Microarray Center

BIOSTATISTICS

Lecture 1

Introduction

Petr Nazarov

petr.nazarov@crp-sante.lu

14-01-2013

Materials: http://edu.sablab.net/sdae2013

Data: http://edu.sablab.net/data/xls

- Data presentation and descriptive statistics
- Discrete and continues distributions
- Sampling distribution and interval estimation for the mean
- Hypotheses about population mean
- Analysis of Variance (ANOVA)
- Linear regression
- Advanced topics

DATA AND STATISTICS

Elements, variables, and observations, data scales and types

Data: Elements, Variables, and Observations

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

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

Can we consider the "Place" as element?

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

Task: Define the Scales

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

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

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

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

?

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

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

TABULAR AND GRAPHICAL PRESENTATION

Frequency distribution, bar and pie charts, histogram, cumulative frequency distribution, scatter plot

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

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

Relative frequency distribution
A tabular summary of data showing the fraction or proportion of data items in each of several nonoverlapping classes. Sum of all values should give 1

Estimation of probability distribution
When number of experiments n $\rightarrow \infty$,
R.F.D. $\rightarrow$ P.D.

`pancreatitis.txt`

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 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

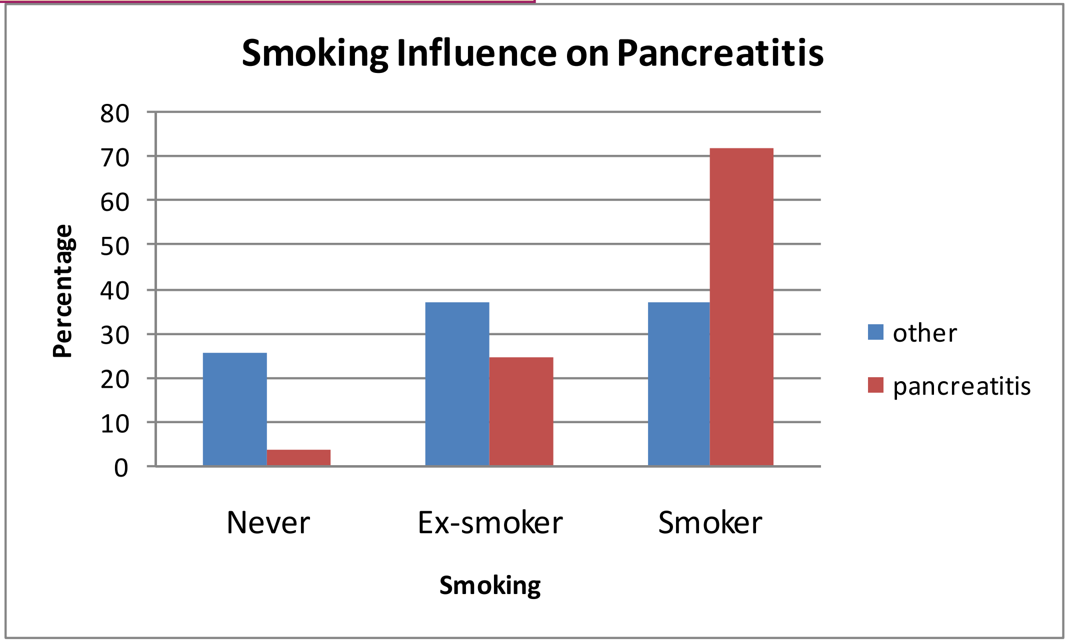
Smoking	*Disease* other	pancreatitis	Total
Ex-smokers	80	13	93
Never	56	2	58
Smokers	81	38	119
Total	217	53	**270**

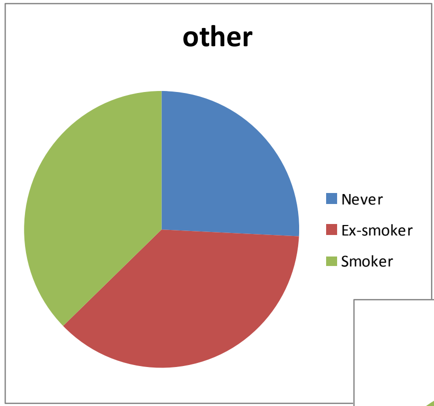
	Disease		
Smoking	other	pancreatitis	Total
Ex-smoker	80	13	**93**
Never	56	2	**58**
Smoker	81	38	**119**
Total	**217**	**53**	**270**

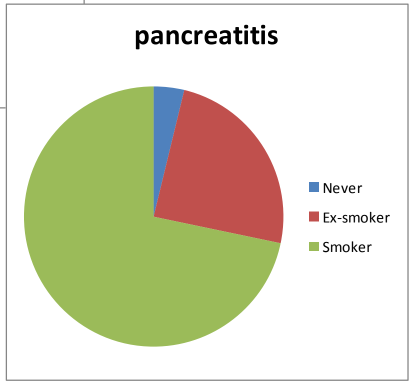
In Excel use the following steps:

- Insert → 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

pancreatitis.xls







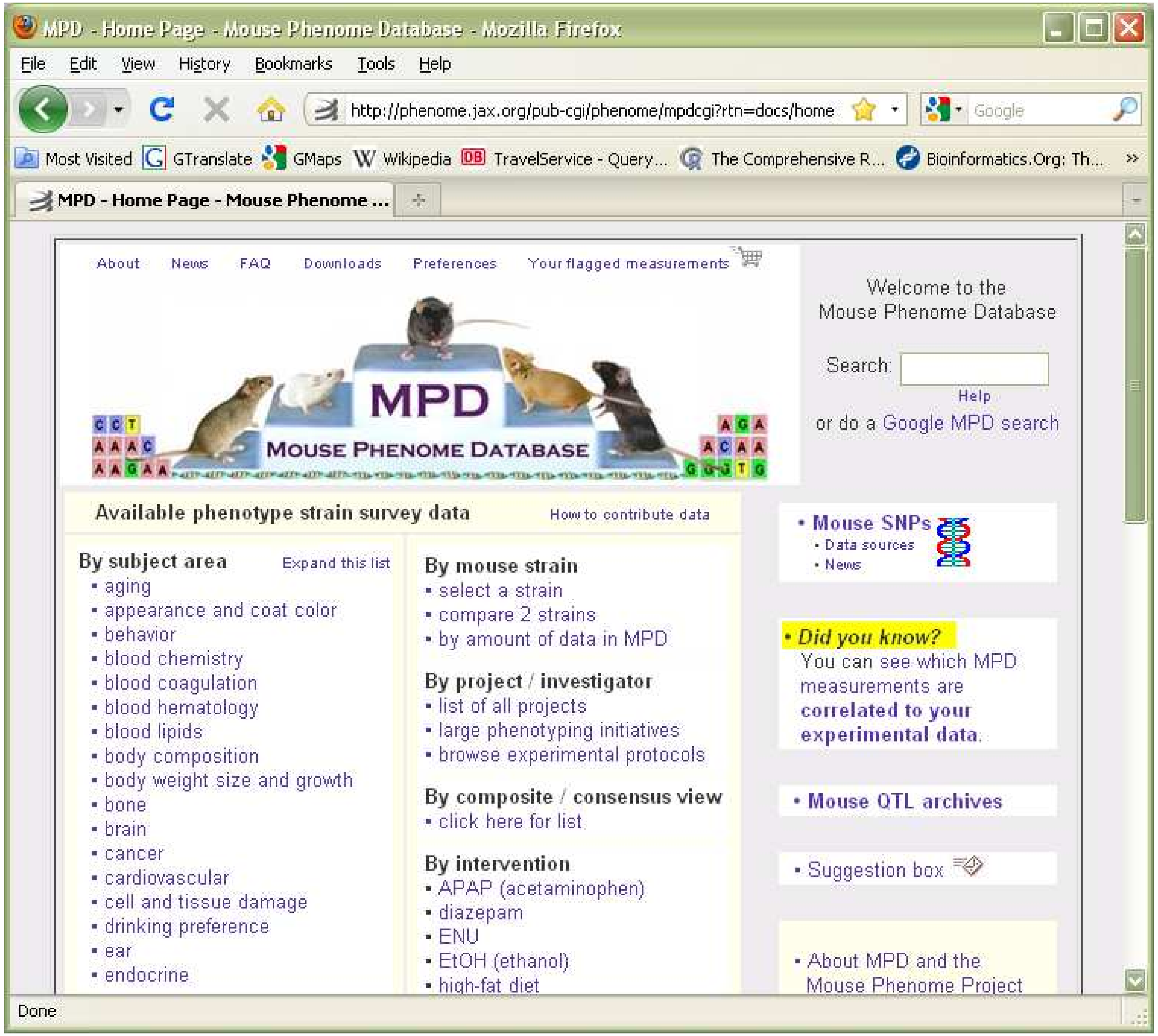
Try to avoid using in scientific reports. For public/business presentations only!

