Rozpoznávanie obrazcov - 3. cvicečenie
Štatistika II.

Viktor Kocur
viktor.kocur@fmph.uniba.sk

DAI FMFI UK

2.3.2020

Random variable

Random variable

A random variable is described as a variable whose values depend on outcomes of a random phenomenon.

Probability mass function

Probability density function - describes probability that a random variable would have a given value.

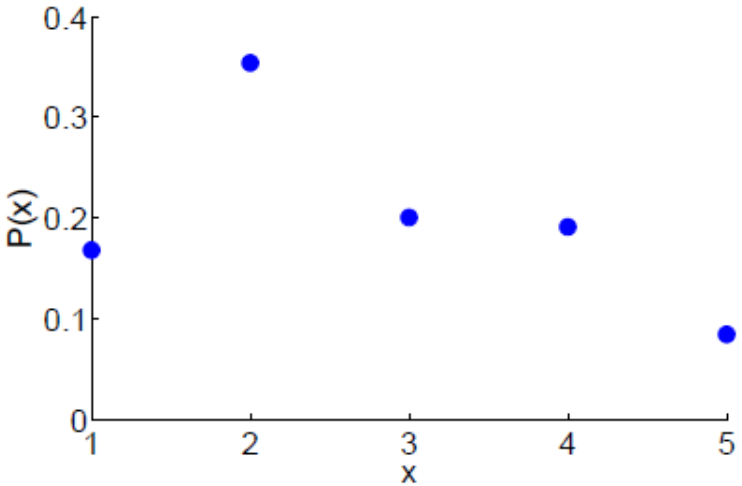
Probability density function

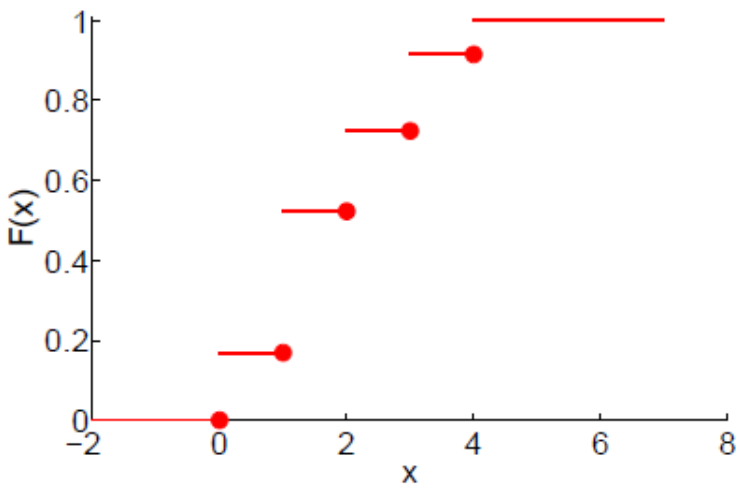
Probability density function - describes probability that a random variable would fall within a given range.

