Introduction to probability and statistics

Lecture 2.

Introduction

Prof. dr hab. inż. Katarzyna Zakrzewska
Katedra Elektroniki, AGH
e-mail: zak@agh.edu.pl
http://home.agh.edu.pl/~zak

References:

- D.C. Montgomery, G.C. Runger, *Applied Statistics and Probability for Engineers*, Third Edition, J. Wiley & Sons, 2003
- A. Plucińska, E. Pluciński, *Probabilistyka, rachunek prawdopodobieństwa, statystyka matematyczna, procesy stochastyczne*, WNT, 2000
- J. Jakubowski, R. Sztencel, *Wstęp do teorii prawdopodobieństwa*, SCRIPT, 2000
- M. Sobczyk, *Statystyka*, Wydawnictwo C.H. Beck, Warszawa 2010
- A. Zięba*, Analiza danych w naukach ścisłych i technice*, PWN, Warszawa 2013, 2014

Outline

- Probability and statistics - scope
- Historical background
- Paradox of Chevalier de Méré
- Statistics – type of data and the concept of random variable
- Graphical representation of data
- The role of probability and statistics in science and engineering

Probabilistic and statistical approach

Theory of probability (also **calculus of probability** or **probabilistics**) – branch of mathematics that deals with ***random events*** and ***stochastic processes***. Random event is a result of random (non-deterministic) experiment.

Random experiment can be repeated many times under identical or nearly identical while its result cannot be predicted.

Frequency of event $\frac{l}{n}$

l – number of times with the given result
n – number of repetitions

When n increases, the frequency tends to some constant value

Probabilistic and statistical approach

Probabilistics studies abstract mathematical concepts that are devised to describe **non-deterministic** phenomena:

1. random variables in the case of single events
2. stochastic processes when events are repeated in time

Big data are considered by statistics

One of the most important achievement of **modern physics** was a discovery of probabilistic nature of phenomena at microscopic scale which is fundamental to quantum mechanics.

Statistics deals with methods of data and information (numerical in nature) acquisition, their analysis and interpretation.

Statistics

DESCRIPTIVE STATISTICS

- *Arrangement of data*
- *Presentation of data*

graphical

numerical

STATISTICAL INFERENCE

Gives methods of formulating conclusions concerning the object of studies (general population) based on a a smaller sample

Historical background

- Theory of probability goes back to 17th century when Pierre de Fermat and Blaise Pascal analyzed games of chance. That is why, initially it concentrated on discreet variables, only, using methods of combinatorics.

- Continuous variables were introduced to theory of probability much later

- The beginning of modern theory of probability is generallly accepted to be axiomatization performed in 1933 by Andriej Kołmogorow.

Gambling

Is based on probability of random events...

...simple, as a coin toss, ...

...complicated, as a poker game..

...and may be analyzed by theory of probability.

...fully random as roulette...

- **Probability of a „tail”**
- **Certain combination of cards held in one hand**

Historical background

Blaise Pascal (1601-1662)

Paris, France

Immortalized Chevalier de Méré and gambling paradox

Pascal's triangle for binomial coefficients

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k \, b^{n-k}$$

Newton's binomial

Pascal's Triangle

Binomial coefficients (read „n choose k")

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}$$

$n=0$	$\binom{0}{0}=1$						
$n=1$	$\binom{1}{0}=1$	$\binom{1}{1}=1$					
$n=2$	$\binom{2}{0}=1$	$\binom{2}{1}=2$	$\binom{2}{2}=1$				
$n=3$	$\binom{3}{0}=1$	$\binom{3}{1}=3$	$\binom{3}{2}=3$	$\binom{3}{3}=1$			
$n=4$	$\binom{4}{0}=1$	$\binom{4}{1}=4$	$\binom{4}{2}=6$	$\binom{4}{3}=4$	$\binom{4}{4}=1$		
$n=5$	$\binom{5}{0}=1$	$\binom{5}{1}=5$	$\binom{5}{2}=10$	$\binom{5}{3}=10$	$\binom{5}{4}=5$	$\binom{5}{5}=1$	
$n=6$	$\binom{6}{0}=1$	$\binom{6}{1}=6$	$\binom{6}{2}=15$	$\binom{6}{3}=20$	$\binom{6}{4}=15$	$\binom{6}{5}=6$	$\binom{6}{6}=1$

Pascal's Triangle

						1							n = 0
					1	+	1						n = 1
				1		2		1					n = 2
			1		3		3		1				n = 3
		1		4		6		4		1			n = 4
	1		5		10		10		5		1		n = 5
1		6		15		20		15		6		1	n = 6

Historical background

Pierre de Fermat (1601-1665)

Touluse, France

Studied properties of prime numbers, theory of numbers, in parallel he developed the concept of coordinates in geometry.

In collaboration with Pascal he laid a base for modern theory of probability.

Historical background

Siméon Denis Poisson (1781-1840)

Paris, France

Friend of Lagrange, student of Laplace at famous École Polytechnique.

Except for physics, he took interest in theory of probability.

Stochastic processes (like Markow's process), Poisson's distribution – cumulative distribution function

Historical background

Carl Frederich Gauss (1777-1855)

Goettingen, Germany

University Professor

Ingenious mathematician who even in his childhood was far ahead of his contemporaries.