In MS Excel use the following steps:

- Insert → Column → Set data range (both columns of Percent freq. distribution)
- Insert→ Pie → Set data range (one columns of Percent freq. distribution)

TABULAR AND GRAPHICAL PRESENTATION

Example: 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

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	…

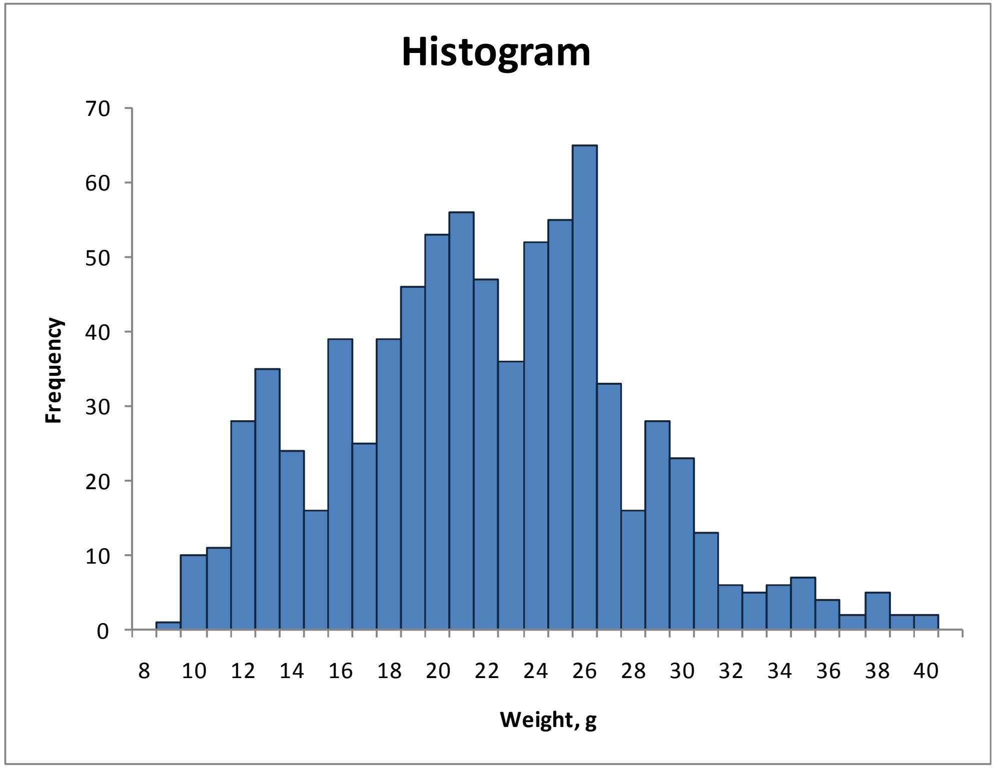
Sorted weights show that the values are in the 10 – 49.6 grams.
Let us divide the weight into the "bins"

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*
8	0
9	1
10	10
11	11
...	...
39	2
40	2
More	0

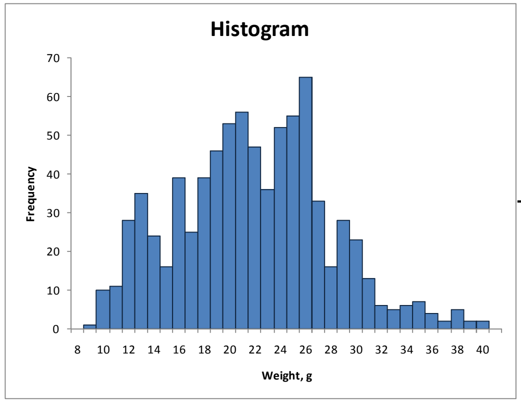


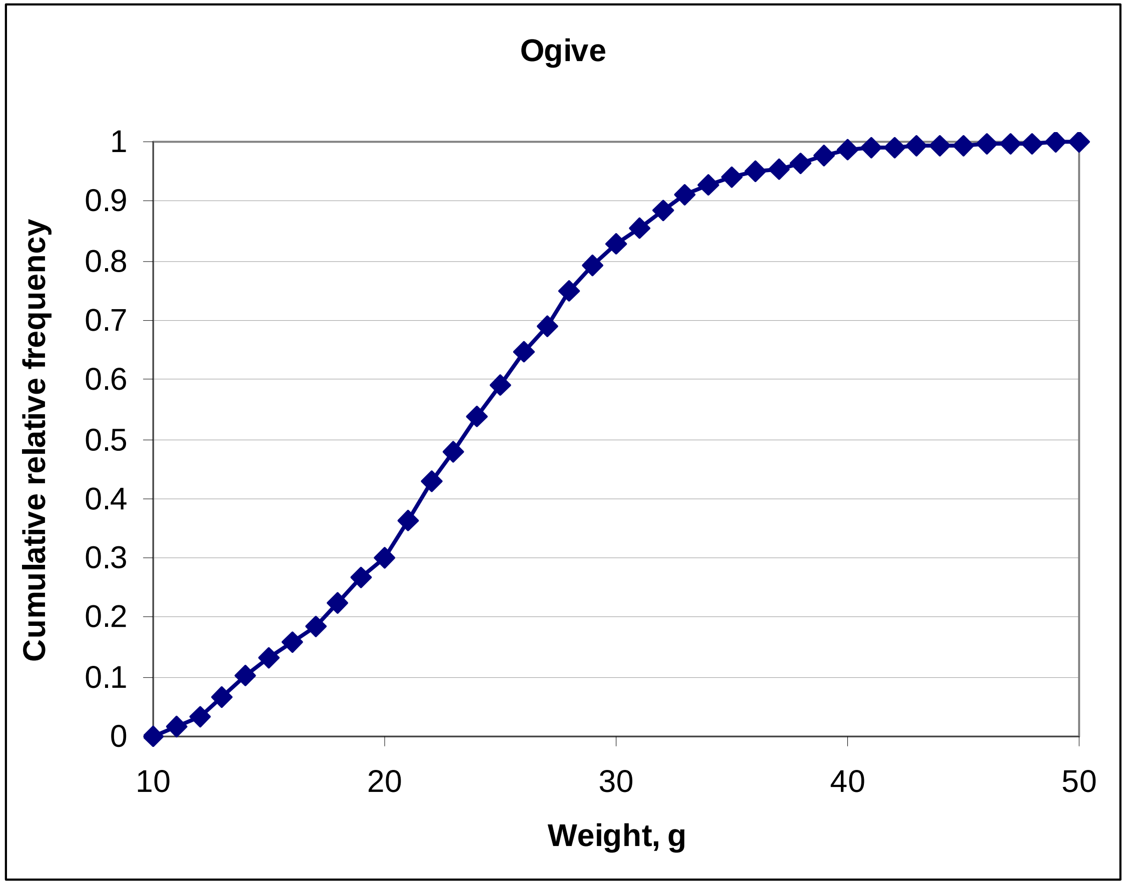
In Excel use the following steps:

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

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.





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 Insert→ XY (Scatter)

NUMERICAL MEASURES

Population and sample, measures of location, quantiles, quartiles and percentiles, measures of variability, z-score, detection of outliers, exploration data analysis, box plot, covariation, correlation

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

$m, \overline{x}$ – 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 m, 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

NUMERICAL MEASURES

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.

