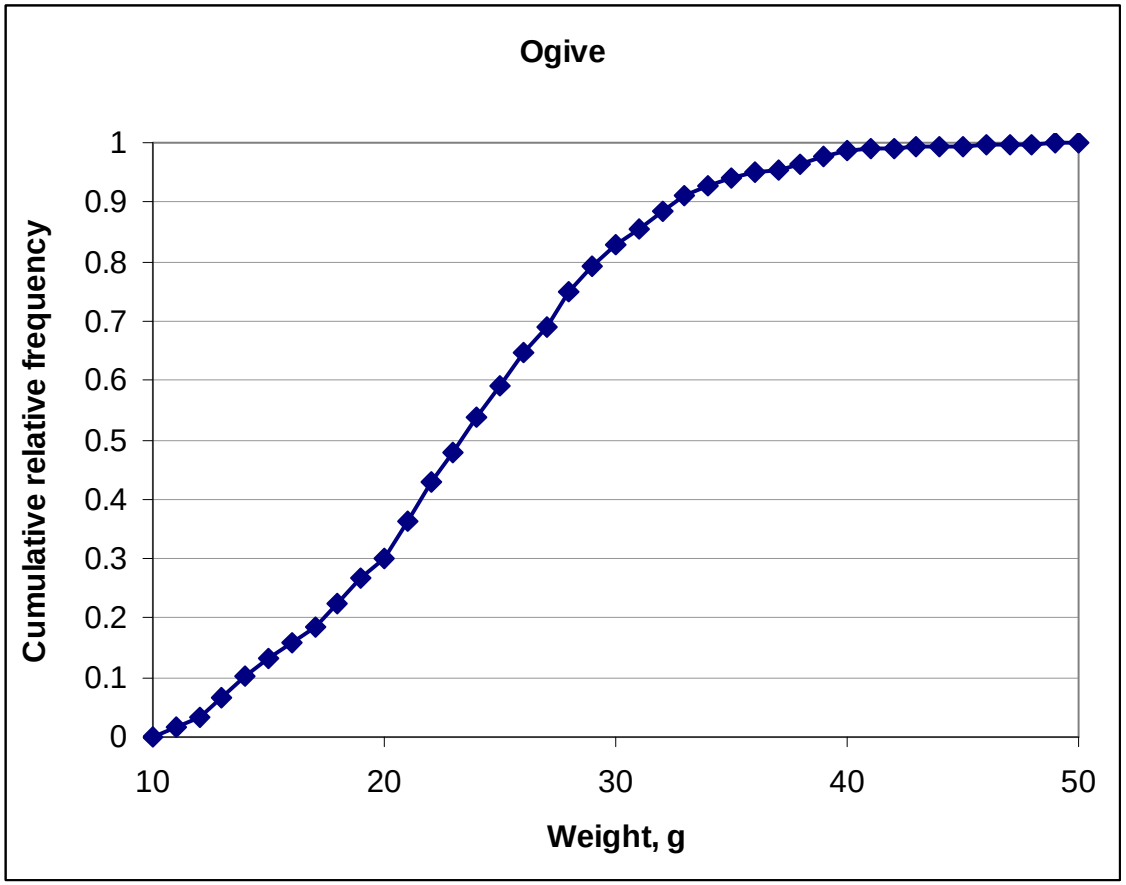
- Specify the column of bins (interval) upper-limits
- Tools → Data Analysis → Histrogram → select the input data, bins, and output (Analysis ToolPak should be installed)
- use Chart Wizard → Columns to visualize the results

TABULAR AND GRAPHICAL PRESENTATION

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.

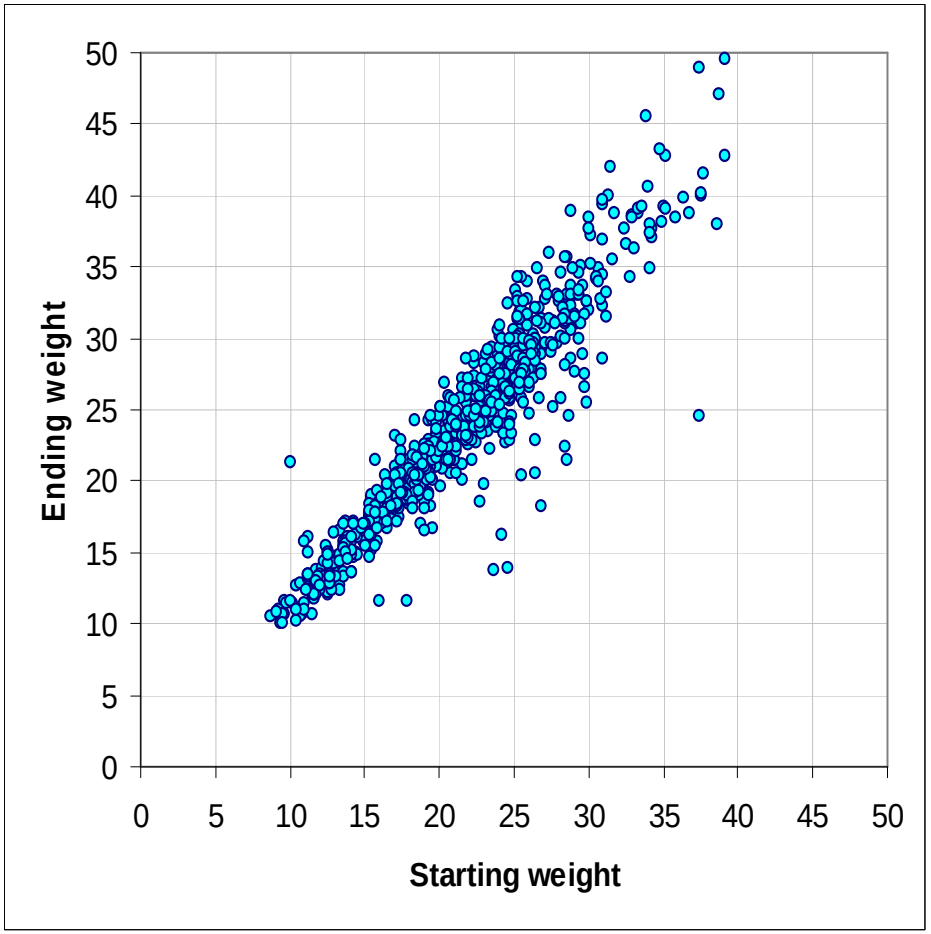




Scatter Plot

mice.xls

Let us look on mutual dependency of the Starting and Ending weights.



