# Biostatistika

Taťána Bržezická, Ondřej Dvorský, Daniela Gachová,
Jan Gečnuk, Natálie Němcová

# Osnova

- Základní pojmy
- Data
- Srovnání průměru a mediánu
- Směrodatná odchylka, chyba průměru
- Test odlehlých hodnot
- Vybraná rozdělení pravděpodobnosti
- Testování hypotéz v biostatistice
  - Parametrické testy
  - Neparametrické testy

# *Úvod*

**Obecně o biostatistice**

Biostatistika primárně vychází ze statistiky

- Na rozdíl od statistiky je více zaměřená do praxe
- Aplikace a vývoj statistických metod pro řešení biologických a klinických problémů

Získání užitečných informací z pozorovaných dat

- prostý popis stavu sledovaného souboru
- identifikace faktorů ovlivňujících chování souboru
- rozhodnutí o nějaké jeho neznámé charakteristice

Důsledek získané informace

- žádný (pouze informace pro hodnotitele)
- výrazná změně lidské činnosti
- např. ke změně metodických a léčebných postupů nebo klinických doporučení

(účinnost a bezpečnost léčivých přípravků v klinických studiích)

# *Úvod*

**Cíl biostatistiky**

Zásadní postavení biostatistiky v dnešní vědě a výzkumu

- statistické zpracování experimentálních výsledků (hlavně biomedicínská data)

Aby při hodnocení experimentů na základě limitovaných dat a údajů nedošlo k nesprávným interpretacím a závěrům

Hlavním cílem je získání informace o tzv. **cílové populaci** (základním souboru)

- Ve většině případů nereálné a sleduje se pouze část cílové populace, tzv. **výběr z cílové populace** (experimentální vzorek)

# *Základní pojmy*

## Obecný postup

- **Cílová populace**
- Předpoklad určitého pravděpodobnostního chování (**model**)
- Vyjádříme **hypotézu**
- Experimentální vzorek (**výběr z cílové populace**)
- Sledované vlastnosti převedeme na číselné vyjádření (**data**)
  - Kvalitativní
  - Kvantitativní
- Platnost hypotézy vyhodnotíme na základě vybraného modelu a pozorovaných dat

# Základní pojmy

**Klíčové body pro korektní hodnocení**

Zkreslení výsledků

- snažíme se vyhnout zkreslení hodnot sledované náhodné veličiny veličinami, které nejsou cílem studie
- tj. zavádějící veličina (př. nošení zapalovače nezpůsobí rakovinu plic, pouze koreluje s kouřením)

Reprezentativnost

- experimentální vzorek odpovídá cílové populaci
- př. odhad střední výšky české dospělé populace (vzorkem by neměla být mužská basketbalová reprezentace – zkreslení výsledku)

Srovnatelnost

- při srovnání dvou a více skupin (srovnávání jablek s jablky a ne jablek s hruškami)
- srovnatelnost zajištěna pomocí tzv. randomizace nebo srovnáváním výsledků experimentu v rámci podskupin

# Základní pojmy

**Klíčové body pro korektní hodnocení**

Spolehlivost

- kvantifikace sledovaného znaku **bodový odhad** (rozdílný odhad u měření 10 a 1000 jedinců), pro úplnost zmínit **intervalový odhad** (interval spolehlivosti)

- interval, který se zvolenou pravděpodobností pokrývá neznámý parametr, který se snažíme odhadnout bodovým odhadem

Významnost

- statistická významnost - na základě pravděpodobnosti hodnotí výsledek experimentu, zda pozorovaný rozdíl mezi dvěma skupinami vznikl náhodou či ne
- praktická významnost - z hlediska experimentátora na základě pozorovaného efektu vedle statistické významnosti zda je biologicky/klinicky podstatný

# *Data*

**Kvalitativní (kategoriální)**

- Binární data - nabývají pouze dvou hodnot, většinou data typu ano/ne (př. osoba s diabetem / osoba bez diabetu)
- Nominální data - více kategorií, které nelze vzájemně seřadit (př. krevní skupina A/B/AB/0)
- Ordinální data - více kategorií, lze vzájemně seřadit (př. stadium maligního onemocnění I/II/III/IV)

**Kvantitativní (numerická)**

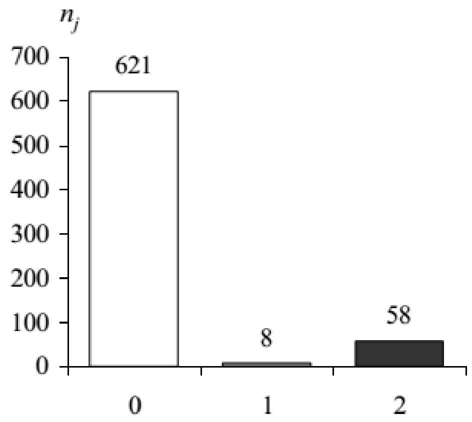
- Spojitá data - jakékoliv hodnoty v určitém intervalu (př. hmotnost osob, velikost nádoru nebo teplota)
- Diskrétní data - nabývají pouze spočetně mnoha hodnot, na reálné ose jsou zobrazena pomocí izolovaných bodů (př. počet krevních buněk v 1 ml krve)

# Popis a vizualizace dat

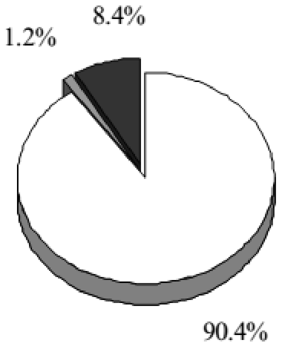
Tabulka četností

| Přítomnost diabetu | $y_i$ | $n_i$ | $n_i / n$ | $n_i / n$ (%) |
|---|---|---|---|---|
| Bez diabetu | 0 | 621 | 0,904 | 90,4 % |
| Diabetes 1. typu | 1 | 8 | 0,084 | 1,2 % |
| Diabetes 2. typu | 2 | 58 | 0,012 | 8,4 % |
| Celkem | | 687 | 1 | 100 % |

Sloupcový graf




Výsečový graf




**Cíl:** pozorovaná data graficky zpřehlednit a poskytnout maximum informací na minimální ploše

**Kvalitativní data**

- Tabulka četností
- Sloupcový graf, výsečový (koláčový) graf

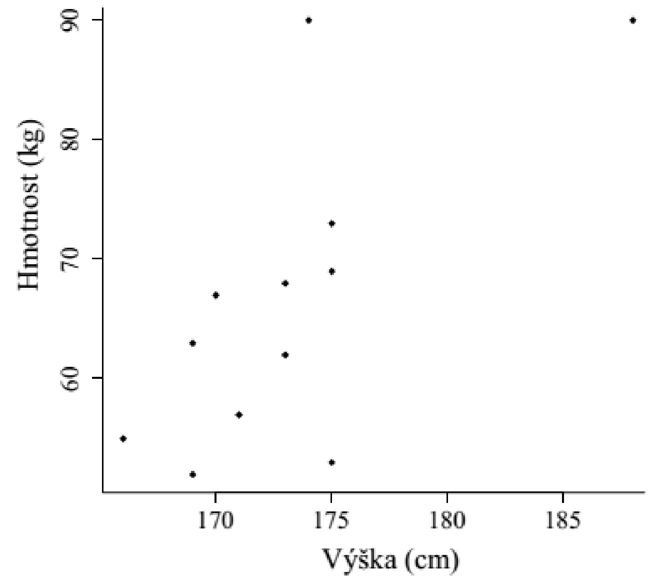
# Popis a vizualizace dat

**Kvantitativní data**

- Míra polohy – shrnuje soubor dat jedním číslem a představuje „typickou hodnotu“ – **průměr, medián**

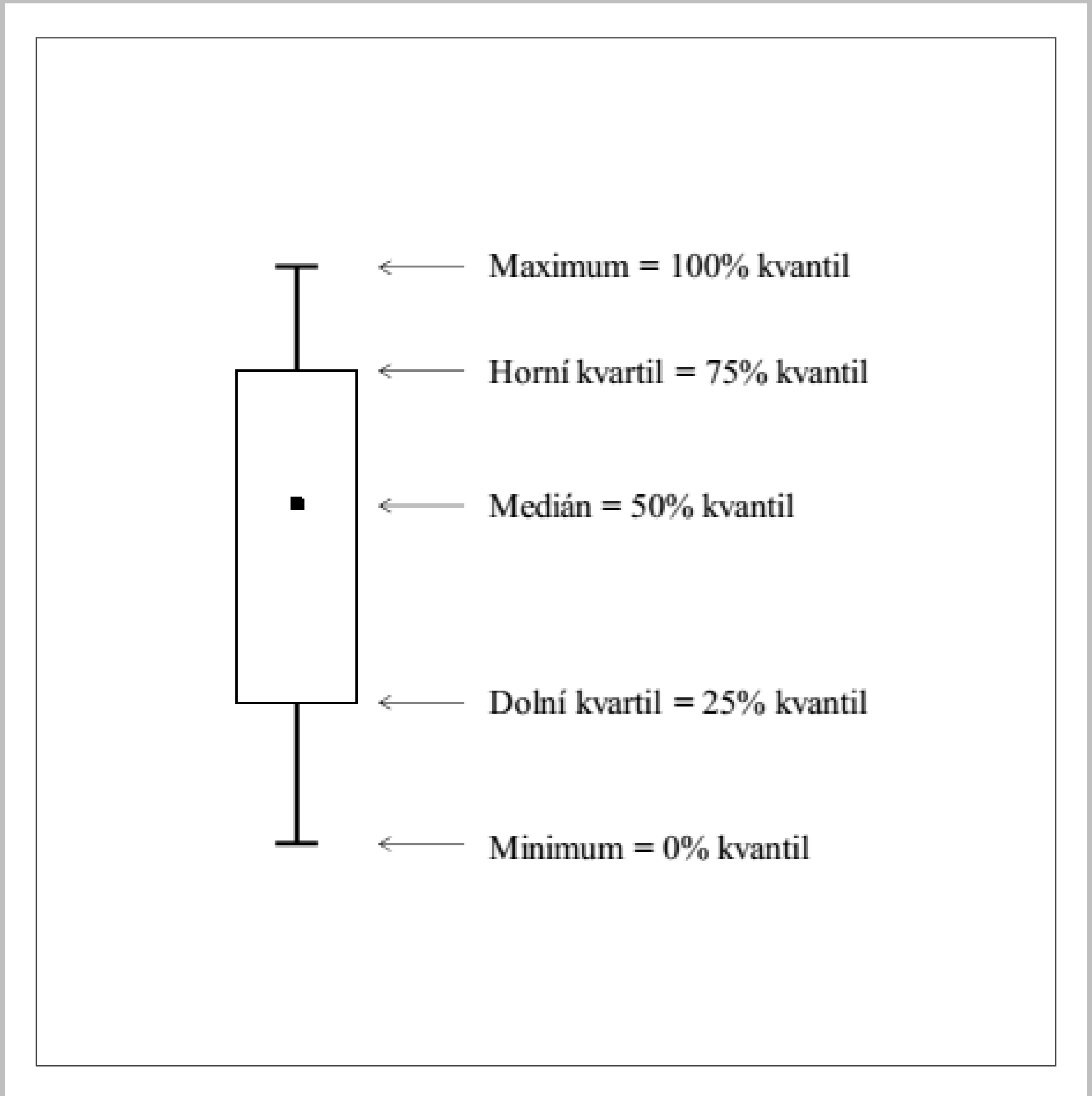
- Míra variability – jak jsou kolem „typické hodnoty“ rozloženy ostatní hodnoty
  – **rozpětí** = rozsah hodnot (interval min a max hodnoty) - náchylnost k odlehlým hodnotám
  – kvantilové rozpětí (interval definovaný hodnotami p% kvantilu a [100 – p]% kvantilu)
  – kvartilové rozpětí (dán dolním a horním kvartilem, pokrývá 50 % pozorovaných hodnot)

- Bodový graf, krabicový graf, histogram




# Bodový graf

- zobrazuje každou měřenou hodnotu jako bod plochy
- použití zejména pro vizualizaci vzájemného vztahu dvou veličin spojitého typu
- hodnoty jedné veličiny jsou zobrazeny na ose x
- hodnoty druhé veličiny jsou zobrazeny na ose y



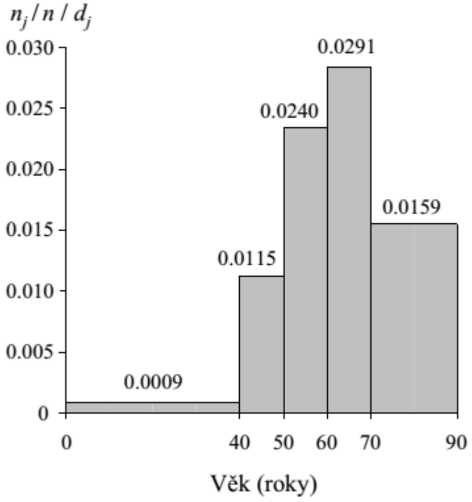

# Krabicový graf

- kvartilové rozpětí – ohraničuje 50% pozorovaných hodnot (hranice horní a dolní kvartil)

