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:

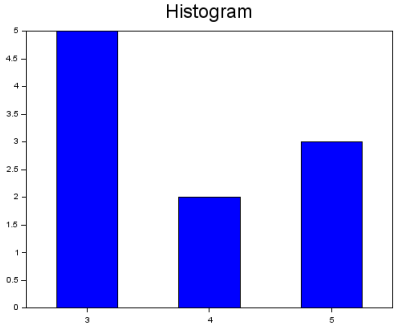
Ordered data

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

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

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

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

Standard deviation

$$\sqrt{0.76} = 0.872$$

Ranks

3	5	3	4	5	3	3	3	4	5
3	3	**3**	3	3	**4**	**4**	5	**5**	5
		3			6.5			9	

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

Mode, median, 0.1 quantile

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

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?

3 Regression

Derivation of simple linear regression

Data: x and y. Define centered data $\tilde{x} = x - \bar{x}$ and $\tilde{y} = y - \bar{y}$. The line with centered variables must go through the origin

$$\tilde{y} = b_1 \tilde{x} + e$$

We derive

$$\sum e_1^2 = \sum \left(\tilde{y}_i - b_1 \tilde{x}_i\right)^2 = \sum \left(\tilde{y}_i^2 - 2b_1 \tilde{x}_i \tilde{y}_i + b_1^2 \tilde{x}_i^2\right) =$$

$$= \underbrace{\sum \tilde{y}_i^2}_{C} - 2b_1 \underbrace{\sum \tilde{x}_i \tilde{y}_i}_{B} + b_1^2 \underbrace{\sum \tilde{x}_i^2}_{A} = Ab_1^2 - 2Bb_1 + C$$

$\cdots$ derivative

$$2Ab_1 - 2B = 0 \qquad b_1 = \frac{B}{A} = \frac{\sum \tilde{x}_i \tilde{y}_i}{\sum \tilde{y}_i^2}$$

About the constant

The equation will now be

$$y = b_1 x + b_0$$

$\cdots$ for all data

$$\sum y_i = b_1 \sum x_i + N b_0 \qquad b_0 = \frac{1}{N} \sum y_i - b_1 \frac{1}{N} \sum x_i = \bar{y} - b_1 \bar{x}$$

The regression line is

$$y = b_1 x + b_0 = b_1 x + \bar{y} - b_1 \bar{x} = b_1 (x - \bar{x}) + \bar{y} =$$

$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (y_i - \bar{y})^2} (x - \bar{x}) + \bar{y}$$

or (expanding the fraction by $1/N$)

$$y = \frac{\overline{(x_i - \bar{x})(y_i - \bar{y})}}{\overline{(y_i - \bar{y})^2}} (x - \bar{x}) + \bar{y}$$

Program

```
x=0:50;
y=2*x+30+1*randn(1,length(x));
xm=x-mean(x);
ym=y-mean(y);
b1=sum(xm.*ym)/sum(xm.^2)
b0=mean(y)-b1*mean(x)
plot(x,y,':')
plot(x,b1*x+b0,'r')
```

4 Important distributions

Constant of standard normal distribution

The basic integral is

$$I = \int_{-\infty}^{\infty} e^{-x^2} dx$$

We take it in square (to be able to use polar coordinates)

$$I^2 = \left(\int_{-\infty}^{\infty} e^{-x^2} dx \right) \left(\int_{-\infty}^{\infty} e^{-y^2} dy \right) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x+y)^2} dxdy$$

transformation to polar coordinates:

$x = r\cos(t), \quad y = r\sin(t), \quad dxdy = r \cdot drdt, \quad$ borders $(0, 2\pi)$ and $(0, \infty)$

$$I^2 = \int_0^{2\pi} \int_0^{\infty} e^{-r^2} r \cdot drdt$$

Evaluations

$i1 = \int_0^\infty e^{-r^2} r \cdot dr dt = \frac{1}{2} \int_0^\infty e^{-u} du = \frac{1}{2}$ (inner integral) $\qquad u = r^2$