In Excel use the following steps:

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

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**

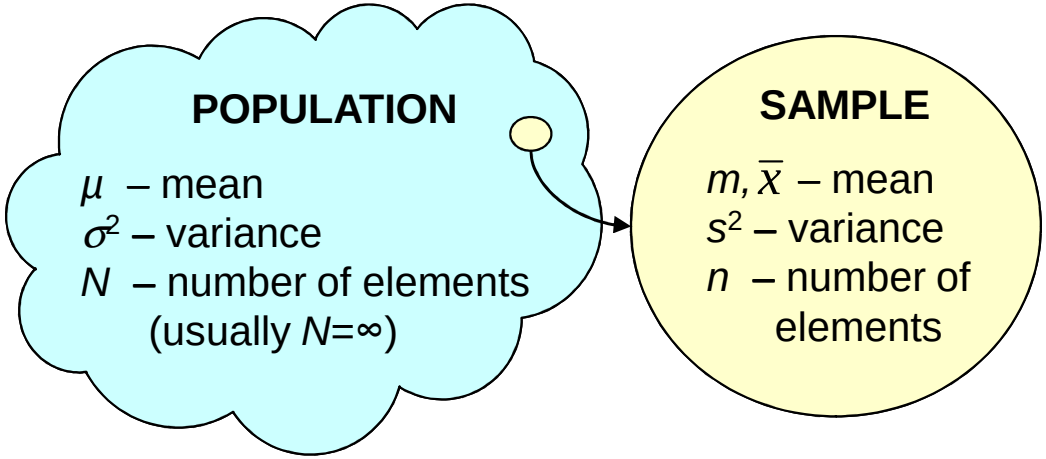
In Excel use the following steps:

- 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

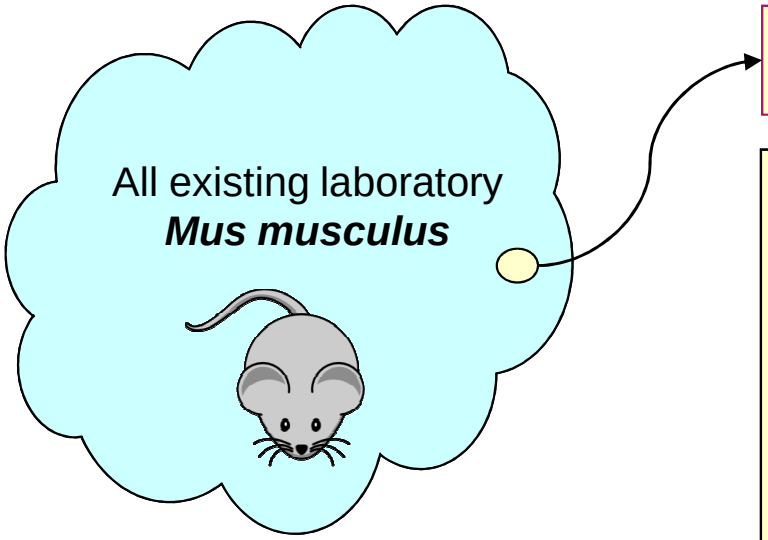
NUMERICAL MEASURES

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)



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



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

Measures of Location

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

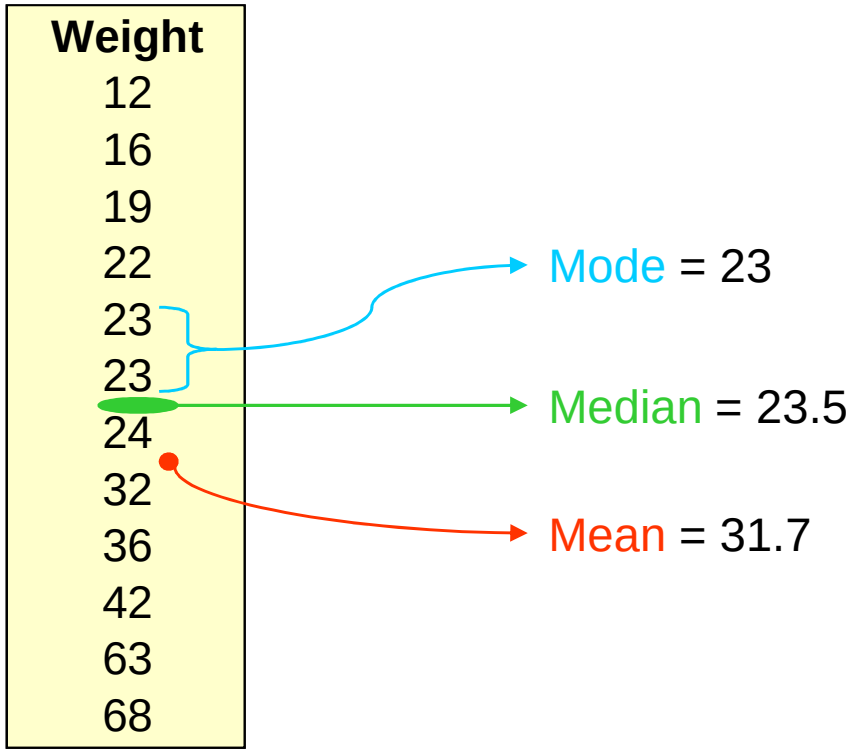
Median
A measure of central location provided by the value in the middle when the data are arranged in ascending order.

