Contents

1 Data

A collection of values measured on the monitored variables.

Discrete data

For data $x = [3, 5, 3, 4, 5, 3, 3, 3, 4, 5]$ determine:

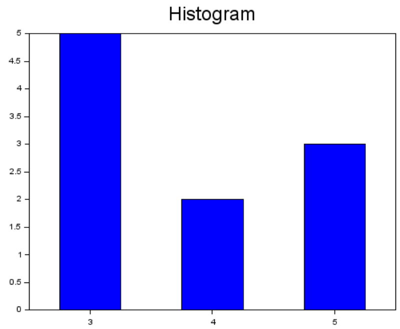
Ordered data

$$ord\,(x) = [3, 3, 3, 3, 3, 4, 4, 5, 5, 5]$$

Frequencies

values	3	4	5
abs. fr.	5	2	3
rel. fr.	0.5	0.2	0.3

Histogram



Average

$$\frac{1}{10}\left(3+5+3+4+5+3+3+3+4+5\right)=$$

$$=\frac{1}{10}\left(3\cdot 5+4\cdot 2+5\cdot 3\right)=3\cdot 0.5+4\cdot 0.2+5\cdot 0.3=3.8$$

Variance

$$\frac{1}{10}\left((3-3.8)^2+(5-3.8)^2+\cdots\right)=$$

$$=(3-3.8)^2\cdot 0.5+(4-3.8)^2\cdot 0.2+(5-3.8)^2\cdot 0.3=0.76$$

Standard deviation

$$\sqrt{0.76} = 0.872$$

Ranks

$$\begin{array}{cccccccccc}
3 & 5 & 3 & 4 & 5 & 3 & 3 & 3 & 4 & 5 \\
3 & 3 & \mathbf{3} & 3 & 3 & \mathbf{4} & \mathbf{4} & 5 & \mathbf{5} & 5 \\
 & & 3 & & & \multicolumn{2}{c}{6.5} & & 9 &
\end{array}$$

$$r = [3,\ 9,\ 3,\ 6.5,\ 9,\ 3,\ 3,\ 3,\ 6.5,\ 9]$$

Mode, median, 0.1 quantile

$$\hat{x} = 3,\ \ \tilde{x} = \frac{3+4}{2} = 3.5,\ \ \zeta_{0.1} = 3$$

Real discrete data - load the file Smart.txt to Statext (data length 656)

Data|Count Data ...

Descriptive|Basic ... N, $Mean$, $Variance$, $Standard\,deviation$, $Range$, Min, $\cdots Max$, $Mode$, and other

Descriptive|Dot Plot ...

Descriptive|Box-and-Whiskers ...

Descriptive|Frequency Table ... (interval = 1)

Descriptive|Histogram ... (interval = 1), (shows only one; for second delete {})

Real continuous data - load data02.txt - speed of the driven car (1000 samples)

Data|Count Data ...

Descriptive|Basic ... N, $Mean$, $Variance$, $Standard\,deviation$, $Range$, Min, $\cdots Max$, $Mode$, and other

Descriptive|Dot Plot ...

Descriptive|Box-and-Whiskers ...

Descriptive|Histogram ... (interval = 10), (shows only one; for second delete {})

2 Probability

Example

We draw a dice. What is the probability of

a) 6?

b) 6, if we know that the number is even (odd)?

c) 6, if we know that the number is greater than 4?

d) even number, if we know that the number is greater than 4?

e) even number, if we know that the number is greater than 3?

The dice

$$1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6$$

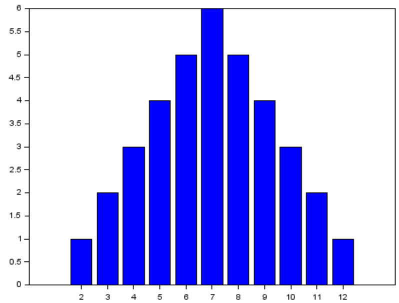
and the condition is a restriction to sample space.

