Teaching Mathematics and Statistics in Sciences, IPA HU-SRB/0901/221/088 - 2011

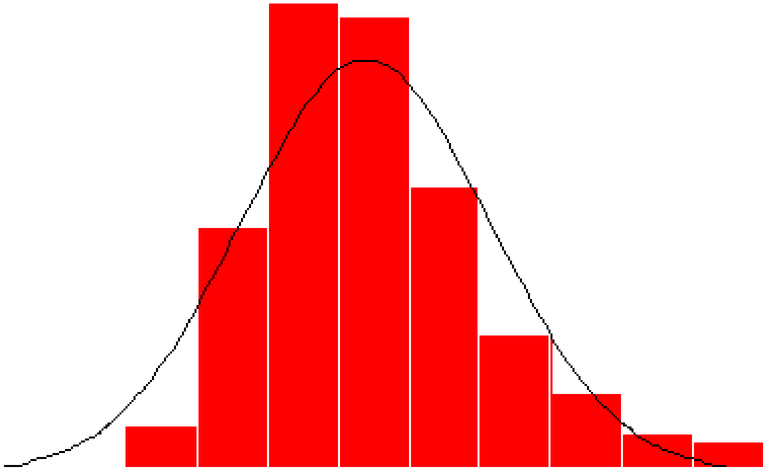
Biostatistics

Author: **Krisztina Boda** PhD

University of Szeged
Department of Medical Physics and Informatics

www.model.u-szeged.hu
www.szote.u-szeged.hu/dmi

The basics of probability theory. Distribution of variables, some important distributions



Random experiment

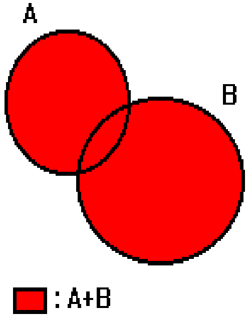
- The outcome is not determined uniquely by the considered conditions.
- For example, tossing a coin, rolling a dice, measuring the concentration of a solution, measuring the body weight of an animal, etc. are experiments.
- Every experiment has more, sometimes infinitely large outcomes

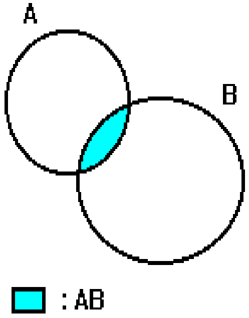
Event: **the result (or outcome) of an experiment**

- elementary events: the possible outcomes of an experiment.
- composite event: it can be divided into sub-events.
- Example. The experiment is rolling a dice.
  - Elementary events are 1,2,3,4,5,6.
  - Composite events:
    - E1={1,3,5} (the result is an odd number).
    - E2={2,4,6} (the result is an even number).
    - E3={5,6} (the result is greater than 4).
    - Ω={1,2,3,4,5,6} (the result is the certain event).

Operations with events

- **The complementary event** of an event A is the event $\overline{A}$ which occurs whenever A fails.
  - Example: $\overline{E_1} = \overline{\{1,3,5\}}$ ={2,4,6}
- The **sum of two events A and B** is the event A+B which occurs if A or B occurs
  - E1+E2={1,3,5}+{2,4,6}={1,2,3,4,5,6}
  - E1+E3={1,3,5}+{5,6}={1,3,5,6}
- The **product of two events A and B** is the event AB which occurs if A and B occur.
  - E1·E2={1,3,5}·{2,4,6}=∅
  - E1·E3={1,3,5}·{5,6}={5}
- If A · B =∅, then we say that A and B are **mutually exclusive** events.





The concept of probability

The concept of probability

- Lets repeat an experiment *n* times under the same conditions. In a large number of *n* experiments the event A is observed to occur *k* times ($0 \leq k \leq n$).
- *k* : frequency of the occurrence of the event A.
- *k/n* : relative frequency of the occurrence of the event A.

$$0 \leq k/n \leq 1$$

If *n* is large, k/n will approximate a given number. This number is called the probability of the occurrence of the event A and it is denoted by P(A).

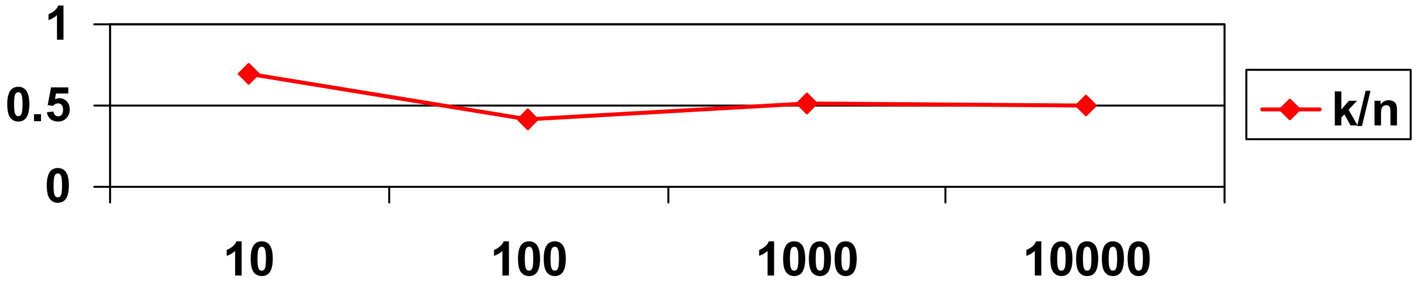
$$0 \leq P(A) \leq 1$$

Example: the tossing of a (fair) coin

k: number of „head”s

n=	10	100	1000	10000	100000
k=	7	42	510	5005	49998
k/n=	0.7	0.42	0.51	0.5005	0.49998

P(„head”)=0.5



Probability facts

- Any probability is a number between 0 and 1.
- All possible outcomes together must have probability 1.
- The probability of the complementary event of A is 1-P(A).

Formula for the calculation of simple exact probabilities

Rules of probability calculus

- Assumption: all elementary events are equally probable

$$P(A) = \frac{F}{T} = \frac{\text{number of favorite outcomes}}{\text{total number of outcomes}}$$

Examples:

- Rolling a dice. What is the probability that the dice shows 5?
  - If we let X represent the value of the outcome, then P(X=5)=1/6.
- What is the probability that the dice shows an odd number?
  - P(odd)=1/2. Here F=3, T=6, so F/T=3/6=1/2.

Population, sample

Population, sample

- **Population**: the entire group of individuals that we want information about.
- **Sample**: a part of the population that we actually examine in order to get information
- A simple random sample of size *n* consists of *n* individuals chosen from the population in such a way that every set of n individuals has an equal chance to be in the sample actually selected.