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

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

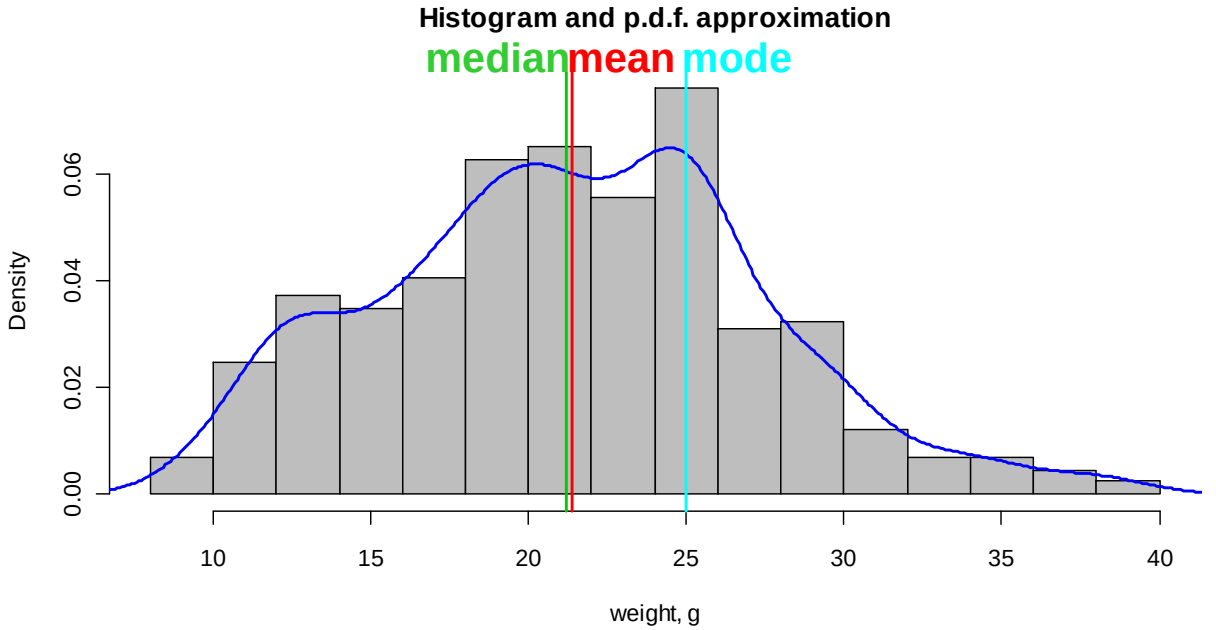
$$\mu = \frac{\sum x_i}{N}$$

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



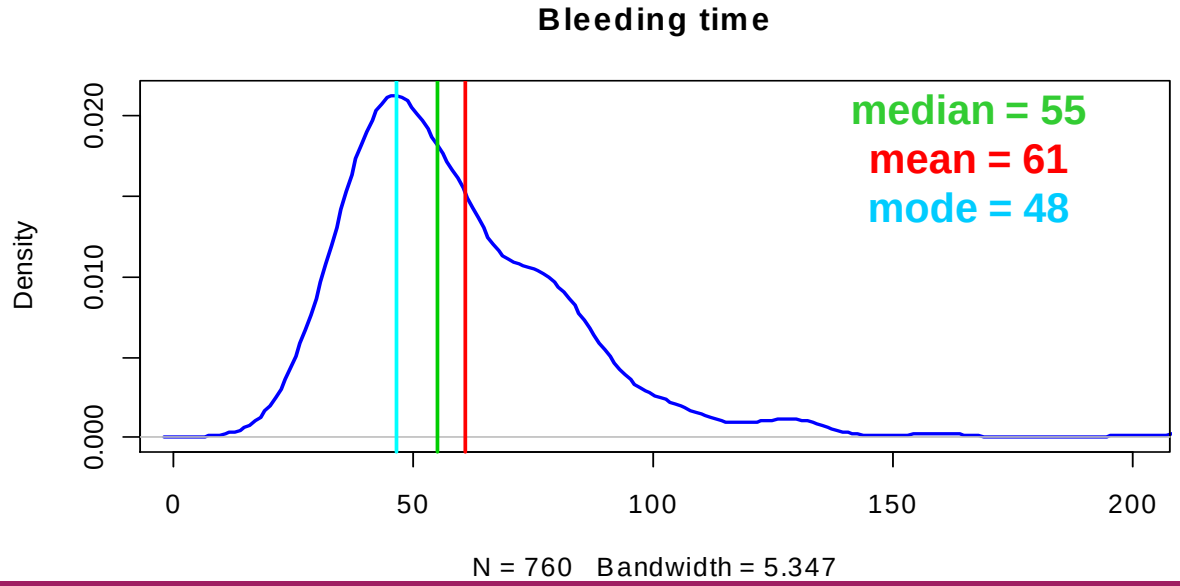
mice.xls

Female proportion
$p_f = 0.501$



In Excel use the following functions:

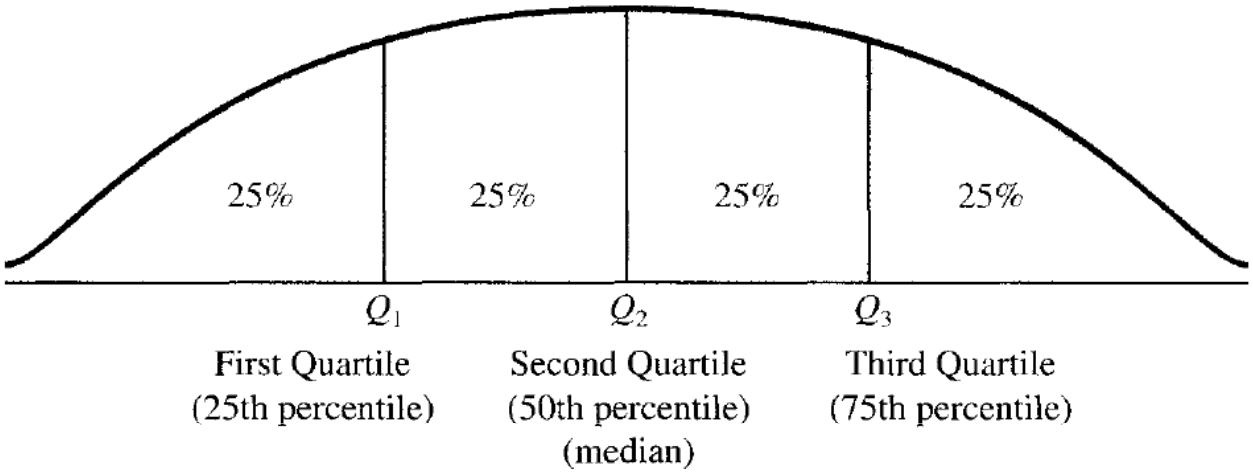
- **= AVERAGE(data)**
- **= MEDIAN(data)**
- **= MODE(data)**



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.

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

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

$$IQR = Q_3 - Q_1$$

population

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

sample

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

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

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

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

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

$CV = 57\%$

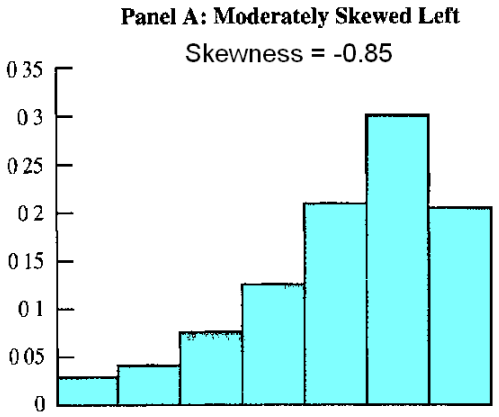
Median absolute deviation (MAD)
MAD is a robust measure of the variability of a univariate sample of quantitative data.

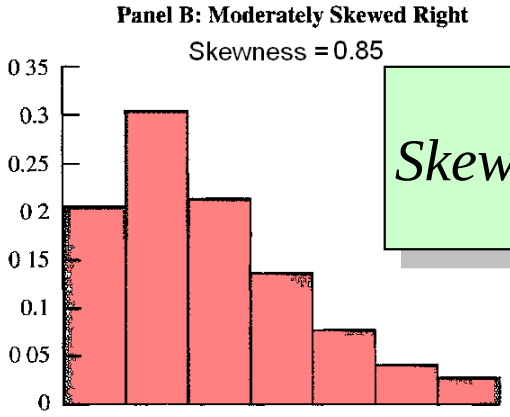
$$MAD = median\left(\left|x_i - median(x)\right|\right)$$

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

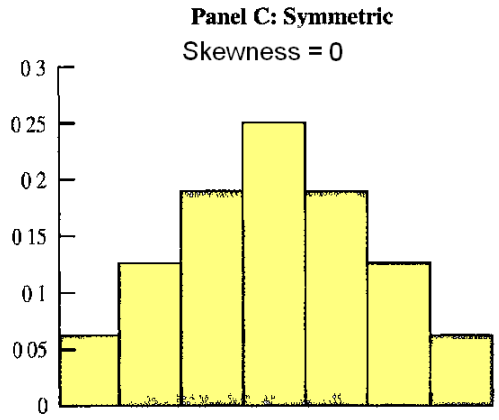
	Set 1	Set 2
Mean	17.3	22.2
Median	18	19
St.dev.	4.23	**18.18**
MAD	5.93	5.93

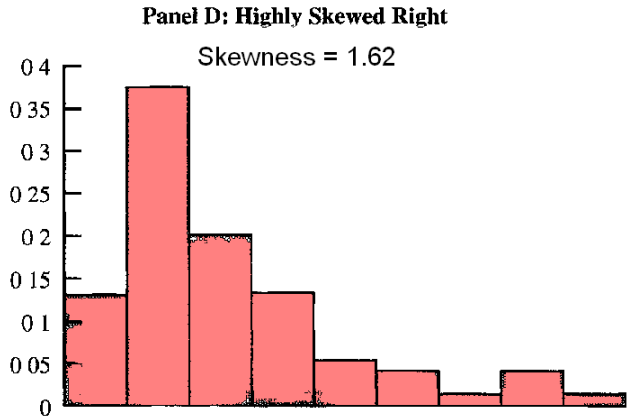
Skewness
A measure of the shape of a data distribution. Data skewed to the left result in negative skewness; a symmetric data distribution results in zero skewness; and data skewed to the right result in positive skewness.