Example

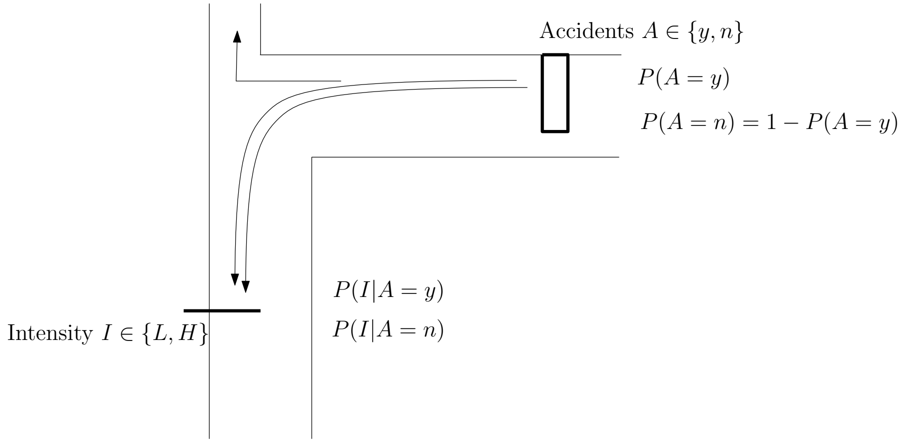
What are results of an experiment: sum on two dices?

2	3	4	5	6	7	8	9	10	11	12
$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Graph



The result is an ordered couple. The shape is because e.g. 3 can be [1, 2] or [2, 1] etc.



1. **Total probability**

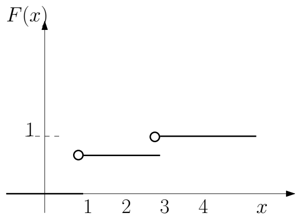
$$P(I) = P(I|A = y)\, P(A = y) + P(I|A = n)\, P(A = n)$$

2. **Bayes rule**

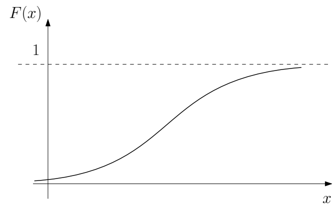
$$P(A = y|I) = \frac{P(I|A = y)\, P(A = y)}{P(I)}$$

3 Random variable

Distribution function

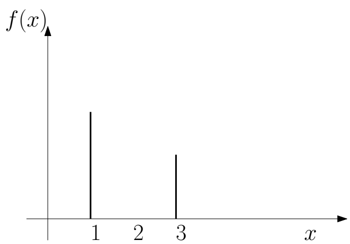


Discrete distribution function

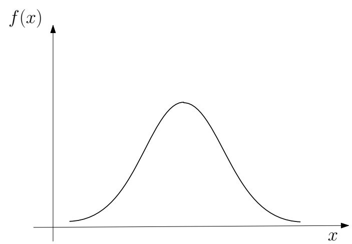


Continuous distribution function

Probability and density function



discrete



continuous

Probability of X on (a, b)

It holds

$$P\left(X \in (a,b)\right) = \int_a^b f\left(x\right) dx$$

Proof

For distribution function it holds

$$F\left(x\right) = P\left(X \leq x\right) \qquad \text{definition of } F,$$

and at the same time

$$F\left(x\right) = \int_{-\infty}^{x} f\left(t\right) dt \qquad \text{definition of } f.$$

From it

$$P\left(X \leq x\right) = \int_{-\infty}^{x} f\left(t\right) dt$$

Now, for $b > a$

$$P\left(X \in (a,b)\right) = P\left(X \leq b\right) - P\left(X \leq a\right) =$$

$$= \int_{-\infty}^{b} f\left(x\right) dx - \int_{-\infty}^{a} f\left(x\right) dx = \int_a^b f\left(x\right) dx$$

Example - discrete rv

Random variable X is defined through the following table

x	*1*	*2*	*3*	*4*	*5*	*6*
$f(x)$	*0.2*	*0.1*	*0.1*	*0.3*	*0.2*	*0.1*

Compute its expectation $E[X]$, variance $D[X]$ and standard deviation.

