Mathematical statistics

The aim of the lectures

Familiarize students with the basics of statistics and the main methods of analyzing data from observation and experimental research. The purpose of laboratory classes is to develop the skills to work independently and freely when describing. analyzing and interpreting data as well as the ability to select statistical methods necessary for proper inference based on existing data set.

Range of the course

Types of random variables (categorical distribution or continuous distribution) and their distributions and cumulative distribution with the focus on the normal distribution and its standardization. Statistical population.

The estimation of distribution parameters: the point estimations and confidence intervals for following population parameters: mean (expected value), fraction, variation, difference between two means, two fraction difference, variances ratio.

Principles of statistical inference. Statistical hypothesis. Testing the statistical hypothesis (Significance tests). Verification of the population hypothesis related to the previously mentioned point estimators.

Range of the course

Statistical analysis of data from factor experiments - one-factor experiment in a completely random system. Method of analysis of variance. Multiple average value comparison procedures (object grouping). Chi-square test of compliance and independence. Relations between quantitative traits - correlation analysis and regression analysis.

Statistics

Statistics is the science of obtaining (data collection. the design of surveys and experiments) organizing, analyzing, interpreting and presenting the data describing reality.

Statistics provides a set of tools i.e. proven methods that will operate on large data sets. The elements of the statistical population are called statistical units or observations. and the tested feature is a statistical variable.

Populations

Due to the size of the collection. the populations can be divided into:

- finite populations - e.g. voivodeships in Poland (a specific number at a given time that does not change)
- infinite populations - in fact they rather do not exist but it is often assumed that for example, plants of a certain species represent an infinite population because theoretically their number can be constantly increased

Random variables

- discrete which take a finite number of values, usually the values are integers from a certain range (e.g. the number of points on the dice, the number of people in the family, the number of flowers on the plant, etc.)
- continuous. i.e. those that take infinitely many values e.g. all real numbers from a certain interval (examples: human height, sugar content in apples, air temperature). Often. such variables are given with a certain accuracy resulting from the limitations of measuring instruments (e.g. a thermometer, balance, etc.)

Distributions

Typical distributions of step random variables:

- two-point distribution.
- binomial (Bernoulli) distribution.
- Poisson distribution.
- Typical distributions of continuous random variable:
- uniform distribution.
- normal distribution.

Empirical distribution

Assigning in ascending order the values of observed variable and defining frequencies of their occurrence

Empirical distribution - example

20 typescript pages were checked with the following error numbers
0, 3, 1, 1, 2, 2, 0, 0, 3, 5, 0, 1, 2, 2, 1, 1, 0, 1, 1, 1
population - 20 type written pages
variable - number of errors on the page
discrete variable

Empirical distribution - example

the number of errors xi	number of pages ni	page frequency fi
0	5	0.25
1	8	0.4
2	4	0.2
3	2	0.1
4	0	0
5	1	0.05
suma	20	1

Empirical distribution

Continuous variables
Class ranges
The interval $h_i = x_{1i} - x_{0i}$
Middle of the interval $x_i = (x_{0i} - x_{1i}) / 2$
Number of intervals. usually $5 \leq k \leq 20$. $k \leq 5 \log n$ - size of the population

Empirical distribution

The service time at the cash register of 25 randomly selected customers was measured. obtaining the following data (time in s):
15 37 34 9 61 24 56 52 6 35 21 46 86 40 74 39 48 55 73 92 43 78 67 30 29

customer service time (s) y0i-y1i	number of clients ni	frequency of clients
0-20	3	0.12
20-40	9	0.36
40-60	6	0.24
60-80	5	0.2
80-100	2	0.08
suma	25	1