While a pupil of primary school he solved a problem of a sum of numbers from 1 to 40 proposing - (40+1)*20

Normal distribution function, Gauss distribution

Paradox of Chevalier a de Méré

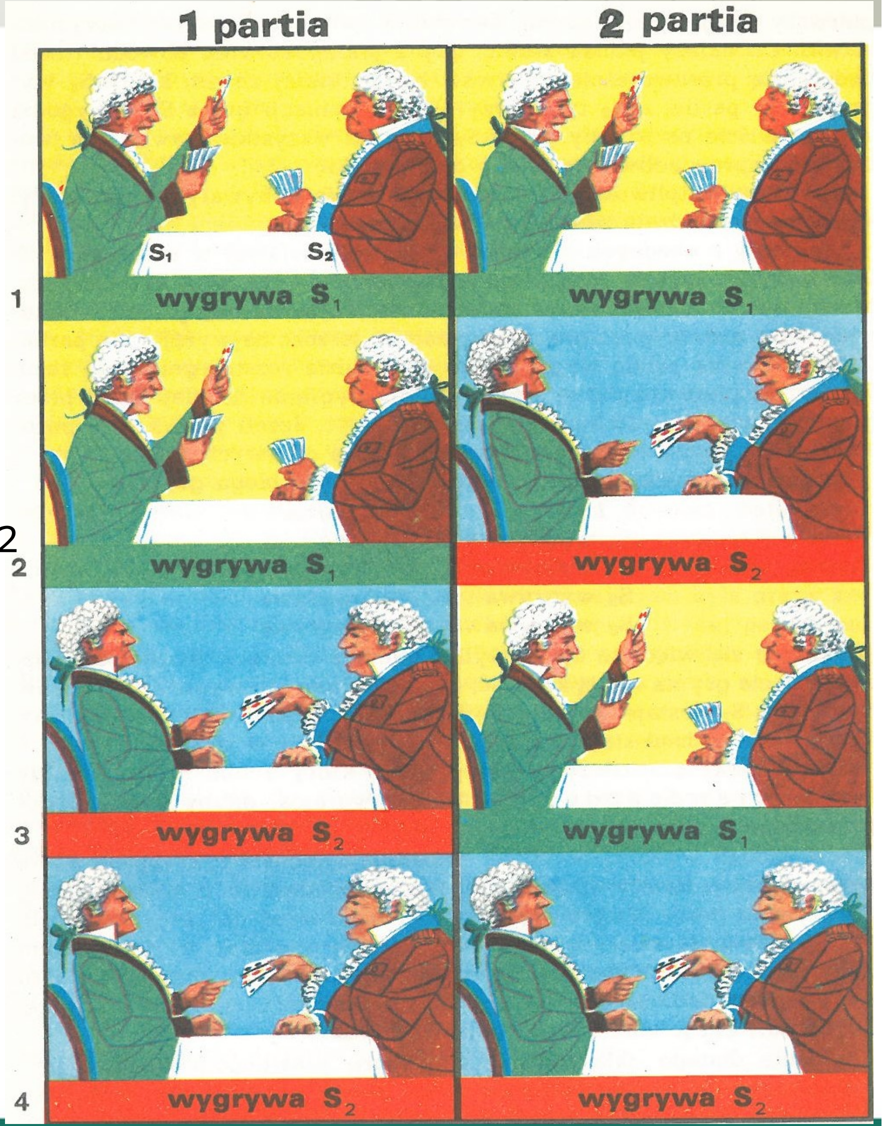
Two gamblers S_1 and S_2 agree to play a certain sequence of sets. The winner is the one who will be the first to gain 5 sets.

What is the score, when the game is interrupted abruptly?

Assume that S_1 wins 4 times and S_2 only 3 times. How to share the stake?

Proposal no. 1: money should be paid in ratio of 4:3

Proposal 2: (5-3):(5-4)=2:1



wg W.R. Fuchs, *Matematyka popularna,* Wiedza Powszechna, Warszawa 1972

Paradox of Chevalier de Méré

Blaise Pascal is believed to have found the solution to this problem quite simply by assuming that the game will be resolved if they play two times more (at the most).

If the first set is won by S_1, the whole game is finished.

If the first set is solved by S_2, the second victory of S_1 makes a deal.

Only in the case both sets are won by S_2 makes him win the score. Then, it is justified to share money as 3:1.

Statistics – types of data

QUANTITATIVE, NUMERICAL

Examples:

- *Set of people*
- *Age*
- *Height*
- *Salary*

Calculations of certain parameters, like averages, median, extrema, make sense.

QUALITATIVE, CATEGORIAL

Examples:

Sex

Marital status

One can ascribe arbitrary numerical values to different categories.

Calculations of parameters do not make sense, only percentage contributions can be given.

The concept of random variable

Random variable is a function X, that attributes a real value x to a certain results of a random experiment.

$$\Omega = \{e_1, e_2, \ldots\}$$

$$X : \Omega \rightarrow R$$

$$X(e_i) = x_i \in R$$

Examples:

1) Coin toss: event 'head' takes a value of 1; event 'tails' - 0.
2) Products: event 'failure' - 0, well-performing – 1
3) Dice: '1' – 1, '2' – 2 etc.…
4) Interval [a, b]– a choice of a point of a coordinate 'x' is attributed a value, e.g. $\sin^2(3x+17)$ etc. ….

Statistics – types of data

Random variable

Discreet

- Toss of a coin
- Transmission errors
- Faulty elements on a production line
- A number of connections coming in 5 minutes

Continuous

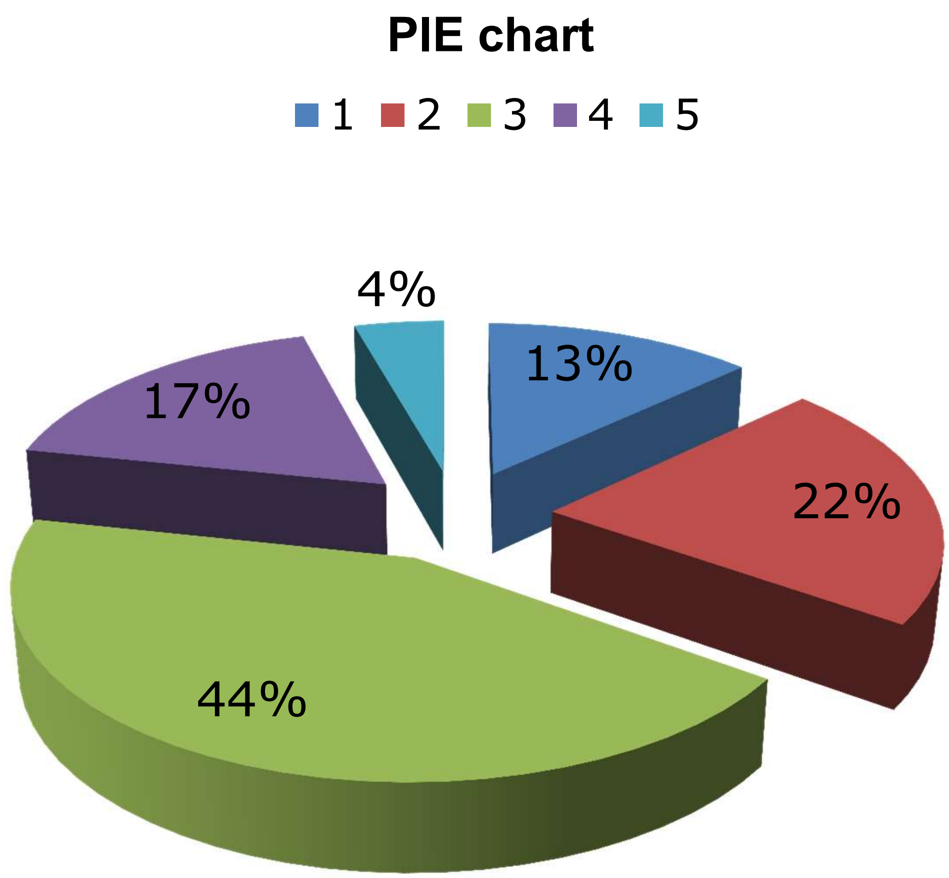
- Electrical current, I
- Temperature, T
- Pressure, p