$i2 = \int_0^{2\pi} dt = 2\pi$ (outer integral)

$$I^2 = \frac{1}{2} 2\pi = \pi$$

$$I = \sqrt{\pi}$$

Normal distribution

$$f(x) = k \cdot e^{-\frac{1}{2}x^2}$$

$$k \int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2} dx = 1$$

substitution

$\left(\frac{x}{2}\right)^2 = z^2 \rightarrow \frac{x}{\sqrt{2}} = z \rightarrow dx = \sqrt{2} dz$

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2} dx = \sqrt{2} \int_{-\infty}^{\infty} e^{-z^2} dz = \sqrt{2} \cdot I = \sqrt{2\pi}$$

So, the constant is

$$k = \frac{1}{\sqrt{2\pi}}$$

Example (of **Poisson distribution**)

Consider a situation where we count the number of passing cars as follows: we observe the road for a minute and register whether there is a car on it or not. The measurements follow each other minute by minute. If the speed of the approaching cars is low and constant and the cars are driving independently of each other, then

- the counts of cars in the end of each minute have Poisson distribution,
- the gaps between two consecutive cars have exponential distribution.

Remark

- *The same task can be formulated for (i) customers coming to a shop, (ii) telephone calls, (iii)radioactive decay.*
- *The process generating such data is called Poisson process.*

Deriving the Poisson Distribution from the Binomial One

The Poisson distribution can be derived from the binomial distribution as its limiting case when the number of trials is large, and the probability of success is small. The derivation follows these steps:

1. **Definition of the Binomial Distribution**

Let X be a random variable that follows the binomial distribution.

This means that the probability of X taking the value k is given by:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \ k = 0, 1, 2, \ldots, n$$

where:

n is the number of independent trials, p is the probability of success in a single trial.

2. **Introducing the Poisson Parameter**

Now, we consider the case where n grows infinitely large ($n \to \infty$), but the probability p decreases in such a way that the expected value λ remains constant:

$$\lambda = np$$

This means that as

- n increases as $(n \to \infty)$,
- the probability of success p decreases proportionally as λ/n.

3. **Taking the Limit**

Substituting

$$p = \lambda/n$$

into the binomial probability formula:

$$P(X = k) = \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

Expanding the binomial coefficient $\binom{n}{k}$:

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!} = \frac{n\,(n-1)\cdots(n-k+1)}{k!}$$

For large n, we approximate:

$$\frac{n(n-1)\ldots(n-k+1)}{n^k} \doteq 1$$

as the highest power at n in the numerator is k.

We also use the well-known limit:

$$\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right) = e^{-\lambda}$$

Thus:

$$\left(1-\frac{\lambda}{n}\right)^{n-k}=\left(1-\frac{\lambda}{n}\right)^{n}\left(1-\frac{\lambda}{n}\right)^{-k}=e^{-\lambda}\cdot 1=e^{-\lambda}$$

This leads to:

$$P\left(X=k\right)=\frac{\lambda^{k}}{k!}e^{-\lambda}$$

which is exactly the Poisson distribution.

Memory-less property of the Exponential Distribution

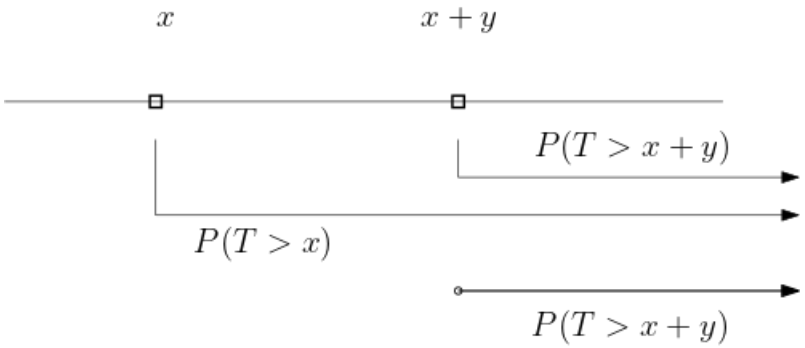
The probability of waiting at least $x+y$ time units given that you've already waited x time units is the same as just waiting for y more units.

Formally:

$$P(T>x+y\mid T>x)=P(T>y)$$

$P\left(T > x + y | T > x\right) = P\left(\left[T > x + y\right] \cap \left[T > x\right]\right) = P\left(T > x + y\right)$

as evident from the picture



For exponential distribution we have $F_T\left(x\right) = 1 - e^{-\lambda x}$ and $P\left(T > x\right) = 1 - F_T\left(x\right) = e^{-\lambda x}$.