Empirical cumulative distribution function

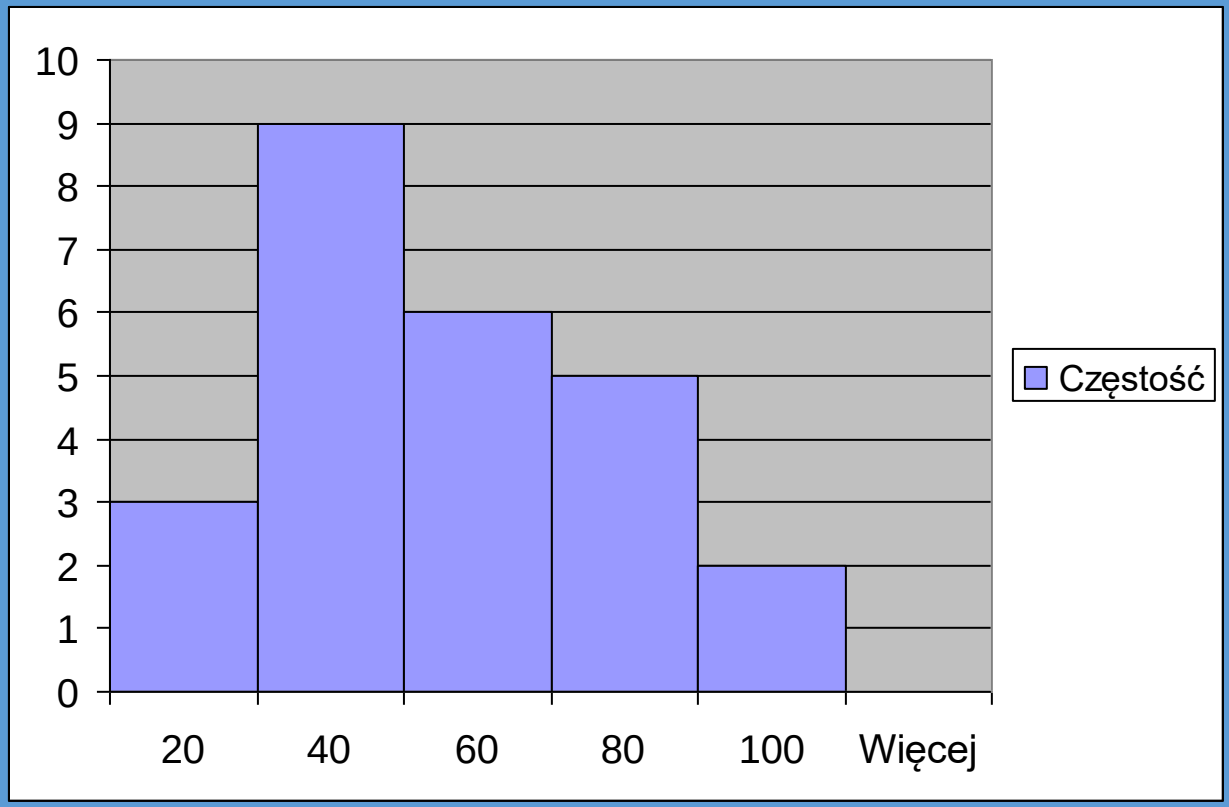
Presentation of the empirical distribution of a feature by means of cumulative relative frequencies

The empirical cumulative distribution function is a non-decreasing function. limited to the <0.1> range

$$F(x) = \left\{ \begin{array}{c} 0, x < x_1 \\ \sum_{s=1}^{i} w_s, x_i \le x < x_{i+1} \\ 1, x \ge x_k \end{array} \right\}$$

the number of errors xi	Cumulative number of pages ni	Empirical cumulative distribution Fn(xi)
0	5	0.25
1	13	0.65
2	17	0.85
3	19	0.95
4	19	0.95
5	20	1

Histogram



Measures of the location of the distribution

Mean

Median of the empirical distribution

Modal value (dominant)

Mean

For $x_1 . \ldots . x_n$ mean is defined as:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Median of the empirical distribution

value separating the higher half from the lower half of a data sample, a population or a probability distribution

$$m_e = \begin{cases} x_{\frac{n+1}{2}} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) \end{cases}$$

Median of empirical distribution - an example

From the general population a $n = 50$-element sample was taken and tested for variable X. Find the median.

3.6	5.6	4.6	5.9	4.7
5.0	3.5	5.1	4.2	6.4
4.0	5.4	4.7	6.4	5.1
4.7	5.2	3.0	5.3	3.4
5.2	4.1	5.5	4.5	5.2
5.9	5.0	6.1	4.9	6.2
4.5	3.1	3.8	4.0	4.4
5.3	5.8	4.9	5.2	4.3
5.5	4.8	5.6	3.3	5.8
3.9	4.4	6.1	5.4	3.7

Median of empirical distribution - an example

The sample in a non-decreasing order:

3.0	3.1	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4.0	4.0	4.1	4.2	4.3	4.4	4.4	4.5	4.5	4.6	4.7	4.7	4.7	4.8	4.9	4.9
5.0	5.0	5.1	5.1	5.2	5.2	5.2	5.2	5.3	5.3	5.4	5.4	5.5	5.5	5.6	5.6	5.8	5.8	5.9	5.9	6.1	6.1	6.2	6.4	6.4

x_{25}=4.9
x_{26}=5.0

$$m_e = \frac{4.9 + 5}{2} = 4.95$$

Mode

The **mode** is the value that appears most often in a set of data values

Mode – example

sample I: 16. 13. 15. 17. 16. 16. 15. 14. 12. 17. 16. 18. 14. 15. 17. 16

sample II: 27. 24. 28. 24. 25. 23. 29. 26. 29. 25

12	1
13	1
14	2
15	2
16	5
17	2
18	1

23	1
24	2
25	2
26	1
27	2
28	1
29	2

mode: sample I – 16. sample II – no mode

Measures of dispersion

Variance

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2$$

Coefficient of variation

$$V = \frac{s}{\bar{x}}$$

Quantile

Quantiles are cut points dividing the range of a probability distribution into continuous intervals with equal probabilities. or dividing the observations in a sample in the same way.

Some q-quantiles have special names

The only 2-quantile is called the median

The 3-quantiles are called tertiles or terciles → T

The 4-quantiles are called quartiles → Q; the difference between upper and lower quartiles is also called the interquartile range. midspread or middle fifty → IQR = Q3 − Q1

The 5-quantiles are called quintiles → QU

The 6-quantiles are called sextiles → S

The 7-quantiles are called septiles

The 8-quantiles are called octiles

The 10-quantiles are called deciles → D

The 12-quantiles are called duo-deciles or dodeciles

The 16-quantiles are called hexadeciles → H

The 20-quantiles are called ventiles. vigintiles. or demi-deciles → V

The 100-quantiles are called percentiles → P

The 1000-quantiles have been called permilles or milliles. but these are rare and largely obsolete

Quantile – example

The intelligence of 20 people was measured:

74, 80, 80, 85. 92, 94, 97, 98, 98, 100, 101, 101, 104, 104, 106, 109, 112, 115, 128, 137

The quantile of 0.25 (i.e. the first quartile) is 92 here. because exactly five samples (i.e. 1/4 of the population of 20 samples) have a value less than or equal to 92. The quantile of 0.75 (i.e. the third quartile) is 106 .

Quantile – example

Find the scattering measures for the 25-element ordered sample:

2.15	2.18	2.19	2.27	2.29	2.31	2.33	2.41	2.43	2.47	2.52	2.54	2.55
2.61	2.62	2.73	2.73	2.75	2.81	2.85	2.97	3.00	3.01	3.08	3.11	

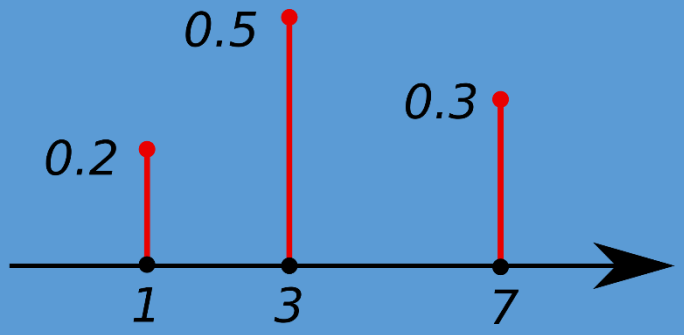
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 2.5964 .\ m_e = x_{13} = 2.55$$

$$Q_1 = x_7 = 2.33$$

$$Q_3 = x_{19} = 2.81$$

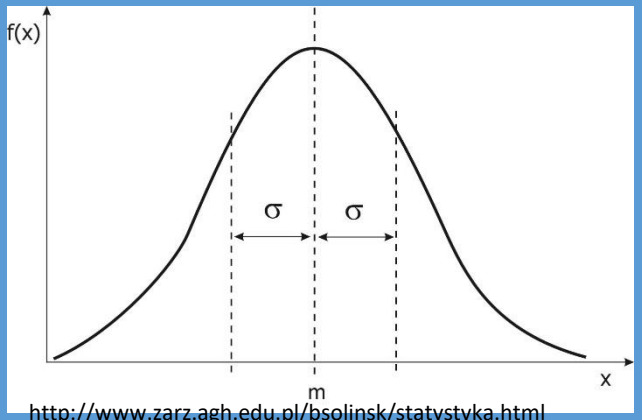
Types of random variables, their distributions and cumulative distributions