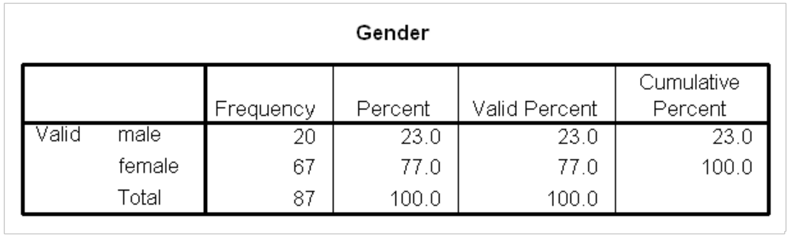
Examples

- Sample data set
  - Questionnaire filled in by a group of pharmacy students
  - Blood pressure of 20 healthy women
  - …

- Population
  - Pharmacy students
  - Students
  - Blood pressure of women (whoever)
  - …

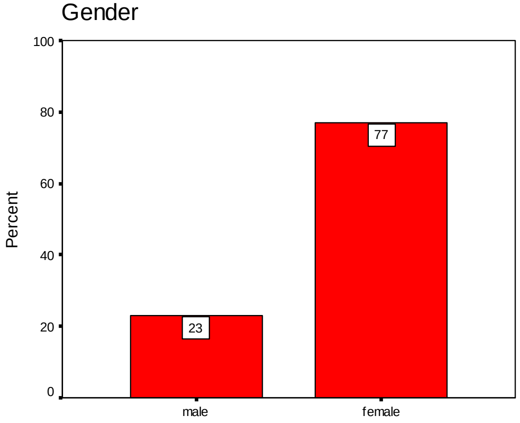
Sample

- **Bar chart of relative frequencies of a categorical variable**



Gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	male	20	23.0	23.0	23.0
	female	67	77.0	77.0	100.0
	Total	87	100.0	100.0	



Population

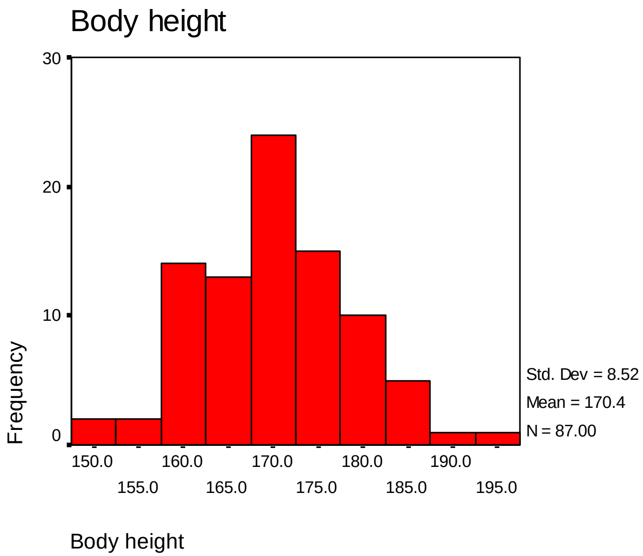
∼ (approximates)

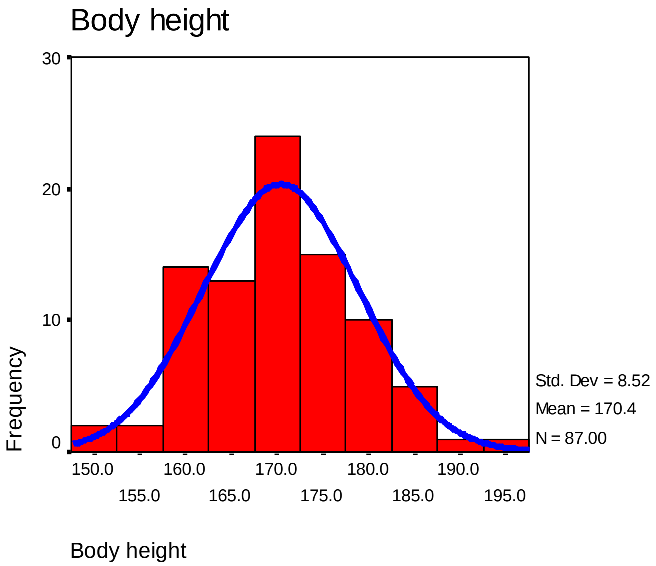
- **Distribution of that variable in the population**

Sample ~ Population

(approximates)

- **Histogram of relative frequencies of a continuous variable**
- **Distribution of that variable in the population**





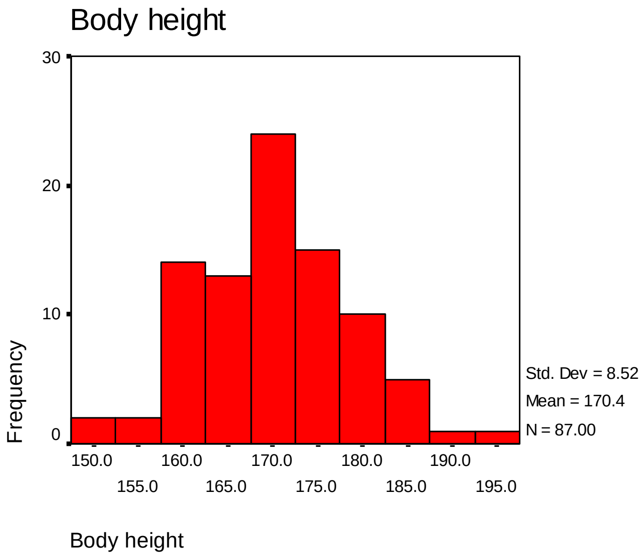
Sample

- **Mean ($\bar{x}$)**
- **Standard deviation (SD)**
- **Median**

∼ (approximates)

Population

- **Mean μ (unknown)**
- **Standard deviation σ (unknown)**
- **Median (unknown)**



Random variables, probability distributions (distribution of the population)

Random variables, probability distributions

- **A random variable** is a variable whose value is a numerical outcome of a random phenomenon.
- Notation: X, Y, ..
- **Examples**
  - The experiment is tossing a coin.
    - X(H)=1 and X(T)=2
    - Y(H)=-10 and Y(T)=10
  - The experiment is rolling a dice. Ω={1,2,3,4,5,6}. Let define the X to be the number shown on the dice.