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





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 - 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.

$$z_i = \frac{x_i - m}{s}$$

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:

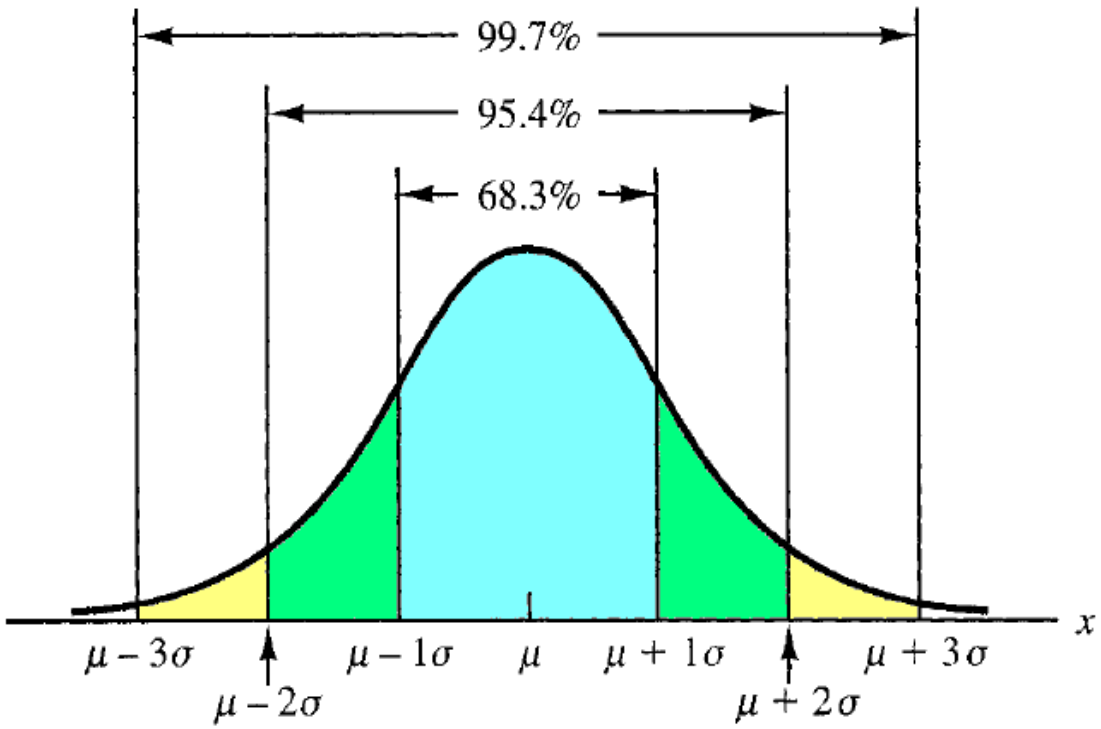
- Approximately 68 % of the values are within 1 st.dev. from mean
- Approximately 95 % of the values are within 2 st.dev. from mean
- Almost all data points are inside 3 st.dev. from mean

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:

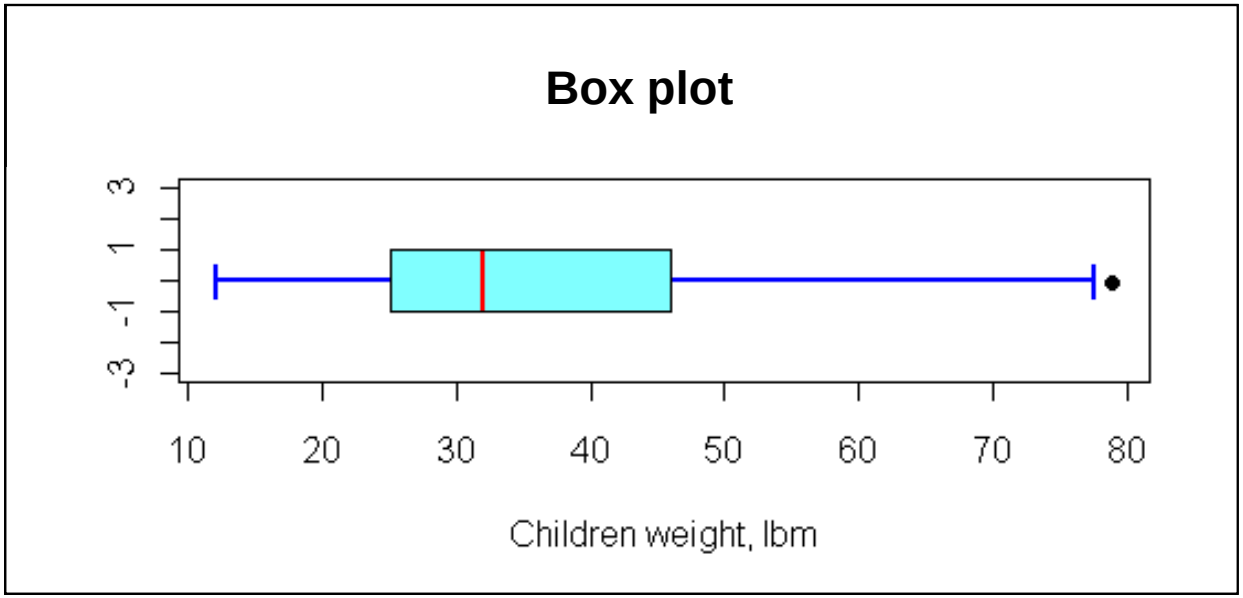
- Tool → Data Analysis → Descriptive Statistics

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

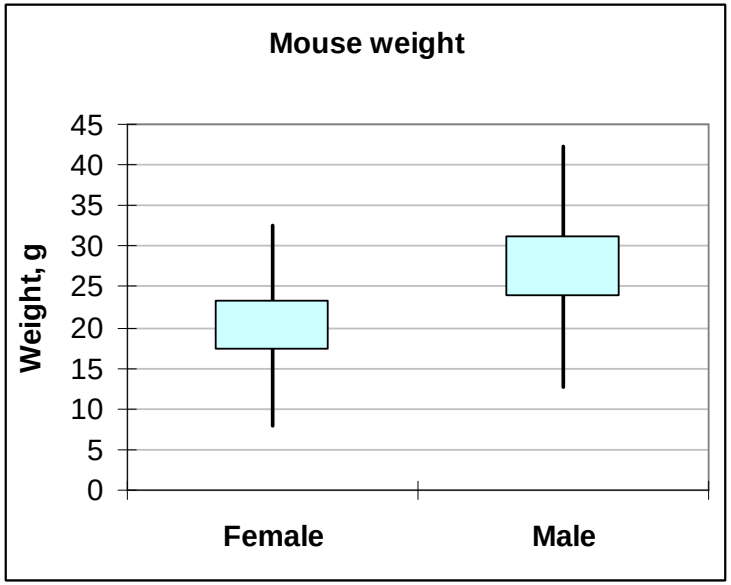
	Female	Male
Min	10.0	12.0
Q1	17.2	23.8
Q2	20.7	27.1
Q3	23.3	31.2
Max	41.5	49.6