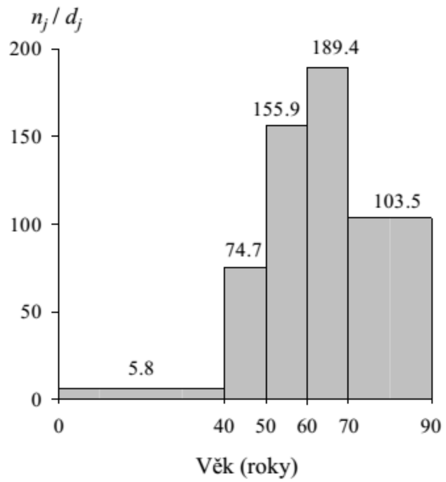
**Příklad pacientek s karcinomem prsu**

| Věkový interval | $d_j$ | $n_j$ | $n_j / n$ | $n_j / d_j$ | $n_j / n / d_j$ |
|---|---|---|---|---|---|
| 0–39 let | 40 | 231 | 0,036 | 5,8 | 0,0009 |
| 40–49 let | 10 | 747 | 0,115 | 74,7 | 0,0115 |
| 50–59 let | 10 | 1559 | 0,240 | 155,9 | 0,0240 |
| 60–69 let | 10 | 1894 | 0,291 | 189,4 | 0,0291 |
| 70 a více let | 20 | 2069 | 0,318 | 103,5 | 0,0159 |
| Celkem | 90 | 6500 | 1 | - | - |

**Relativní četnost**




**Absolutní**




**n** … celkový počet hodnot sledované veličiny
**$d_j$** … šířka intervalu
**$n_j$** … počet pozorovaných hodnot v intervalu

# Histogram

- pro vizualizaci poměrových a intervalových dat
- připomíná sloupcový graf
  - Rozdíl: každý sloupec v histogramu odráží absolutní nebo relativní četnost na jednotku sledované veličiny na vodorovné ose
- seřazení hodnot dle velikosti
- rozdělení do vzájemně disjunktních intervalů
- standardizace na šířku intervalu (70+ vypadá četnější než 60-69 viz tabulka, ale není, viz graf)

# SROVNÁNÍ PRŮMĚRU A MEDIÁNU

- ukazatele středu souboru dat
- jedno číslo, které představuje „typickou hodnotu“, kolem které mají ostatní pozorované hodnoty tendenci kolísat

## PRŮMĚR

- vypočten ze všech pozorovaných hodnot
- symetrická data, neobsahují odlehlé hodnoty

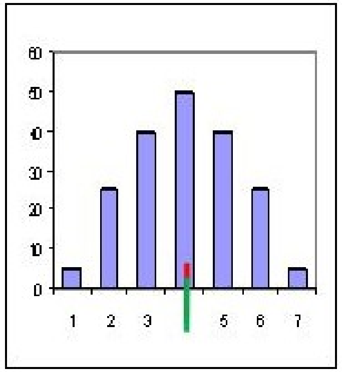
$$X = \frac{1}{n}\sum_{i=1}^{n} x_i$$

# MEDIÁN

- prostřední hodnota – dělí celý soubor na dvě poloviny
- asymetrická data, přítomnost odlehlé hodnoty
- např. výpočet průměrného platu

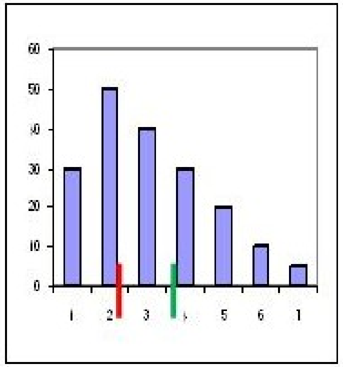
- zvážit, co použiju, uvést obě hodnoty

**Symetrická data**

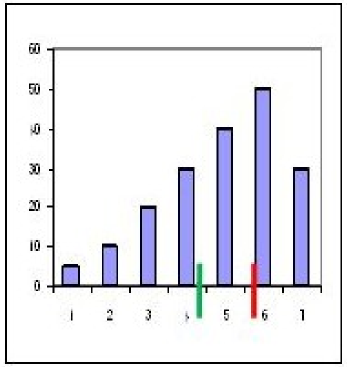



Průměr = medián

**Asymetrická data**



Průměr > medián



Průměr < medián

# SMĚRODATNÁ ODCHYLKA, CHYBA PRŮMĚRU

- ukazatele šířky rozložení
  - např. větší variabilita u hodnot 0-100 než u 40-60

ROZPTYL $s^2 = \frac{\sum(x_i - X)^2}{n - 1}$

- získaný na základě odchylky jednotlivých hodnot od průměru
- nejvyšší vypovídající schopnost v případě symetrického rozdělení

SMĚRODATNÁ ODCHYLKA (SD) – variabilita pozorované proměnné

- druhá odmocnina z rozptylu
- pro n < 7:

$$s = k_n R$$

X …. průměr
R …. rozpětí
n …. počet měření

## odhad směrodatné odchylky – tabelované hodnoty $k_n$

$$s = k_n \cdot R$$

| *n* | $k_n$ |
|---|---|
| 2 | 0,886 |
| 3 | 0,591 |
| 4 | 0,486 |
| 5 | 0,430 |
| 6 | 0,395 |
| 7 | 0,370 |
| 8 | 0,351 |
| 9 | 0,337 |
| 10 | 0,325 |

RELATIVNÍ SMĚRODATNÁ ODCHYLKA (RSD; %)

- vhodné pro srovnání variability více souborů, které se liší úrovní hodnot

$$s_r = \frac{s}{X} \times 100$$

STŘEDNÍ CHYBA PRŮMĚRU

- směrodatná odchylka rozložení průměru
- měří se rozptýlenost vypočteného X v různých výběrových souborech vybraných z 1 velkého základního souboru → s menší chybou

INTERVAL SPOLEHLIVOSTI

- interval, v němž s danou pravděpodobností leží správná hodnota μ
- pro n <7: $\mu = X \pm K_n R$
- pro n > 7: $\mu = X \pm s\frac{t_\alpha}{\sqrt{n}}$

# TEST ODLEHLÝCH HODNOT

- Q-test a T-test
- zjišťujeme, zda se krajní hodnoty souboru statisticky významně liší od ostatních paralelních měření

## DEAN-DIXONŮV Q-TEST – pro n<7

$$Q_1 = \frac{|x_2 - x_1|}{R} \qquad Q_n = \frac{|x_n - x_{n-1}|}{R}$$

- nalezené hodnoty $Q_1$ nebo $Q_n$ se srovnávají s tabelovanou hodnotou $Q_k$
- když $Q_1$ nebo $Q_n < Q_k$ → výsledek NENÍ odlehlý a zůstane součástí souboru dat
- když $Q_1$ nebo $Q_n > Q_k$ → výsledek JE odlehlý a vyloučí se ze souboru dat

# GRUBSŮV T-TEST – pro n>7

$$T_1 = \frac{|X - x_1|}{s} \qquad T_n = \frac{|x_n - X|}{s}$$

- nalezené hodnoty $T_1$ nebo $T_n$ se srovnávají s tabelovanou hodnotou $T_k$
- když $T_1$ nebo $T_n < T_k$ → výsledek NENÍ odlehlý a zůstane součástí souboru dat
- když $T_1$ nebo $T_n > T_k$ → výsledek JE odlehlý a vyloučí se ze souboru dat

Postup při Q- nebo T-testu:

- výsledky daného stanovení se seřadí podle velikosti hodnot → $x_1 < x_2 < x_3 < x_4 \ldots < x_n$
- otestují se pouze krajní hodnoty souboru dat podle Q- nebo T- testu
- pokud ani jedna z krajních hodnot není odlehlá, tak počítáme se všemi hodnotami souboru dat
- pokud je alespoň jedna z hodnot odlehlá, vyloučíme ji ze souboru dat a pokračujeme opět krokem 1)

- Př. Ve vzorku multivitaminového přípravku byl pomoci metody AAS stanovován obsah Zn. Opakovaným měřením byly získány tyto obsahy: 164,1 mg/l, 165,0 mg/l, 166,9 mg/l, 157,2 mg/l, 166,9 mg/l, 163,0 mg/l. Výsledky otestujte na odlehlost a vypočtěte průměrnou hodnotu a směrodatnou odchylku měření.

157,2 mg/l, 163,0 mg/l, 164,1 mg/l, 165,0 mg/l, 166,9 mg/l, 166,9 mg/l

R = 166,9 – 157,2 = 9,7 mg/l

$$Q_1 = \frac{|x_2 - x_1|}{R} = \frac{|157{,}2 - 163{,}0|}{9{,}7} = 0{,}598 \qquad Q_6 = \frac{|x_6 - x_5|}{R} = \frac{|166{,}9 - 166{,}9|}{9{,}7} = 0$$

$Q_1 = 0{,}598 > 0{,}560\ (Q_{k6})$ …. **JE** odlehlé

$Q_6 = 0 < 0{,}560\ (Q_{k6})$ …. **NENÍ** odlehlé

163,0 mg/l, 164,1 mg/l, 165,0 mg/l, 166,9 mg/l, 166,9 mg/l

R = 166,9 – 163,0 = 3,9 mg/l

$$Q_1 = \frac{|x_2 - x_1|}{R} = \frac{|164{,}1 - 163{,}0|}{9{,}7} = 0{,}282$$

$Q_1 = 0{,}282 < 0{,}642\ (Q_{k5})$ … **NENÍ** odlehlé

$Q_5 = 0 < 0{,}642\ (Q_{k5})$ …. **NENÍ** odlehlé

X = **165,2 mg/l**

$s = k_n R = 0{,}430 \times 3{,}9 =$ **1,7 mg/l**

# test odlehlosti výsledků – tabelované hodnoty $Q_k$ a $T_k$

| | $T_k$ | | $Q_k$ | |
|---|---|---|---|---|
| n | $\alpha$=0,05 | $\alpha$=0,01 | $\alpha$=0,05 | $\alpha$=0,01 |
| 3 | 1,412 | 1,416 | 0,941 | 0,988 |
| 4 | 1,689 | 1,723 | 0,765 | 0,889 |
| 5 | 1,869 | 1,955 | 0,642 | 0,760 |
| 6 | 1,996 | 2,130 | 0,560 | 0,698 |
| 7 | 2,093 | 2,265 | 0,507 | 0,637 |
| 8 | 2,172 | 2,374 | 0,468 | 0,590 |
| 9 | 2,237 | 2,464 | 0,437 | 0,555 |
| 10 | 2,294 | 2,540 | 0,412 | 0,527 |
| 11 | 2,343 | 2,606 | | |
| 12 | 2,387 | 2,663 | | |

# TEST PRAVDIVOSTI VÝSLEDKU

- Lordův a Studentův test
- naměřený výsledek se srovnává se správnou hodnotou

## LORDŮV TEST – pro n < 7

$$u = \frac{|X - \mu|}{R}$$

- nalezená hodnot u se srovnává s tabelovanou hodnotou $u_k$
- $u < u_k$ → výsledek JE pravdivý

## STUDENTŮV TEST – pro n > 7

$$t = \frac{|X - \mu|}{s} \sqrt{n}$$

- nalezená hodnot t se srovnává s tabelovanou hodnotou $t_k$
- $t < t_k$ → výsledek JE pravdivý