Then

$$P(T > x + y \mid T > x) = \frac{P\left(\left[T > x + y\right] \cap \left[T > x\right]\right)}{P\left(T > x\right)} =$$

$$= \frac{P\left(\left[T > x + y\right]\right)}{P\left(T > x\right)} = \frac{e^{-\lambda(x+y)}}{e^{-\lambda x}} = \frac{e^{-\lambda x} e^{-\lambda y}}{e^{-\lambda x}} = e^{-\lambda y} = P\left(T > y\right)$$

Q.E.D

5 Population and Sample

Intro to sampling

Random variable

We have random variable (random experiment) X

Its description is

$$f(x|\theta)$$

with some unknown parameter θ which needs to be estimated.

It shows where data fall

Random sample

$$\vec{X} = [X_1, X_2, \cdots X_n]$$

Sample realization

$$\vec{x} = [x_1, x_2, \cdots x_n]$$

Statistics

Statistics - function of the sample (estimator of θ) $\quad T = T_\theta\left(\vec{X}\right)$

Estimate of θ

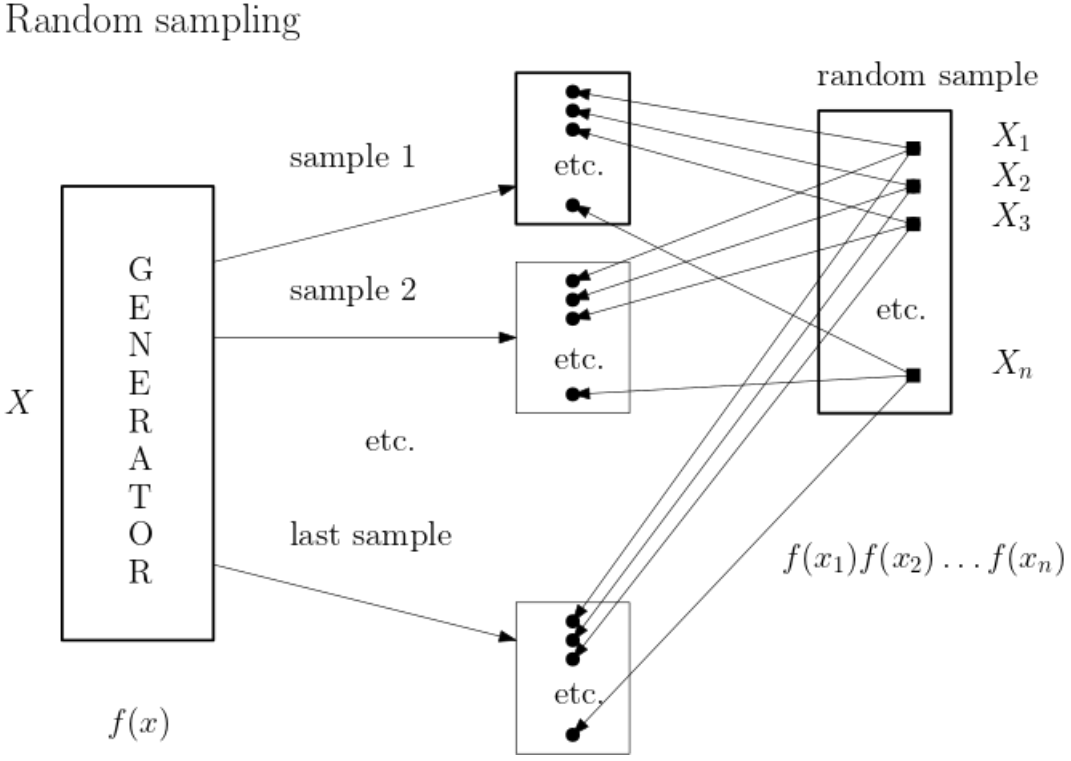
$$T_\theta\left(\vec{x}\right) = \hat{\theta}$$

Its description

$$f\left(T\right)$$

It shows where parameter estimates fall

Repetitive sampling



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

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

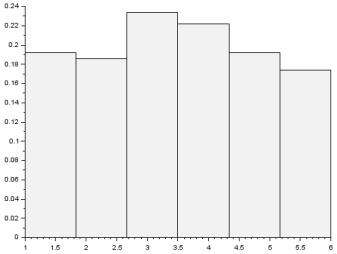
! ! ! ! !