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



Measure of Association between 2 Variables

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

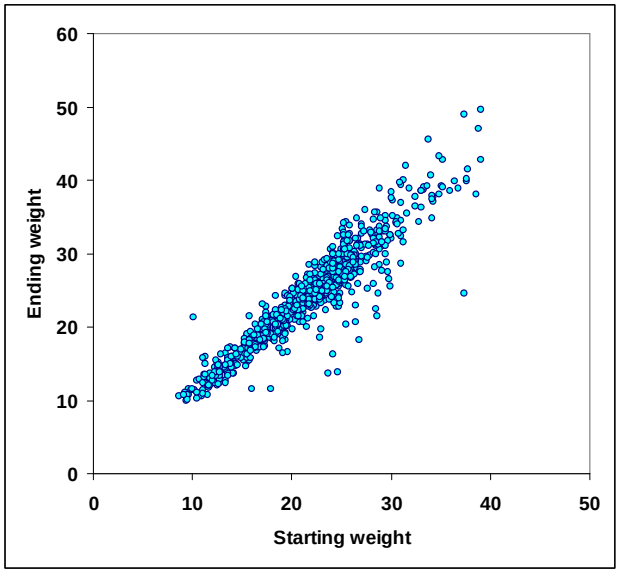
population

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

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

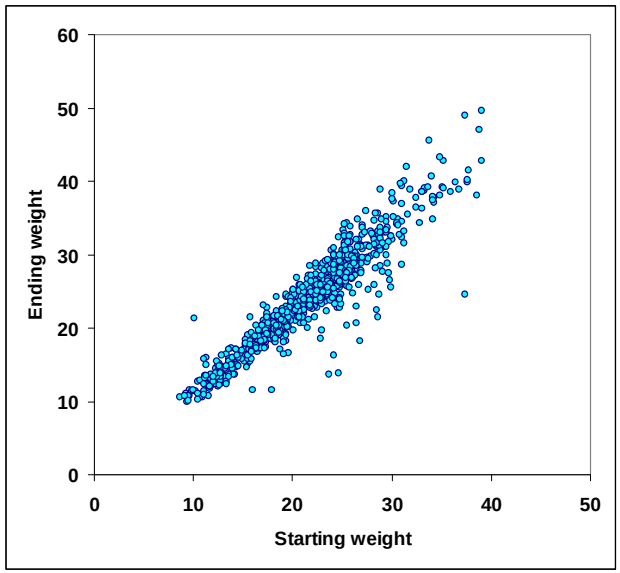
Correlation (Pearson product moment correlation coefficient)
A measure of linear association between two variables that takes on values between -1 and +1. Values near +1 indicate a strong positive linear relationship, values near -1 indicate a strong negative linear relationship; and values near zero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

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



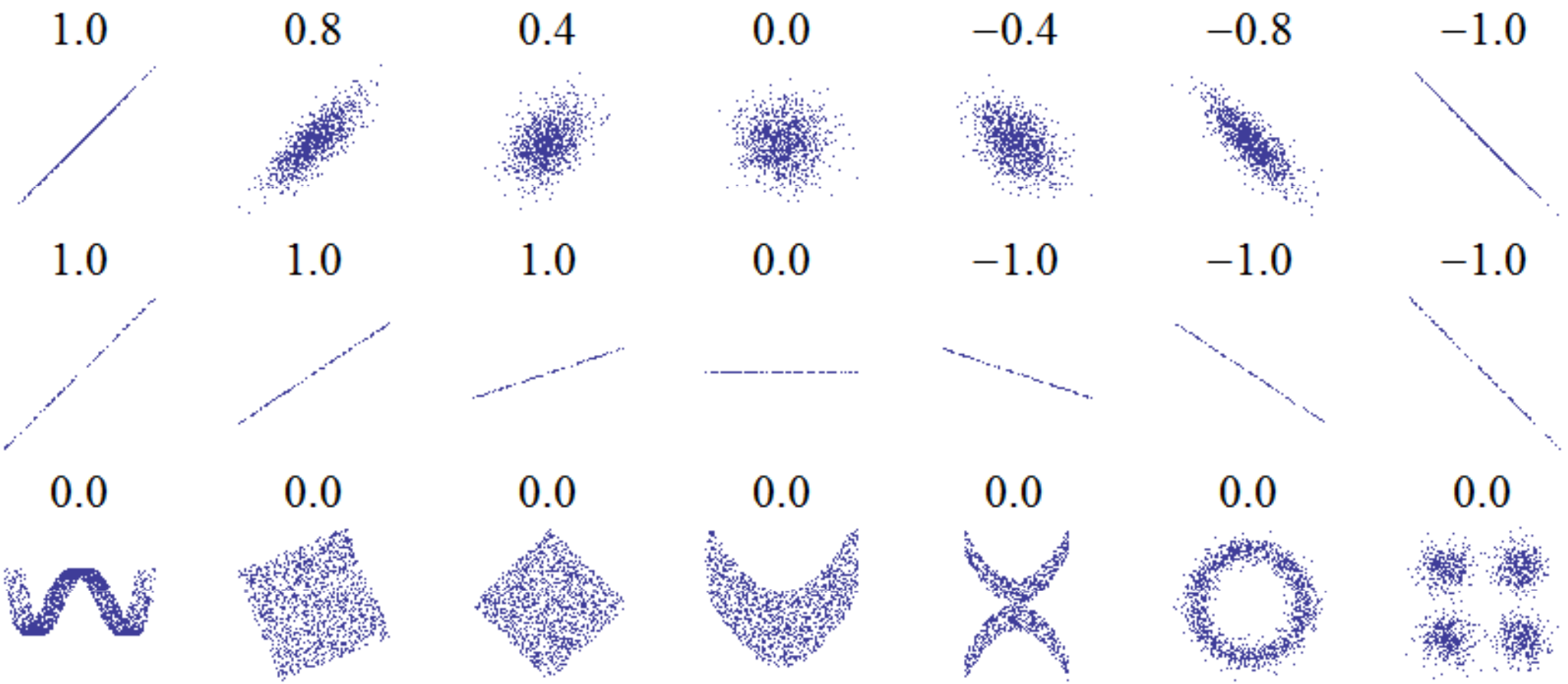
In Excel use function:

- `=CORREL(data)`

r_{xy} = 0.94

`mice.xls`

Correlation Coefficient



Wikipedia

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

- **Discrete and continuous probability distributions**
  - discrete probability distribution
  - continuous probability distribution
  - normal probability distribution

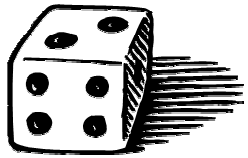
Random variable
A numerical description of the outcome of an experiment.