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

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

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

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

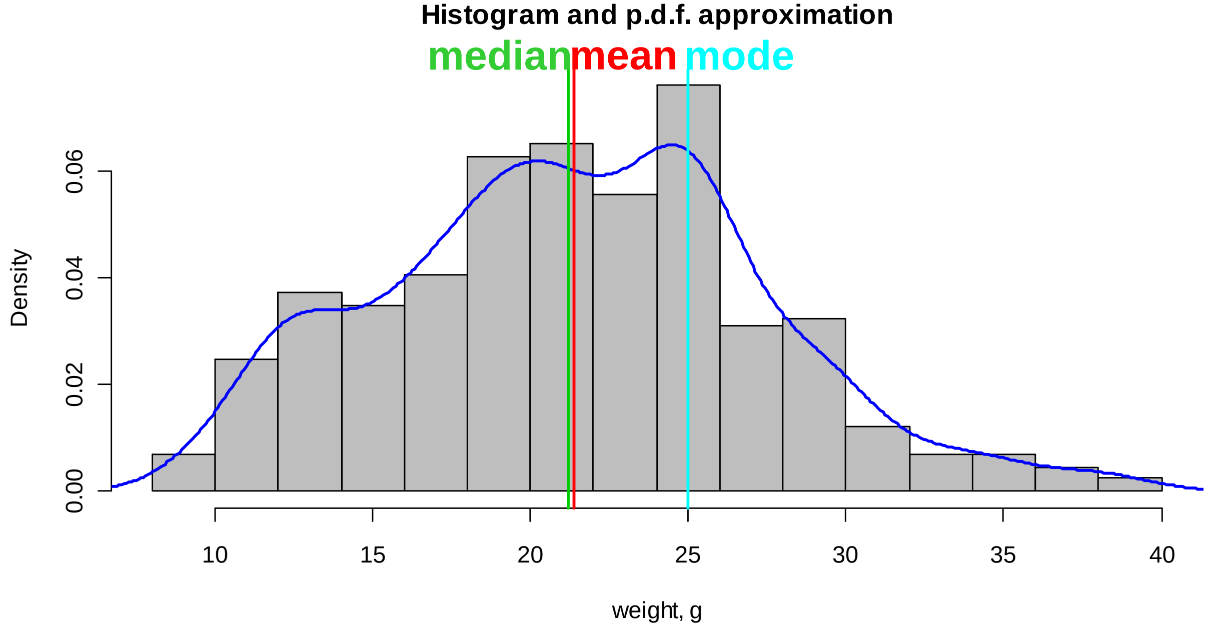
Mode = 23

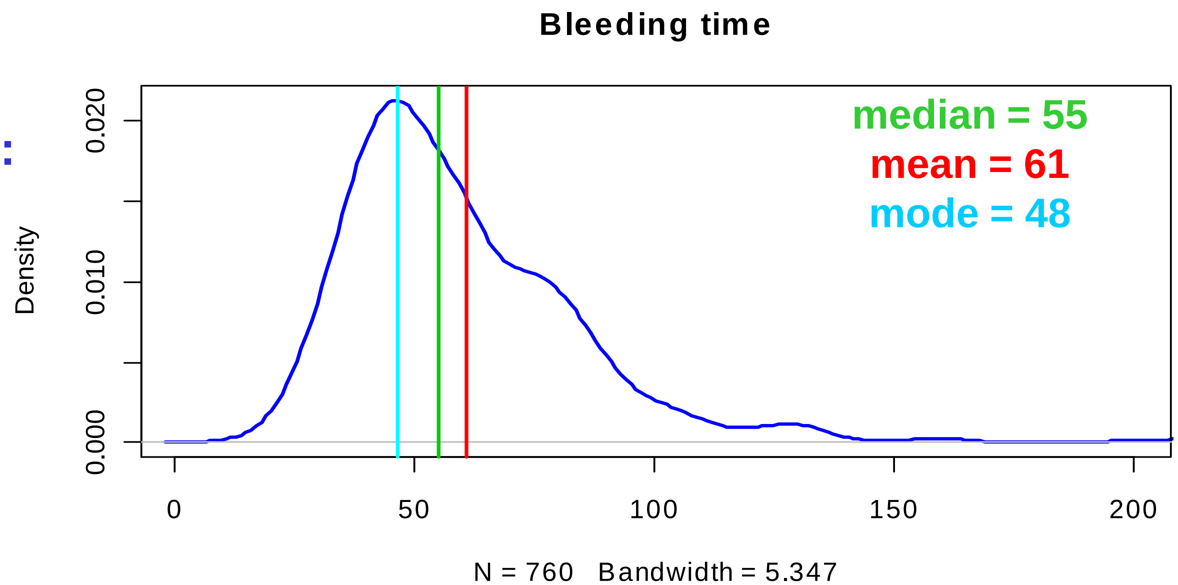
Median = 23.5

Mean = 31.7

mice.xls

Female proportion
$p_f = 0.501$





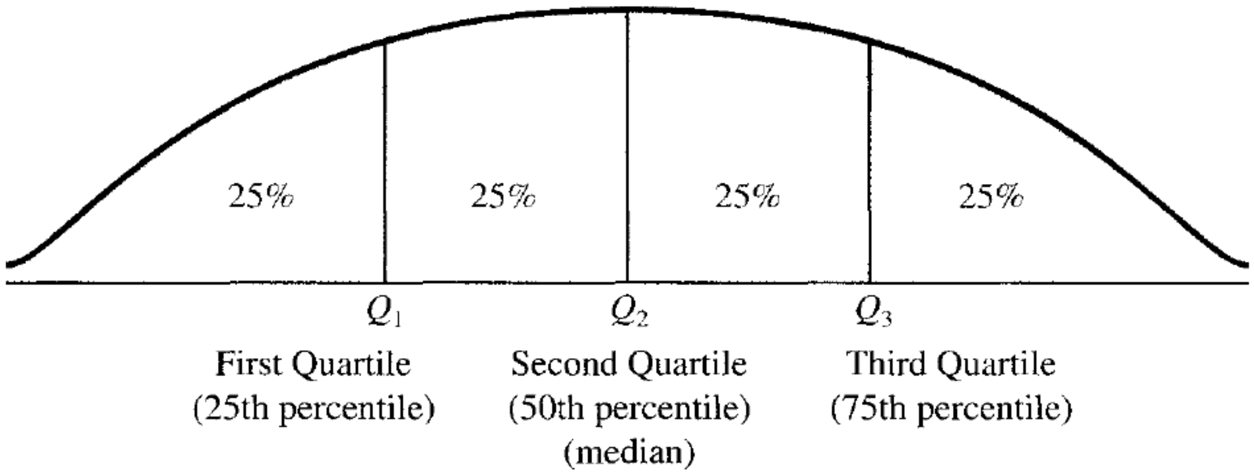
In Excel use the following functions:

- **= 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

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

NUMERICAL MEASURES

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

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)`

In Excel 2010 use the following functions:

- `=VAR.S(data), =STDEV.S(data)`

`(for a sample)`

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.

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

Online: http://www.miniwebtool.com/median-absolute-deviation-calculator/

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

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

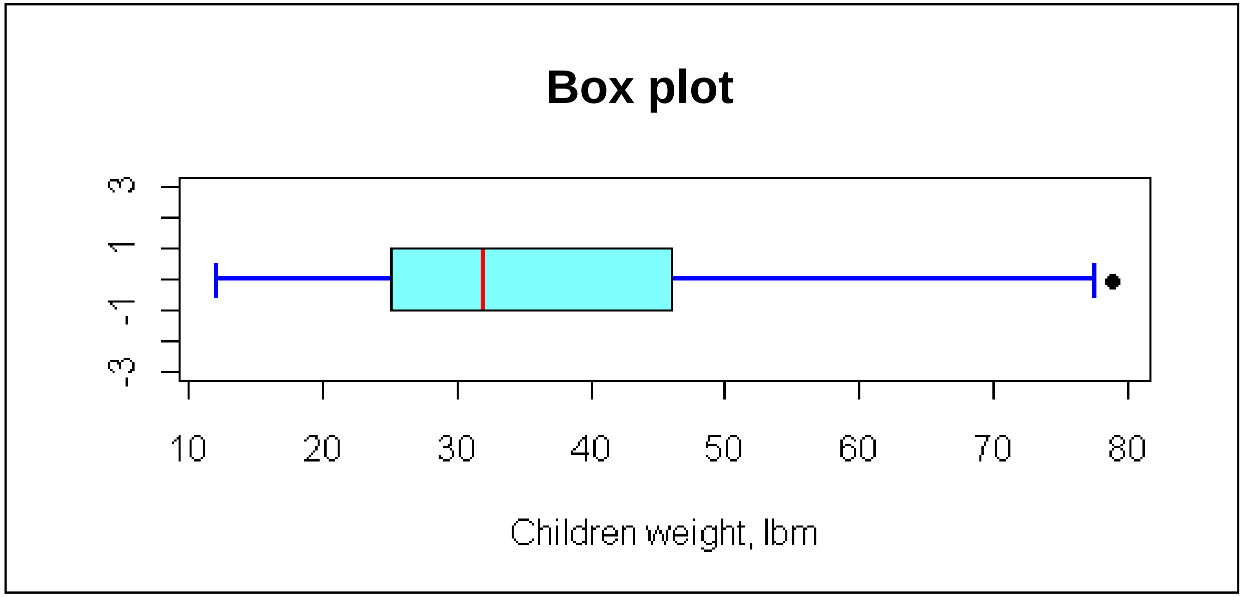
In Excel use:

- Tool → Data Analysis → Descriptive Statistics

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

In Excel use (indirect):

http://www.youtube.com/watch?v=s8ZW4PVarwE



http://peltiertech.com/WordPress/excel-box-and-whisker-diagrams-box-plots/

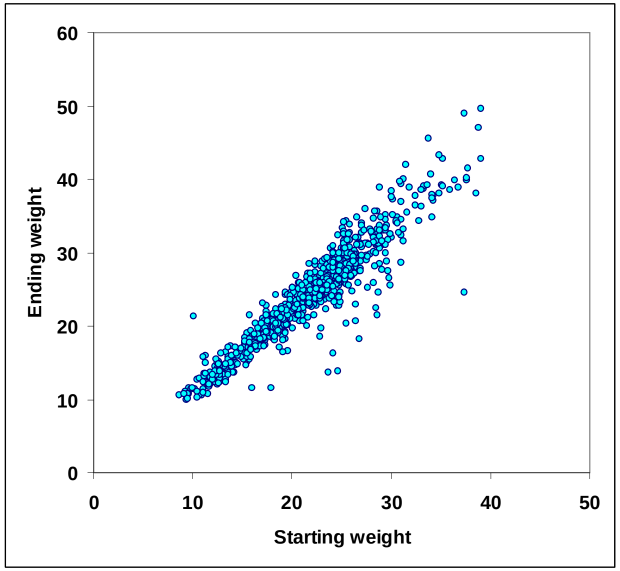
Correlation (Pearson product moment correlation coefficient)
A measure of linear association between two variables that takes on values between -1 and +1. Values near +1 indicate a strong positive linear relationship, values near -1 indicate a strong negative linear relationship; and values near zero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

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



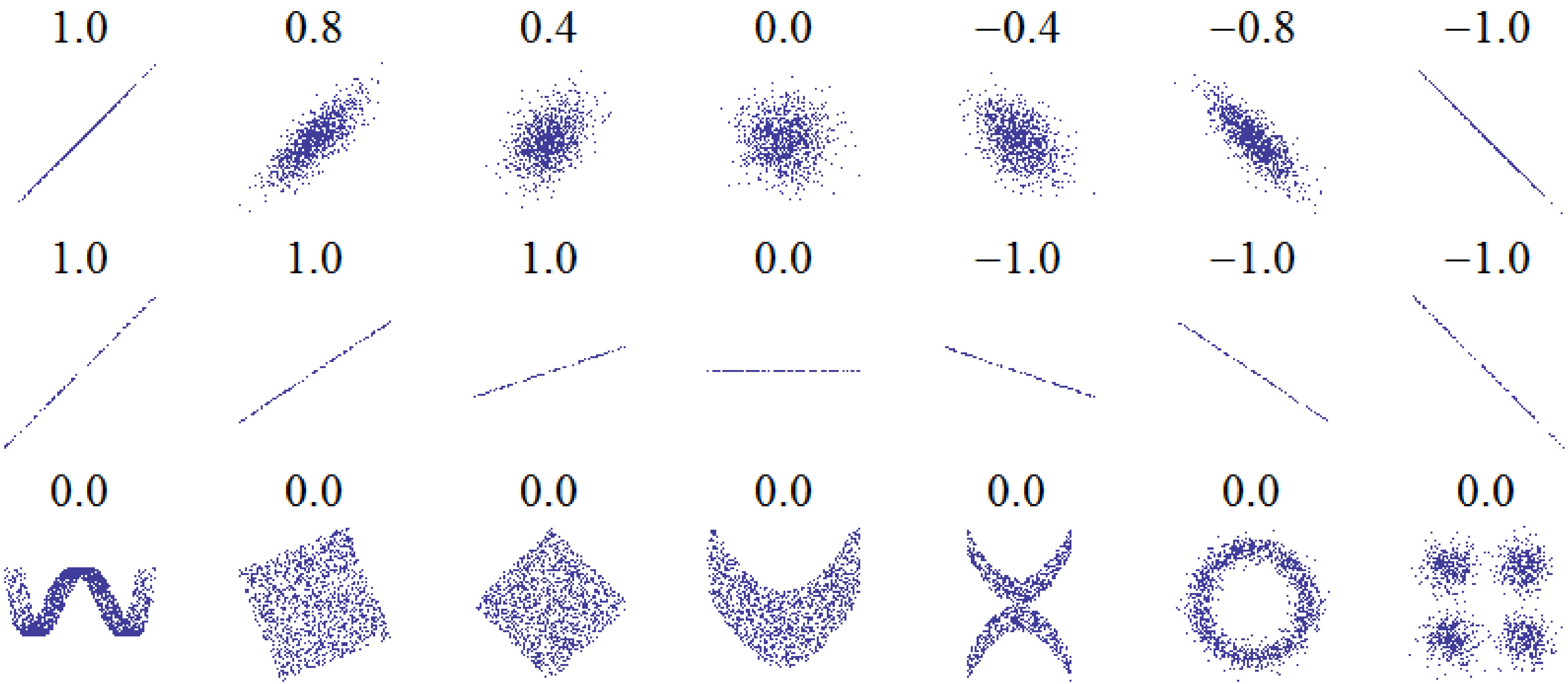
In Excel use function:

- `=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 ?