Graphical presentation of data

x	Number of outcomes	Frequency
1	3	3/23 = 0,1304
2	5	5/23 = 0,2174
3	10	10/23 = 0,4348
4	4	4/23 = 0,1739
5	1	1/23 = 0,0435
Sum:	23	1,0000

Graphical presentation of data

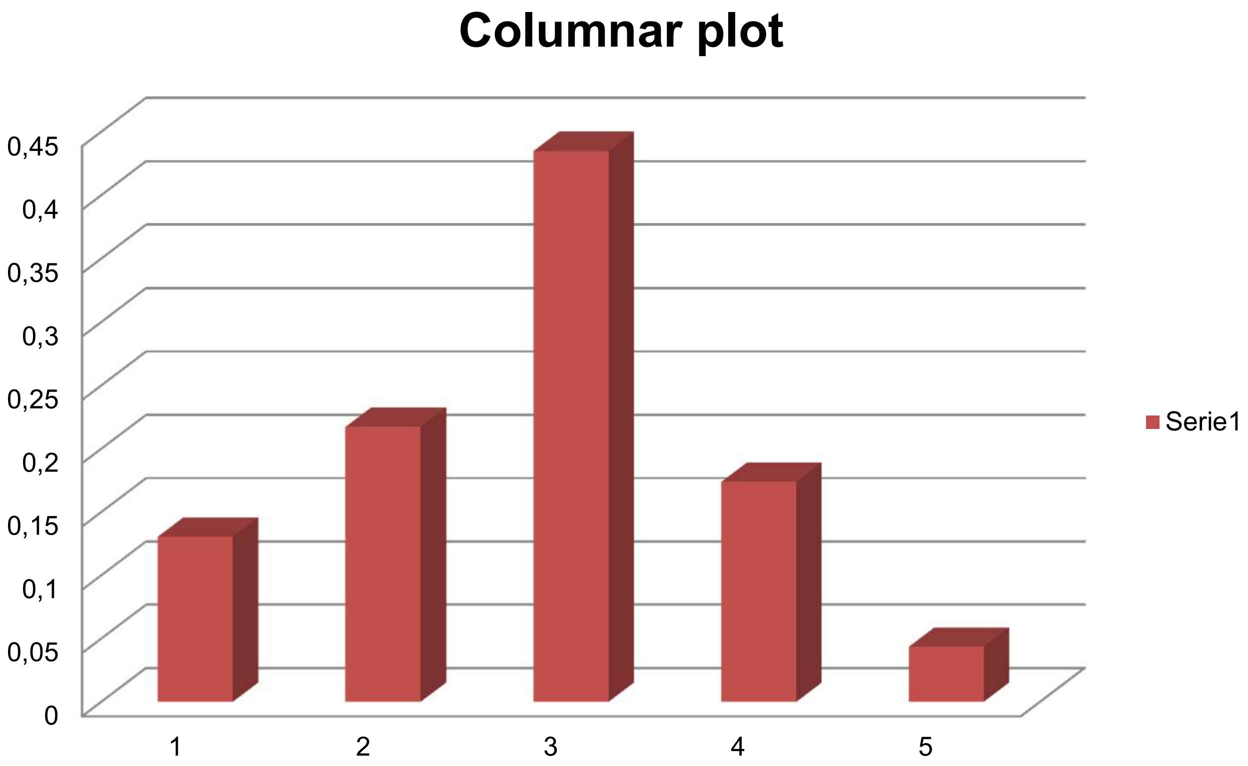
1	0,13043478
2	0,2173913
3	0,43
4	0,17391
5	0,04347826



graf1

Graphical presentation of data

1	0,13043478
2	0,2173913
3	0,43
4	0,17391
5	0,04347826



Numerical data

Results of 34 measurements (e.g. grain size in [nm], temperature in consequitive days at 11:00 in [deg. C], duration of telephone calls in [min], etc.

3,6	13,2	12	12,8	13,5	15,2	4,8
12,3	9,1	16,6	15,3	11,7	6,2	9,4
6,2	6,2	15,3	8	8,2	6,2	6,3
12,1	8,4	14,5	16,6	19,3	15,3	19,2
6,5	10,4	11,2	7,2	6,2	2,3	

These data are difficult to deal with!