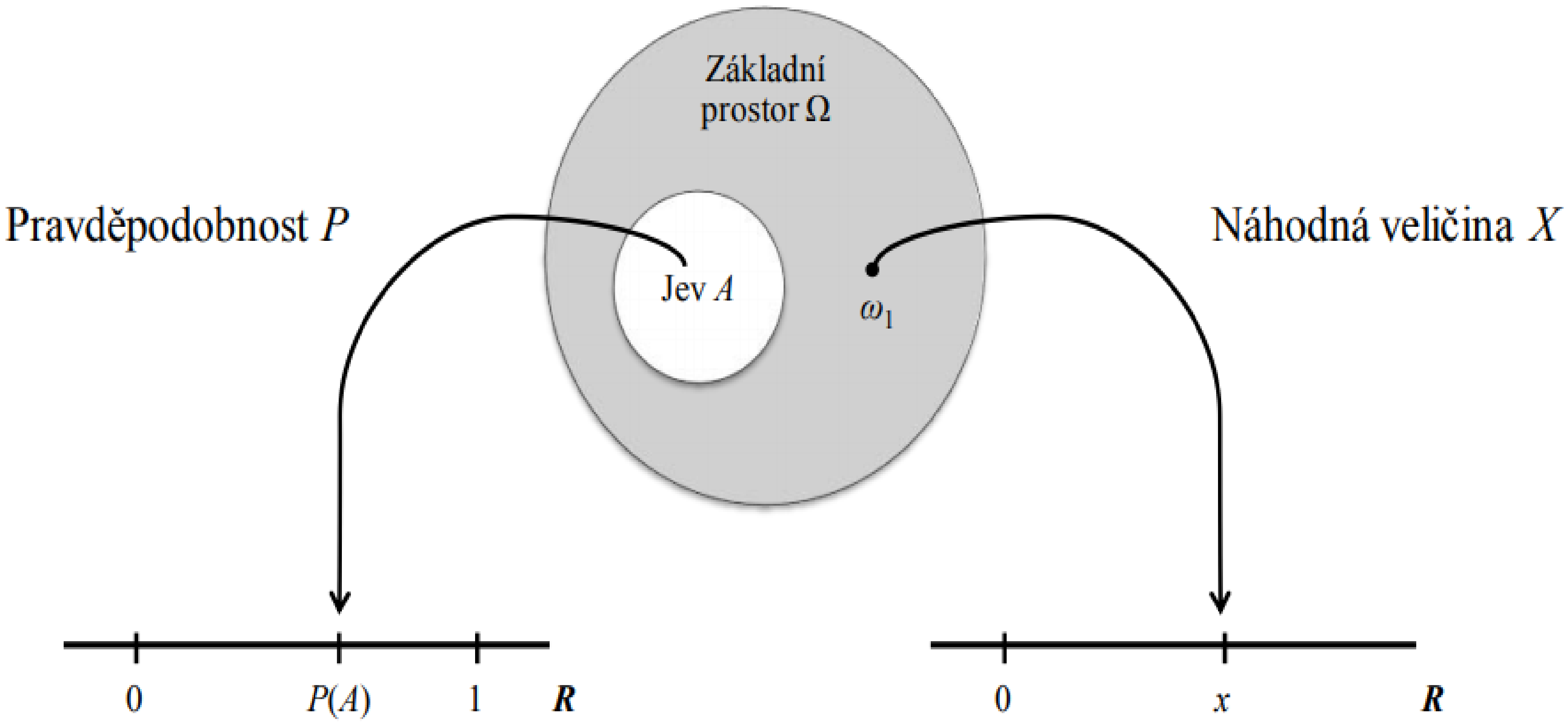
# test pravdivosti výsledků – tabelované hodnoty $u_k$ a $t_k$

| n | $u_{krit}$ | | $t_{krit}$ | |
|---|---|---|---|---|
| | $\alpha$=0,05 | $\alpha$=0,01 | $\alpha$=0,05 | $\alpha$=0,01 |
| 2 | 1,714 | 3,958 | 12,706 | 63,657 |
| 3 | 0,636 | 1,046 | 4,303 | 9,925 |
| 4 | 0,406 | 0,618 | 3,182 | 5,841 |
| 5 | 0,306 | 0,448 | 2,776 | 4,604 |
| 6 | 0,250 | 0,357 | 2,571 | 4,032 |
| 7 | 0,213 | 0,300 | 2,447 | 3,707 |
| 8 | 0,186 | 0,260 | 2,365 | 3,499 |
| 9 | 0,167 | 0,232 | 2,306 | 3,355 |
| 10 | 0,152 | 0,210 | 2,262 | 3,250 |
| 11 | | | 2,228 | 3,169 |

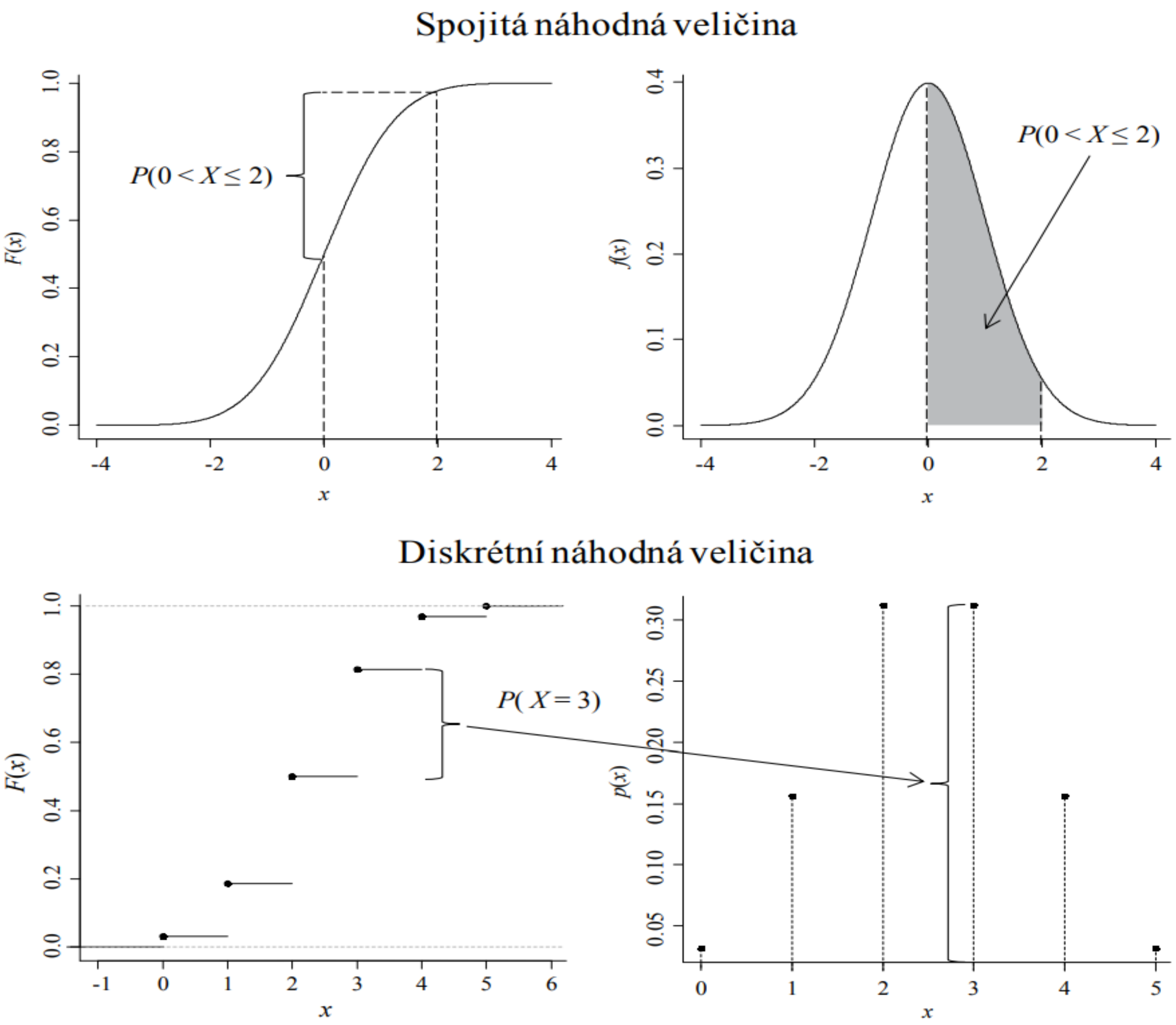
# Vybraná rozdělení pravděpodobnosti a Kaplan-Meierův odhad funkce přežití

Bc. Ondřej Dvorský

# Schematické vyjádření konceptu náhodné veličiny



# Spojité a diskrétní náhodné veličiny



# Statistické metody

- **Neparametrické** - nevyžadují specifikaci konkrétního rozdělení pozorovaných hodnot
- **Parametrické** - vyžadují specifikaci

# Spojitá rozdělení pravděpodobnosti

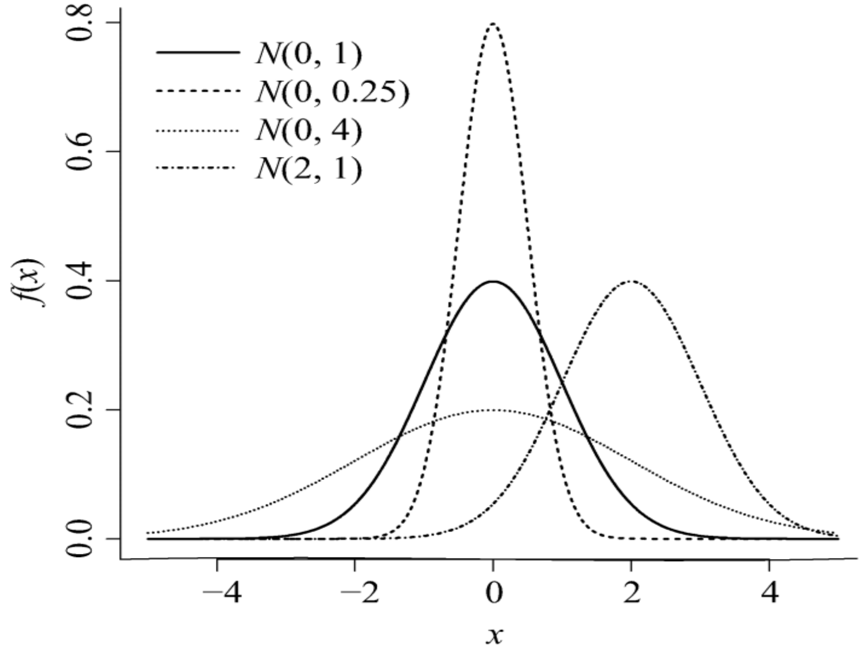
- **Normální**
- Rovnoměrně spojité - f(x) na intervalu (a, b) konstantní a mimo tento interval nulová
- Chí-kvadrát - při konstrukci intervalu spolehlivosti pro rozptyl náhodné veličiny
- **Studentovo *t***
- **Logaritmicko-normální**
- Exponenciální - popisuje délku časových intervalů mezi jednotlivými událostmi, když se tyto události vyskytují vzájemně nezávisle a s konstantní intenzitou
- Fisherovo *F* - sestrojení intervalu spolehlivosti pro podíl dvou rozptylů normálního rozdělení; testování hypotézy o rovnosti středních hodnot veličiny *X*

# Normální rozdělení

- Pro veličiny, jejichž hodnoty se symetricky shlukují kolem střední hodnoty a vytvářejí tak charakteristický tvar hustoty pravděpodobnosti, která je známá také pod pojmem **Gaussova křivka**
- **Parametry $\boldsymbol{\mu}$** a **$\boldsymbol{\sigma^2}$ -** první z nich představuje střední hodnotu normálního rozdělení a druhý představuje rozptyl normálního rozdělení
- $X \sim N(\mu,\sigma^2)$. Hustota náhodné veličiny $\boldsymbol{X}$ má tvar:

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

# Ukázky hustot náhodných veličin s normálním rozdělením




| Interval | Pravděpodobnost realizace uvnitř intervalu | Pravděpodobnost realizace vně intervalu |
|---|---|---|
| $\mu \pm 1\sigma$ | 0,683 | 0,317 |
| $\mu \pm 2\sigma$ | 0,954 | 0,046 |
| $\mu \pm 3\sigma$ | 0,997 | 0,003 |

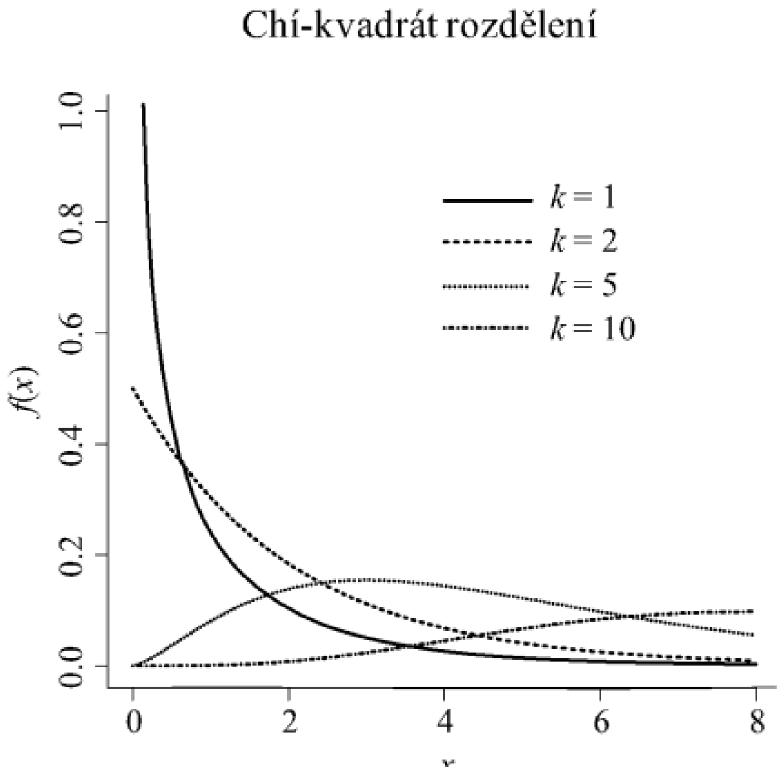
# <u>Studentovo *t* rozdělení – *t*(*k*)</u>