Distribution of a discrete (categorical) random variable

- A discrete random variable *X* has finite number of possible values

- **The probability distribution of *X* lists the values and their probabilities:**

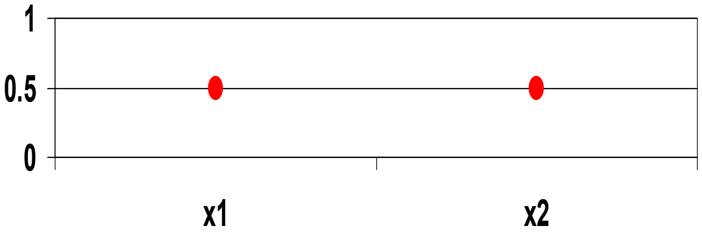
Value of X:	x_1	x_2	x_3	…	x_n
Probability:	p_1	p_2	p_3	…	p_n

$$p_i \geq 0,\ \ p_1 + p_2\ \ + p_3 \ldots + p_n = 1$$

Examples

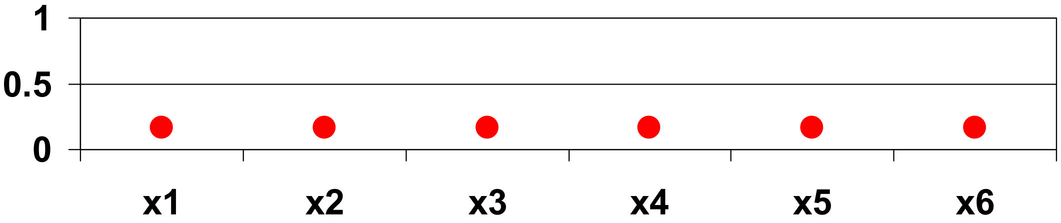
- The experiment is tossing a coin.

  $p_1 = 0.5, p_2 = 0.5$



- The experiment is rolling a dice.

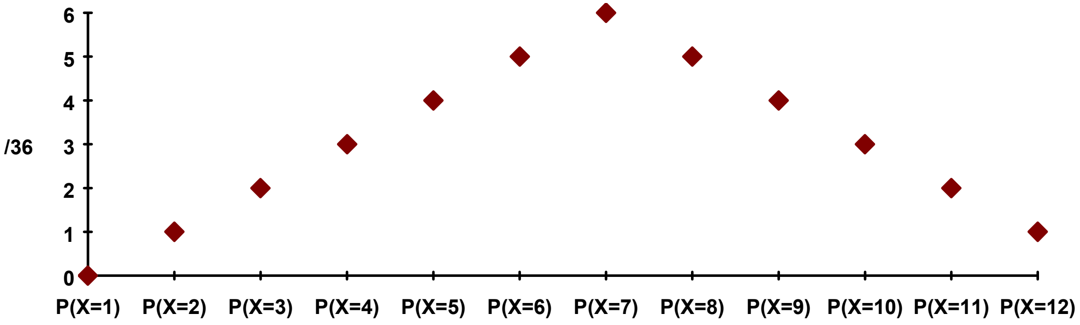
  $p_1 = 1/6, p_2 = 1/6, \ldots, p_6 = 1/6$



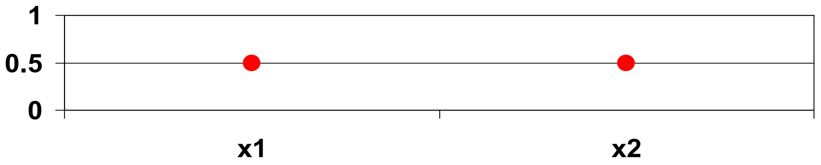
Example: rolling two dices

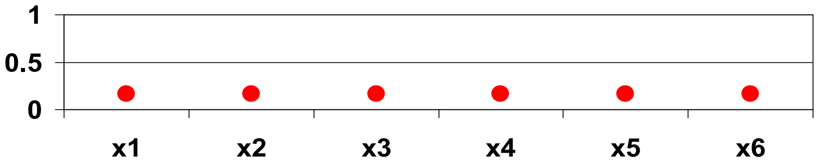
- Let random variable X be the sum of the two numbers shown on the two dices.
- P(X=1)=0, (X=1 is impossible)
- P(X=2)=1/36 (the only favourable event is (1,1), and the number of all possible event is 36.)
-

	j=1	j=2	j=3	j=4	j=5	j=6
i=1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
X	**2**	**3**	**4**	**5**	**6**	**7**
i=2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
X	**3**	**4**	**5**	**6**	**7**	**8**
i=3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
X	**4**	**5**	**6**	**7**	**8**	**9**
i=4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
X	**5**	**6**	**7**	**8**	**9**	**10**
i=5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
X	**6**	**7**	**8**	**9**	**10**	**11**
i=6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)
X	**7**	**8**	**9**	**10**	**11**	**12**