Histogram

How to prepare a histogram:

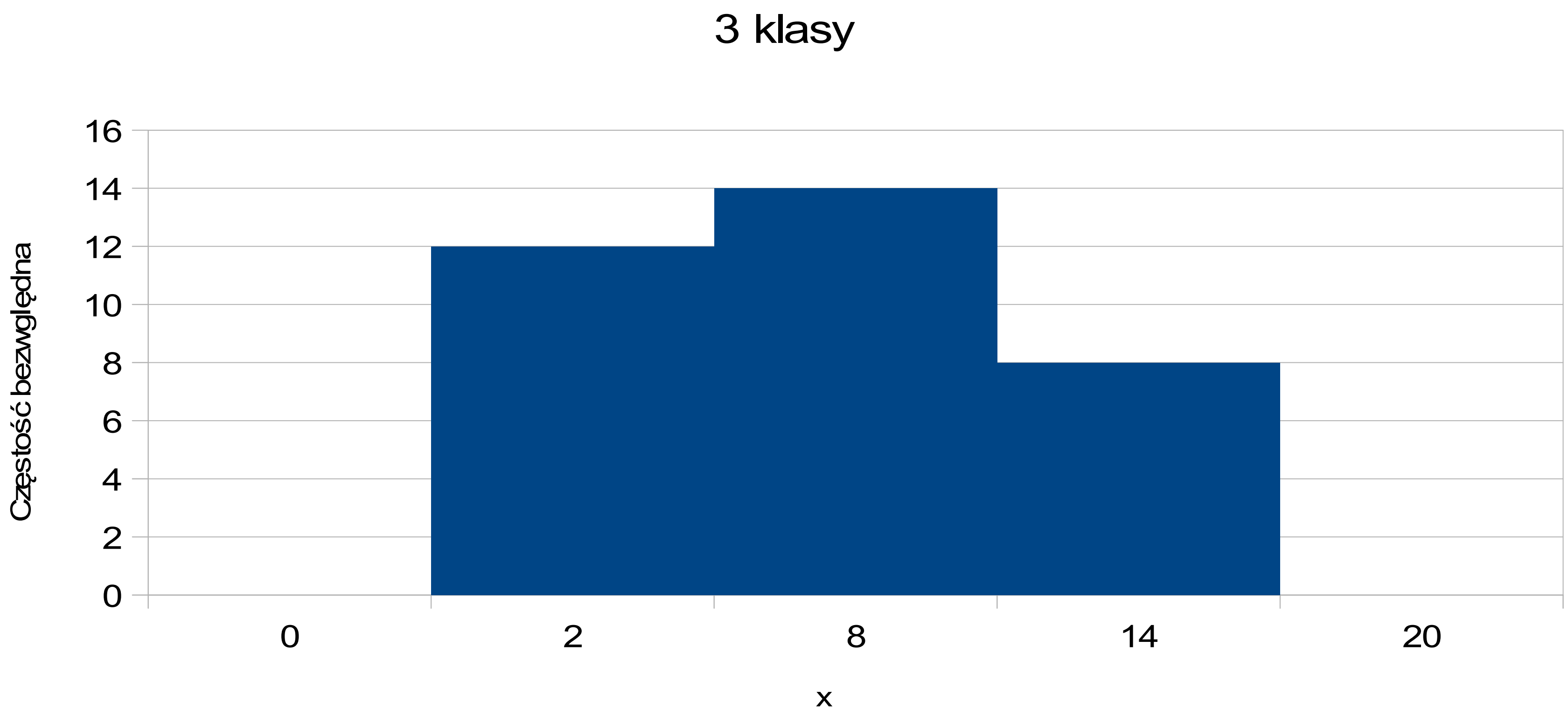
1. Order your data (increasing or decreasing values – program Excel programme has such an option.

2. Results of experiments (a set of n numbers) can contain the same numerical values. We divide them into classes.

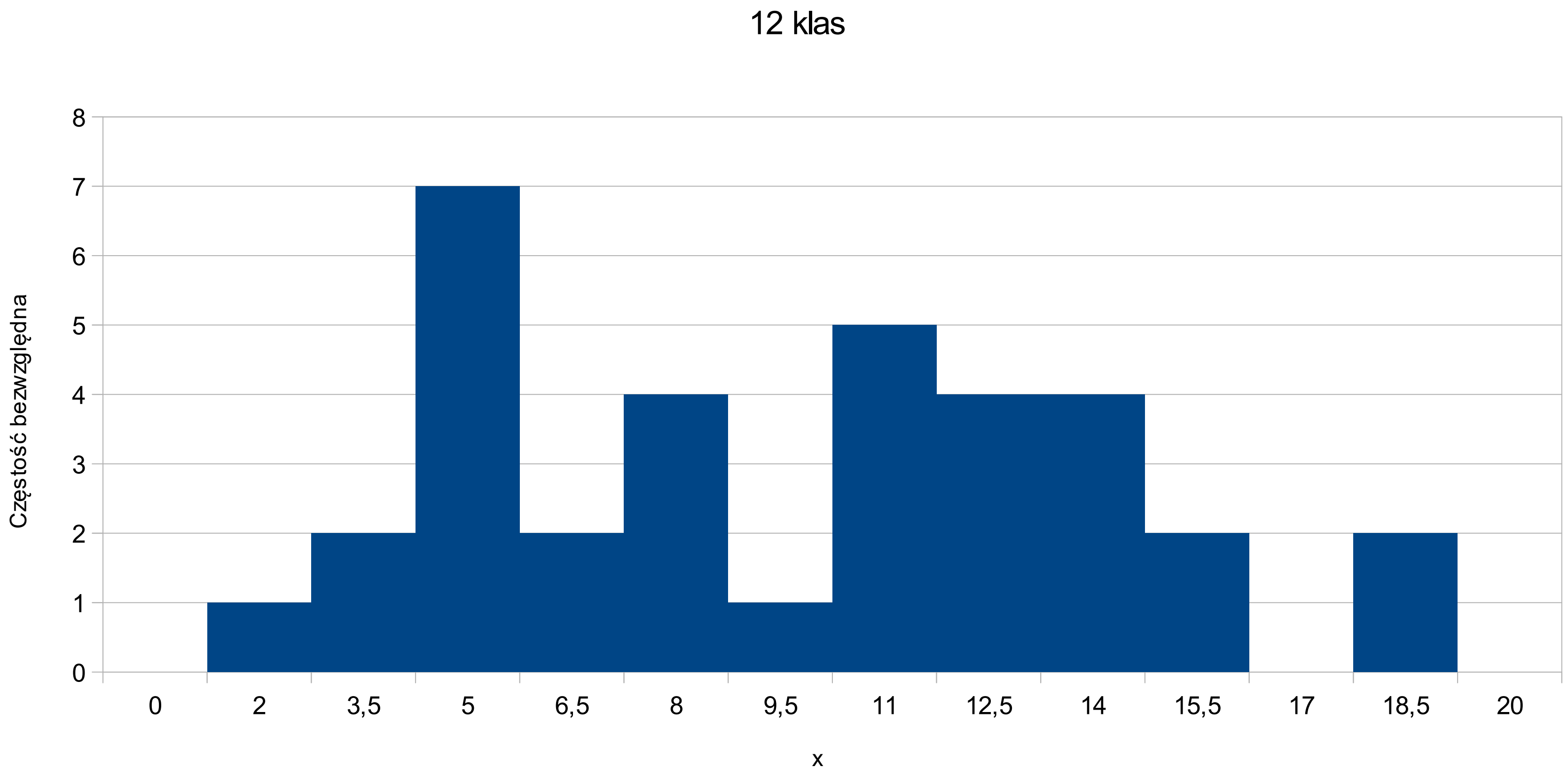
3. The width of a class is not necessarily constant but usually it is chosen to be the same.