Discrete probability distribution



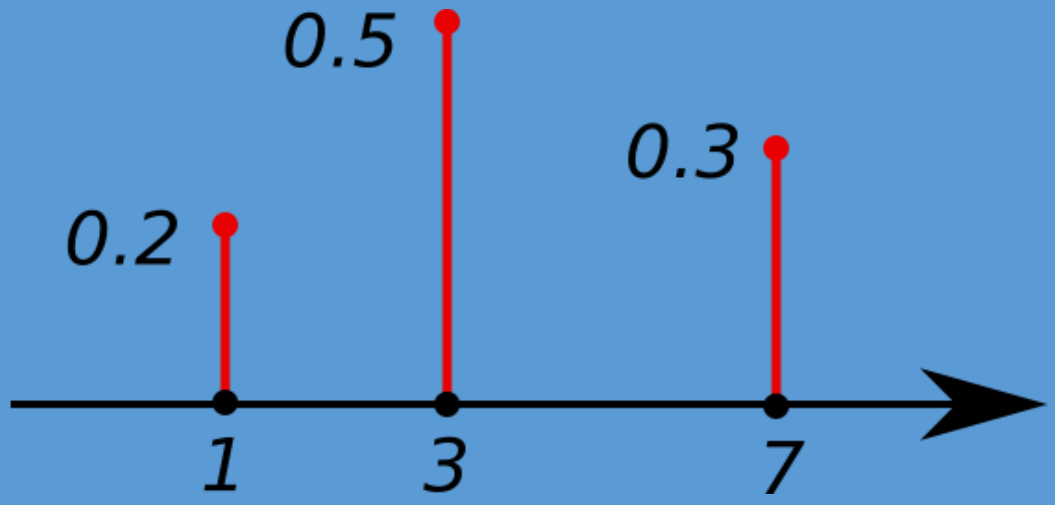
https://en.wikipedia.org/wiki/Probability_distribution

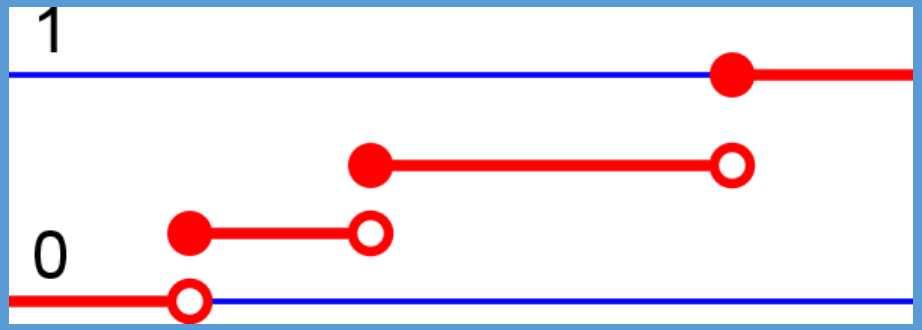
Continuous probability distribution



http://www.zarz.agh.edu.pl/bsolinsk/statystyka.html

Discrete probability distribution





http://en.wikipedia.org/wiki/Probability_distribution#Discrete_probability_distribution

Expected value and variance of a random descrete variable X

$$E(X) = x_1 p_1 + x_2 p_2 + \cdots + x_k p_k$$

$$p_1 + p_2 + \cdots + p_k = 1$$

$$E(X) = \frac{x_1 p_1 + x_2 p_2 + \cdots + x_k p_k}{p_1 + p_2 + \cdots + p_k}$$

$$Var(X) = E[(X - E[X])^2] = E[X^2 - 2XE[X] + (E[X])^2] =$$

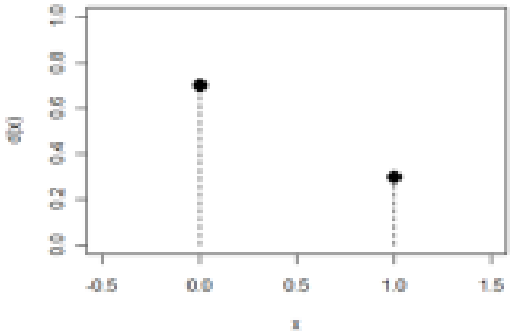
$$= E[X^2] - 2E[X]E[X] + (E[X])^2] = E[X^2] - (E[X])^2]$$