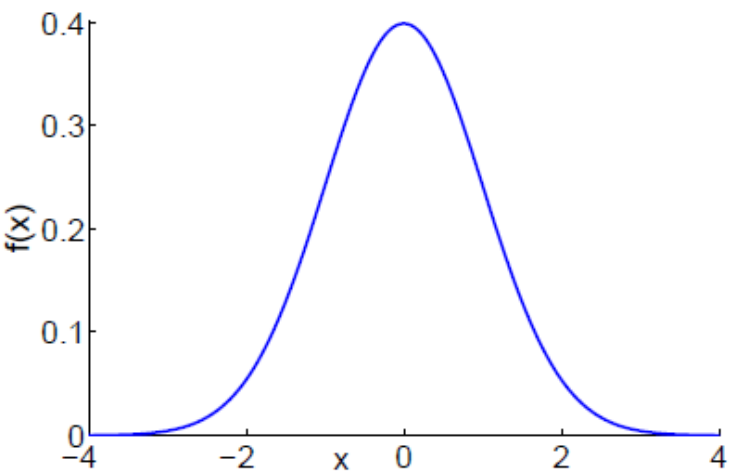
Cummulative distribution function

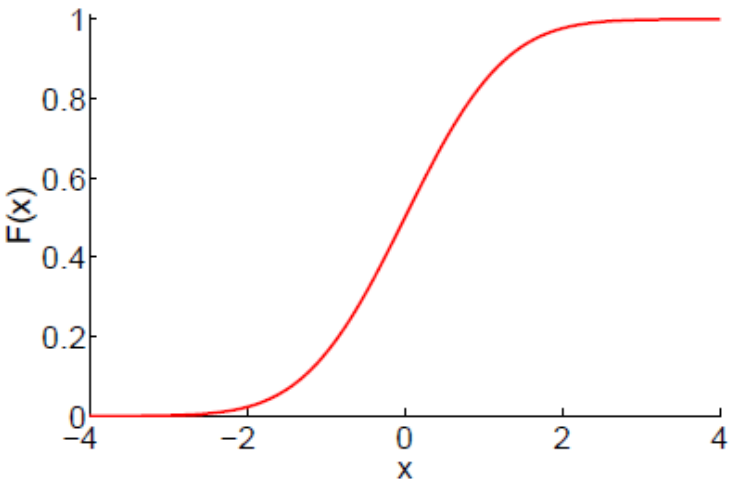
A function which for each value X determines the probability $P(x < X)$.

PMF, PDF and CDF









Bernoulli scheme

Bernoulli schéma

Let us consider n independent experiments. The probability that each experiment succeeds is p. Then for the variable X which determines the number of successful experiments we get:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{1}$$

12. Exercise