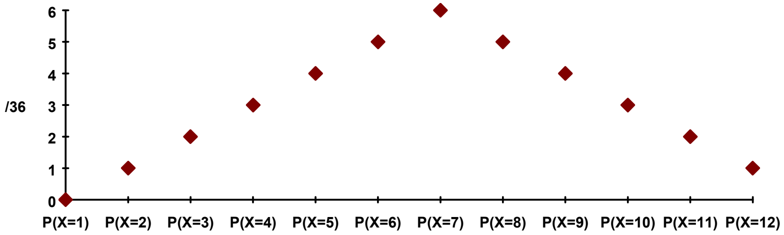


Uniform discrete distributions: all p_i-s are equal





Not uniform

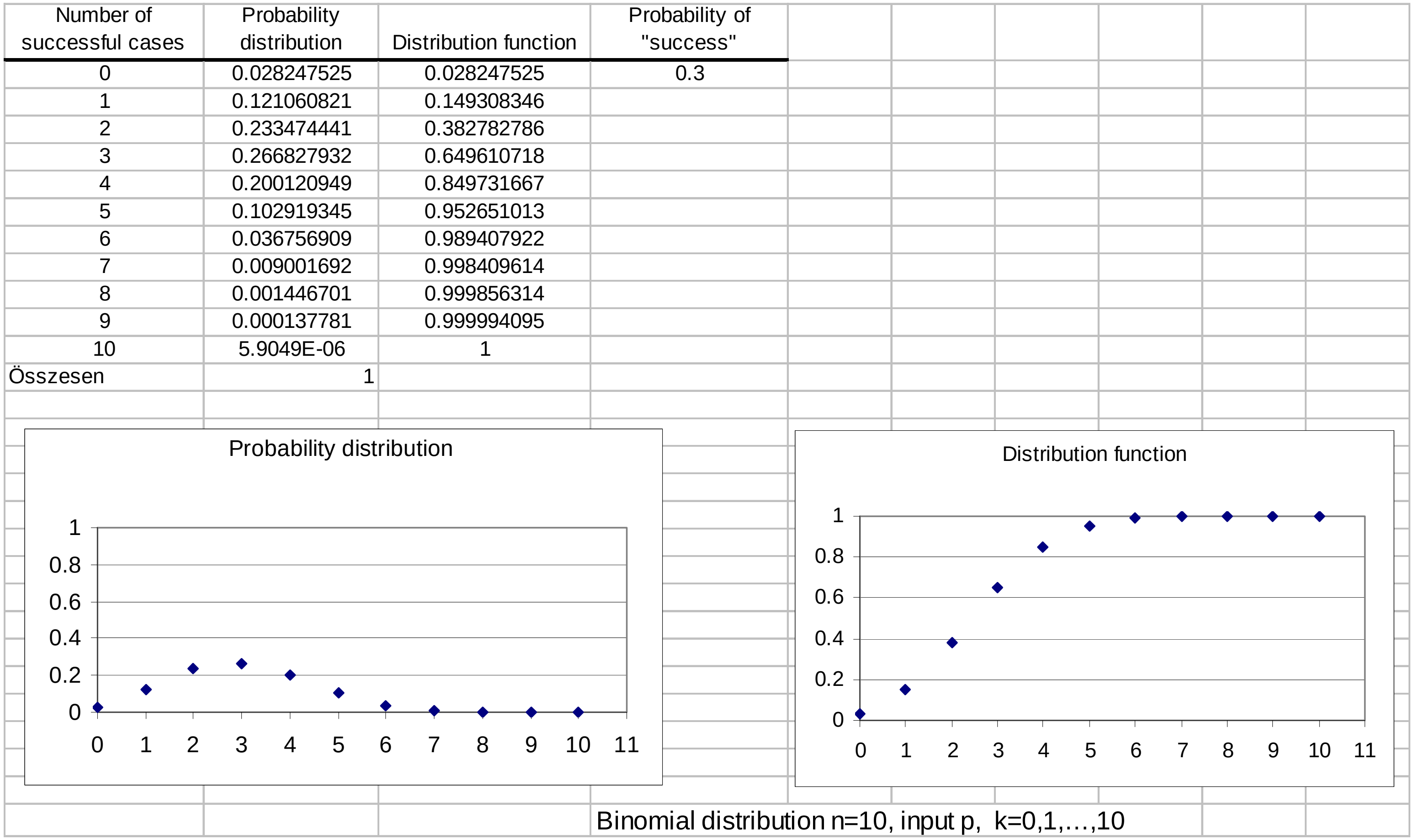


The binomial distribution

- Let's consider an experiment ***A*** that may have only two possible mutually exclusive outcomes (success, failure)
- Let P(***A***)=*p*
- We now repeat the experiment *n* times, let X denote the absolute frequency of the event ***A***.
- The probability that ***X*** will assume any given possible value *k* is expressed by the binomial formula

$$P_k = P(X = k) = \binom{n}{k} p^k q^{n-k}, \quad k = 0,1,\ldots,n$$

Example



Number of successful cases	Probability distribution	Distribution function	Probability of "success"
0	0.028247525	0.028247525	0.3
1	0.121060821	0.149308346	
2	0.233474441	0.382782786	
3	0.266827932	0.649610718	
4	0.200120949	0.849731667	
5	0.102919345	0.952651013	
6	0.036756909	0.989407922	
7	0.009001692	0.998409614	
8	0.001446701	0.999856314	
9	0.000137781	0.999994095	
10	5.9049E-06	1	
Összesen	1		

Binomial distribution n=10, input p, k=0,1,…,10

In a certain population the occurrence of some disease is p=0.3.
What is the probability that examining n=10 patients, there will be exactly k=4 diseased?
According to the formula, *P*(*X*=4)=10!/(4!6!)(0.3)4(0.7)6=210 0.0081 0.117649= 0.200120949.

Continuous random variable

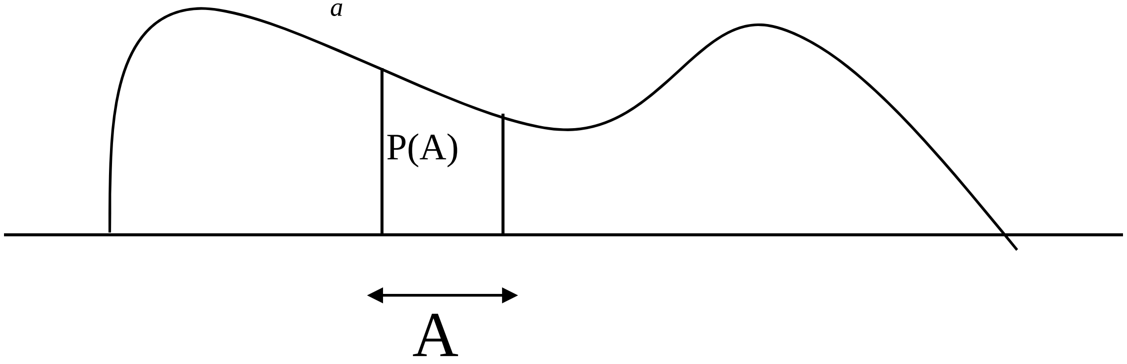
- A continuous random variable *X* has takes all values in an interval of numbers.

- **The probability distribution of *X*** is described by a **density curve.**
- The density curve
  - is on the above the horizontal axis, and
  - has area exactly 1 underneath it.

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

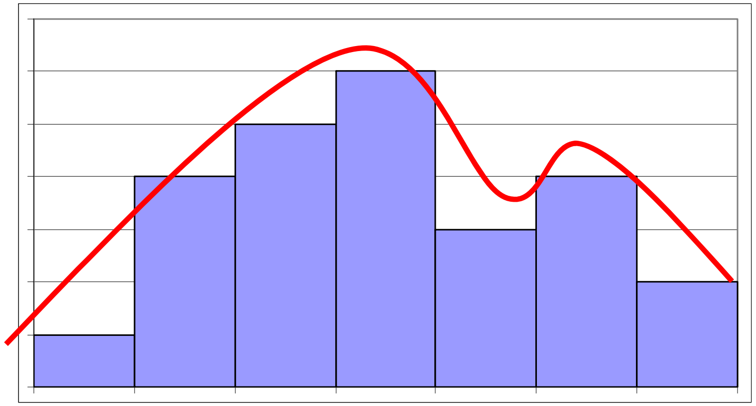
- The probability of any event is the area under the density curve and above the values of X that make up the event.