4. Number of classes should not be to small or to big. The optimum number of classes 'k' is given by Sturge formula.

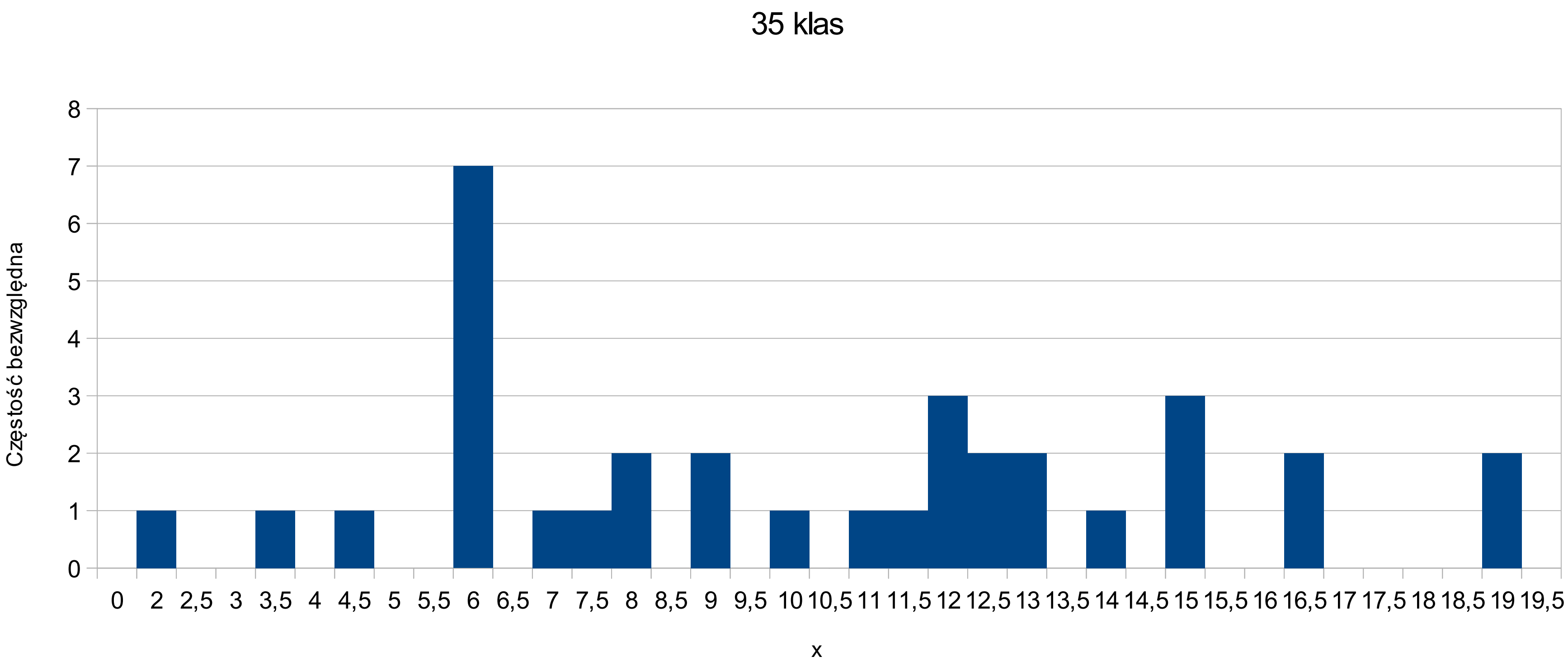
Histogram



Histogram


12 klas
Częstość bezwzględna
8
7
6
5
4
3
2
1
0
0
2
3,5
5
6,5
8
9,5
11
12,5
14
15,5
17
18,5
20
x

Histogram



35 klas
Częstość bezwzględna
8
7
6
5
4
3
2
1
0
0 2 2,5 3 3,5 4 4,5 5 5,5 6 6,5 7 7,5 8 8,5 9 9,5 10 10,5 11 11,5 12 12,5 13 13,5 14 14,5 15 15,5 16 16,5 17 17,5 18 18,5 19 19,5
x