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

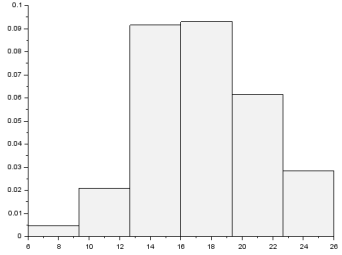
Central limit theorem

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

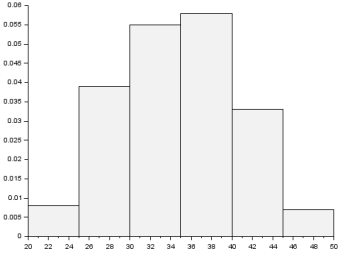
Example: 200 throws of dice gives the histogram



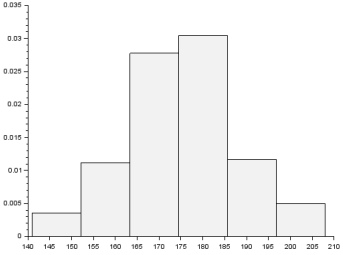
200 samples of the sum of 5 throws



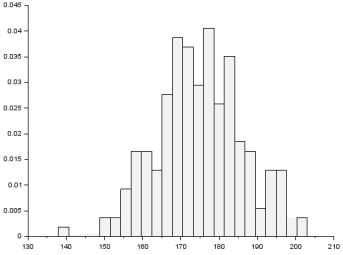
Sum of 10 throws



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 $\bar{x} = 2.2$

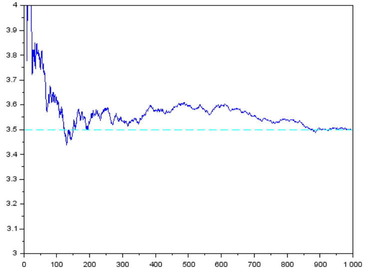
Sample with 10 entries $\bar{x} = 2.7$

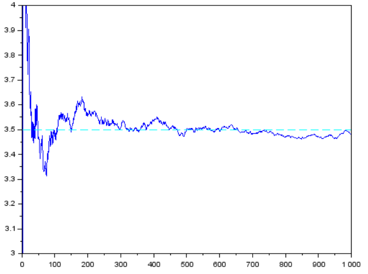
Sample with 30 entries $\bar{x} = 4.27$

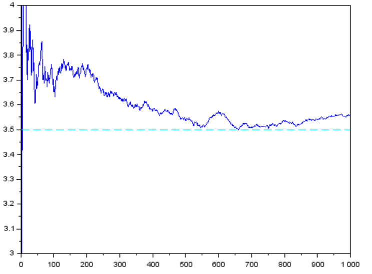
Sample with 100 entries $\bar{x} = 3.56$

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

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



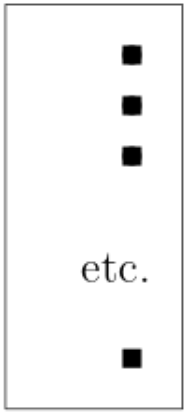




6 Estimation

Distribution of the statistic

random sample



T is the statistics for estimation θ

$X = [x_1, x_2, \ldots x_n]$ $\theta = T(X)$

T is transformation $X \to \theta$

$\downarrow$

$f(x) \to f(\theta)$ with the transformation T

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.

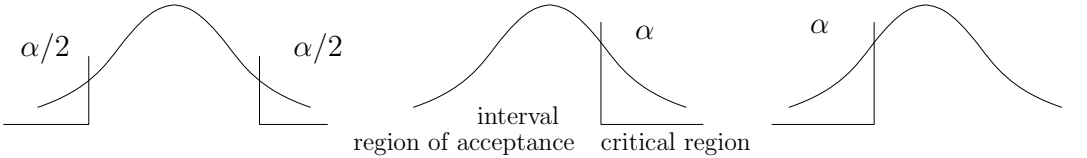
Interval estimate is based on the probability function of the Statistics.

- 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%.