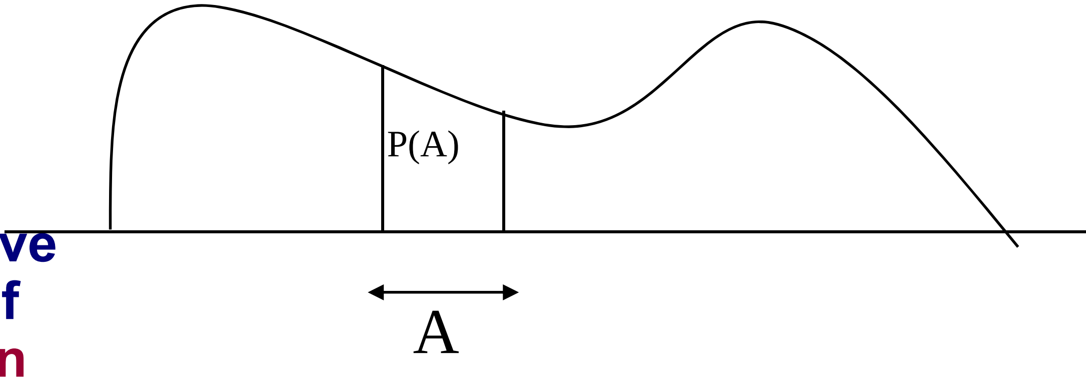
$$\int_{a}^{b} f(x)dx = P(a < x < b)$$



The density curve

- The density curve is an idealized description of the overall pattern of a distribution that smoothes out the irregularities in the actual data.
- The density curve
  - is on the above the horizontal axis, and
  - has area exactly 1 underneath it.
  - The area under the curve and above any range of values is the proportion of all observations that fall in that range.





The density curve

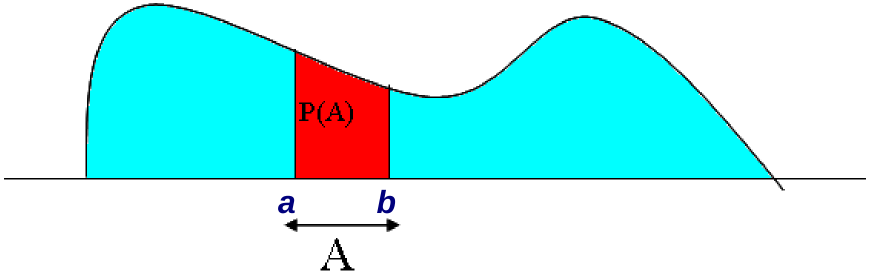
- The density curve

1. is on the above the horizontal axis: f(x) ≥0

2. has area exactly 1 underneath it.

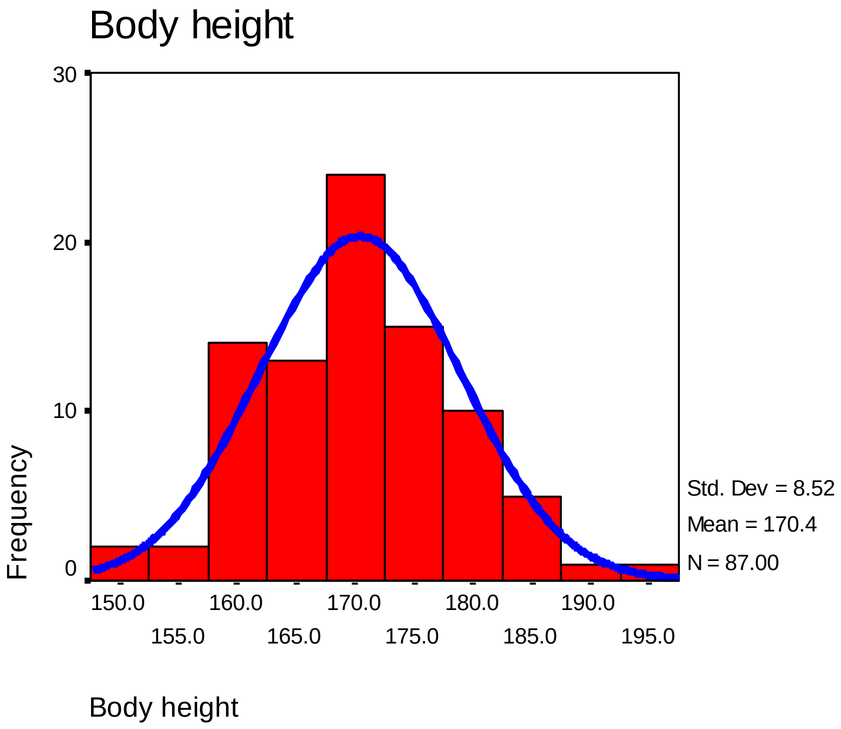
$$\int_{-\infty}^{\infty} f(x)dx = 1$$

3. The area under the curve and above any range of values is the proportion of all observations that fall in that range.



$$\int_{a}^{b} f(x)dx = P(a \le x < b)$$

Special density curve: Normal distribution

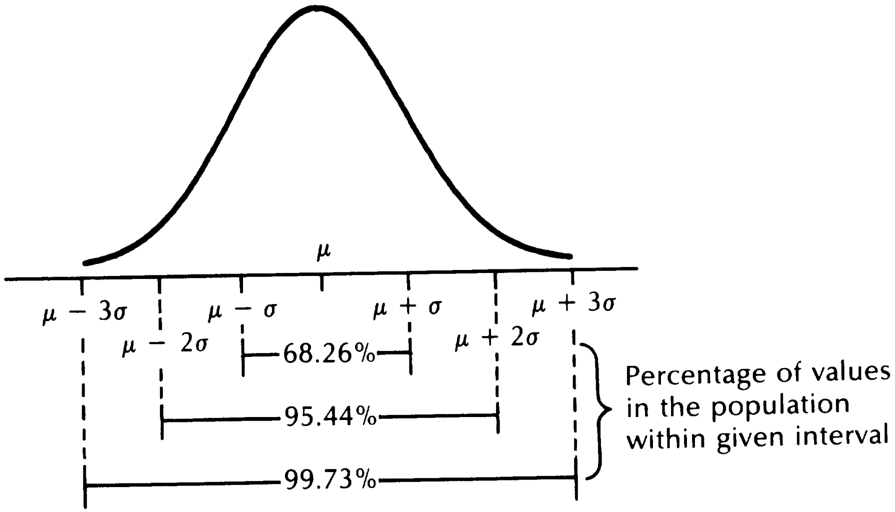


Special distribution of a continuous variable:
The normal distribution

- The normal curve was developed mathematically in 1733 by DeMoivre as an approximation to the binomial distribution. His paper was not discovered until 1924 by Karl Pearson. Laplace used the normal curve in 1783 to describe the distribution of errors. Subsequently, Gauss used the normal curve to analyze astronomical data in 1809.
- The normal curve is often called the Gaussian distribution.
- The term bell-shaped curve is often used in everyday usage.