A random variable is always a numerical measure.

Discrete random variable
A random variable that may assume either a finite number of values or an infinite sequence of values.

Continuous random variable
A random variable that may assume any numerical value in an interval or collection of intervals.

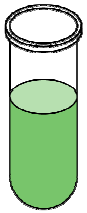
Roll a die



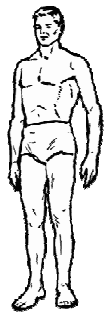
Number of calls to a reception per hour

Time between calls to a reception

Volume of a sample in a tube



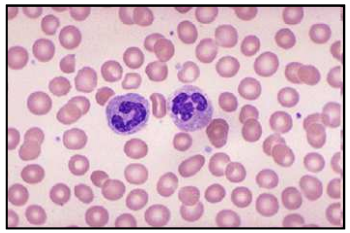
Weight, height, blood pressure, etc



Probability distribution
A description of how the probabilities are distributed over the values of the random variable.

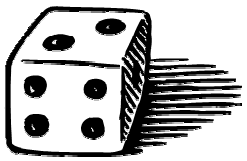
Probability function
A function, denoted by *f(x),* that provides the probability that *x* assumes a particular value for a discrete random variable.

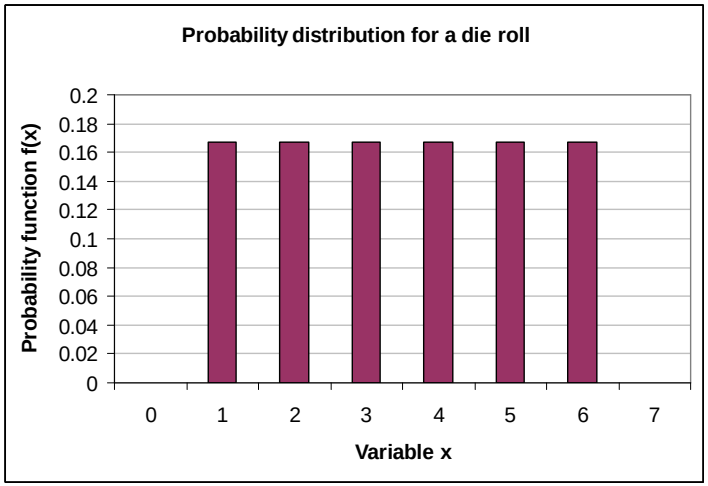
Number of cells under microscope
Random variable X:
x = 0
x = 1
x = 2
x = 3
…

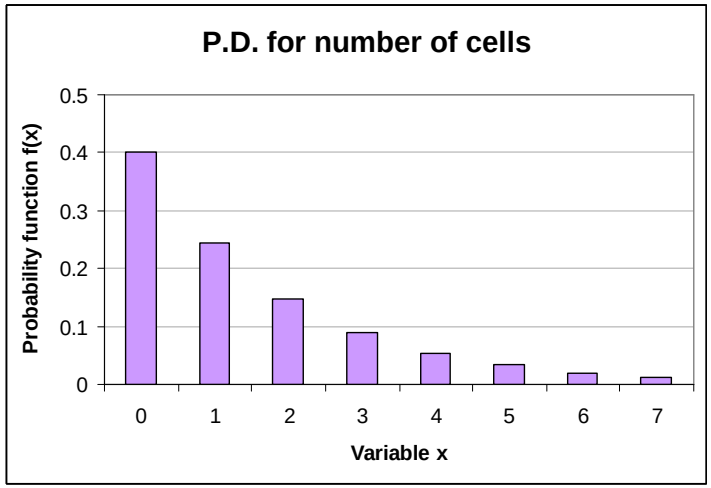


$$f(x) \geq 0$$
$$\sum f(x) = 1$$

Roll a die
Random variable X:
x = 1
x = 2
x = 3
x = 4
x = 5
x = 6

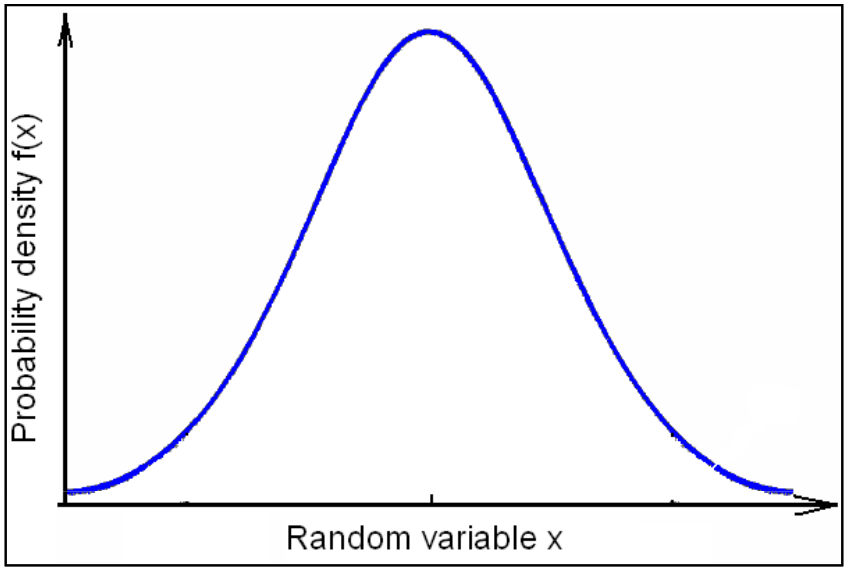


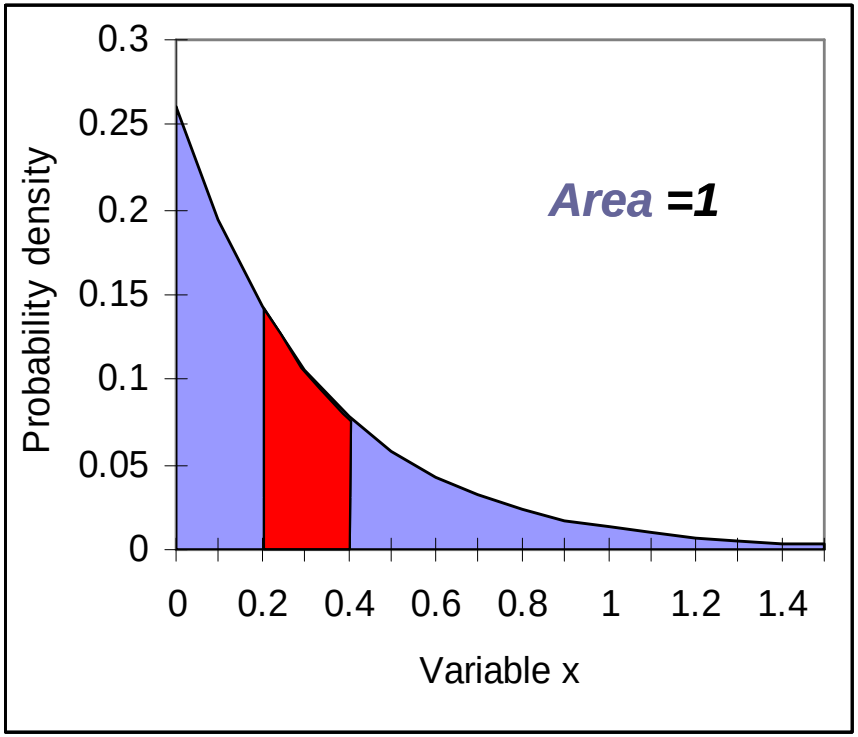




Probability Density

Probability density function
A function used to compute probabilities for a continuous random variable. The area under the graph of a probability density function over an interval represents probability.