Expectation

$$E[X] = 1 \cdot 0.2 + 2 \cdot 0.1 + 3 \cdot 0.1 + 4 \cdot 0.3 + 5 \cdot 0.2 + 6 \cdot 0.1 = 3.5$$

Variance

$$D[X] = (1 - 3.5)^2 \cdot 0.2 + (2 - 3.5)^2 \cdot 0.1 + (3 - 3.5)^2 \cdot 0.1 +$$

$$+ (4 - 3.5)^2 \cdot 0.3 + (5 - 3.5)^2 \cdot 0.2 + (6 - 3.5)^2 \cdot 0.1 = 2.65$$

Standard deviation

$$\sqrt{D[X]} = \sqrt{2.65} = 1.628$$

Example - continuous rv

Determine distribution function $F(x)$ of random variable with density function

$$f(x) = \frac{1}{2}x, \text{ on } x \in (0,2)$$

For the distribution function on $x \in (0,2)$ it holds

$$F(x) = \int_0^x f(t)\,dt = \int_0^x \frac{1}{2}t\,dt = \frac{1}{2}\left[\frac{t^2}{2}\right]_0^x = \frac{1}{4}x^2$$

The whole distribution function for $x \in R$ it holds

$$F(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ \frac{1}{4}x^2 & \text{for } x \in (0,2) \\ 1 & \text{for } x \geq 2 \end{cases}$$

4 Regression

Open the file data1.txt in Statext and compare various types of regression.

1. Check Show Graphic Result to see the result plotted in a graph

2. Find the regression equation (at the beginning of the Result window)

3. Compare the p-value (it should tend to zero - its meaning will be explained later)

5 Population and Sample

Example 1 - Checking the quality of minimarkets in ČR

Population: All minimarkets in ČR

Sample: A set of selected minimarkets that have been checked.

Here, the number of minimarkets is finite, but very large. We cannot check them all, that is why we use a sample.

Example 2 - Monitoring the speed of cars at a given point.

Population: All possible speeds that a car can go at the point.

Sample: The set of speeds that have been measured.

Here, the number of possible speeds is infinite. As if there is a generator producing cars according according to some its fixed probabilistic rules.

Sample and its realization

Let us have 10 minimarkets with qualities (in per cent) with $E[X] = 72.6$ and $D[X] = 206$

minimarket	1	2	3	4	5	6	7	8	9	10
quality	89	43	69	75	94	62	81	75	66	72

We want to take a sample of three minimarkets and check their average quality. Randomly we select minimarkets 3, 7 and 9. Then the average quality (sample average) is equal to

$$\frac{69 + 81 + 66}{3} = 72$$

The number 72, however, depends on our selection into the sample. We can do an experiment with repetitive sampling - even if in practice we work only with one sample. Let us obtain the following table

sample no.	sample	average
1	69, 81, 66	72
2	75, 62, 72	69.67
3	43, 69, 62	58
4	89, 62, 66	72.33
5	43,62, 81	62
6	89, 94, 66	83
7	69, 94, 66	76.33
average from averages		70.48
...	...	...
average from all averages		72.6

Now, the population expectation is 72.6 and average of sample averages is 70.48 what is closer to expectation than individual averages.