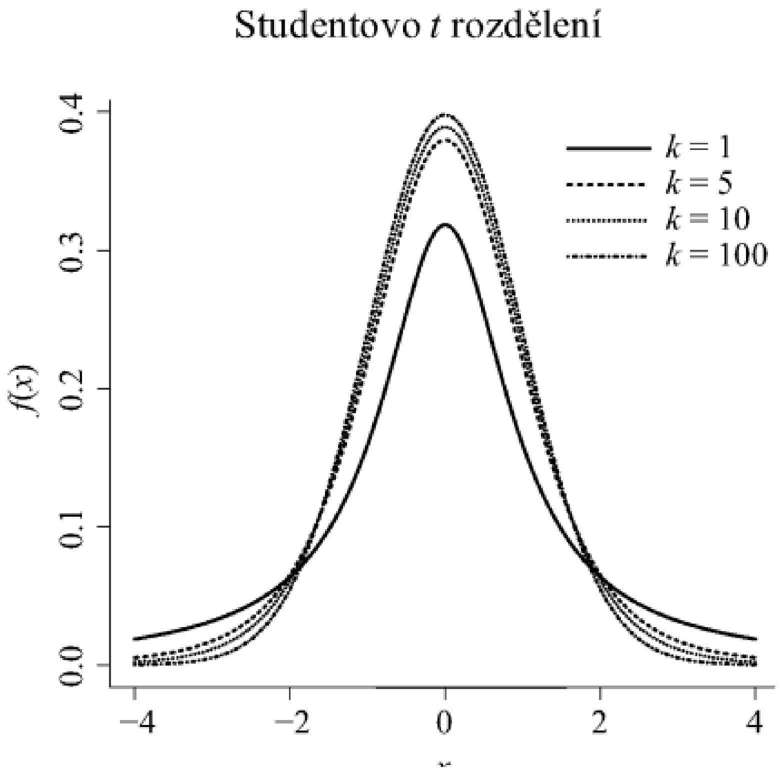
- Charakterizuje rozdělení pravděpodobnosti průměru jako odhadu střední hodnoty veličiny s normálním rozdělením v případě, že **neznáme přesnou hodnotu rozptylu** (což je v praktickém životě téměř vždy). Studentovo *t* rozdělení vzniká jako **podíl dvou nezávislych náhodnych veličin, jedné s rozdělením *N*(0,1) a druhé s rozdělením χ2(*k*)**
- Platí tedy:

$$Z \sim N(0,1), K \sim \chi^2(k) \rightarrow T = \frac{Z}{\sqrt{K/k}} \sim t(k)$$

Počet stupňů volnosti ***k***, ktery přebírá od **rozdělení chí-kvadrát**

# Ukázky hustot náhodných veličin s chí-kvadrát rozdělením a Studentovým *t* rozdělením



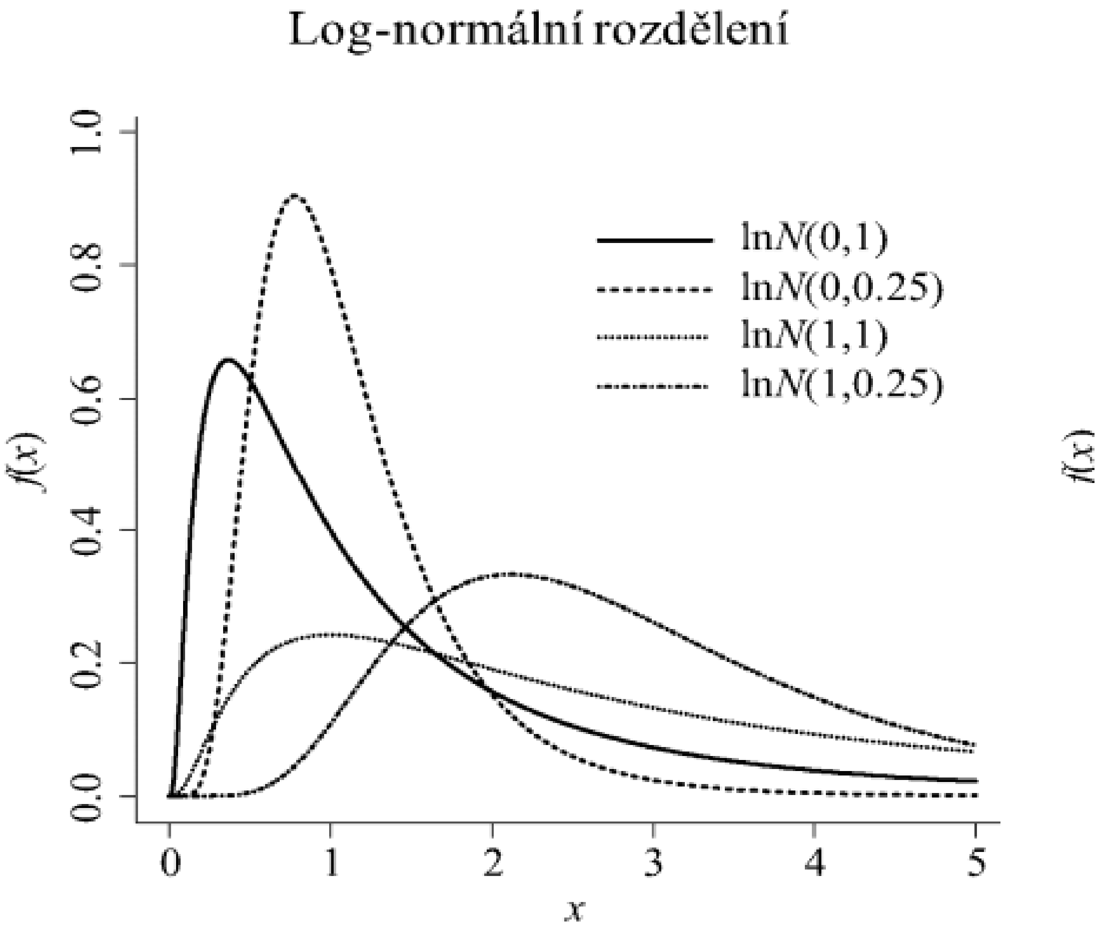



# Logaritmicko-normální rozdělení – lnN(μ,σ²)

- Náhodná veličina ***X*** má logaritmicko-normální rozdělení právě tehdy, když veličina ***Y*** = ln(X) má normální rozdělení. To samé platí i naopak, když veličina ***Y*** má normální rozdělení, pak náhodná veličina ***X*** = exp(Y) má rozdělení logaritmicko-normální.
- Hustota je dána vztahem:

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

# Ukázka hustot náhodných veličin s log-normálním rozdělením



# Diskrétní rozdělení pravděpodobnosti

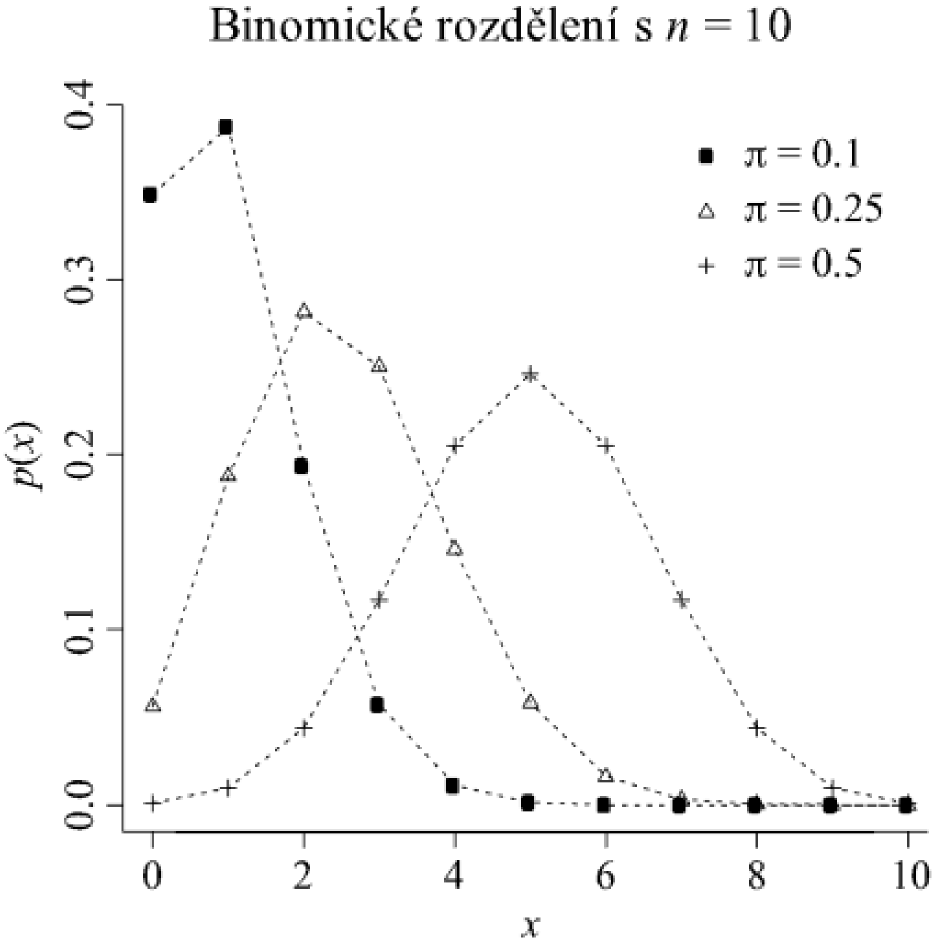
- Binomické
- Poissonovo

# <u>Binomické rozdělení – Bi(n,π)</u>

- Popisuje počet výskytů sledovaného znaku nebo události (ve formě ano/ne, nastala/nenastala) v sérii **n** nezávislých experimentů, kdy v každém experimentu **máme stejnou pravděpodobnost výskytu** daného znaku (události), označenou ***π***.
- Funkce má tvar:

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

# Pravděpodobnostní funkce náhodné veličiny s binomickým rozdělením pro n=10

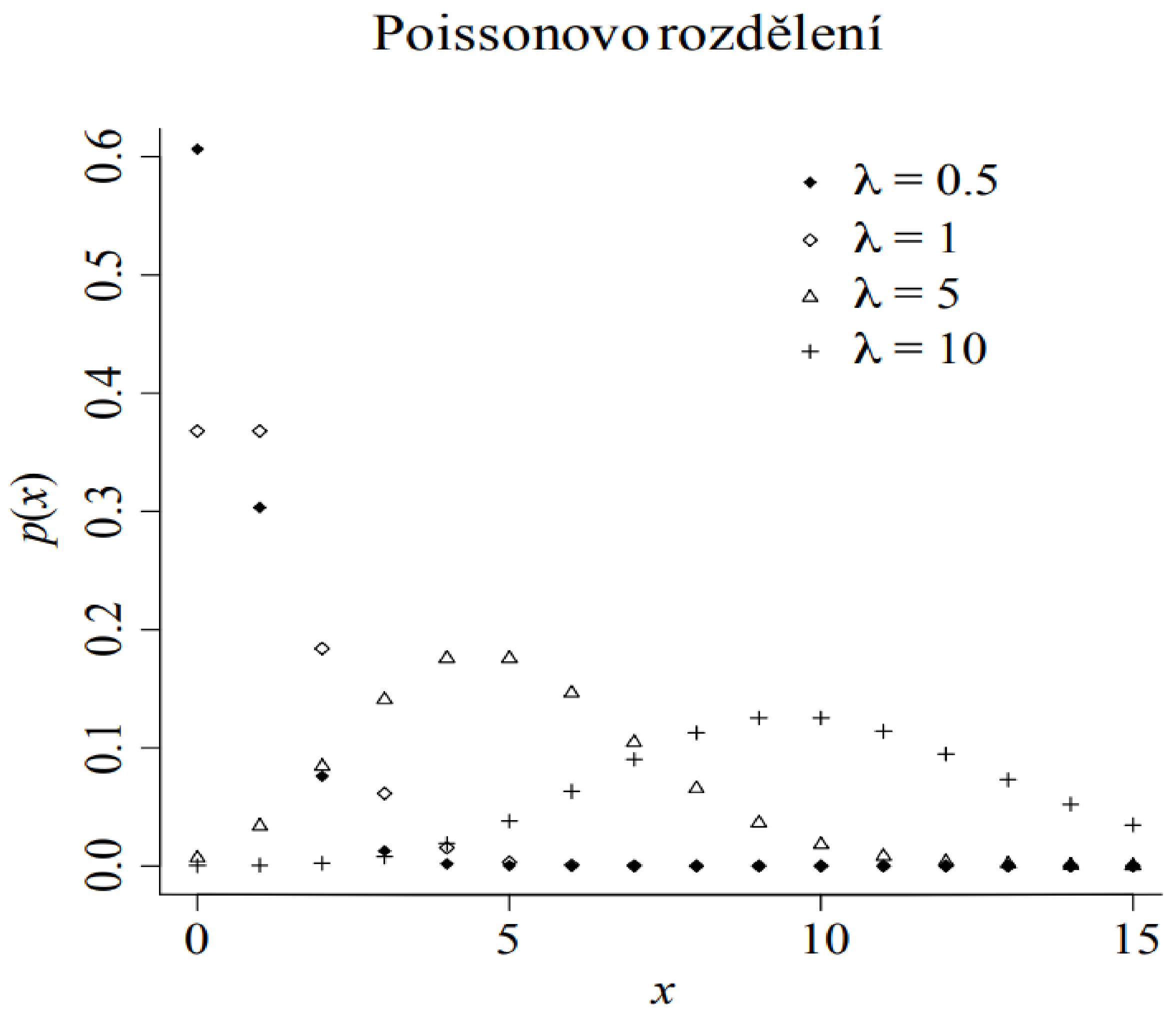


# Poissonovo rozdělení – Po(λ)

- Popisuje počet výskytů sledovaného znaku nebo události na danou jednotku času, plochy, případně objemu s tím, že se tyto události vyskytují vzájemně nezávisle a s konstantní intenzitou (tu popisuje jediný parametr tohoto rozdělení, intenzita ***λ***.
- Při počtu opakování n→∞;pravděpodobnosti výskytu jednotlivé události *π*→0
- součin *nxπ* přechází v intenzitu *λ*
- Platí tedy:

$$p(x;\lambda) = P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, x = 0, 1, 2, \ldots$$