Sturge formula

$$k = 1 + 3{,}3\log_{10} n$$

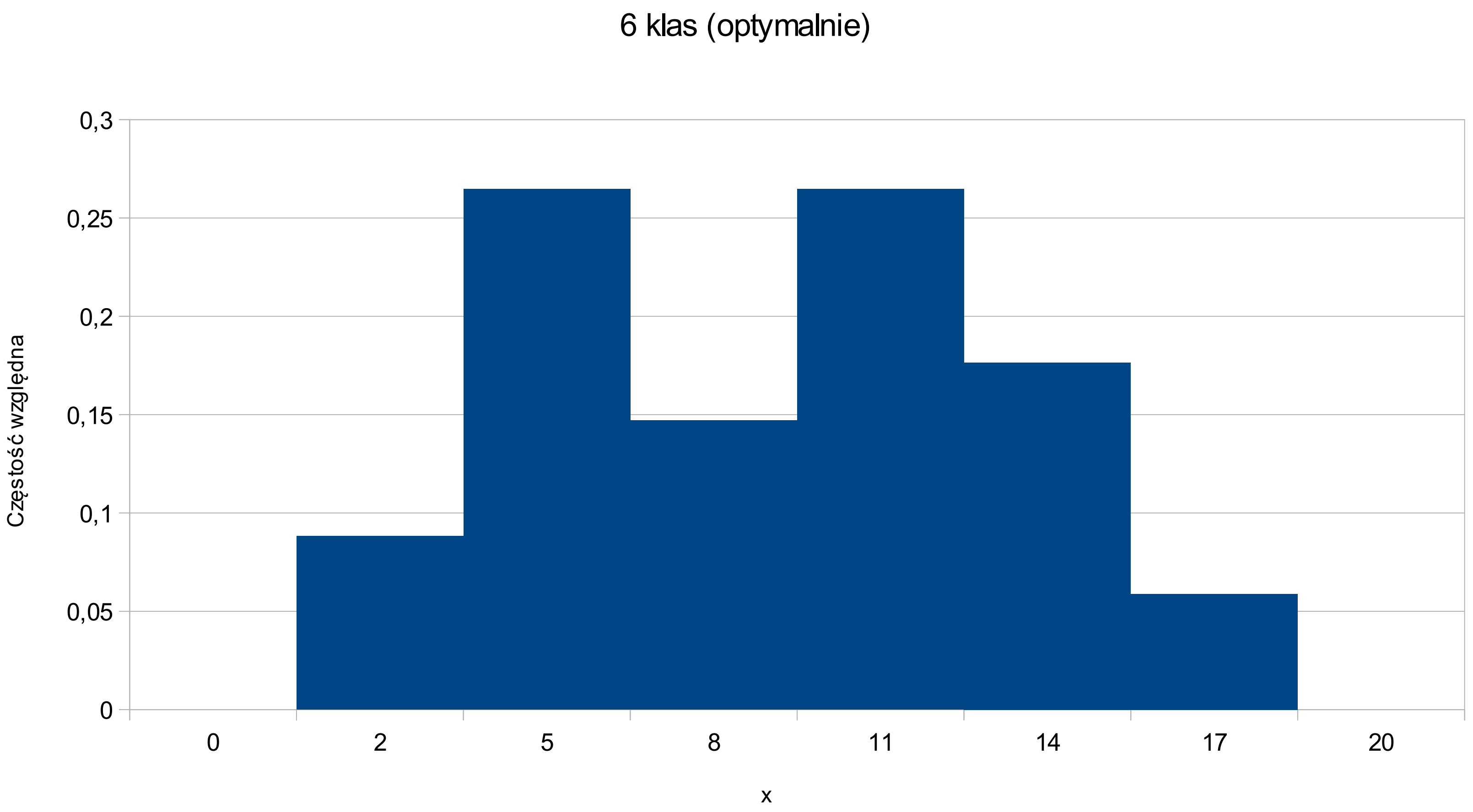
In our case:

Sample count, n	Number of classes, k
< 50	5 – 7
50 – 200	7 – 9
200 – 500	9 – 10
500 – 1000	10 -11
1000 – 5000	11 – 13
5000 – 50000	13 – 17
50000 <	17 – 20

$$n = 34$$

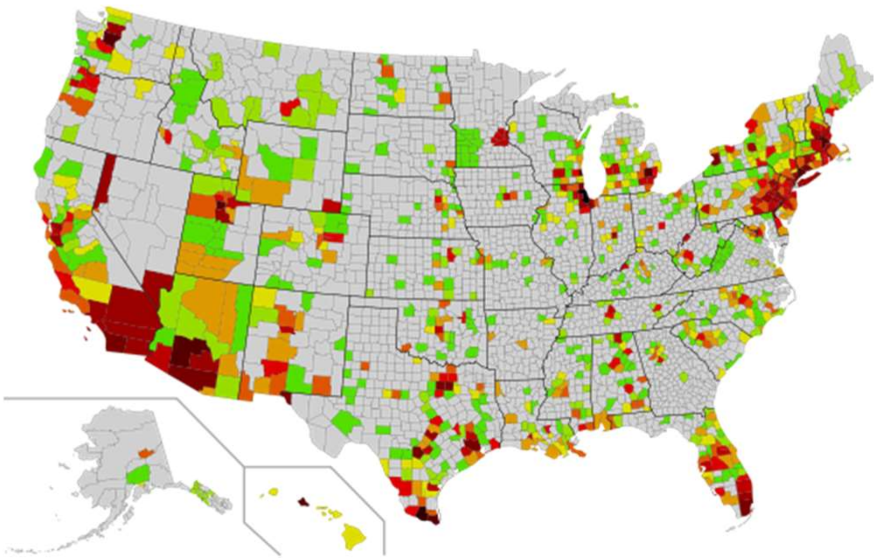
$$k = 5.59 \approx 6$$

Optimum histogram



The role of probability and statistics in science and engineering

Statistics allows us to analyze and perform modelling of development of diseases with the aim to prevent epidemics.