DETECTION OF OUTLIERS

z-score, detection of outliers

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

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 by z-score

For bell-shaped distributions:

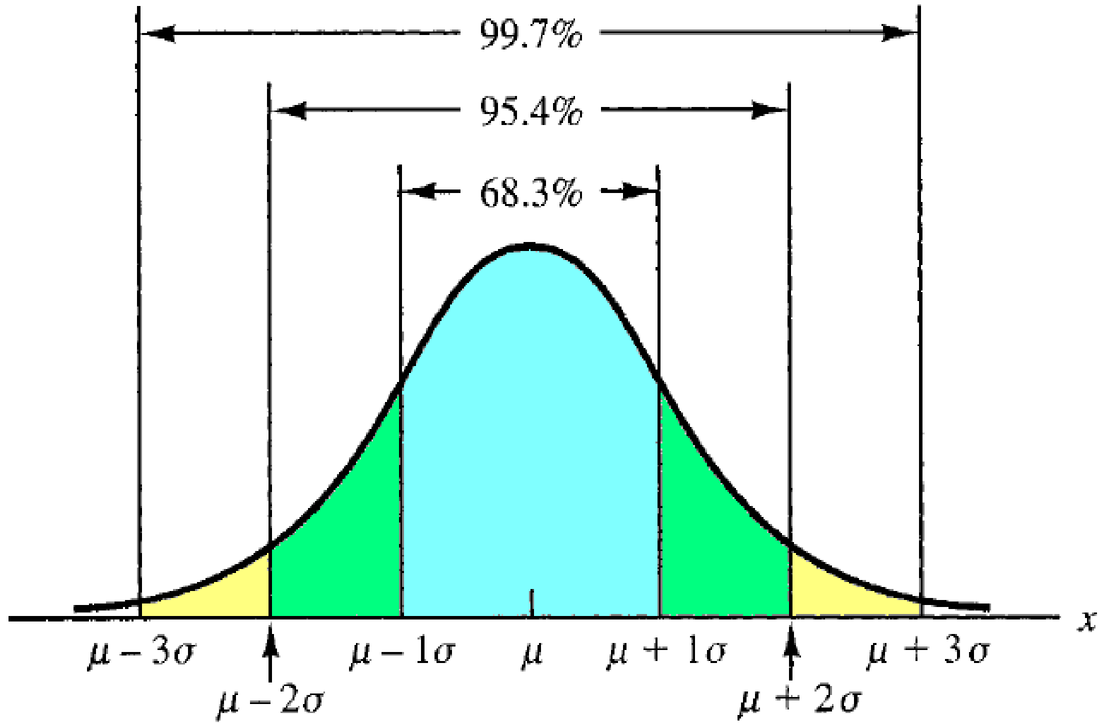
- Approximately 68 % of the values are within 1 st.dev. from mean
- Approximately 95 % of the values are within 2 st.dev. from mean
- Almost all data points are inside 3 st.dev. from mean

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



mice.xls

Using Excel, try to identify outlier mice on the basis of ***Weight change*** variable

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

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

In Excel use the following functions:

- **`= AVERAGE(data)`** **- mean, m**
- **`= STDEV.S(data)`** **- standard deviation, s**
- **`= ABS(data)`** **- absolute value**
- **sort by z-scale to identify outliers ☺**

More advanced is Grubbs' test for outliers (only works for reasonably normal data).

Online tool: **http://www.graphpad.com/quickcalcs/Grubbs1.cfm**

Task: Detection of Outliers

Iglewicz-Hoaglin method: modified Z-score

$$z_i = 0.6745 \frac{x_i - median(x)}{MAD(x)}$$

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

These authors recommend that modified Z-scores with an absolute value of greater than 3.5 be labeled as potential outliers.

|z|>3.5 $\Rightarrow$ outlier

Boris Iglewicz and David Hoaglin (1993), "Volume 16: How to Detect and Handle Outliers", The ASQC Basic References in Quality Control: Statistical Techniques, Edward F. Mykytka, Ph.D., Editor

More methods are at:

http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm

Grubbs' test is an iterative method to detect outliers in a data set assumed to come from a normally distributed population.

Grubbs' statistics at step k+1:

$$G_{(k+1)} = \frac{\max\left|x_i - m_{(k)}\right|}{s_{(k)}} = \max\left|z_i\right|$$

(k) – iteration k
m – mean of the rest data
s – st.dev. of the rest data

The hypothesis of no outliers is rejected at significance level α if

$$G > \frac{N-1}{\sqrt{N}}\sqrt{\frac{t^2}{N-2+t^2}}$$

$$\text{where } t^2 = t^2_{a/(2N),N-2}$$

More methods are at:

http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm

Let's perform Grubb's test for "Weight change" of mice.xls

Weight change	abs(x-m)/s		N	790
0	9.847692462		t^2	17.51895
2.109	8.91981		G.Crit.	4.139802
0.565	4.819888341			
0.578	4.704204352			
0.642	4.134683177			
0.658	3.992302884			

Step 1. Generate critical value

N: =COUNTIF(A:A,">=0")

t^2: =**TINV**(0.05/(~~2~~*E1),E1-2)^2

=**T.INV**(0.05/(**2***E1),E1-2)^2

G_{Crit} = (E1-1)/SQRT(E1)* SQRT(E2/(E1-2+E2))

$$G_{Crit} = \frac{N-1}{\sqrt{N}}\sqrt{\frac{t^2}{N-2+t^2}}$$

$$\text{where } t^2 = t^2_{a/(2N),N-2}$$

Step 2. Build |z| and sort in descending order

Step 3. If the first |z| value is > G_{Crit} – remove it and go to step 2, else finish.

Better Tool: http://graphpad.com/quickcalcs/grubbs2/

PROBABILITY DISTRIBUTIONS

Discrete and Continuous

- **Random variables**
- **Discrete probability distributions**
  - discrete probability distribution
  - expected value and variance
  - discrete uniform probability distribution
  - binomial probability distribution
  - hypergeometric probability distribution
  - Poisson probability distribution

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

A random variable is always a numerical measure.

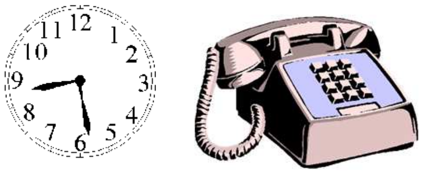
Discrete random variable
A random variable that may assume either a finite number of values or an infinite sequence of values.

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

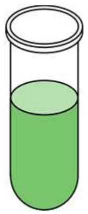
Roll a die

Number of calls to a reception per hour

Time between calls to a reception



Volume of a sample in a tube



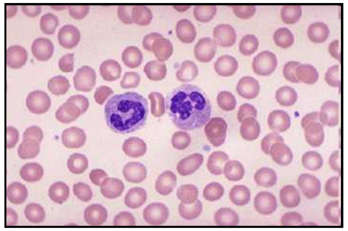
Weight, height, blood pressure, etc

Discrete Probability Distribution

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

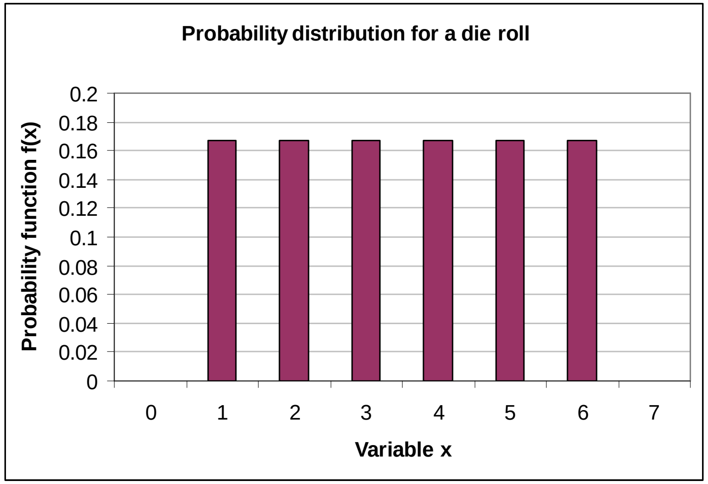
Probability function
A function, denoted by *f(x),* that provides the probability that *x* assumes a particular value for a discrete random variable.

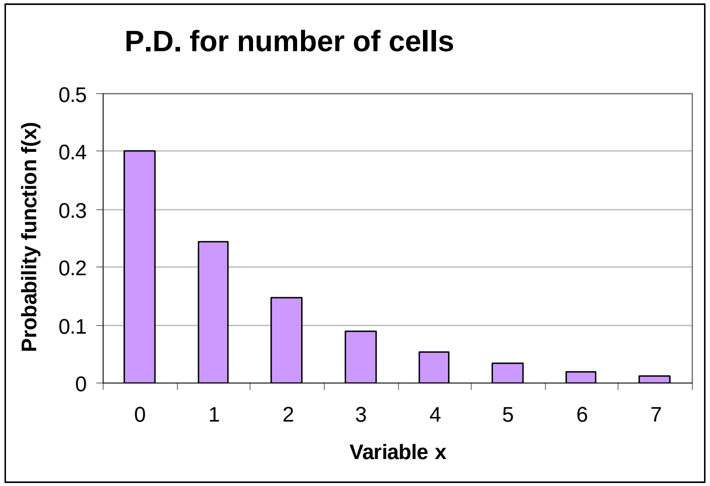
Number of cells under microscope
Random variable X:
x = 0
x = 1
x = 2
x = 3
…



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

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





Discrete Probability Distribution

Expected value
A measure of the central location of a random variable, mean.