# Ukázky pravděpodobnostní funkce náhodných veličin s Poissonovým rozdělením



# Kaplan-Meierův odhad funkce přežití

- Neparametrický odhad funkce přežití
- Aby byl subjekt v čase ***t*** bez sledované události (aby se např. pacient s nádorovým onemocněním dožil času ***t***), nesmí se u něj událost vyskytnout v žádném čase takovém, pro nějž platí, že ***t>t****. Pravděpodobnost přežití daného času můžeme vyjádřit pouze s pomocí údajů o úmrtí v daném čase. Funkci přežití pak můžeme odhadnout pomocí vztahu:

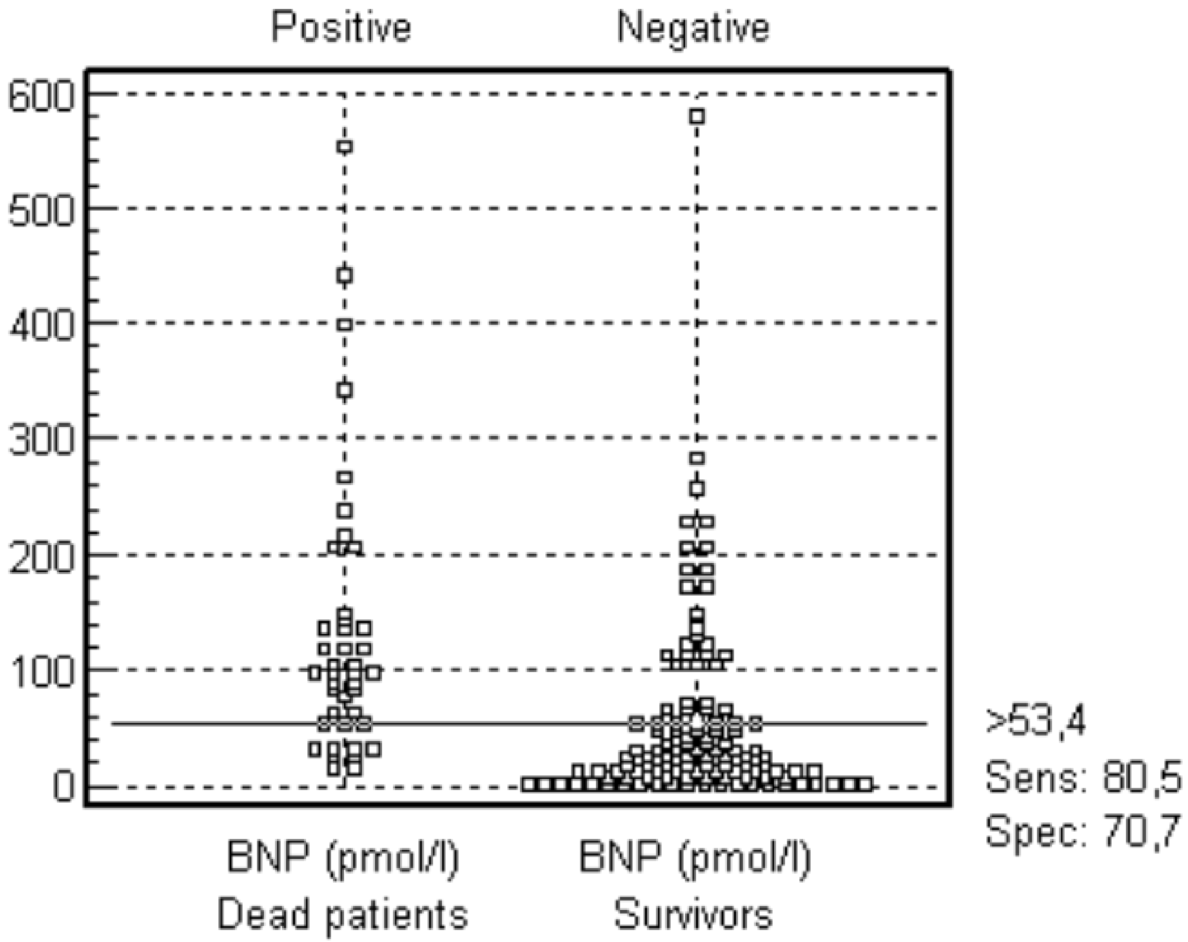
$$\hat{S}(t) = \hat{p}_1 \hat{p}_2 \dots \hat{p}_n = \frac{R_1 - d_1}{R_1} \frac{R_2 - d_2}{R_2} \dots \frac{R_{n-1} - d_{n-1}}{R_{n-1}} \frac{R_n - d_n}{R_n}$$

kde **d** je počet sledovaných událostí zaznamenaných v čase a **R** je počet subjektů v riziku výskytu sledované události v čase

# Efektivita BNP ve vztahu k celkové mortalitě

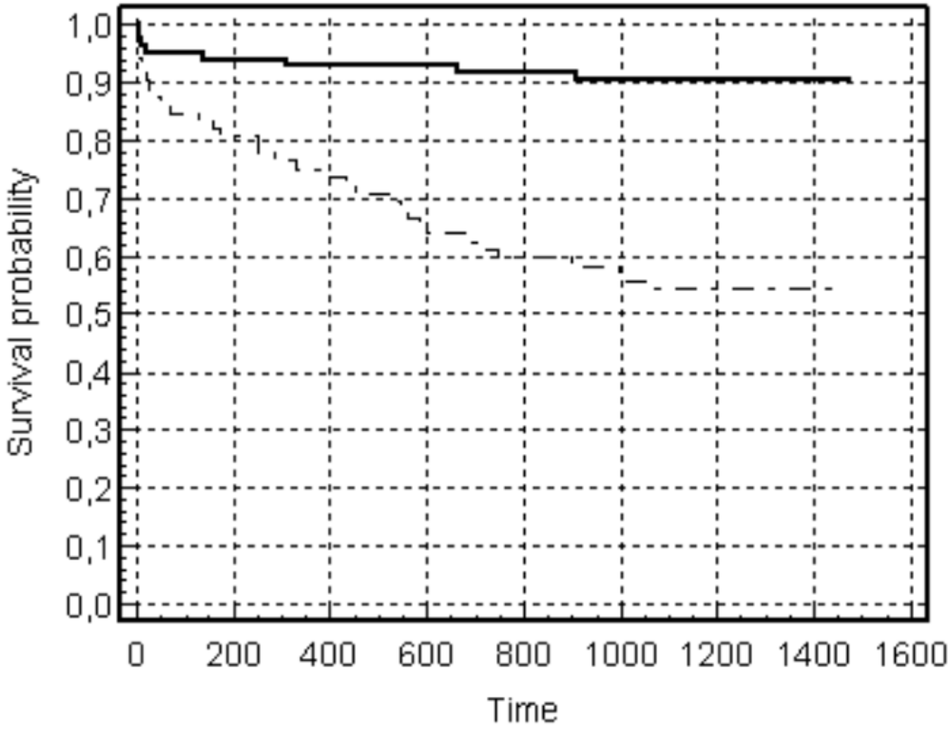
- BNP - natriuretický peptid typu B
- 157 osob
- 4 roky
- 116 přežilo

| Úmrtí | Doba přežití | BNP (pmol/l) | log (BNP) |
|---|---|---|---|
| 0 | 1471 | 1,3 | 0,1139433523 |
| 0 | 1471 | 8,9 | 0,9493900066 |
| 1 | 712 | 99,8 | 1,9991305413 |
| 0 | 1469 | 2,4 | 0,3802112417 |
| 0 | 1466 | 3,1 | 0,4913616938 |
| 0 | 1466 | 24,3 | 1,3856062736 |
| 1 | 896 | 266,3 | 2,4253711664 |
| 0 | 1464 | 29,7 | 1,4727564493 |
| 0 | 1464 | 5,7 | 0,7558748557 |
| 0 | 1464 | 55,8 | 1,7466341989 |
| 0 | 1463 | 9,9 | 0,9956351946 |
| 1 | 450 | 217,2 | 2,3368598209 |
| 0 | 1459 | 32,4 | 1,5105450102 |


Positive
Negative
600
500
400
300
200
100
0
>53,4
Sens: 80,5
Spec: 70,7
BNP (pmol/l)
Dead patients
BNP (pmol/l)
Survivors

# Výsledky

- Mez BNP - 50 pmol/l
- Pod – 84 pacientů; 90,5 %
- Nad – 73 pacientů; 54,3 %
- p < 0,0005



# Statistika

Jan Gečnuk

# Obsah

Hypotézy

Formulace hypotézy (příklad)

Rozdělení testů para x neparametrické

F- test (model, vstup)

T- test (model, vstup)

# Hypotézy

Hypotéza = domněnka

Testování hypotéz = potvrzení, či vyvrácení domněnky

Možnosti hypotéz

Změna sledované veličiny na vlivu vnějšího zásahu

Srovnání nezávislosti dvou proměnných veličin

Srovnání odhadu střední hodnoty se známým údajem

# Formulace hypotézy

$H_0$ = nulová hypotéza, začátek

je platná, dokud nenajdeme silný důkaz, který ji vyvrátí (presumpce neviny)

$H_1$ = alternativní hypotéza, vyvrací první

hypotézy jsou pouze dvě

Pokud je jedna vyvrácena, pravděpodobně platí ta druhá

# Antibiotikum a délka hojení rány

nulová hypotéza $H_0$: $\nu_0 = \nu_1$

Alternativní hypotéza $H_1$: $\nu_0 \neq \nu_1$ (místo ≠ případně: $<$;$>$)

$\nu_0$ = standartní doba léčby

$\nu_1$ = doba léčby za použití antibiotik

Stanovení hladiny významnosti α (nejčastěji 0,05)

ovlivňuje chybu I. Druhu (přijeme H1, ale platí H0)

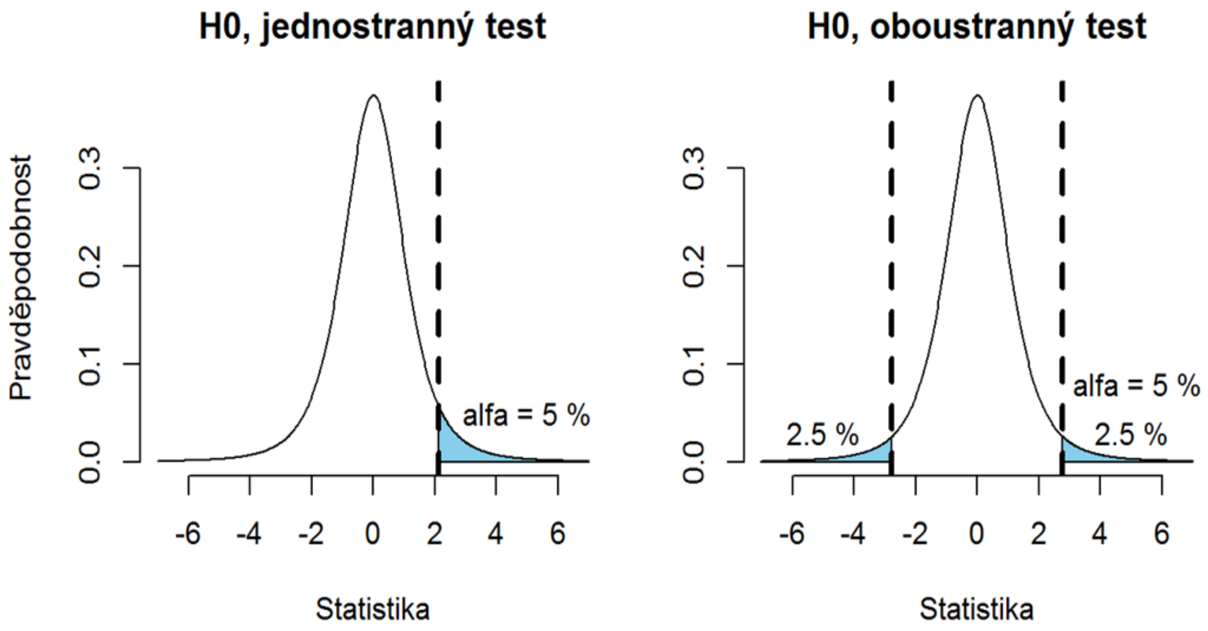
Výběr testového kritéria (vzorec T,t)

Rozdělení výběrového prostoru (oboru hodnot)

1. Obor přijetí(V): zamítnuta $H_1$

2. kritický obor(W): přijememe $H_1$, $H_0$ zamítnuta,

Obory od sebe odlišuje kritická hodnota, úsek závislý na hladině významnosti a počtu měření (tabulková hodnota)