A student has to finish and exam with 10 questions. Each question has 4 possible answer and only one of them is correct. What is the probability that student who is guessing completely randomly will a) guess at least 5 questions correctly b) at most 5 questions correctly

Solution a)

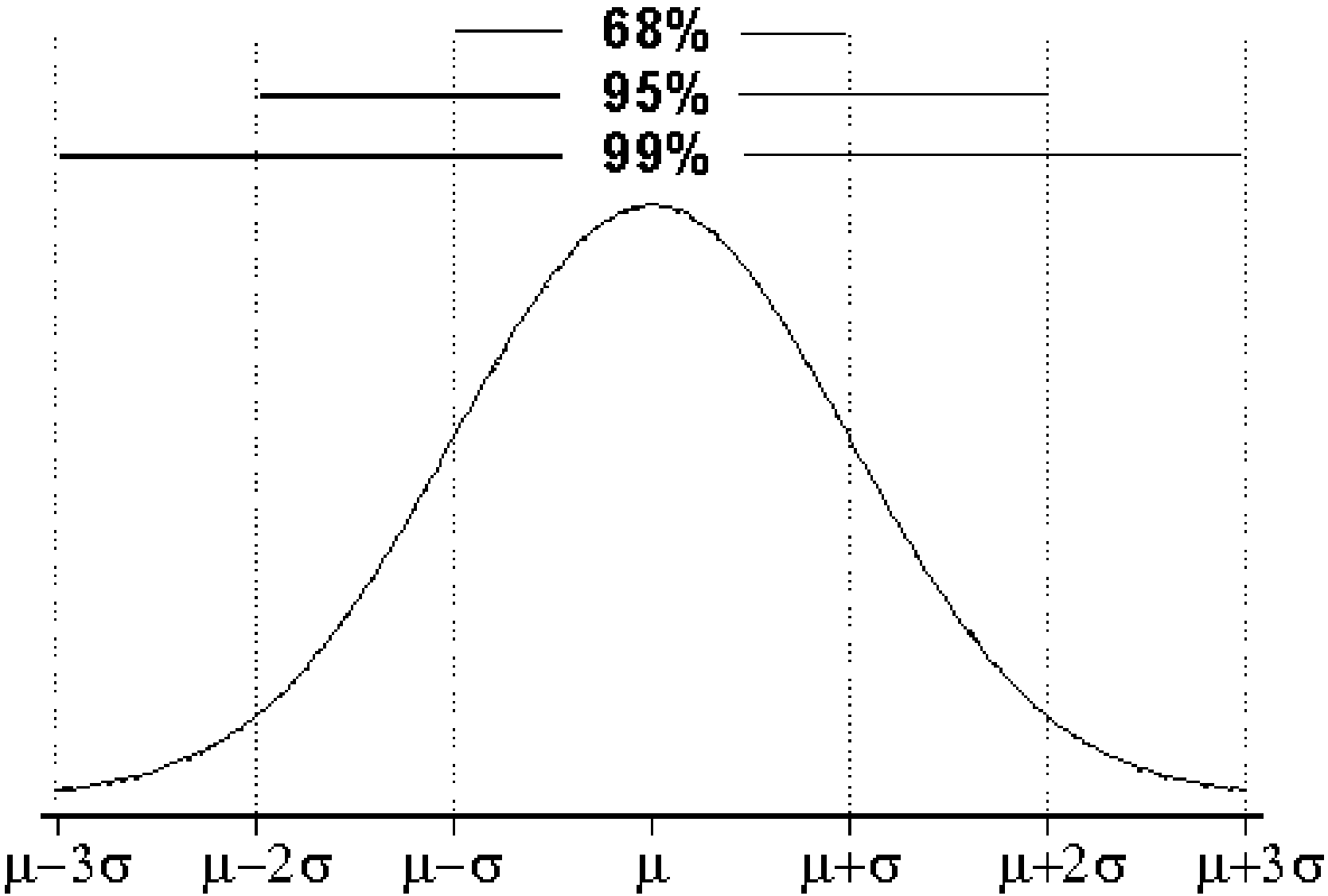
- $P(A) = P(X = 5) + P(X = 6) + P(X = 7) + P(X = 8) + P(X = 9) + P(X = 10)$
- $P(X = 5) = \binom{10}{5} 0.25^5 \cdot 0.75^5$
- $P(A) = \sum_{k=5}^{10} \binom{10}{k} 0.25^k \cdot 0.75^{10-k}$