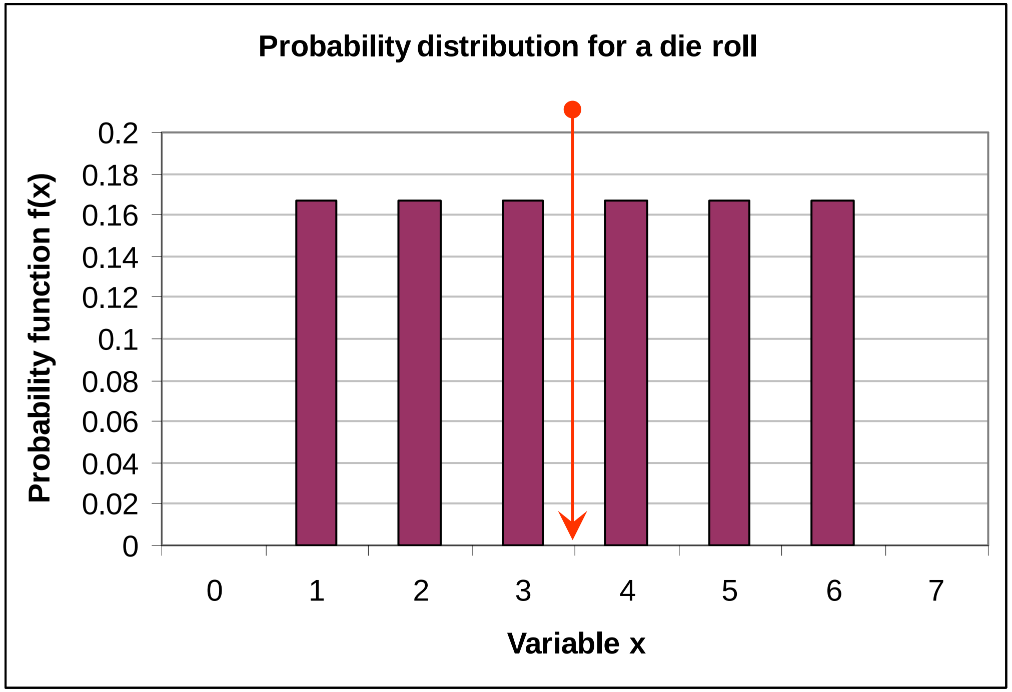
$$E(x)=\mu=\sum xf(x)$$

Variance
A measure of the variability, or dispersion, of a random variable.

$$\sigma^2=\sum(x-\mu)^2 f(x)$$

Roll a die
Random variable X:

x = 1
x = 2
x = 3
x = 4
x = 5
x = 6



DISCRETE PROBABILITY DISTRIBUTIONS

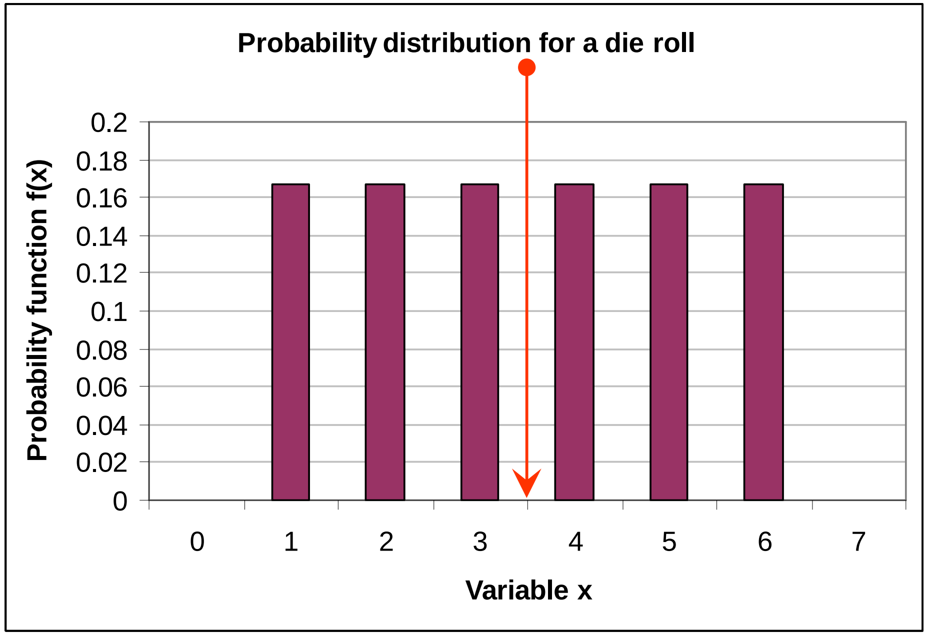
Discrete Uniform Probability Function

Discrete uniform probability distribution
A probability distribution for which each possible value of the random variable has the same probability.

$$f(x) = \frac{1}{n}$$

n – number of values of *x*

x	*f(x)*
1	0.1667
2	0.1667
3	0.1667
4	0.1667
5	0.1667
6	0.1667



$\mu = \sum(x_i / n) = \sum(x_i) / n$

$\mu = 3.5$
$\sigma^2 = 2.92$
$\sigma = 1.71$

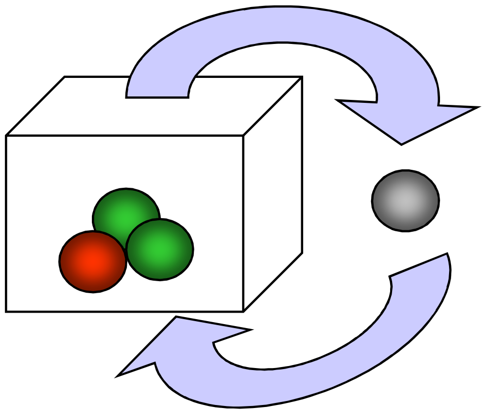
Example

Assuming that the probability of a side effect for a patient is 0.1. What is the probability that in a group of 3 patients none, 1, 2, or all 3 will get side effects after treatment?

Binomial experiment

An experiment having the four properties:

1. The experiment consists of a sequence of ***n* identical trials**.
2. **Two outcomes** are possible on each trial, one called success and the other failure.
3. The probability of a success ***p*** does not change from trial to trial. Consequently, the probability of failure, **1–*p***, does not change from trial to trial.
4. The trials are independent.



n trials

Binomial probability distribution

A probability distribution showing the probability of ***x*** successes in ***n*** trials of a binomial experiment, when the probability of success ***p*** does not change in trials.

Probability distribution for a binomial experiment

$$\boxed{f(x) = C_x^n p^x (1-p)^{(n-x)}}$$

$$\boxed{E(x) = \mu = np}$$

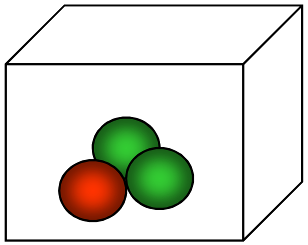
$$\boxed{Var(x) = \sigma^2 = np(1-p)}$$

$$C_x^n \equiv \binom{n}{x} \equiv \frac{n!}{x!(n-x)!}$$

$$n! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot n$$

$$0! = 1$$

Probability of red p(red)=1/3, 3 trials are given. Random variable = number of “red” cases



$$f(2) = \frac{3!}{2!(3-2)!}\left(\frac{1}{3}\right)^2\left(1-\frac{1}{3}\right)^{(3-2)}$$

f(0) = 8/27 = 0.296
f(1) = 4/9 = 0.444
f(2) = 2/9 = 0.222
f(3) = 1/27 = 0.037

Test: $\sum$f(x) = 1

Example: Binomial Experiment

Example

Assuming that the probability of a side effect for a patient is 0.1.

1. What is the probability to get none, 1, 2, etc. side effects in a group of 5 patients?
2. What is the probability that not more than 1 get a side effect
3. What is the expected number of side effects in the group?

$$f(x) = C_x^n p^x (1-p)^{(n-x)}$$

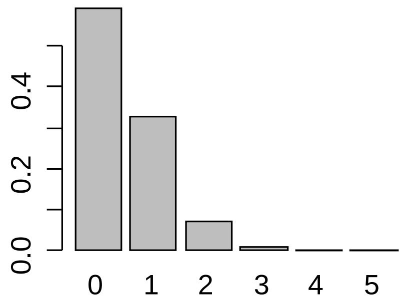
p = 0.1
n = 5

In Excel use the function:

- `= BINOMDIST(x,n,p,false)`

In Excel 2010 use the function:

- `= BINOM.DIST(x,n,p,false)`



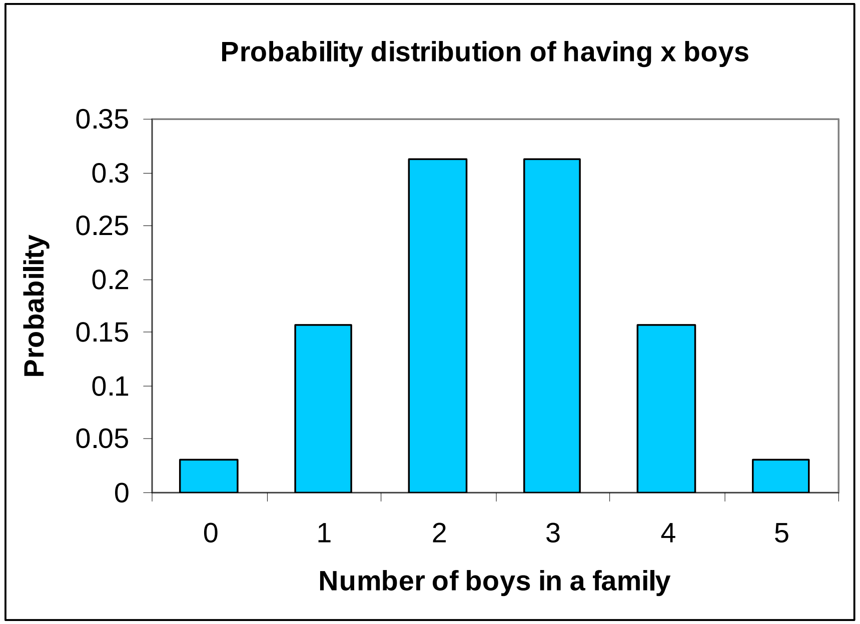
DISCRETE PROBABILITY DISTRIBUTIONS

Practical : Binomial Experiment

Assume the probability of getting a boy or a girl are equal.