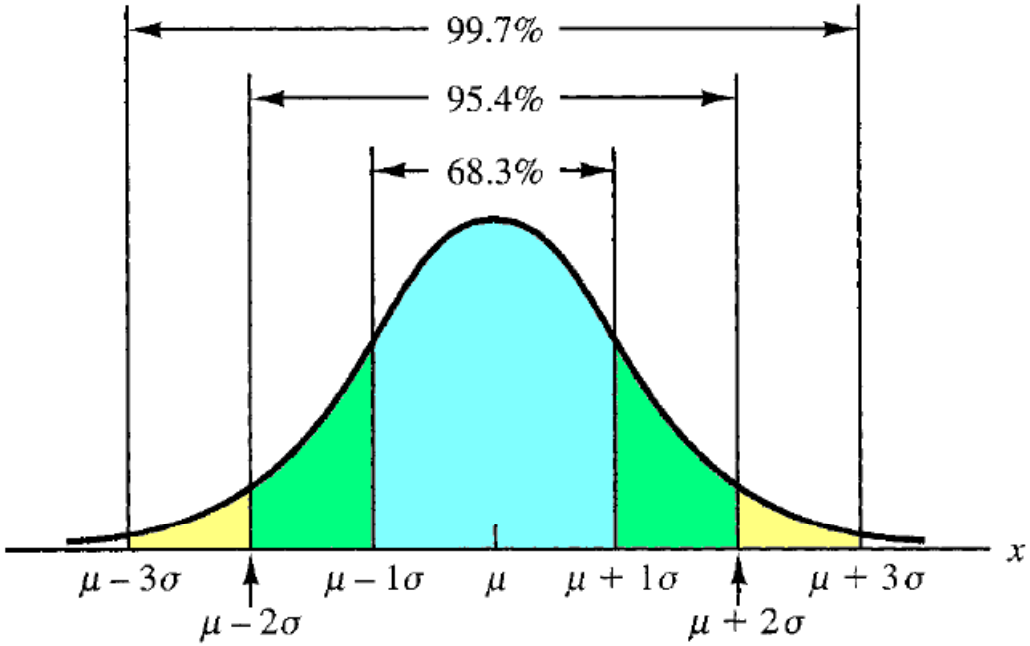


$$\int_x f(x) = 1$$

Normal Probability Distribution

Normal probability distribution
A continuous probability distribution. Its probability density function is bell shaped and determined by its mean μ and standard deviation σ.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



In Excel use the function:

- **= NORMDIST(x,m,s,false)** for probability density function
- **= NORMDIST(x,m,s,true)** for cumulative probability function of normal distribution (area from left to x)

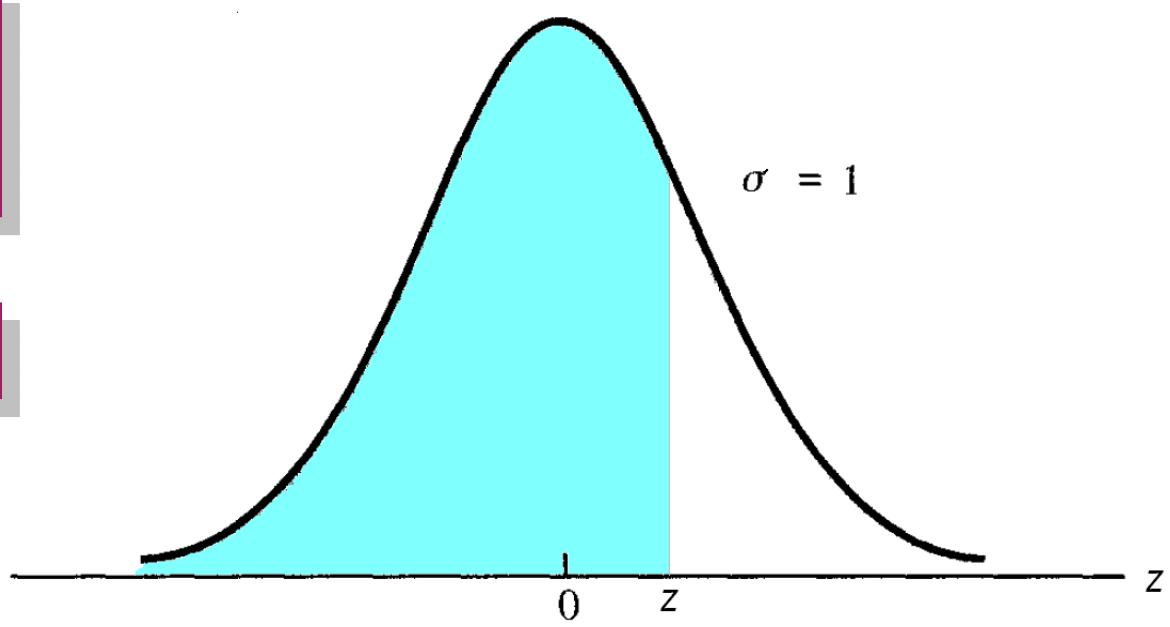
Standard Normal Probability Distribution

Standard normal probability distribution
A normal distribution with a mean of zero and a standard deviation of one.

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

$$z = \frac{x - \mu}{\sigma}$$

$$x = z\sigma + \mu$$



In Excel use the function:

- `= NORMSDIST(z)`

Example

Assume that you have developed an extremely efficient chemical treatment for glioblastoma. During tests on animal models it was found that the substance X, which you use, is able to kill all tumor cells (theoretically), but being given at high concentration it leads to the death of a patient due to intoxication. As the survived cancer cells fast evolve into resistant form, the efficiency of the treatment is significantly reduced if the second course is given. Therefore the treatment should be performed in one injection.