# Antibiotkum a délka hojení rány

Máme stanoveno

Nulová hypotéza $H_0$: $\nu_0 = \nu_1$

Alternativní hypotéza $H_1$: $\nu_0 \neq \nu_1$

množství pacientu, n= 40

Hladina významnosti $\alpha$=0.95

Průměrná doba hojení bez přídavku antibiotika $\nu_0$ je 17 dny

Průměrná doba hojení s přídavkem antibiotika $\nu_1$ je 16 dny

Známe i rozptyl $\sigma^2$ = 4 dny

$$= \frac{\nu_0 - \nu_1}{\sigma} \times \sqrt{n}$$

$$= \frac{17 - 16}{2} \times \sqrt{40}$$

# Antibiotikum a délka hojení rány

t-test

kritická hodnota $T_{krit}$ = 1,684

tabulka: Kvantily $t_{1-\alpha/2}$ (ν) Studentova t rozdělení

ν = n-1 (počet stupňů volnosti = n-1)

Vypočtená T = 3,16

je vyšší než kritická hodnota

alternativní hypotézu zamítáme

rozdíl není statisticky významný α=0.95

Tabulky

**Tab. č. 3 Kvantily $t_{1-\alpha/2}$ (ν) Studentova *t* rozdělení**

| St. volnosti ν | 0,80 | 0,90 | 0,95 | 0,975 | 0,9875 | 0,995 |
|---|---|---|---|---|---|---|
| 1 | 1,376 | 3,078 | 6,314 | 12,706 | 25,452 | 63,657 |
| 2 | 1,061 | 1,886 | 2,920 | 4,303 | 6,205 | 9,925 |
| 3 | 0,978 | 1,638 | 2,353 | 3,182 | 4,176 | 5,841 |
| 4 | ,941 | 1,533 | 2,132 | 2,776 | 3,495 | 4,604 |
| 5 | ,920 | 1,476 | 2,015 | 2,571 | 3,163 | 4,032 |
| 6 | ,906 | 1,440 | 1,943 | 2,447 | 2,969 | 3,707 |
| 7 | ,896 | 1,415 | 1,895 | 2,365 | 2,841 | 3,499 |
| 8 | ,889 | 1,397 | 1,860 | 2,306 | 2,752 | 3,355 |
| 9 | ,883 | 1,383 | 1,833 | 2,262 | 2,685 | 3,250 |
| 10 | ,879 | 1,372 | 1,812 | 2,228 | 2,634 | 3,169 |
| 11 | ,876 | 1,363 | 1,796 | 2,201 | 2,593 | 3,106 |
| 12 | ,873 | 1,356 | 1,782 | 2,179 | 2,560 | 3,055 |
| 13 | ,870 | 1,350 | 1,771 | 2,160 | 2,533 | 3,012 |
| 14 | ,868 | 1,345 | 1,761 | 2,145 | 2,510 | 2,977 |
| 40 | ,851 | 1,303 | 1,684 | 2,021 | 2,329 | 2,704 |

# Rozdělení testů

- Parametrické testy
- Testují parametr ze základního souboru (střední hodnota, rozptyl)
- mají vyšší sílu, (schopnost správně zamítnout ve skutečnosti neplatnou nulovou hypotézu) než testy neparametrické
- „přísnější na data“
- Vyžadují podmínky - nepřítmnost extérmů, normální rozdělení, rozsah výběru
- Používají se prioritně

# Neparametrické testy

- Testuje hypotézu o základním souboru, jinou, než je parametr ($\mu$, $\sigma$)
- nižší síla (ty data v $H_0$, které by parametrický test zamítl, neparametrický nezamítne)
- Nezávsilý na extrémech a rozdělení
- Širší použití

# Rozdělení testů - výběry

Testy pro jeden výběr – porovnáváme parametr odhadnutý z měřených dat se zadanou konstantou (energetický příjem oproti známému průměru)

Testy pro dva výběry – porovnávání parametrů ze dvou výběrů, testování shody, nezávislosti

Podmnožiny

Dvouvýběrový test pro nezávislé výběry (výběry jsou navzájem nezávislé)

Dvouvýběrový test pro závislé výběry (párový test – 1 jedinec, dva testované přístroje)

# Rozdělení testů

Testy pro více výběrů

metodika – analýza rozptylů ANOVA

# Fisherův, F-test

sledování rozdílu dvou rozptylů

Používáme k testování hypotéz týkající se σ, μ

U populací, které odpovídají Gaussovu norm. rozdělení

U populací se využívají výběrové rozptyly k ověření nulové hypotézy

Nulová hypotéza a následující alternativní hypotézy zapíšeme

$H_0\colon \sigma_1^2 = \sigma_2^2$ $\quad$ $H_0\colon \sigma_1^2 \neq \sigma_2^2$ $\quad$ $H_0\colon \sigma_1^2 > \sigma_2^2$ $\quad$ $H_0\colon \sigma_1^2 < \sigma_2^2$

# Fisherovo rozdělení

**Tab. č. 7 Kvantily $F_{0,975}$ ($\nu_V$, $\nu_M$) Fisher-Snedecorova rozdělení ($\alpha$ = 0,05)**

| $\nu_M$ | Počet stupňů volnosti čitatele ($\nu_V$) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | 647,79 | 799,50 | 864,16 | 899,58 | 921,85 | 937,11 | 948,22 | 956,66 | 963,28 |
| 2 | 38,506 | 39,000 | 39,165 | 39,248 | 39,298 | 39,331 | 39,355 | 39,373 | 39,387 |
| 3 | 17,443 | 16,044 | 15,439 | 15,101 | 14,885 | 14,735 | 14,624 | 14,540 | 14,473 |
| 4 | 12,218 | 10,649 | 9,979 | 9,605 | 9,365 | 9,197 | 9,074 | 8,980 | 8,905 |
| 5 | 10,007 | 8,434 | 7,764 | 7,388 | 7,146 | 6,978 | 6,853 | 6,757 | 6,681 |
| 6 | 8,813 | 7,260 | 6,599 | 6,227 | 5,988 | 5,820 | 5,696 | 5,600 | 5,523 |
| 7 | 8,073 | 6,542 | 5,890 | 5,523 | 5,285 | 5,119 | 4,995 | 4,899 | 4,823 |
| 8 | 7,571 | 6,060 | 5,416 | 5,053 | 4,817 | 4,652 | 4,529 | 4,433 | 4,357 |
| 9 | 7,209 | 5,715 | 5,078 | 4,718 | 4,484 | 4,320 | 4,197 | 4,102 | 4,026 |

# Příklad F-testu

Dva soubory, pokusný ($\mu_1$ a $\sigma_1^2$) kontrolní ($\mu_2$ a $\sigma_2^2$)

Určíme výběrový rozptyl

Testovací statistika:

Pokud platí $H_0$, pak má Fisherovo rozdělení s parametry ($n_1$-1) a ($n_2$-2),

Možnosti zamítnutí $H_0$ hypotézy

$$' = \frac{(větší\ z\ rozptylů\ \ s_1^2, s_2^2)}{(menší\ z\ rozptylů\ \ s_1^2, s_2^2)}$$

$$F = \frac{(s_1^2)}{(s_2^2)}$$

$$2 = \frac{\sum_N^{i-1}(x_i - x)^2}{n-1}$$

*Tabulka 7.6 Pravidla pro zamítnutí $H_0$ pro F-test dle zvolené alternativy.*

| | | | |
|---|---|---|---|
| Alternativa | $H_1 : \sigma_1^2 \neq \sigma_2^2$ | □ Zamítáme $H_0$, když | $F < F_{\alpha/2}^{(n_1-1,n_2-1)}$ nebo $F > F_{1-\alpha/2}^{(n_1-1,n_2-1)}$ |
| Alternativa | $H_1 : \sigma_1^2 > \sigma_2^2$ | □ Zamítáme $H_0$, když | $F > F_{1-\alpha}^{(n_1-1,n_2-1)}$ |
| Alternativa | $H_1 : \sigma_1^2 < \sigma_2^2$ | □ Zamítáme $H_0$, když | $F < F_{\alpha}^{(n_1-1,n_2-1)}$ |

# Příklad F-test

Děti s hyperthyreózou

Výpočet testové statistiky

$$F = \frac{(s_1^2)}{(s_2^2)} = \frac{(14{,}22_1^2)}{(37{,}48_2^2)} = 0{,}144$$

V tabulkách najdeme kvantil, pro stupně volnosti

$$= 0{,}144 < 0{,}279 = F_{0{,}05}^{8,6} = F_{\alpha}^{(n_1-1, n_2-1}$$

Závěr – zamítáme nulovou hypotézu

Obě skupiny se statisticky významně

liší ve variabilitě tyroxinu v séru

| Hladina tyroxinu v séru (nmol/l) | Mírné symptomy ($n_1$ = 9) | Výrazné symptomy ($n_2$ = 7) |
|---|---|---|
| Průměr | 56,4 | 42,1 |
| Směrodatná odchylka | 14,22 | 37,48 |

# T-test

Nejčastěji používaný parametrický test

testování rozdílu dvou středních hodnot $\mu$

Výpočet vychází z odhadu parametrů $\mu$ a $s^2$ u výběrových souborů $x$, $\sigma^2$

Testovací kritérium, porovnáme s tabulkovou kritickou hodnotou ($1-\alpha/2$ kvantil Studentova t-rozdělení pro dané $\upsilon$ a zvolené $\alpha$)

Podmínka: stejný rozptyl obou měření

# T-test - jednovýběrový

Jednovýběrový – známe střední hodnotu základního souboru, pak nulová hypotéza je $H_o$: $\mu$= konst.

Stanovíme aritmetický průměr, rozptyl výběrového souboru

$$t = \frac{|\bar{x} - \mu|}{s \times \sqrt{n}}$$

Testovací kritérium (statistika) = t

Stanovení $\alpha$, $\nu$ $\quad \nu = n - 1$

Kritická hodnota: $t_{1-\alpha}/2(\nu)$

Tabulky: Kvantily $t_{1-\alpha}/2(\nu)$ Studentova t rozdělení

t=testovací kritérium
x = průměr výběrového souboru
μ = střední hodnota základního souboru
s = rozptyl výběrového souboru
n = počet členů VS

# T-test závěry

Výsledek: <u>$t \le t_{1-\alpha}/2(\nu)$, statisticky nevýznamný rozdíl</u> mezi střední hodnotou μ a známou konstantou při zvolené α (nezamítáme $H_0$, tzn. výběrový soubor pochází z populace se známou střední hodnotou m = konst.).

Závěr: <u>pokusný zásah byl neúčinný</u>, protože nebyla ovlivněna střední hodnota souboru při aplikaci zásahu ($p > 0{,}05$).

<u>$t > t_{1-\alpha}/2(\nu)$, statisticky významný rozdíl</u> mezi střední hodnotou μ a známou konstantou ($\alpha = 0{,}05$) nebo statisticky vysoce významný rozdíl (při $\alpha = 0{,}01$) (zamítáme $H_0$, tzn. výběrový soubor nepochází σ populace se známou střední hodnotou a pochází z jiné populace, kde $\mu \neq$ konst.).

Závěr: <u>pokusný zásah byl účinný</u>, protože způsobil změnu střední hodnoty u pokusného souboru ve srovnání se známou konstantní střední hodnotou ($p < 0{,}05$ resp. $p < 0{,}01$).

# T-test dvouvýběrový párový

Neznáme střední hodnotu základního souboru

porovnáváme 2 soubory dat

1 soubor – měření před zásahem

2 soubor – měření po zásahu

sledujeme rozdíly v „párech“ v řadách hodnot

testujeme hypotézu, že střední hodnota před a po pokusu se rovná

Př. slepice ve velkochovu – hmotnost vajec, upravení světelného režimu, opětovné měření, zjišťování významnosti

# T-test dvouvýběrový párový

1. vypočítáme rozdíly v párových hodnotách

2. vezmeme rozdíly před a po zásahu, z těchto stanovíme arimtetický průměr, x ze zjištěných rozdílů a $s^2$ (nebo s = směrodatnou odchylku)

3. vypočítáme testovácí kritérium

$$\nu = n - 1$$

4. stanovení α, ν

$$t = \frac{|\overline{x_1} \times \overline{x_2}|}{s \times \sqrt{n}}$$

6. Kvantily t1-α/2(ν) Studentova t rozdělení

7. t porovnáváme s $t_{krit}$

# T-test dvouvýběrový párový

Výsledek : $t \leq t_{1-\alpha}/2(\nu)$, statisticky nevýznamný rozdíl mezi $\mu_1$ a $\mu_2$ na dané hladině výynamnosti

Závěr: pokusný zásah byl neúčinný, protože nebyla ovlivněna střední hodnota měření provedeného po aplikaci zásahu ($p > 0{,}05$).

$t > t_{1-\alpha}/2(\nu)$, statisticky významný rozdíl $\mu_1$ a $\mu_2$ ($\alpha = 0{,}05$) nebo statisticky vysoce významný rozdíl (při $\alpha = 0{,}01$)

(zamítám H0, tzn. střední hodnota měření před pokusem se liší od střední hodnoty měření po pokusu).

Závěr: pokusný zásah byl účinný, protože způsobil změnu střední hodnoty u měření provedeného po aplikaci pokusného zásahu ve srovnání se střední hodnotou zjištěnou před aplikací zásahu ($p < 0{,}05$ resp. $p < 0{,}01$).