13. Exercise

Approximately 75% of tourists like bryndzové halušky. What is the probability that from 20 tourists a) at least 17 will like halušky b) all of them would like halušky.

- $P(A) = P(X = 17) + P(X = 18) + P(X = 19) + P(X = 20)$
- $P(X = 20) = \binom{20}{20} 0.75^{20} \cdot 0.25^{0} = 0.75^{20}$
- $P(A) = \sum_{k=17}^{20} \binom{10}{k} 0.75^{k} \cdot 0.25^{10-k}$

Standard deviation



Approxiamting distribution parameters

Sample mean

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

Sample standard deviation

$$S = \sqrt{S^2}$$

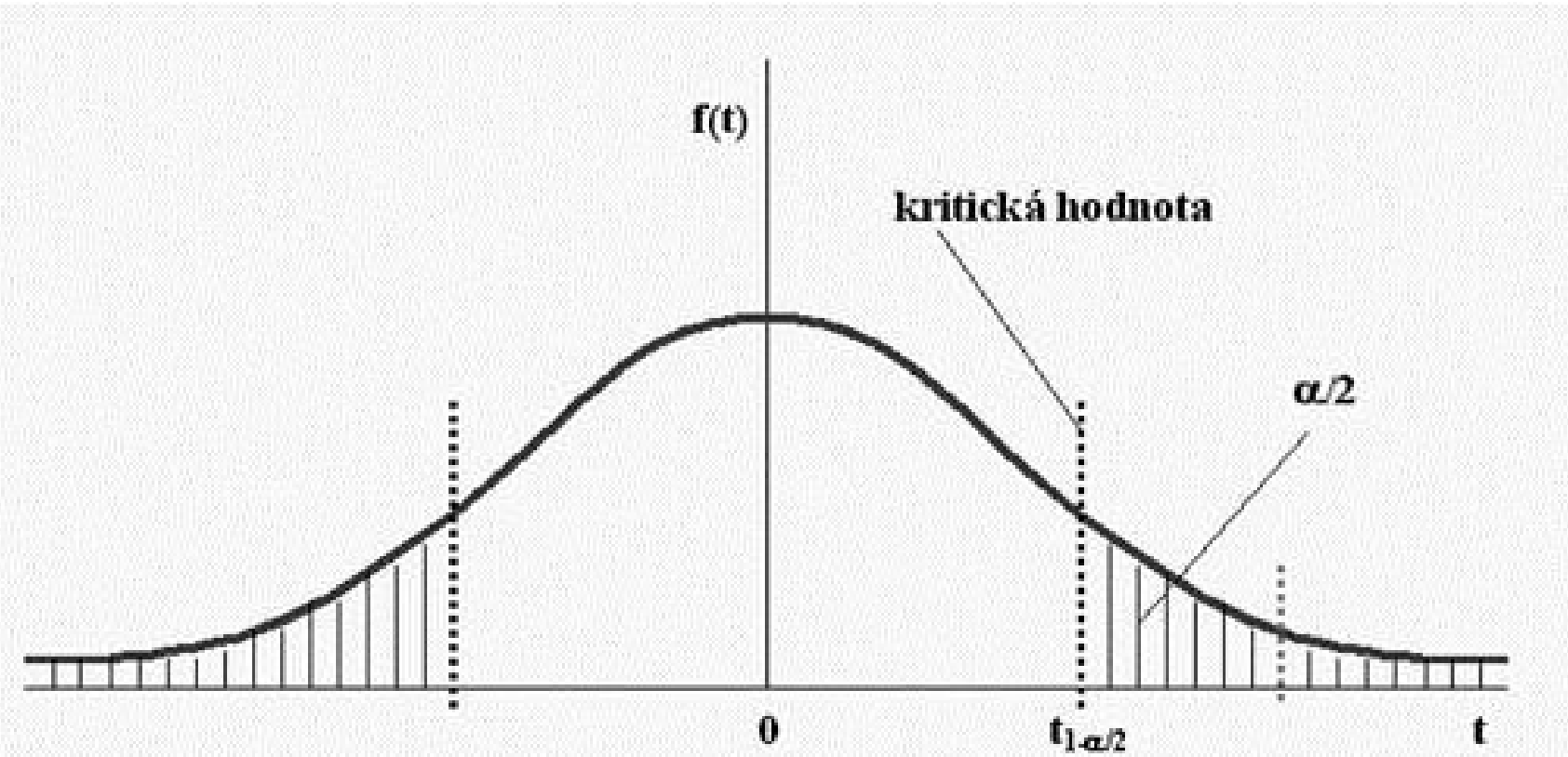
Sample covariance

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$$

Approxiamting distribution parameters

Confidence intervals

$P(G_D < \theta < G_H) = 1 - \alpha$



Approxiamting distribution parameters

α	0.01	0.02	0.05	0.1	0.2
$u_{\alpha/2}$	2.5758	2.3263	1.9599	1.6448	1.299

$$X \sim N(0,1) P(|X| > u_{\alpha/2}) = \alpha$$

$$\begin{aligned} 1-\alpha &= P(-u_{\alpha/2} < U < u_{\alpha/2}) \\ &= P(-u_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma} < u_{\alpha/2}) \\ &= P(\overline{X} - u_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + u_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}) \end{aligned}$$

Test statistics

If we know the σ value of the original distribution we can use normal distribution:

$$u = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

If we do not know it and we have $n > 30$ we will use:

$$u = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Otherwise we use the Student distribution:

$$t = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Test statistics - Matlab

Values of u_α and t_α

For a given confidence interval we can find values of u_α and t_α from tables. Today we will use Matlab functions.

norminv

norminv(alpha) - returns critical value for given alpha for normal distribution

tinv

tinv(alpha, n) - returns critical value for given alpha for Students distribution with n degrees of freedom

Note

If we desire to obtain a centered confidence interval of 0.95 we need to use $\alpha = 0.975$ (or $[0.025, 0.0975]$).

14. Exercise

Let us assume that the height of boys of ages 9-10 is distributed normally with unknown mean and standard deviation$\sigma^2 = 39.112$. We measured a height of 15 boys and calculated the sample mean as 139.13 cm. Determine the 99% confidence interval for this value.