1. Calculate the distribution of boys/girl in a family with **5 children**.
2. Plot the probability distribution
3. Calculate the probability of having all 5 children of only one sex

x	f(x)
0	0.03125
1	0.15625
2	0.3125
3	0.3125
4	0.15625
5	0.03125



Q3.
P(0 or 5) = P(0) + P(5)
= 0.03 + 0.03 = **0.06**

? Assume that a family has 4 girls already. What is the probability that the 5th will be a girl?

Hypergeometric Distribution

Example

There are 12 mice, of which 5 have an early brain tumor. A researcher randomly selects 3 of 12. What is the probability that none of these 3 has a tumor? What is the probability that more then 1 have a tumor?

Hypergeometric experiment

A probability distribution showing the probability of *x* successes in *n* trials from a population *N* with *r* successes and *N-r* failures.

$$E(x) = \mu = n\left(\frac{r}{N}\right)$$

$$Var(x) = \sigma^2 = n\left(\frac{r}{N}\right)\left(1 - \frac{r}{N}\right)\left(\frac{N-n}{N-1}\right)$$

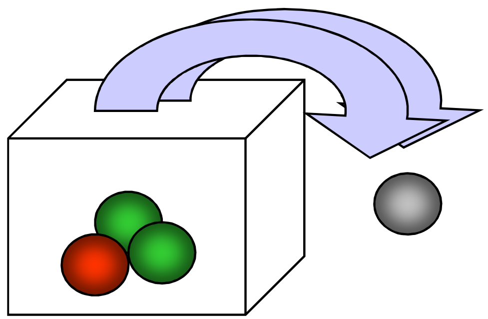
$$f(x) = \frac{C_x^r C_{n-x}^{N-r}}{C_n^N}, \qquad for\ 0 \le x \le r$$

In Excel use the function:

- `= HYPGEOMDIST (x,n,r,N)`

In Excel 2010 use the function:

- `= HYPGEOM.DIST (x,n,r,N)`



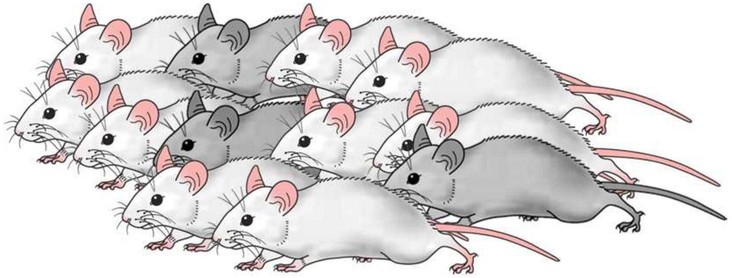
n trials

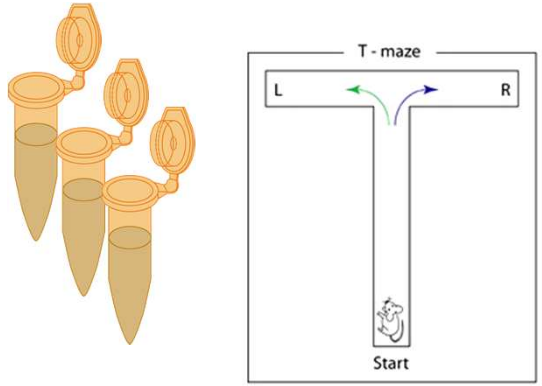
Example: Hypergeometric Distribution for Mice

Example

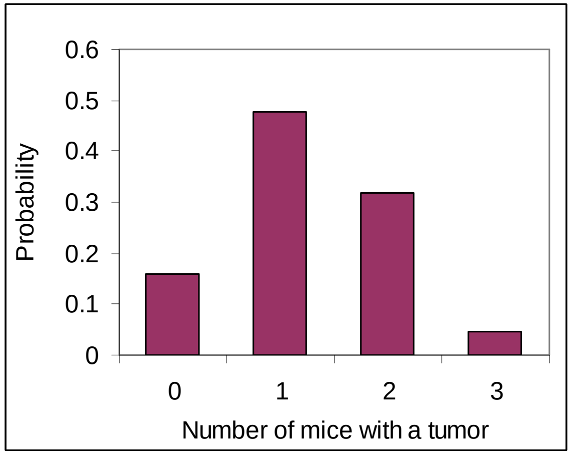
There are 12 mice, of which 5 have an early brain tumor. A researcher randomly selects 3 of 12.

1. What is the probability that none of these 3 has a tumor?
2. What is the probability that more than 1 have a tumor?





x	f(x)
0	0.159
1	0.477
2	0.318
3	0.045



Q1.
P(0) =0.159

Q2.
P(>1) =P(2)+P(3)=0.364

In Excel use the function:

- **= HYPGEOM.DIST (x,n,r,N)**

Example

Number of calls to an Emergency Service is on average 3 per hour b/w 2 a.m. and 6 a.m. of working days. What are the probabilities to have 0, 5, 10 calls in the next hour?

Poisson probability distribution
A probability distribution showing the probability of *x* occurrences of an event over a specified interval of time or space.

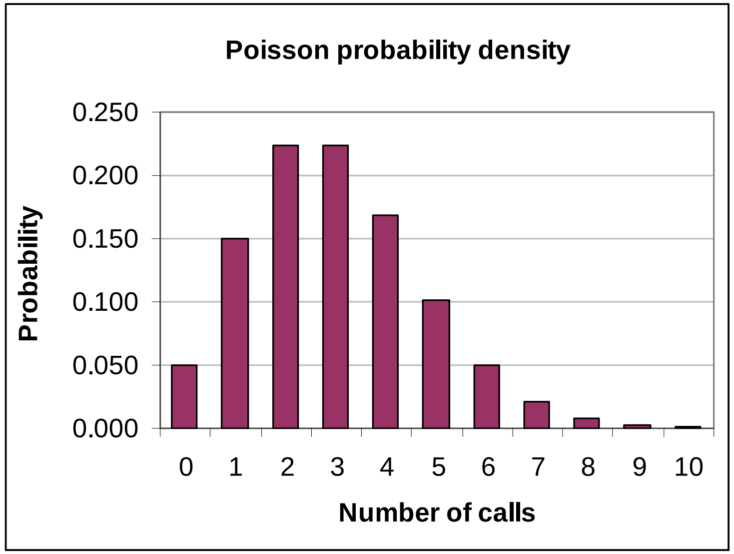
Poisson probability function
The function used to compute Poisson probabilities.

$$f(x) = \frac{\mu^x e^{-\mu}}{x!}$$

$$\mu = \sigma^2$$

where μ – expected value (mean)

x	f(x)
0	0.050
1	0.149
2	0.224
3	0.224
4	0.168
5	0.101
6	0.050
7	0.022
8	0.008
9	0.003
10	0.001



In Excel use the function:

- `= POISSON(x,mu,false)`
- `= POISSON.DIST(…)`

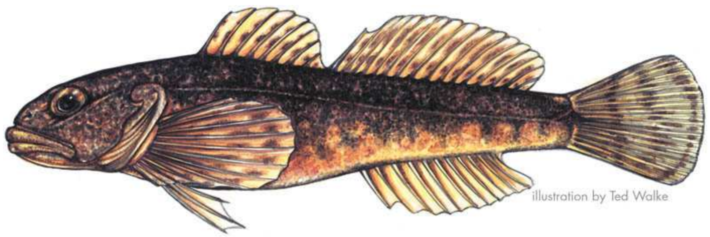
Example: Poisson Distribution for Fish Counting

Example

An ichthyologist studying the *spoonhead sculpin* catches specimens in a large bag seine that she trolls through the lake. She knows from many years experience that on averages she will catch 2 fish per trolling.

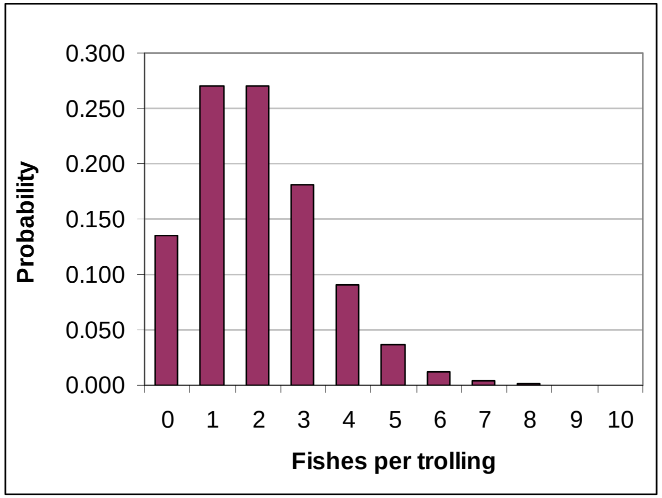
Find the probabilities of catching:

1. *No fish;*
2. *Less than 4 fishes;*
3. *More then 1 fish.*



In Excel use the function:

- **= POISSON.DIST(x,mu,false)**



Q1.
P(0) = 0.135

Q2.
P(<4) = P(0)+P(1)+P(2)+P(3)=0.857

Q3.
P(>1) =1-P(0)-P(1)=0.594

Glover , Mitchell, An Introduction to Biostatistics

- **Continuous probability distribution**
  - a continuous probability distribution
  - uniform probability distribution
  - normal probability distribution
  - exponential probability distribution

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

A random variable is always a numerical measure.

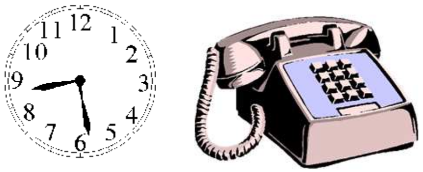
Discrete random variable
A random variable that may assume either a finite number of values or an infinite sequence of values.