# T-test dvou výběrový nepárový

Porovnávání dat ze dvou nezávislých výběrů – typicky testovaná a kontrolní skupina

Typická $H_0 : \mu_1 = \mu_2$

Vychází z odhadu $\mu$ a $\sigma^2$ kontrolního a pokusného souboru

neznáme rozptyl hodnot sledované veličiny

nutno provést F-test $H_0 : \sigma^2{}_1 = \sigma^2{}_2$

$$F = \frac{(\text{větší z rozptylů } s_1^2, s_2^2)}{(\text{menší z rozptylů } s_1^2, s_2^2)}$$

(Příklad – podání léčiva testované skupině pacientů, kontrolní skupině placebo)

# T-test dvou výběrový nepárový

Stanovení stupňů volnosti čitatele i jmenovatele

$$F = \frac{(\textit{větší z rozptylů } s_1^2, s_2^2)}{(\textit{menší z rozptylů } s_1^2, s_2^2)}$$

ν (větší) = $n_{(1,2)}$-1

ν (menší) = $n_{(1,2)}$-1

Pro stanovení kritických hodnot se používají tabulky: Fisher-Snedecorova rozdělení

vyhledáme kritickou hodnotu $F_{krit}$ = 1 - α/2 kvantil F-rozdělení o (νV,νM) stupních volnosti pro zvolenou hladinu významnosti

# T-test dvou výběrový nepárový

F- test:

potvrdil, že rozptyly jsou shodné, poté

$$t = \frac{/\,\bar{x}_1 - \bar{x}_2\,/}{\sqrt{(n_1 - 1).s_1^2 + (n_2 - 1).s_2^2}} . \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}$$

pokud se $n_1 = n_2$ pak

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2 + s_2^2}{n}}} \qquad \upsilon = 2n - 2$$

- F-test nevyšel pozitivně, pak

$$t = \frac{/\,\bar{x}_1 - \bar{x}_2\,/}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \qquad \upsilon \approx \frac{n_1 + n_2 - 2}{2}$$

Pokud $n_1 = n_2 \Rightarrow \upsilon = n - 1$

$$t^*_{krit.} = \frac{t_1 . \frac{s_1^2}{n_1} + t_2 . \frac{s_2^2}{n_2}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

# Závěr

Nulová hypotéza = presumpce nevinny

F-test = srovnání dvou rozptylů (Fisher Snedecorovo rozdělení)

$$F = \frac{(s_1^2)}{(s_2^2)}$$

T-test srovnání: (Studentovo, T-rozdělení)

$$t = \frac{|\bar{x} - \mu|}{s \times \sqrt{n}}$$

1. střední hodnota s danou hodnotou

$$t = \frac{|\overline{x_1} \times \overline{x_2}|}{s \times \sqrt{n}}$$

2. dvě střední hodnoty ze stejného souboru dat (párový)

3. dvě střední hodnoty z různých souborů (nepárový)

$$\nu = 2n - 2$$

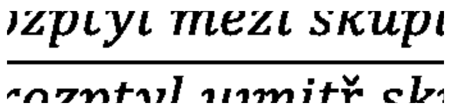
$$= \left( \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2 + s_1^2}{n}}} \right)$$

# ANOVA (analysis of variance; analýza rozptylu)

- Testování hypotéz o středních hodnotách <u>více než dvou výběrů</u>
- Využívá srovnání <u>variability mezi výběry a variability uvnitř výběrů (hodnotí rozptyly)</u>
- Předpoklady: nezávislost měření, normalita dat (Q-Q diagram, Shapirův-Wilkův test, Kolmogorovův-Smirnovův test), alespoň přibližná shoda rozptylů ve skupině
- Př.: Sledování pacienta v několika stádiích onemocnění (II, III, IV)

# ANOVA – jednofaktorová (one-way)

- $H_0 : \mu_k = \mu_1 = \mu_2 = \mu_3 = \mu_4$ $H_1$: nejméně jedno $\mu_i$ je odlišné od ostatních
- Sledujeme účinek jednoho faktoru na zkoumanou proměnnou
- Př. Sledujeme změny váhy u zvířat s různými druhy krmiva (krmivo A, krmivo B a kontrola)
- Testovací kritérium F - zda se průměry ve skupinách odlišují více než na základě přirozeného kolísání
- $F > F_{krit}$ –> zamítám $H_0$ -> testy pro mnohonásobné testování (Tukey-test, Sheffe-test, Student-Neuman-Keuls-test (SNK test), Dunnett-test)



- Postup:

1. Průměr v rámci výběru + celkový průměr všech výběrů
2. Celkový součet čtverců (= celková variabilita) $S_T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{y}_{..})^2$
3. Skupinový součet čtverců (= variabilita mezi skupinovými průměry; variabilita příslušná vlivu sledované proměnné) $S_A = \sum_{i=1}^{k} n_i (\bar{y}_{i\cdot} - \bar{y}_{..})^2$
4. Reziduální součet čtverců (= variabilita v rámci skupin) $S_e = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{y}_{i\cdot})^2$
5. Výběrové odhady $\frac{S_e}{df_e} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{y}_{i\cdot})^2}{n-k}$ a $\frac{S_A}{df_A} = \frac{\sum_{i=1}^{k} n_i (\bar{y}_{i\cdot} - \bar{y}_{..})^2}{k-1}$
6. Testová statistika F:

$$F = \frac{\frac{\sum_{i=1}^{k} n_i (\bar{y}_{i\cdot} - \bar{y}_{..})^2}{k-1}}{\frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{y}_{i\cdot})^2}{n-k}} = \frac{S_A/df_A}{S_e/df_e}$$

*Tab. 8.2: Fiktivní datový soubor se třemi srovnávanými skupinami.*

| Léčba | Pozorovaná hodnota | Skupinový průměr | Skupinový průměr *mínus* celkový průměr | Pozorovaná hodnota *mínus* skupinový průměr | Pozorovaná hodnota *mínus* celkový průměr |
|---|---|---|---|---|---|
| A | 10 | 12 | -4 | -2 | -6 |
| A | 12 | 12 | -4 | 0 | -4 |
| A | 14 | 12 | -4 | 2 | -2 |
| B | 19 | 20 | 4 | -1 | 3 |
| B | 20 | 20 | 4 | 0 | 4 |
| B | 21 | 20 | 4 | 1 | 5 |
| C | 14 | 16 | 0 | -2 | -2 |
| C | 16 | 16 | 0 | 0 | 0 |
| C | 18 | 16 | 0 | 2 | 2 |
| | Celkový průměr = 16 | | Součet čtverců = 96 | Součet čtverců = 18 | Součet čtverců = 114 |

# ANOVA - vícefaktorová

- Sledování více faktorů: vliv krmení a plemene, vliv léku v různých stádiích onemocnění, apod.
- Rozlišujeme: „hlavní efekt“ – přímý efekt faktoru na závisle proměnnou<br>„interakční efekt“ – spojení dvou nebo více faktorů na závisle proměnnou

Nejjednodušší: **ANOVA dvojného třídění (two-way)**

- Sledujeme vliv dvou faktorů na závisle proměnnou
- Faktory: A – plánovitě měníme<br>B – „rušivý vliv“; snažíme se ho oddělit od A

-> pacienty rozdělíme např. podle úrovně B (dávka antibiotika)
-> pacientům náhodně přiřadit úroveň A (druh antibiotika)
--> sledujeme rozdíly účinku jednotlivých druhů antibiotik podávaných v různých dávkách

# Zdroje

https://www.iba.muni.cz/res/file/ucebnice/pavlik-biostatistika-v2.pdf

http://user.mendelu.cz/drapela/Statisticke_metody/Prezentace/zakladni/testy.pdf

https://cit.vfu.cz/statpotr/POTR/Teorie/tabulky.htm#Ftest

http://www.fsps.muni.cz/~novotny/Statistika.htm

https://cit.vfu.cz/statpotr/POTR/prednasky.htm

http://user.mendelu.cz/drapela/Statisticke_metody/teorie%20text%20I.pdf

https://k101.unob.cz/~neubauer/pdf/testy_hypotez1.pdf

## Tabulky

# Neparametrické testy pro spojitou náhodnou veličinu

- Spojitá náhodná veličina (jakékoli hodnoty v určitém rozmezí)
- Nemusíme znát typ rozdělení
- Často založené na pořadí hodnot (není potřeba mít přesné hodnoty)
- Obecnější ale při srovnání s  parametrickými testy mají menší sílu

# PŘEHLED TESTŮ

| rozdělení | normální | spojité | alternativní / diskrétní |
|---|---|---|---|
| **populační parametr (o čem je hypotéza)** | **populační průměr** | **populační medián (distribuční funkce)** | **pravděpodobnost jevu / (ne)závislost / poměr šancí** |
| jeden výběr | jednovýběrový t-test | jednovýběrový Wilcoxonův test<br>znaménkový test | test proporcí |
| výběr dvojic | párový t-test | párový Wilcoxonův test<br>znaménkový test | McNemarův test |
| dva nezávislé výběry (třídění na dvě kategorie / závislost na binární proměnné) | dvouvýběrový t-test | Mann-Whitney<br>(dvouvýběrový Wilcoxon)<br>Kolmogorov-Smirnov | Fisherův exaktní test<br>$\chi^2$-test |
| *k* nezávislých výběrů (závislost na kategoriální proměnné) | analýza rozptylu (jednoduché třídění, F-test) | Kruskal-Wallis | Fisherův exaktní test<br>$\chi^2$-test |
| závislost na spojité proměnné | lineární regrese | zobecněná lineární regrese<br>(speciální případy)<br>jádrová regrese (kernel smoothing)<br>... | logistická regrese<br>multinomiální logistická regrese |
| opakovaná měření | smíšená lineární regrese (mixed models) | smíšená zobecněná lineární regrese | |

# Wilcoxonův test pro jeden výběr

- Neparametrická alternativa t-testu
- Symetrie rozdělení kolem mediánu
- Získané hodnoty jednoho výběru srovnávány se správnou hodnotou
- $n \geq 6$

- Př. Denní energetický příjem srovnáván s doporučenou hodnotou 7725 kJ.

  $H_0$: $x = 7725$ $H_1$: $x \neq 7725$ (x … medián – chybí vlnka)

***Tabulka 7.3** Denní energetický příjem skupiny 11 žen ve věku 22 – 30 let.*

| Žena | Denní energetický příjem v kJ | Diference od hodnoty 7725 kJ | Pořadí absolutní hodnoty diference |
|---|---|---|---|
| 1 | 5260 | -2465 | 11 |
| 2 | 5470 | -2255 | 10 |
| 3 | 5640 | -2085 | 9 |
| 4 | 6180 | -1545 | 8 |
| 5 | 6390 | -1335 | 7 |
| 6 | 6515 | -1210 | 6 |
| 7 | 6805 | -920 | 4 |
| 8 | 7515 | -210 | 1,5 |
| 9 | 7515 | -210 | 1,5 |
| 10 | 8230 | 505 | 3 |
| 11 | 8770 | 1045 | 5 |

1. Diference od hodnoty, se kterou srovnáváme ($x_1$-$x_0$)
2. Určení pořadí absolutních hodnot (stejným hodnotám přiřazujeme průměrnou hodnotu)
3. Součet kladných pořadí, součet záporných pořadí

$S^+ = 8$; $S^- = 58$ -> min ($S^+$ , $S^-$ ) = **8** $w_{11}$ (0,05) = **10** -> 8 < 10 -> **<u>zamítám $H_0$</u>**

# Párový Wilcoxonův test

- Podobný jako W. test pro jeden výběr
- Párová data jsou na sobě závislá (2 měření na jednom subjektu)
- Např. stav pacienta před léčbou a po léčbě, míra stisku pravé a levé ruky

1. Dva výběry -> diference mezi párovými hodnotami -> seřadit vzestupně -> pořadí (nulové hodnoty vyřadíme)
2. Postup jako W. test pro jeden výběr

# Wilcoxonův test pro dva výběry (Mann-Whitney test)

- Obdoba testu shodnosti středních hodnot dvou výběrů ($X_1$, $X_2$)
- Ne střední hodnoty ale rozložení funkce -> $H_0$: $F(x) = F(y)$ $H_1$: $F(x) \neq F(y)$

1. Prvky z obou souborů seřadit do neklesající posloupnosti -> součet pořadí $X_1$ a $X_2$ (označit $T_1$, $T_2$)
2. Výpočet statistiky: $$U_1 = T_1 - \frac{N_1(N_1+1)}{2} \qquad U_2 = T_2 - \frac{N_2(N_2+1)}{2}$$
3. Menší číslo porovnáváme s kritickou hodnotou ($U_{min} \leq w(N_1, N_2; 0{,}05)$ -> zamítám $H_0$)

- Př. Pole hnojené 2 různými způsoby (výnos v tunách na hektar)

$X_1$: 5.7, 5.5, 4.3, 5.9, 5.2, 5.6, 5.8, 5.1

$X_2$: 5.0, 4.5, 4.2, 5.4, 4.4

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4.20 | 4.30 | 4.40 | 4.50 | 5.00 | 5.10 | 5.20 | 5.40 | 5.50 | 5.60 | 5.70 | 5.80 | 5.90 |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |

$T_1 = 70$ $U_1 = 34$

$T_2 = 21$ $U_2 = 6$ -> $w(5, 8; 0{,}05) = 6$ -> $6 = 6$ -> zamítám $H_0$

# Kruskallův-Wallisův test (neparametrická alternativa ANOVA)

- Zobecnění M.-W. testu pro více než 2 srovnávané skupiny

  $H_0$: $F_1(x) = F_2(x) = \ldots = F_k(x)$ $H_1$: nejméně jedna $F_i$ je odlišná od ostatních

- Prvky z všech souborů seřadit do neklesající posloupnosti

  -> součet pořadí $X_1, X_2, .. X_k$ (označit $T_1, T_2, .. T_k$)

- Výpočet statistiky: $$KW = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3 \cdot (n+1)$$

- $KW > \chi^2_{\alpha\,(k-1)}$ -> zamítám $H_0$

- Pokud jsou shodné hodnoty v pořadí -> průměrné hodnoty pořadí

  -> korekční faktor K

  -> opravný výpočet: $KW_{opr} = KW/K$

  $$K = 1 - \frac{\sum_{i=1}^{p} (t_i^3 - t_i)}{n^3 - n}$$

  (p … počet tříd se stejným pořadím, $t_i$ … počet pořadí v i-té třídě)

- Pokud zamítám $H_0$ -> neparametrické metody mnohonásobného porovnávání (Neményiho nebo Dunnova metoda)

## Příklad

V polním pokusu byly ověřovány čtyři varianty hnojení silážní kukuřice rozdílnou dávkou NPK v hnojivech, označené jako $H_1$ až $H_4$. Každá varianta byla ověřována na 8 parcelách s výnosem sklizené hmoty v tunách z parcely, uvedeným v tabulce dat. Ověřte, zda se výnosy u jednotlivých variant hnojení průkazně liší.