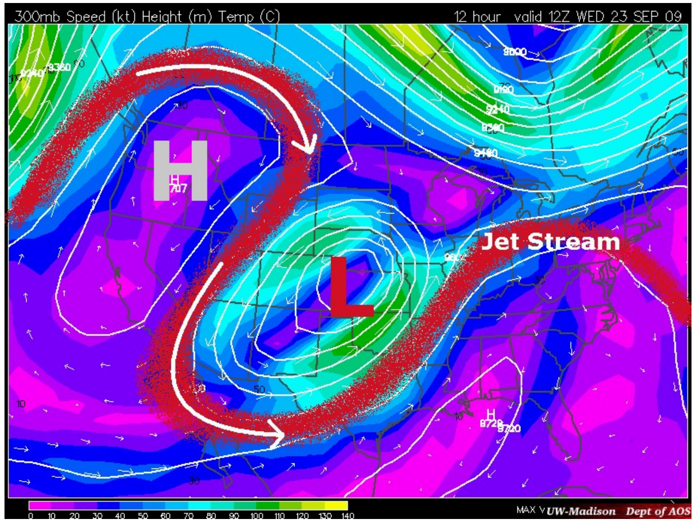


Incidence of swine flu in 2009,USA
(Source: http://commons.wikimedia.org)

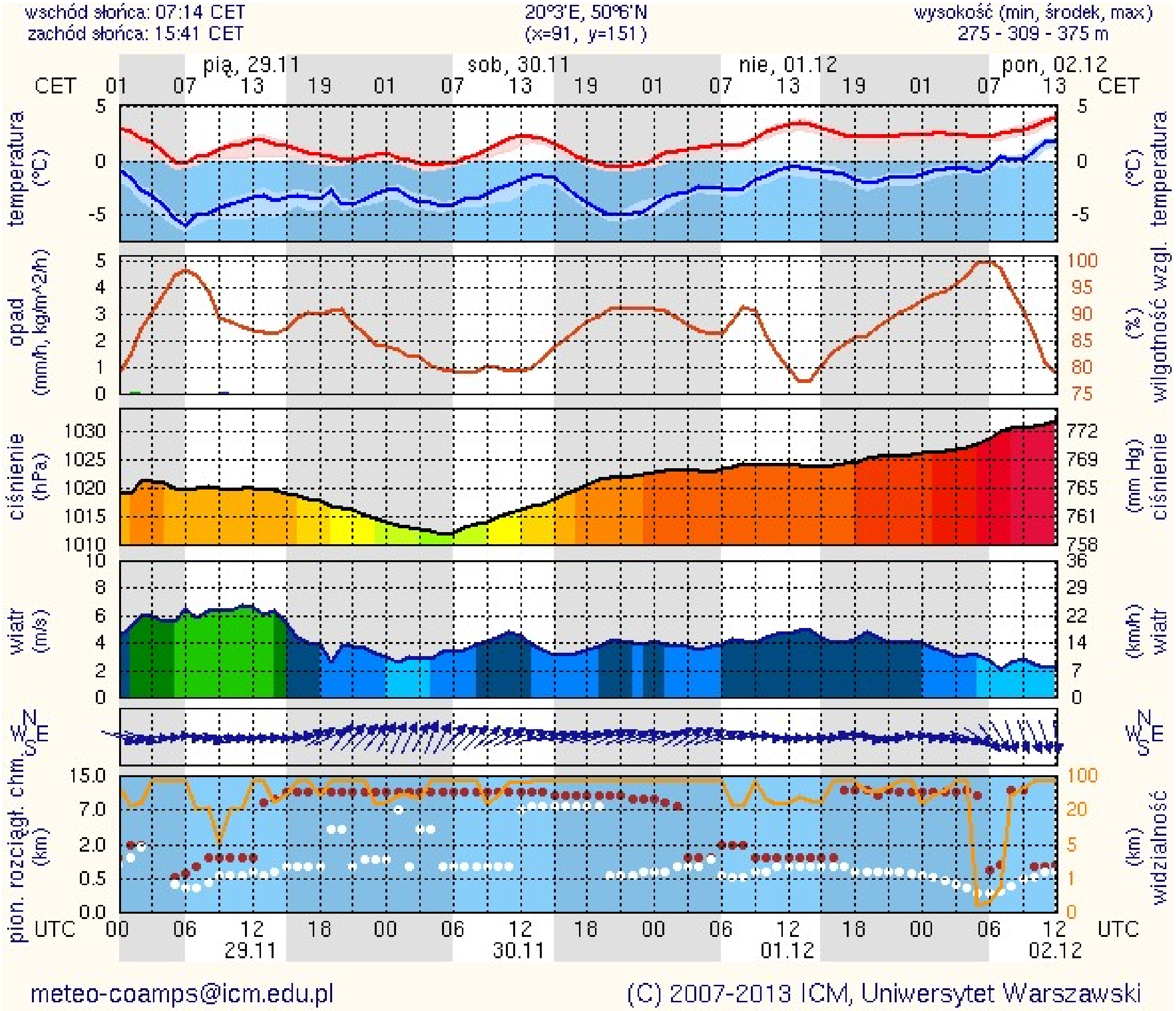
- Medical statistics , e.g. the average number of cases (incidence of influenza) in a certain region
- Social statistics, e.g. density of population
- Industrial statistics, e.g. GDP (gross domestic product), expenses for medical care

Metrology

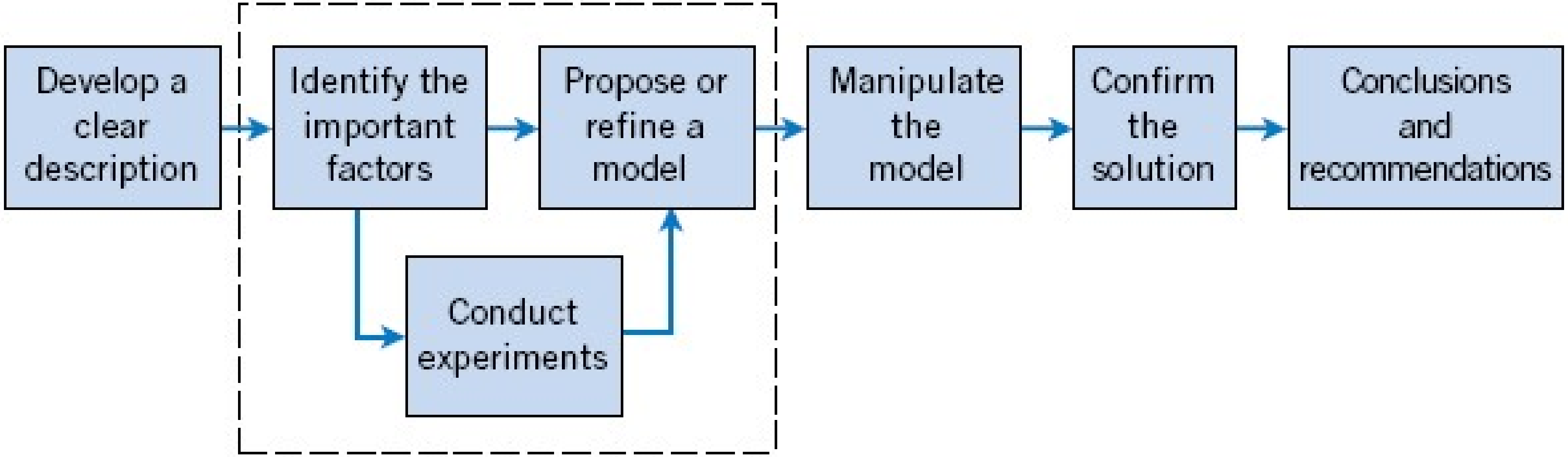
Weather forecast models enable to predict potential disasters like storms, tornados, tsunami, etc.