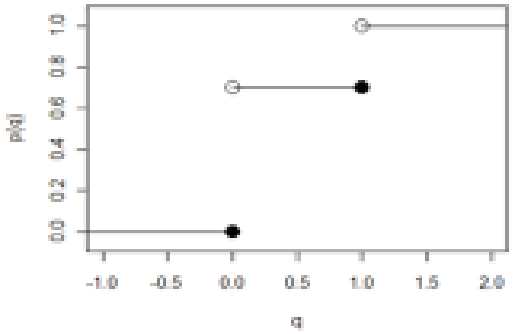
expected value and variance of a random variable *X*

X_i	1	3	7
P_i	0.2	0.5	0.3

- E(X)=3.80
- Var(X)=4.96

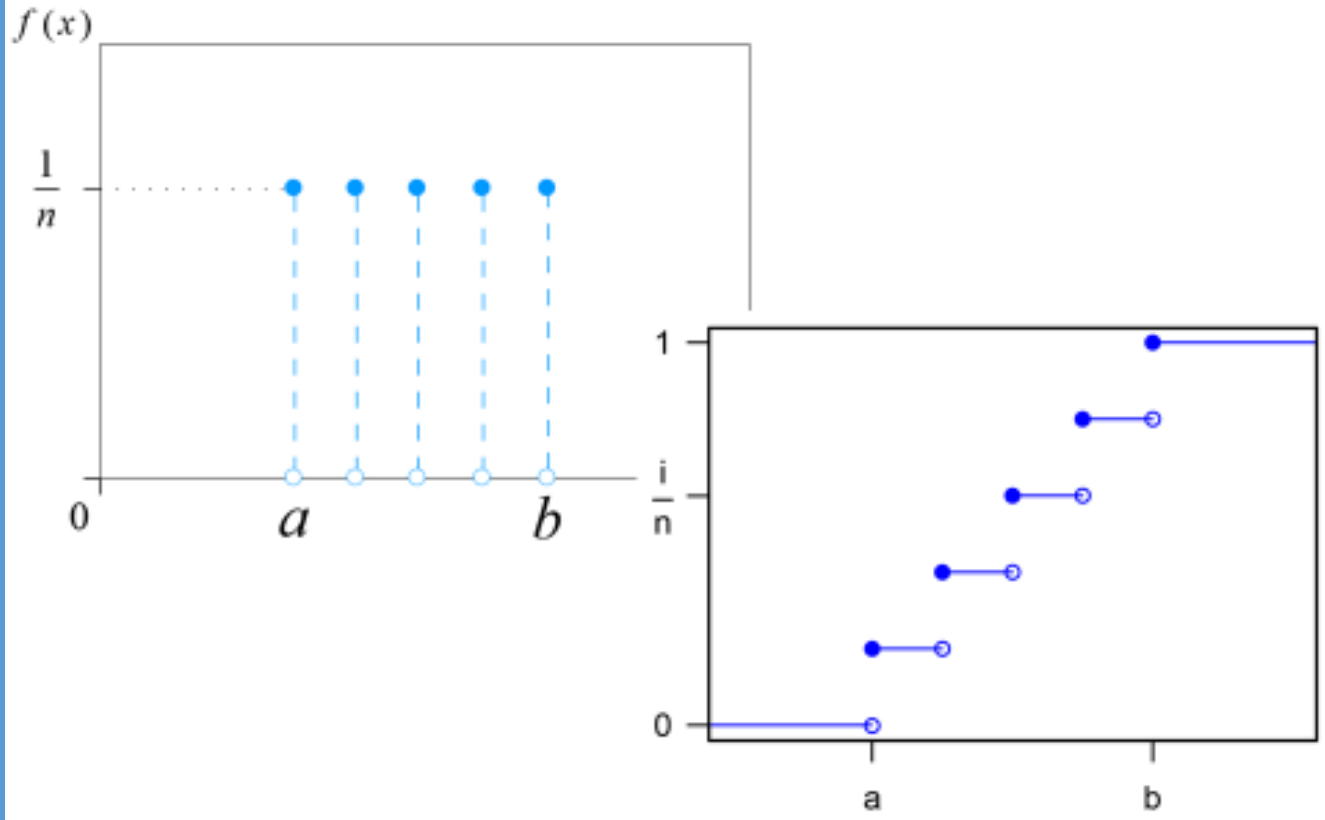
Two-point distribution



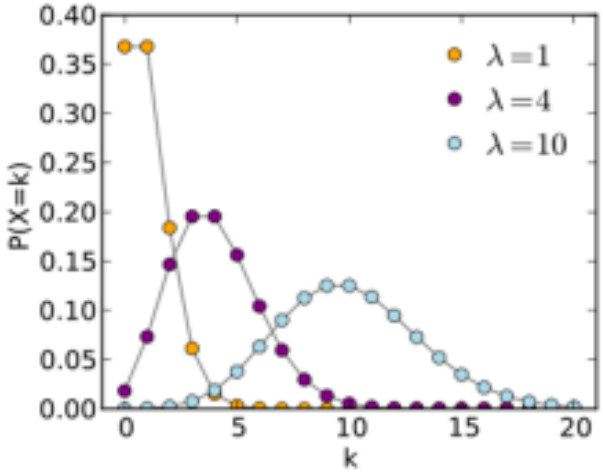


http://stat.ue.katowice.pl/stat11/dwupunktowy.html

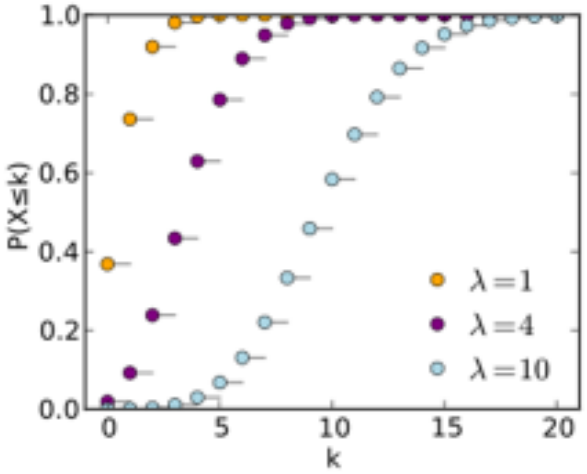