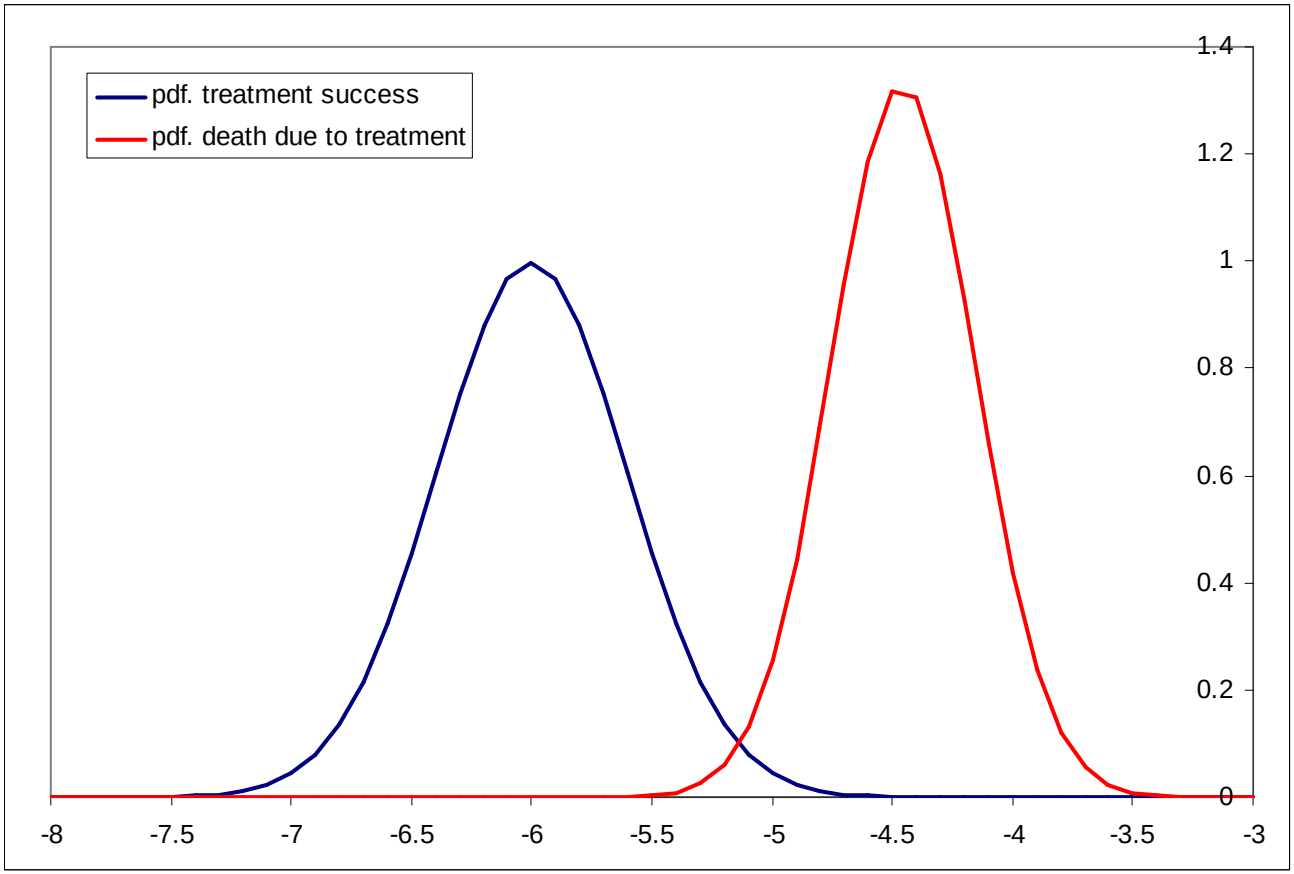
The experimental data suggest that the average concentration needed for the positive treatment is 1 μg/kg. The concentration needed for effective treatment is, of course, a random variable. Being presented in log10 scale and in g/kg, it can be approximated by a normal random variable with mean of –6 and standard deviation of 0.4.

The 50% lethal dose for human is 35 μg/kg. And the tests on animals suggest that in log10 scale it has a normal distribution as well with the standard deviation of 0.3.

parameter	**ug/kg**	**log scale**
mean positive treatment	1	-6
std positive treatment	x	0.4
mean lethal dose	35	-4.456
std lethal dose	x	0.3

Dose Selection

parameter	ug/kg	log scale
mean positive treatment	1	-6
std positive treatment	x	0.4
mean lethal dose	35	-4.456
std lethal dose	x	0.3



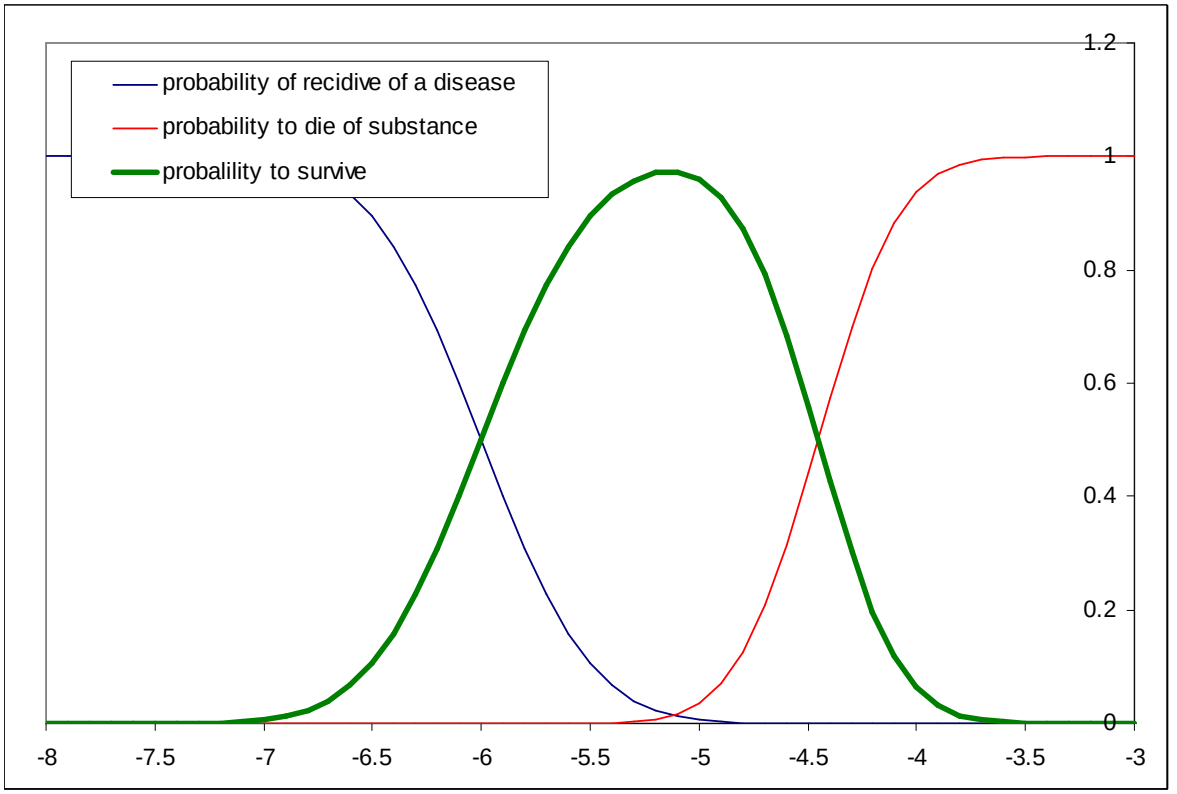
In Excel use the function:

- `= NORMDIST(x,mean,std,FALSE)`

Dose Selection

- Probability to die from disease = inverse probability to treat
- Over-dose and disease behaviors are independent =>

$$P_{survive} = (1 - P_{lethal\ disease}) \cdot (1 - P_{lethal\ treatment})$$



Thank you for your attention

to be continued…