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

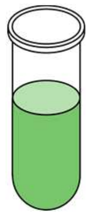
Roll a die

Number of calls to a reception per hour

Time between calls to a reception



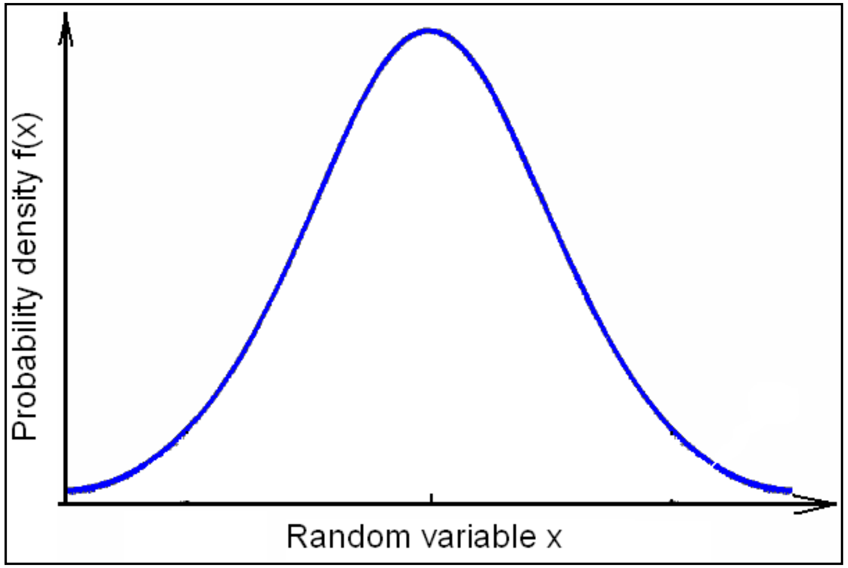
Volume of a sample in a tube

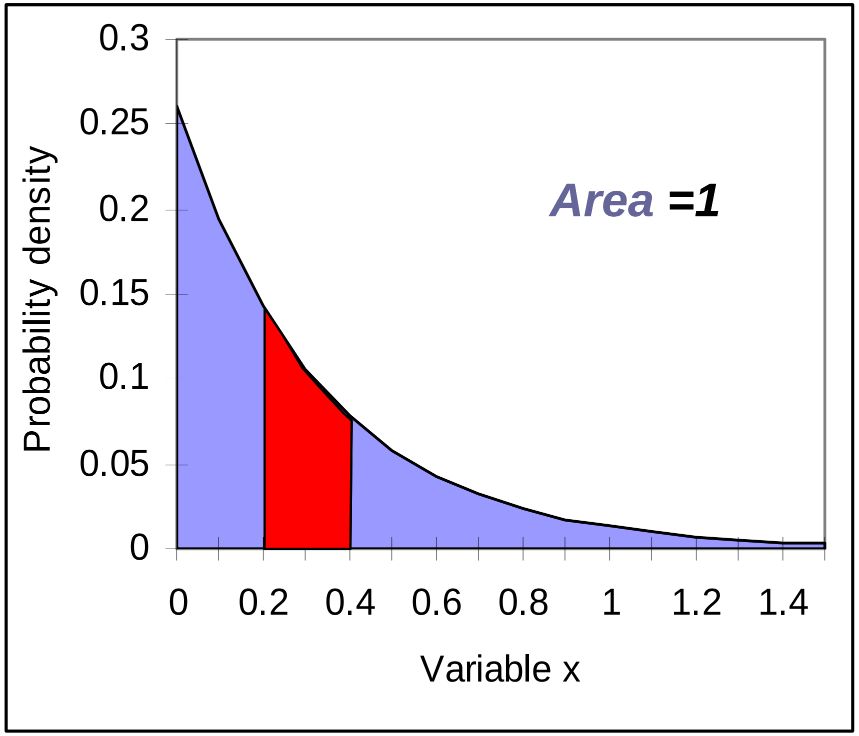


Weight, height, blood pressure, etc

Probability 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.

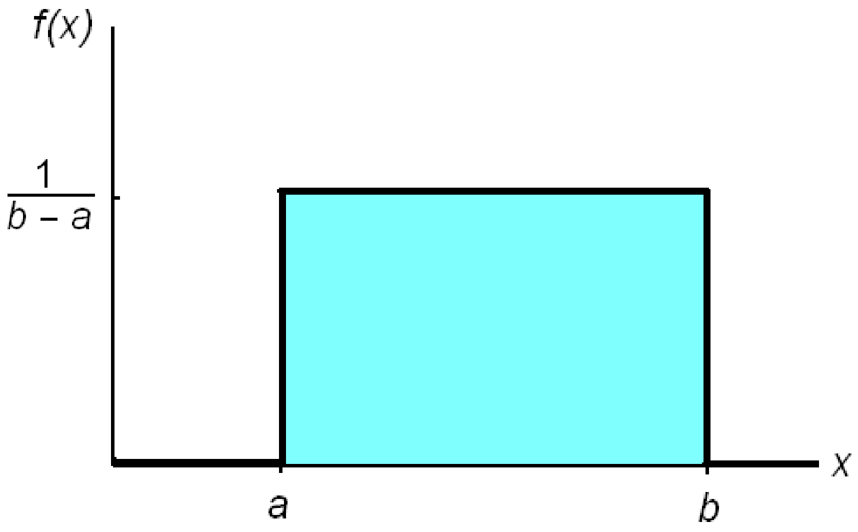




$$\int_x f(x) = 1$$

Uniform probability distribution

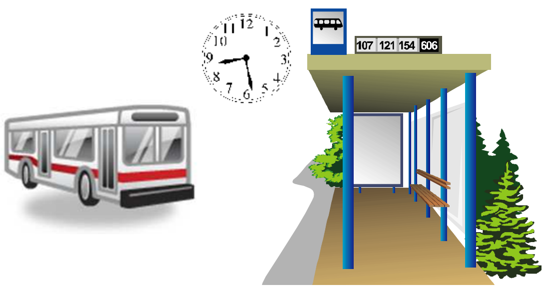
A continuous probability distribution for which the probability that the random variable will assume a value in any interval is the same for each interval of equal length.



$$f(x) = \begin{cases} \dfrac{1}{b-a}, & \text{for } a \le x \le b \\ 0, & \text{elsewhere} \end{cases}$$

$$E(x) = \mu = \frac{a+b}{2}$$

$$Var(x) = \sigma^2 = \frac{(b-a)^2}{12}$$

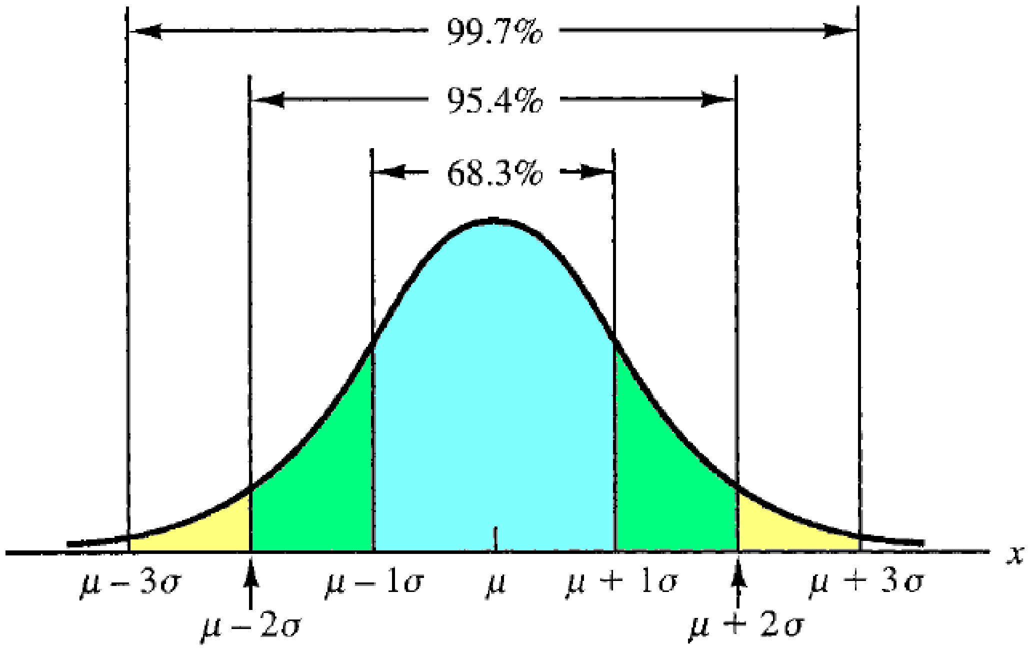


Example

The bus 22 goes every 7 minutes. You are coming to CHL bus station, having no idea about precise timetable. What is the distribution for the time, you may wait there?

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:

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

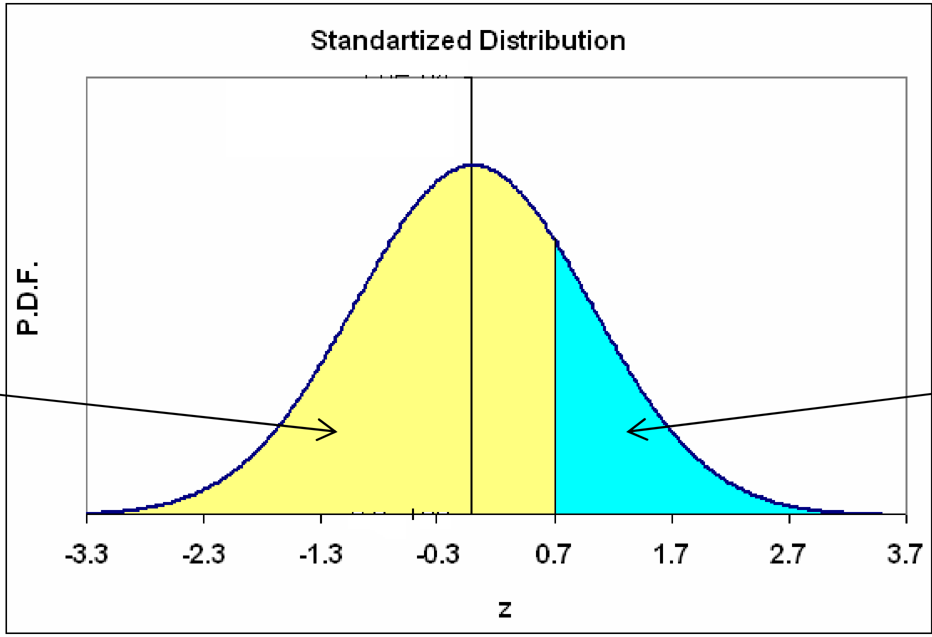
Standard Normal Probability Distribution