Uniform distribution



Poisson distribution

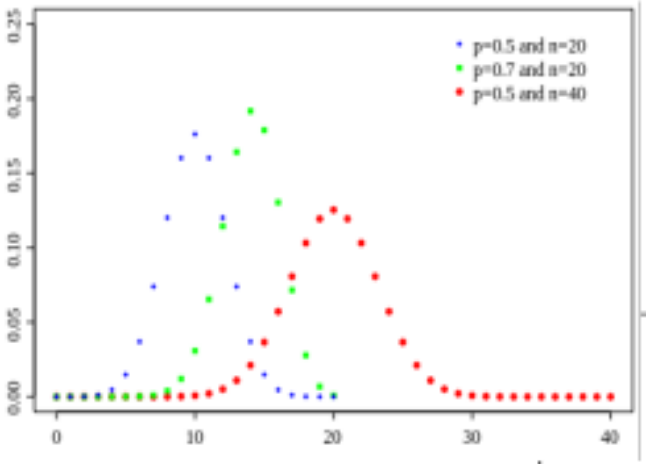


$$f(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

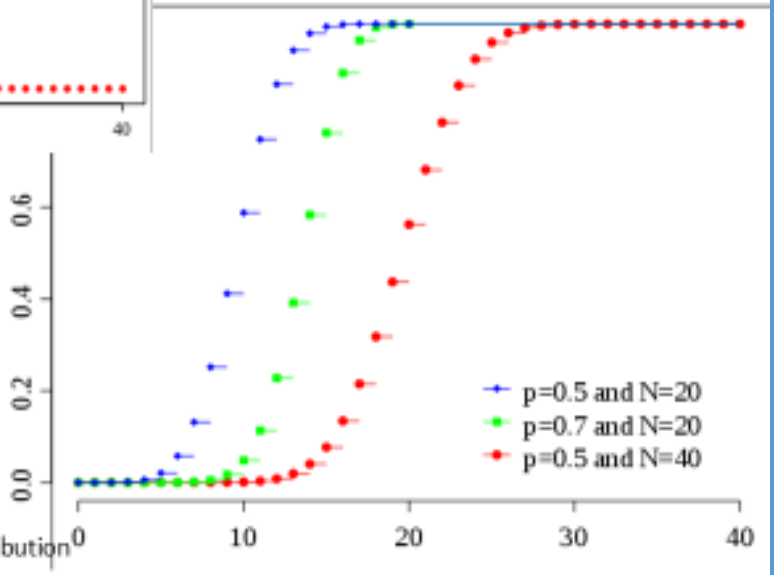


http://en.wikipedia.org/wiki/Poisson_distribution

Binomial distribution

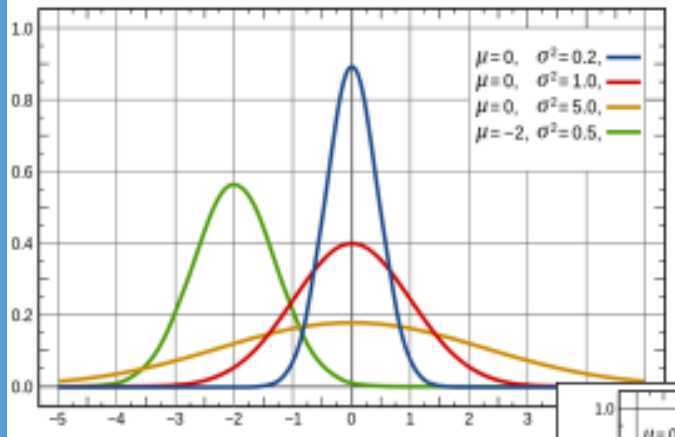