If the table would include all possible samples - whose number is $\binom{10}{3} = 120$, then the average of sample averages would be exactly the population expectation, i.e.

$$E\left[\bar{X}\right] = E\left[X\right] = \mu$$

Variance of the population is $D[X] = 206$. Sample variance is $D\left[\bar{X}\right] = 70.85$. Approximately it holds

$$D\left[\bar{X}\right] = \frac{D[X]}{n} = \frac{\sigma^2}{n}.$$

Remark

$$E\left[\bar{X}\right] = \int_{-\infty}^{\infty} \frac{1}{n} \sum_i x_i f(x_i)\, dx_i = \frac{1}{n} \sum_i \int x_i f(x_i)\, dx = \frac{1}{n} \sum_i E[X_i] = \frac{1}{n} \sum_i \mu = \frac{1}{n} n\mu = \mu$$

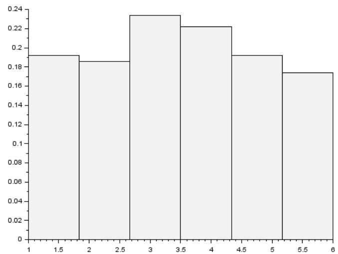
! ! ! ! !

So we have:	distribution of data	distribution of averages (statistics)
	$f\left(x \mid \mu, \sigma^2\right) = N_x\left(\mu, \sigma^2\right)$	$f\left(\bar{x} \mid \mu, \sigma^2\right) = N_{\bar{x}}\left(\mu, \frac{\sigma^2}{n}\right)$

Central limit theorem

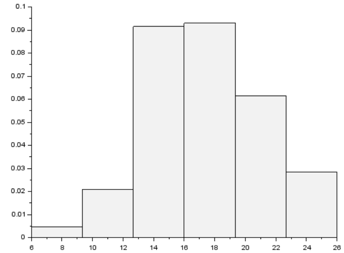
For $N \to \infty$ the sum characteristics tends to normal distribution.

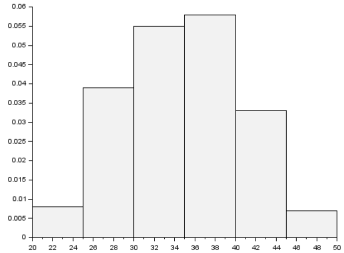
Example: 200 throws of dice gives the histogram



200 samples of the sum of 5 throws

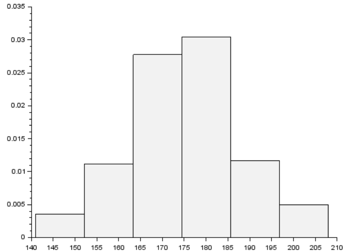
Sum of 10 throws

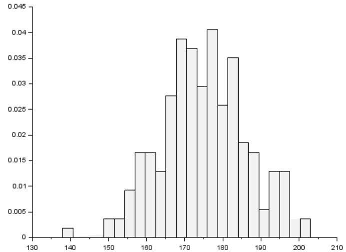




Sum of 50 throws

$\cdots$ and more detailed view.





which approaches the normal distribution.

The law of large numbers

Again throwing a dice.

Expectation is $E[X] = (1 + 2 + \cdots + 6)/6 = 3.5$

Sample with 5 entries $\bar{x} = 2.2$

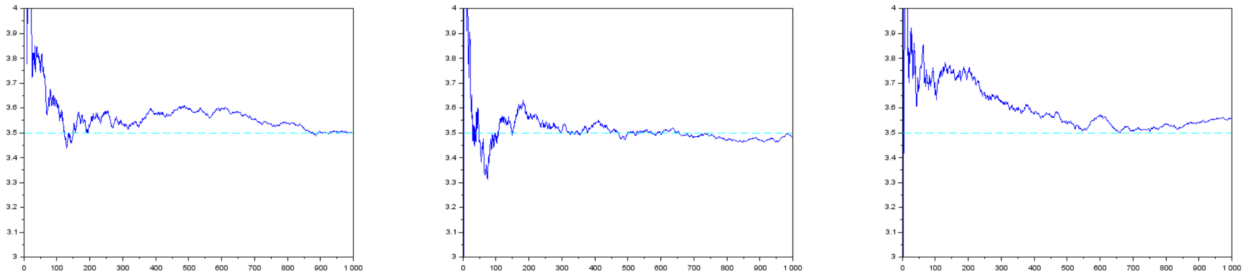
Sample with 10 entries $\bar{x} = 2.7$

Sample with 30 entries $\bar{x} = 4.27$

Sample with 100 entries $\bar{x} = 3.56$

Sample with 1000 entries $\bar{x} = 3.502$

Graphical result for 1 ... 1000 samples (three different experiments)



6 Estimation

Point estimate is the value of the Statistics with the sample realization inserted.