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

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

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

= NORM.S.DIST(z)



= 1-NORM.S.DIST(z)

In Excel use the function:
- **`= NORMSDIST(z)`**

In Excel 2010 use the function:
- **`= NORM.S.DIST(z)`**

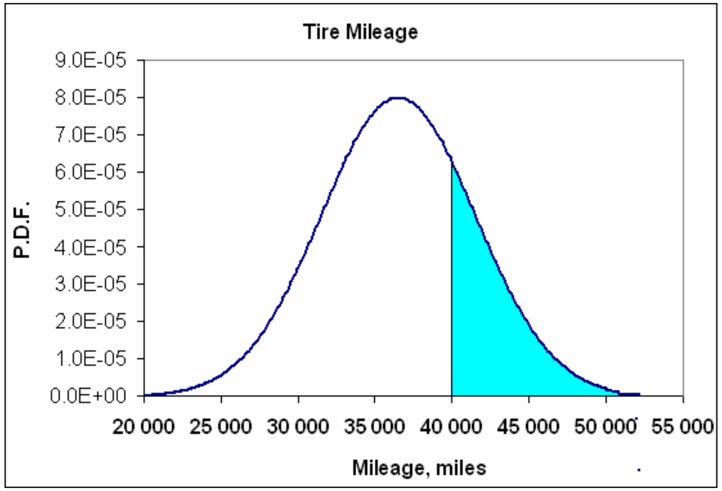
Example: Gear Tire Company

Example

Suppose the Grear Tire Company just developed a new steel-belted radial tire that will be sold through a chain of discount stores. Because the tire is a new product, Grear's managers believe that the mileage guarantee offered with the tire will be an important factor in the acceptance of the product. Before finalizing the tire mileage guarantee policy, Grear's managers want probability information about the number of miles the tires will last.

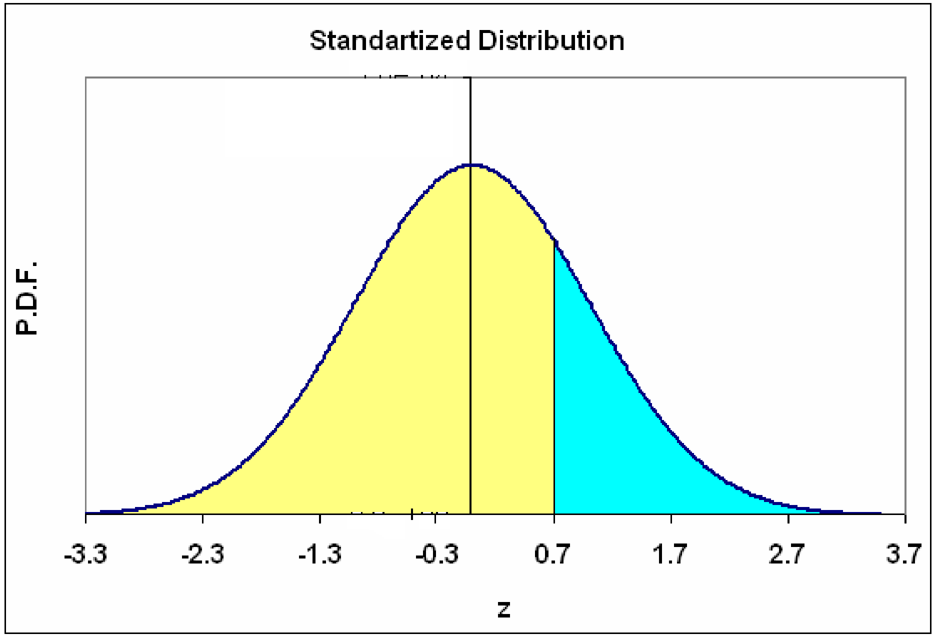
From actual road tests with the tires, Grear's engineering group estimates the mean tire mileage is **μ = 36 500** miles with a standard deviation of **σ = 5 000**. In addition, data collected indicate a normal distribution is a reasonable assumption.

What percentage of the tires can be expected to last more than 40 000 miles? In other words, what is the probability that a tire mileage will exceed 40 000?



Anderson et al Statistics for Business and Economics

Example: Gear Tire Company



1. Let's transfer from Normal distribution to Standard Normal, then z, corresponding to 40000 will be

$$z = \frac{40000 - 36500}{5000} = 0.7$$

2. Calculate the "blue" area P(z >0.7) using the table:

P(z>0.7) = 1 – P(z<0.7) = 1 – 0.5 – **P(0<z<0.7)** = 1 – 0.5 – 0.258 = **0.242**

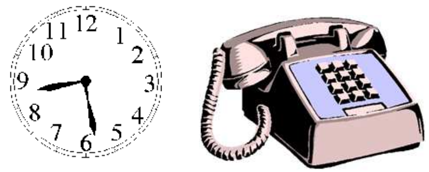
Alternatively in Excel

```
=1-NORM.DIST(40000,36500,5000,true)
```

Example

Number of calls to an Emergency Service is on average 3 per hour b/w 2.00 and 6.00 of working days. What are the distribution of the time between the calls?

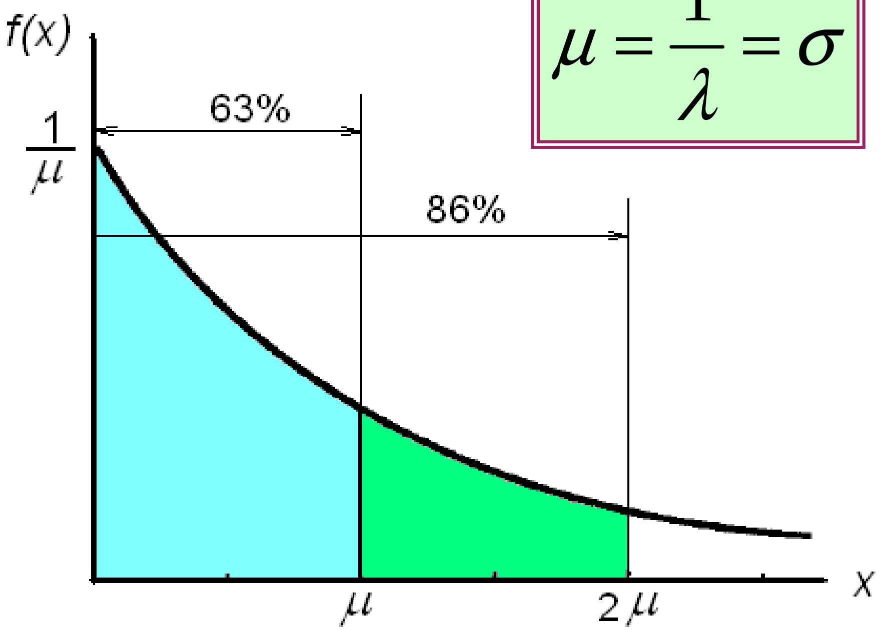
Time between calls to a reception



Exponential probability distribution

A continuous probability distribution that is useful in computing probabilities for the time between independent random events.

$$\mu = \frac{1}{\lambda} = \sigma$$



$$f(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}} \qquad for\ x \geq 0, \mu > 0$$

$$f(x) = \lambda e^{-\lambda x}$$

Cumulative probability function

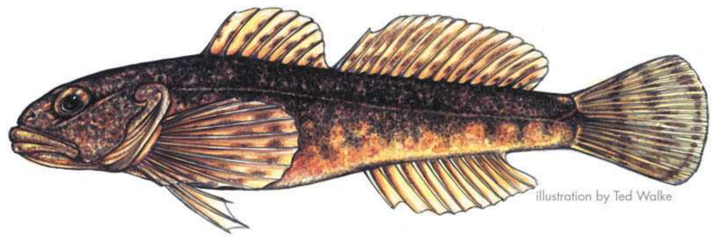
$$P(x \leq x_0) = F(x_0) = 1 - e^{-\frac{x_0}{\mu}}$$

Example: Exponential Distribution for Fish Counting

Example

An ichthyologist studying the *spoonhead sculpin* catches specimens in a large bag seine that she trolls through the lake. She knows from many years experience that on averages she will catch **2 fishes per trolling**. Each trolling takes **~30 minutes**.

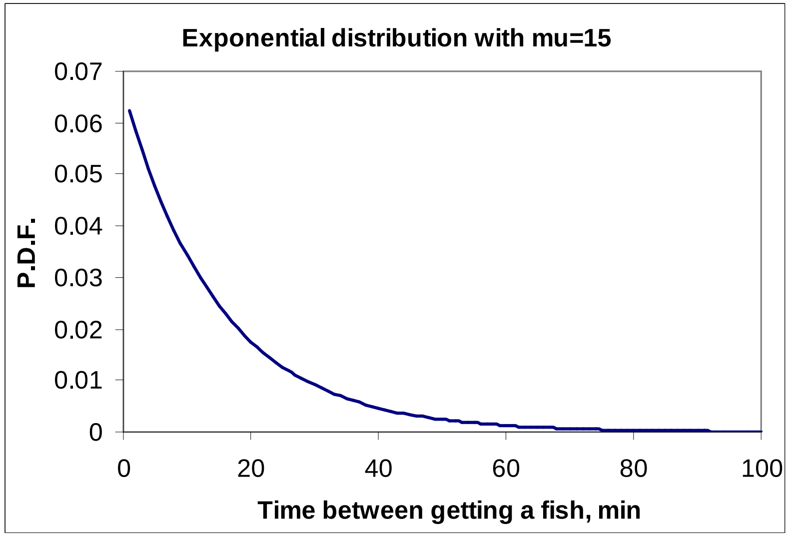
Find the probability of catching no fish in the next hour



In Excel use the function:

- `= EXPON.DIST(x,1/mu,false)`

1. Let's calculate μ for this situation: $\mu = 30 / 2 = 15$ minutes



2. Use either a cumulative probability function or Excel to calculate:

$$P(x \geq 60) = 1 - P(x \leq 60) = 1 - F(60) = e^{-\frac{60}{15}} \approx 0.02$$

Thank you for your attention

to be continued…