$$f(k;n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

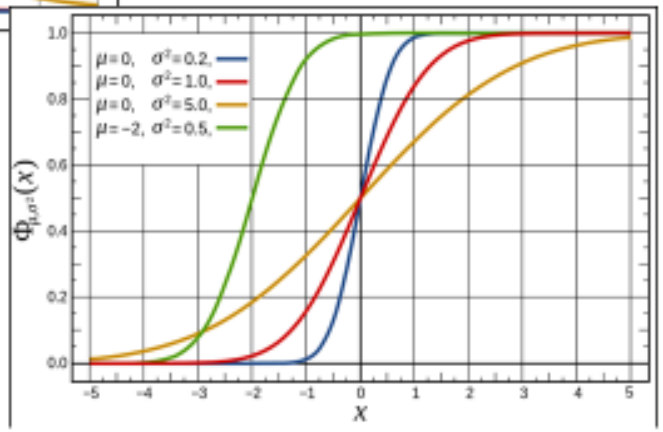


http://en.wikipedia.org/wiki/Binomial_distribution

Normal distribution

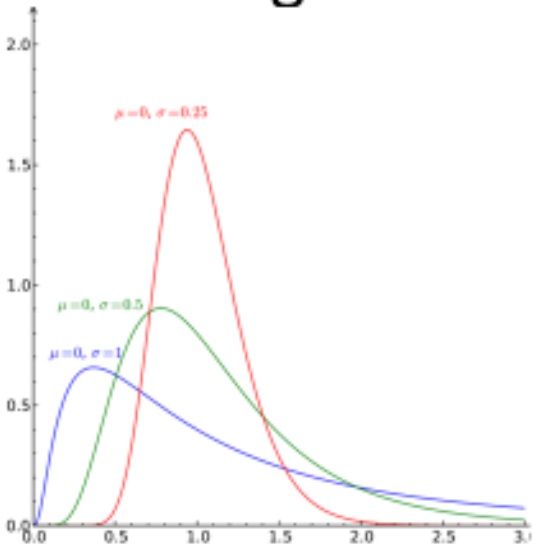


$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



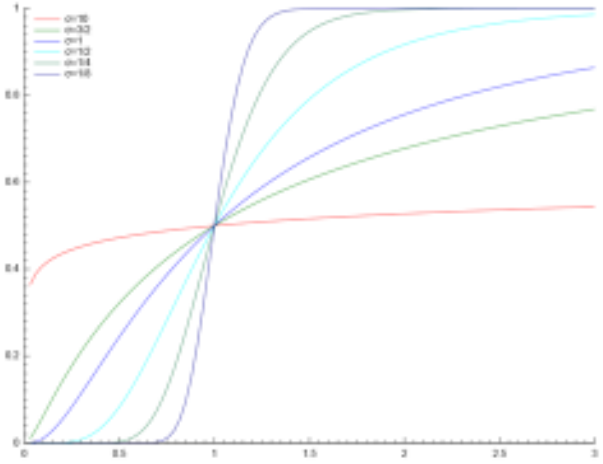
http://en.wikipedia.org/wiki/Normal_distribution