- E.g. an average of measured data.
- It does not take into account the uncertainty of data which makes statistics to be random.
- If we take new sample, the average will be slightly different.

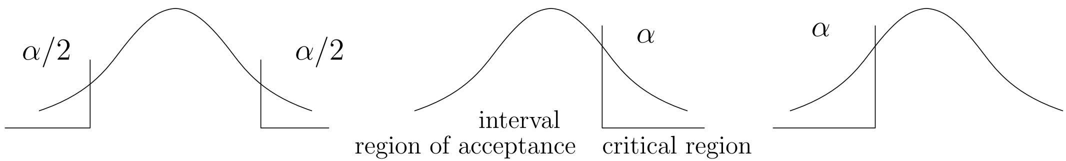
Interval estimate is based on the probability function of the Statistics.

- For example for normal population with μ and σ^2 (known variance) the Statistics will be normal with expectation $\bar{x}$ and variance $\frac{\sigma^2}{N}$ (N is length of the sample).
- For other parameters e.g. two expectations, variance, type of distribution, independence etc., the derivation of the distribution in much more complex and it is a component of the particular interval or test.
- For the derivation of the interval, the **density of the Statistics** is used (not the density of the population).

7 Testing

Sides of intervals / tests are defined as follows

$f(T)$ for interval and $f(T|H0)$ for test



Both sided interval / test   Right sided interval / test   Left sided interval / test

Side of a test for two expectations

Let us have samples S_A and S_B from two random variables A and B with expectations μ_A and μ_B, respectively. We want to test $H_0 : \mu_A = \mu_B$ against $H_A : \mu_B < \mu_A$.

Solution

1. It is the test for two expectations.

2. We will assign: A is first, B is second (in the order how they are treated)

3. $H_A : \mu_A > \mu_B$ (in the order decided above)

4. $H_A : \mu_A - \mu_B > 0 \cdots$ the test will be right-sided.

!!! H_A decides about the side; we must keep the order; > 0 right-sided, < 0 left-sided. !!!

Lambda coefficient

For discrete variables x and y. Its value says, how much the knowledge of x improves improves the prediction of y.

Example

The prediction of y with x is given by the frequency table (after normalization in rows it is conditional probability function $f(y|x)$)

$x\backslash y$	1	2	3
1	21	13	$\underline{25}$
2	8	$\underline{22}$	11
3	6	12	$\underline{18}$
4	$\underline{27}$	3	11

where the maxima in rows are underlined. Now, for given x we predict y with the highest frequency in the tow. Thus, for $x = 1$ we always predict $y = 3$ but actually there were 21 cases, where y were 1 and 13 cases, where y were 2. That means that we do 34 errors in prediction. Similarly, for $x = 2$ we predict $y = 2$ and do 19 errors. For $x = 3$ the prediction is $y = 3$ with 18 errors and for $x = 4$ it is $y = 1$ with 14 errors. So, all in all, we do $E_c = 34 + 19 + 18 + 14 = 85$ errors.

Without knowledge of x we have only frequencies of y sums of the table over columns

y	1	2	3
fr.	62	50	$\underline{65}$

As the maximum is in the third column, we always predict $y = 3$ and we do $E_u = 62 + 50 = 112$ errors.

We define lambda as

$$\Lambda = \frac{E_u - E_c}{E_u}$$

which is decrease in errors when considering x in relation to the number of errors when not knowing x.

In our example it is

$$\Lambda = \frac{112 - 85}{112} = 0.24$$

which means, that the decrease of errors is 24%.