(Source:stormdebris.net/Math_Forecasting.html)



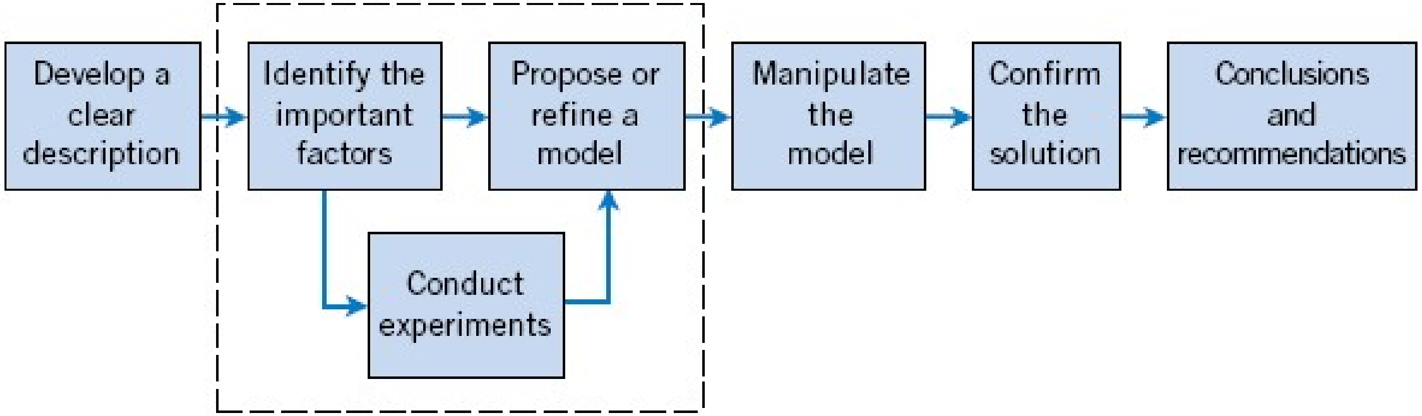
How to solve an engineering problem?



Problem description

Example: Suppose that an engineer is designing a nylon connector to be used in an automotive engine application. The engineer is considering establishing the design specification on wall thickness at 3/32 inch but is somewhat uncertain about the effect of this decision on the connector pull-off force. If the pull-off force is too low, the connector may fail when it is installed in an engine.

How to solve an engineering problem?



Identification of the most important factors

Eight prototype units are produced and their pull-off forces measured, resulting in the following data (in pounds): 12.6, 12.9, 13.4, 12.3, 13.6, 13.5, 12.6, 13.1. As we anticipated, not all of the prototypes have the same pull-off force. We say that there is **variability** in the pull-off force measurements. Because the pull-off force measurements exhibit variability, we consider the pull-off force to be a **random variable.**

How to solve an engineering problem?

Proposed model

A convenient way to think of a random variable, say *X*, that represents a measurement, is by using the model

$$X = \underbrace{\mu}_{\text{constant}} + \underbrace{\epsilon}_{\text{disturbance}}$$

The constant remains the same with every measurement, but small changes in the environment, test equipment, differences in the individual parts themselves, and so forth change the value of disturbance. If there were no disturbances, *X* would always be equal to the constant . However, this never happens in the real world, so the actual measurements *X* exhibit variability. We often need to describe, quantify and ultimately **reduce variability**.

Experiments

Figure 1-2 presents a **dot diagram** of these data. The dot diagram is a very useful plot for displaying a small body of data—say, up to about 20 observations. This plot allows us to see easily two features of the data; the **location,** or the middle, and the **scatter** or **variability.** When the number of observations is small, it is usually difficult to identify any specific patterns in the variability, although the dot diagram is a convenient way to see any unusual data features

Figure 1-2 Dot diagram of the pull-off force data when wall thickness is 3/32 inch.

The average pull-off force is **13.0** pounds.

Model modification

The need for statistical thinking arises often in the solution of engineering problems. Consider the engineer designing the connector. From testing the prototypes, he knows that the average pull-off force is 13.0 pounds. However, he thinks that this may be too low for the intended application, so he decides to consider an alternative design with a greater wall thickness, 1/8 inch. Eight prototypes of this design are built, and the observed pull-off force measurements are 12.9, 13.7, 12.8, 13.9, 14.2, 13.2, 13.5, and 13.1. Results for both samples are plotted as dot diagrams in Fig. 1-3.

Figure 1-3 Dot diagram of pull-off force for two wall thicknesses.

The average pull-off force is **13.4** pounds.

Confirmation of the solution

This display gives the impression that increasing the wall thickness has led to an increase in pull-off force.

Figure 1-3 Dot diagram of pull-off force for two wall thicknesses.

Is it really the case?

Conclusions and recommendations

Statistics can help us to answer the following questions:

- How do we know that another sample of prototypes will not give different results?
- Is a sample of eight prototypes adequate to give reliable results?
- If we use the test results obtained so far to conclude that increasing the wall thickness increases the strength, what risks are associated with this decision?
- Is it possible that the apparent increase in pull-off force observed in the thicker prototypes is only due to the inherent variability in the system and that increasing the thickness of the part (and its cost) really has no effect on the pull-off force?