Log-normal distribution



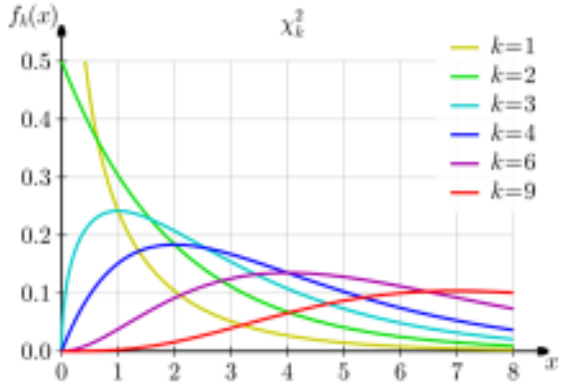
$$x \in (0, +\infty)$$

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

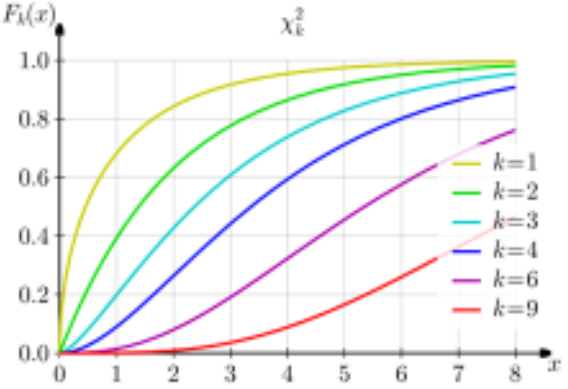


http://en.wikipedia.org/wiki/Log-normal_distribution

chi-square or χ^2-distribution

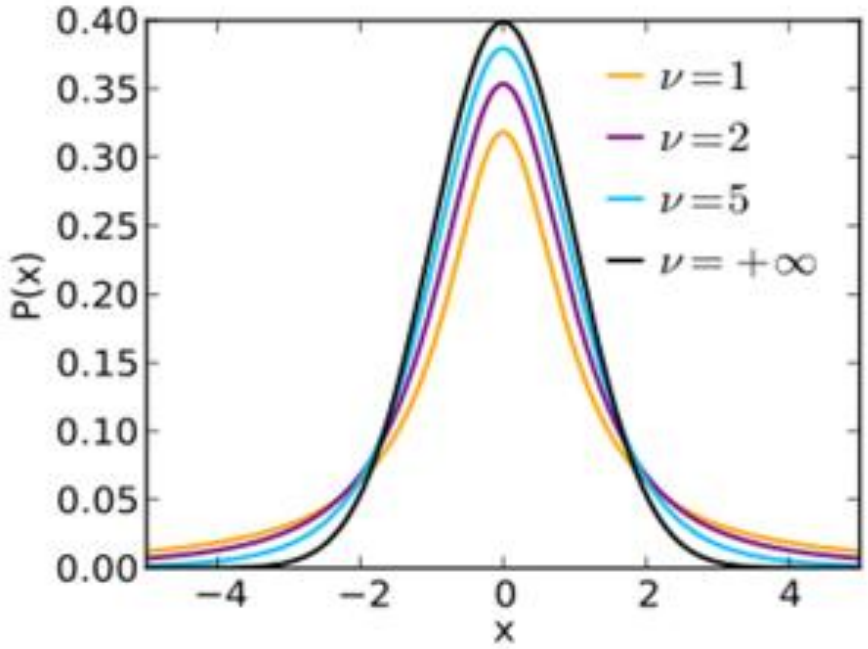


The chi-squared distribution with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables.



http://en.wikipedia.org/wiki/Chi-squared_distribution

Student's t-distribution



http://en.wikipedia.org/wiki/Student%27s_t-distribution#mediaviewer/File:Student_t_pdf.svg

Central limit theorem

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

Thank you!