- $n = 15, \sigma = 6.253, \overline{X} = 139.13$
- $1 - \alpha = P(\overline{X} - u_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + u_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}})$
- $139.13 \pm 2.5758 \cdot \frac{6.253}{\sqrt{15}}$
- $134.97 \leq \mu \leq 143.28$

15. Exercise

An airliner estimates the average number of travelers. In the last 20 days the average number of travelers was 112 with sample variance of 25. Determine the 95% confidence interval for the mean of number of travelers μ.

- $n = 20, S = 5, \overline{X} = 112$
- $1 - \alpha = P(\overline{X} - t_{\alpha/2,n-1} \cdot \frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{\alpha/2,n-1} \cdot \frac{S}{\sqrt{n}})$
- $112 \pm 2.093 \cdot \frac{5}{\sqrt{20}}$
- $109.65 \leq \mu \leq 114.34$

16. Exercise

A random variable X has a normal distribution. The mean and variance are unknown. We measured the following values of X: 27, 15, -3, -6, 12, 20, 13, 0, 7, 10. Determine the 95% confidence interval for the distribution mean.

- $n = 10, S = 10.319, \overline{X} = 9.5$
- $1 - \alpha = P(\overline{X} - t_{\alpha/2,n-1} \cdot \frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{\alpha/2,n-1} \cdot \frac{S}{\sqrt{n}})$
- $9.5 \pm 2.262 \cdot \frac{10.319}{\sqrt{10}}$
- $2.118 \leq \mu \leq 16.881$

17. Exercise

We picked a sample from a normal distribution with known variance $\sigma^2 = 0.66$. The picked values are: 1.3, 1.8, 1.4, 1.2, 0.9, 1.5, 1.7. Determine the 95% confidence interval for the mean μ of the distribution.

- $n = 7, \sigma = 0.245, \overline{X} = 1.4$
- $1 - \alpha = P(\overline{X} - u_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + u_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}})$
- $1.4 \pm 1.9599 \cdot \frac{0.245}{\sqrt{7}}$
- $1.218 \leq \mu \leq 1.581$

Hypothesis testing - Matlab

ztest

[h, p, ci] = ztest(X, m, sigma, 'Alpha', alpha) - returns a test result for a hypothesis that data in vector X are from a normal distribution with mean of m and standard deviation sigma. h contains 1 if the hypothesis is not confirmed for a given critical value of alpha, otherwise it is zero. ci will contain the confidence interval.

ttest

[h, p, ci] = ttest(X, m, 'Alpha', alpha) - returns a test result for a hypothesis that data in vector X are from a normal distribution with mean of m and unknown standard deviation. h contains 1 if the hypothesis is not confirmed for a given critical value of alpha, otherwise it is zero. ci will contain the confidence interval.

18. Exercise

We claim that bearing made with an automatic lathe have a diameter mean of 10mm. Using a test with critical values $\alpha = 0.05$ test the hypothesis that if we pick 16 random bearings then their mean is 10.3mm for a) $\sigma^2 = 1$ b)$S^2 = 1.21$
Solution a)

- $n = 16, \sigma = 1, \overline{X} = 10.3$
- $1 - \alpha = P(\overline{X} - u_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + u_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}})$
- $10.3 \pm 1.9599 \cdot \frac{1}{\sqrt{16}}$
- $9.81 \leq \mu \leq 10.789$
- We do not reject the hypothesis

18. Exercise

We claim that bearing made with an automatic lathe have a diameter mean of 10mm. Using a test with critical values $\alpha = 0.05$ test the hypothesis that if we pick 16 random bearings then their mean is 10.3mm for a) $\sigma^2 = 1$ b)$S^2 = 1.21$
Solution b)

- $n = 16, S = 1.1, \overline{X} = 10.3$
- $1 - \alpha = P(\overline{X} - t_{\alpha/2,n-1} \cdot \frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{\alpha/2,n-1} \cdot \frac{S}{\sqrt{n}})$
- $10.3 \pm 2.131 \cdot \frac{1.1}{\sqrt{16}}$
- $9.71 \leq \mu \leq 10.88$
- We do not reject the hypothesis