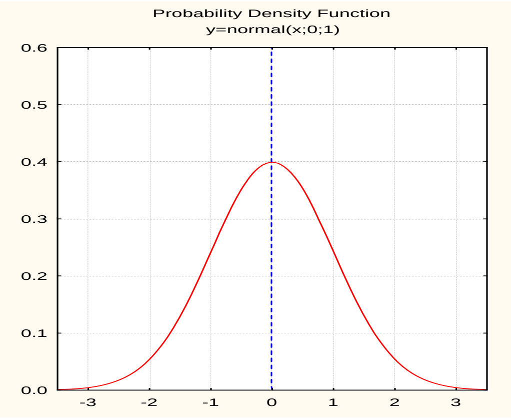
Special distribution of a continuous variable: The normal distribution $N(\mu,\sigma^2)$

- The density curves are symmetric, single-peaked and bell-shaped
- The 68-95-99.7 rule . In the normal distribution with mean μ and standard deviation σ:
  - 68% of the observations fall within σ of the mean μ
  - 95% of the observations fall within 2σ of the mean μ
  - 99.7% of the observations fall within 3σ of the mean μ

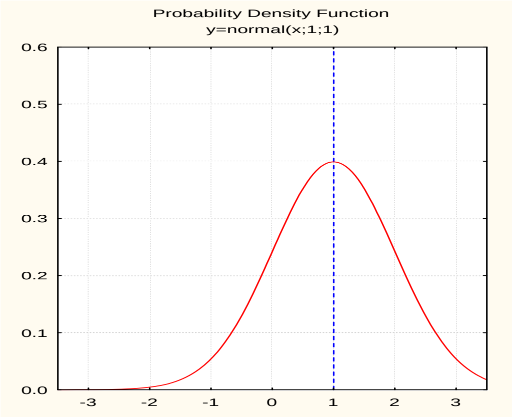


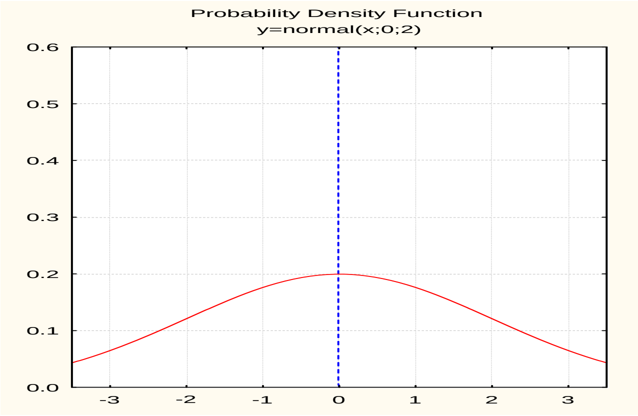
Normal distributions N(μ, σ ^{2})

N(0,1)



N(1,1)



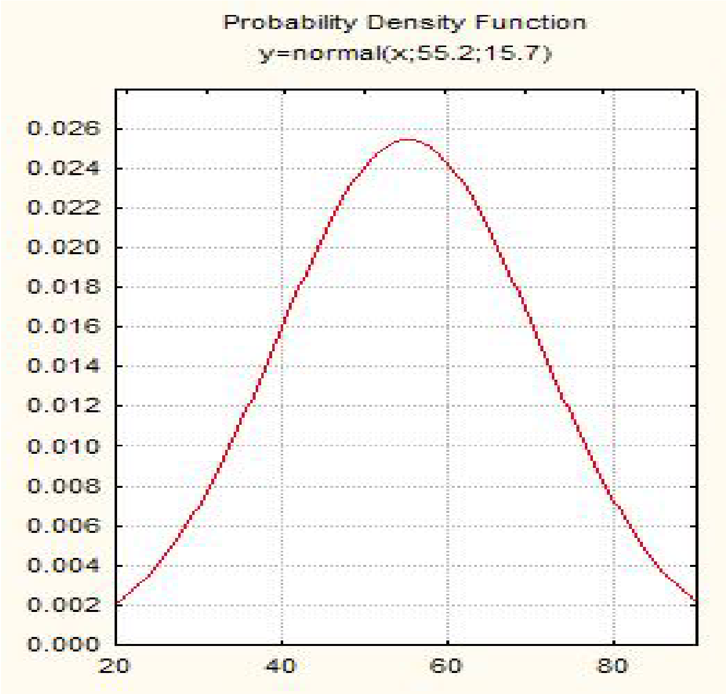


N(0,2)

μ, σ : parameters
(a **parameter** is a number that describes the distribution)

„Imagination” of the distribution given the sample mean and sample SD – supposing normal distribution

- In the papers generally sample means and SD-s are published. We use these value to know the original distribution
- For example,
  age in years: 55.2 ± 15.7



39.5

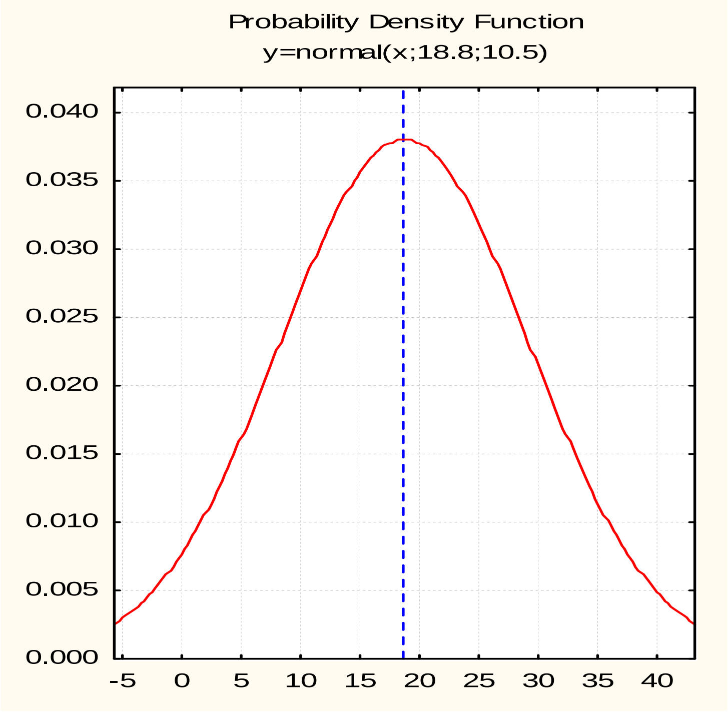
70.9

68.26% of the data are supposed to be in this interval

95.44% of the data are supposed to be in the interval (23.8, 86.6)

Stent length per lesion (mm): 18.8 ± 10.5

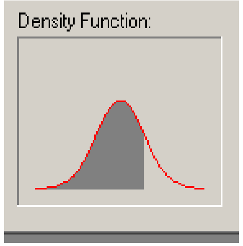
Using these values as parameters, the „normal" distribution would be like this:

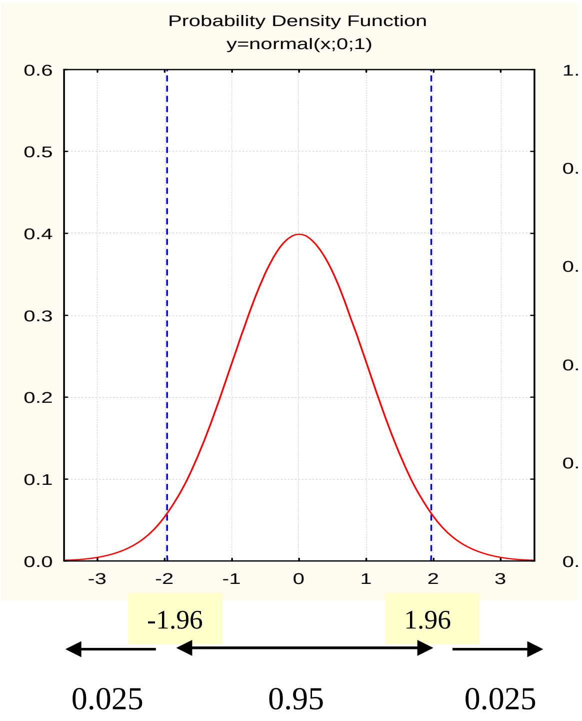


The original distribution of stent length was probably skewed, or if the distribution was normal, there were negative values in the data set.

Standard normal probabilities

x	(x): proportion of area to the left of x
-4	0.00003
-3	0.0013
-2.58	0.0049
-2.33	0.0099
-2	0.0228
-1.96	0.0250
-1.65	0.0495
-1	0.1587
0	0.5
1	0.8413
1.65	0.9505
1.96	0.975
2	0.9772
2.33	0.9901
2.58	0.9951
3	0.9987
4	0.99997

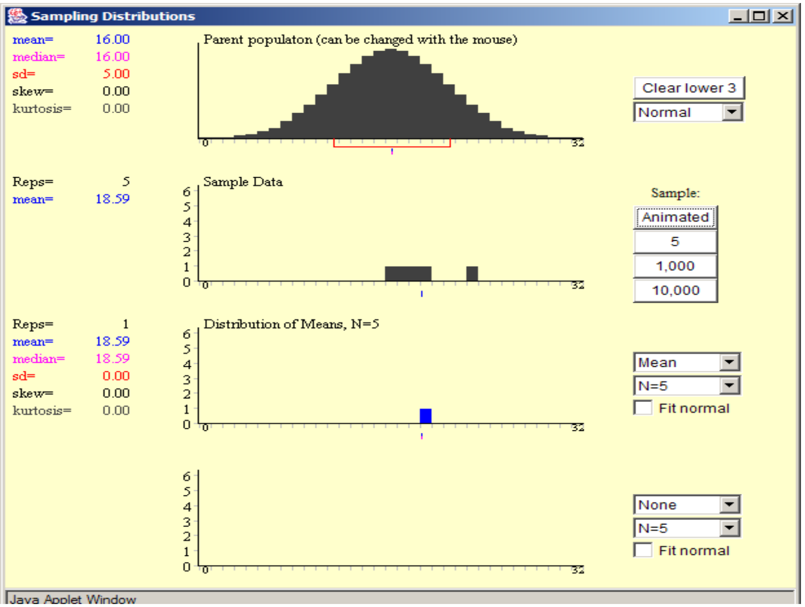




The central limit theorem

Distribution of sample means

http://onlinestatbook.com/stat_sim/sampling_dist/index.html

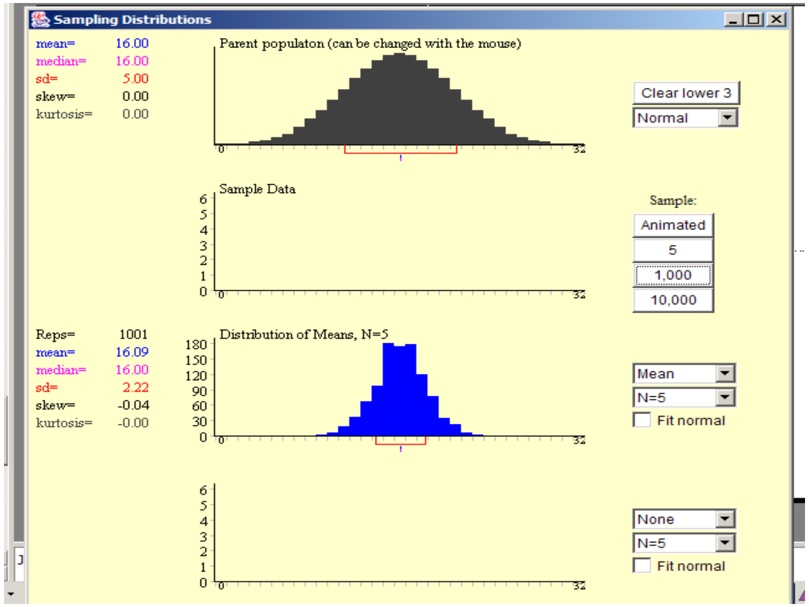


```
mean=      16.00
median=    16.00
sd=         5.00
skew=       0.00
kurtosis=   0.00

Reps=          5
mean=      18.59

Reps=          1
mean=      18.59
median=    18.59
sd=         0.00
skew=       0.00
kurtosis=   0.00
```

Normal

Mean ▼ N=5 ▼ □ Fit normal

None ▼ N=5 ▼ □ Fit normal



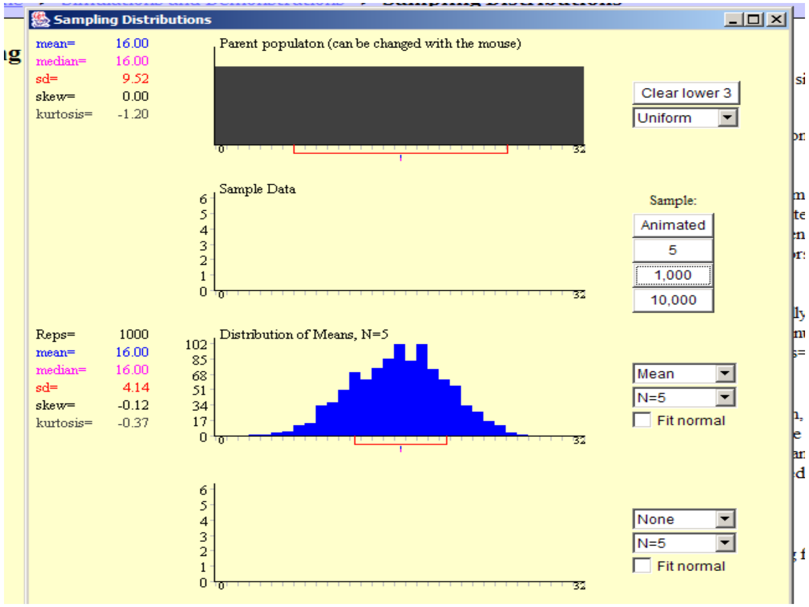
```
mean=      16.00
median=    16.00
sd=         5.00
skew=       0.00
kurtosis=   0.00

Reps=       1001
mean=      16.09
median=    16.00
sd=         2.22
skew=      -0.04
kurtosis=  -0.00
```

Normal

Mean ▼ N=5 ▼ □ Fit normal

None ▼ N=5 ▼ □ Fit normal

The population is not normally distributed



Sampling Distributions

mean= 16.00
median= 16.00
sd= 9.52
skew= 0.00
kurtosis= -1.20

Parent populaton (can be changed with the mouse)

Uniform

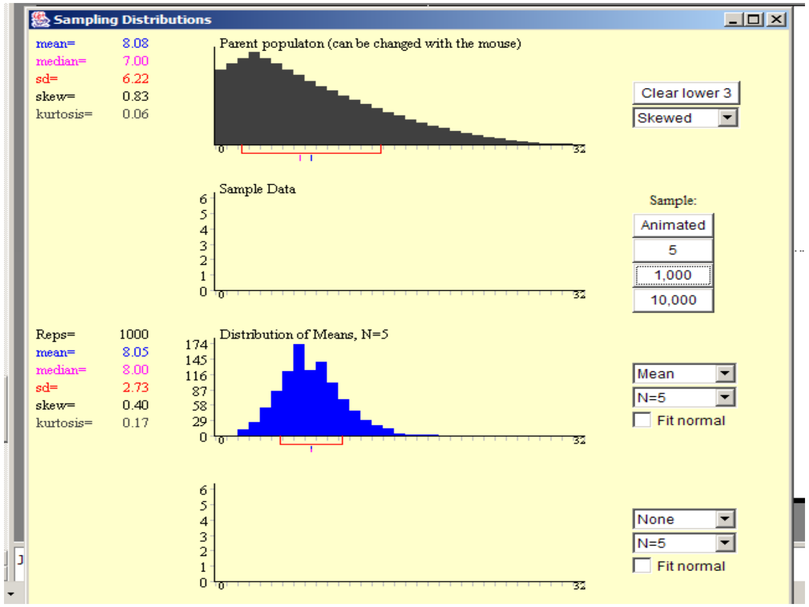
Sample Data

Reps= 1000
mean= 16.00
median= 16.00
sd= 4.14
skew= -0.12
kurtosis= -0.37

Distribution of Means, N=5

Mean
N=5
□ Fit normal

None
N=5
□ Fit normal



Sampling Distributions

mean= 8.08
median= 7.00
sd= 6.22
skew= 0.83
kurtosis= 0.06

Parent populaton (can be changed with the mouse)

Skewed

Sample Data

Reps= 1000
mean= 8.05
median= 8.00
sd= 2.73
skew= 0.40
kurtosis= 0.17

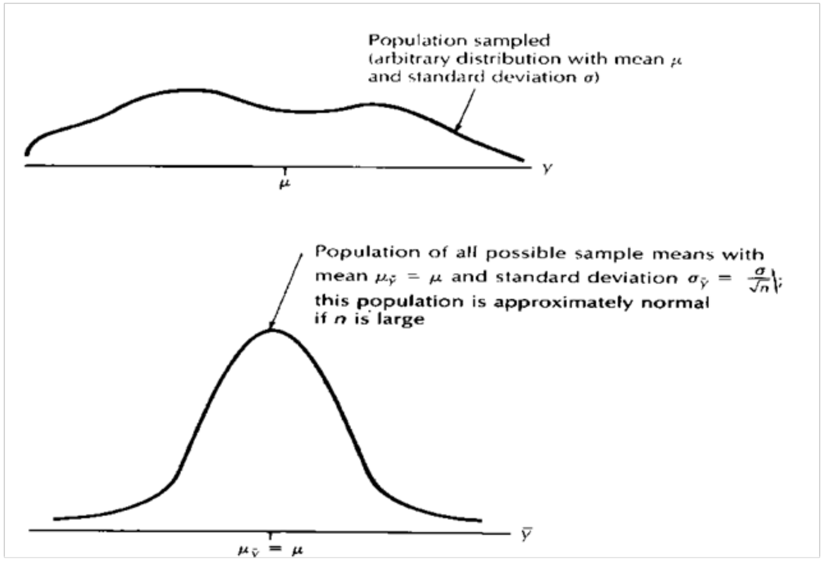
Distribution of Means, N=5

Mean
N=5
□ Fit normal

None
N=5
□ Fit normal

The central limit theorem

- If the sample size *n* is large (say, at least 30), then the population of all possible sample means approximately has a normal distribution with mean µ and standard deviation $\frac{\sigma}{\sqrt{n}}$ no matter what probability describes the population sampled



Population sampled
(arbitrary distribution with mean μ
and standard deviation σ)

Population of all possible sample means with mean $\mu_{\bar{y}} = \mu$ and standard deviation $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$;
this population is approximately normal
if *n* is large

Tha standard error of mean (SE or SEM)

- $\frac{\sigma}{\sqrt{n}}$ is called the standard error of mean
- Meaning: the dispersion of the sample means around the (unknown) population mean.

Calculation of the standard error from the standard deviation when σ is unknown

- Given $x_1, x_2, x_3, \dots, x_n$ statistical sample, the stadard error can be calculated by

$$SE = \frac{SD}{\sqrt{n}} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n(n-1)}}$$

- It expresses the dispersion of the sample means around the (unknown) population mean.

Review questions

- What is the concept of probability?
- What is the formula for computing simple exact probabilities? When can it be used?
- What is the distribution of a discrete variable?
- What are the properties of a discrete distribution?
- What is the distribution of a continuous variable?
- What are the properties of the density function?
- What is the uniform distribution?
- What is the binomial distribution?
- What is the normal distribution?
- Parameters of the normal distribution.
- Properties of the normal distribution.

Problems

1. If we roll a dice, there are 6 possible outcomes. If X represents the value of the outcome, find the following probabilities: a) $P(X=1)$;b) $P(X>1)$; c) $P(1<X<4)$
2. A fair coin is tossed twice. List the possible outcomes? What is the probability of getting two tails?
3. For a standard normal distribution, find the following probabilities using the standard normal table:
   - $P(X<0)$=….
   - $P(X>0)$=…..
   - $P(X<1)$=….
   - $P(X>1)$=…..
   - $P(X<-1)$=…..
   - $P(-1<X<1)$=………

Useful WEB pages

- http://onlinestatbook.com/rvls/index.html
- http://my.execpc.com/~helberg/statistics.html
- http://www.regentsprep.org/Regents/math/algtrig/ATS2/NormalLesson.htm