$H_0$: výnosy jednotlivých variant jsou shodné
$H_1$: výnosy jednotlivých variant jsou rozdílné

| Varianta | Výnos kukuřice v t z parcely číslo | | | | | | | | Součet $T_i$ |
|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | |
| $H_1$ | 1,29 | 1,19 | 1,23 | 1,33 | 1,27 | 1,29 | 1,31 | 1,20 | |
| pořadí | 23,5 | 8,5 | 15 | 28,5 | 20 | 23,5 | 27 | 11,5 | 157,5 |
| $H_2$ | 1,30 | 1,33 | 1,29 | 1,37 | 1,35 | 1,25 | 1,38 | 1,29 | |
| pořadí | 26 | 28,5 | 23,5 | 31 | 30 | 18,5 | 32 | 23,5 | 213 |
| $H_3$ | 1,20 | 1,24 | 1,25 | 1,24 | 1,20 | 1,21 | 1,28 | 1,17 | |
| pořadí | 11,5 | 16,5 | 18,5 | 16,5 | 11,5 | 14 | 21 | 7 | 116,5 |
| $H_4$ | 1,03 | 1,14 | 1,09 | 1,20 | 1,07 | 1,19 | 1,01 | 1,05 | |
| pořadí | 2 | 6 | 5 | 11,5 | 4 | 8,5 | 1 | 3 | 41 |

$$KW = \frac{12}{32 \cdot 33}\left(\frac{157{,}5^2}{8} + \frac{213^2}{8} + \frac{116{,}5^2}{8} + \frac{41^2}{8}\right) - 3 \cdot 33 = 22{,}3473$$

Vzhledem k výskytu stejných údajů (bylo použito průměrné pořadí) je vhodné opravit testové kritérium korekčním faktorem:

$$K = 1 - \frac{(2^3 - 2) + (4^3 - 4) + (2^3 - 2) + (2^3 - 2) + (4^3 - 4) + (2^3 - 2)}{32^3 - 32} = 0{,}9956$$

$$KW_{opr.} = \frac{22{,}3473}{0{,}9956} = 22{,}446 \qquad \chi^2_{0{,}05\,(3)} = 7{,}815 \qquad \chi^2_{0{,}01\,(3)} = 11{,}34$$

Opravené testové kritérium KW = 22,446 > $\chi^2$ => na hladině významnosti $\alpha = 0{,}05$ i 0,01 přijímáme alternativní hypotézu, podle které se průkazně liší hodnoty výnosu nejméně ve 2 třídách.

# Znaménkový test

- Zjednodušený Wilcoxonův test pro dva závislé výběry (párový test)
- Spojité binomické rozdělení
- $n \geq 20$
- Často pro orientační hodnocení (sledovanou veličinu nedokážeme přesně změřit)
- Př. Účinek jedné terapie nad druhou, výskyt proměnné, pravolevost
- 3 případy:

  A > B …… **"+"** (*terapie A je účinnější než B, A se vyskytlo a B ne*)

  A < B …… **"–"** (*terapie A je méně účinná než B, B se vyskytlo a A ne*)

  A = B …… vyřadíme (*terapie A je stejně účinná jako B, současný výskyt či nepřítomnost A i B*)

- Postup:

  Vyhodnotit, co je + a - -> sečíst + a – (= $m_+$, $m_-$) -> nižší hodnota porovnána s kritickou hodnotou-> $m \leq m_{krit}(n; 0{,}05)$ -> zamítám $H_0$

- Př. Moč 15 pacientů – Furantoin x Penicilin – zkoumáme počet bakterií

1. Rozdíly: 13x – méně bakterií u F než u P -> F < P

   1x – více bakterií u F než u P -> F > P

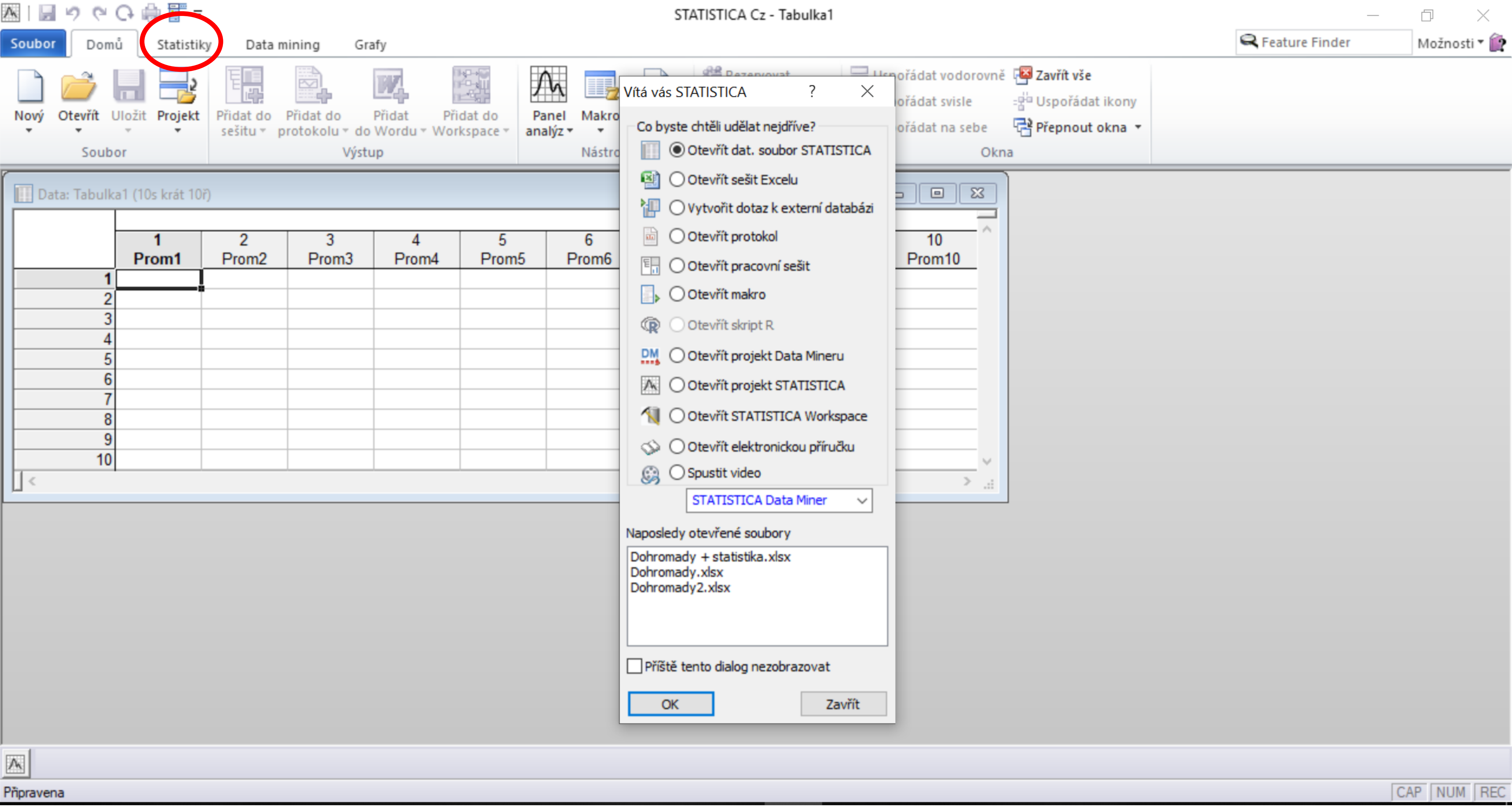
   1x – nelze rozhodnout -> F = P (vylučujeme)

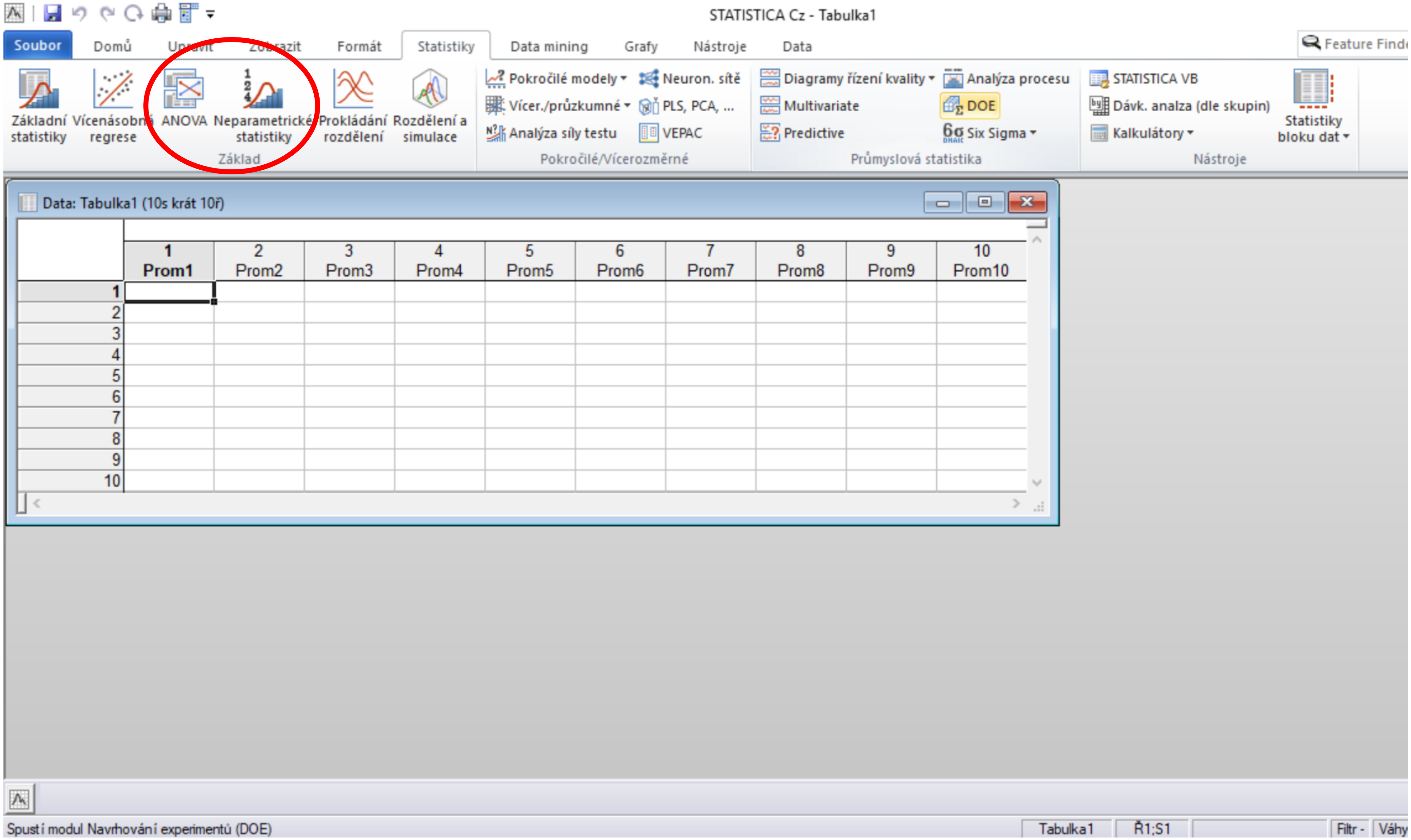
2. Součet + a -: $m_- = 13$ $m_+ = 1$ -> $n = 14$

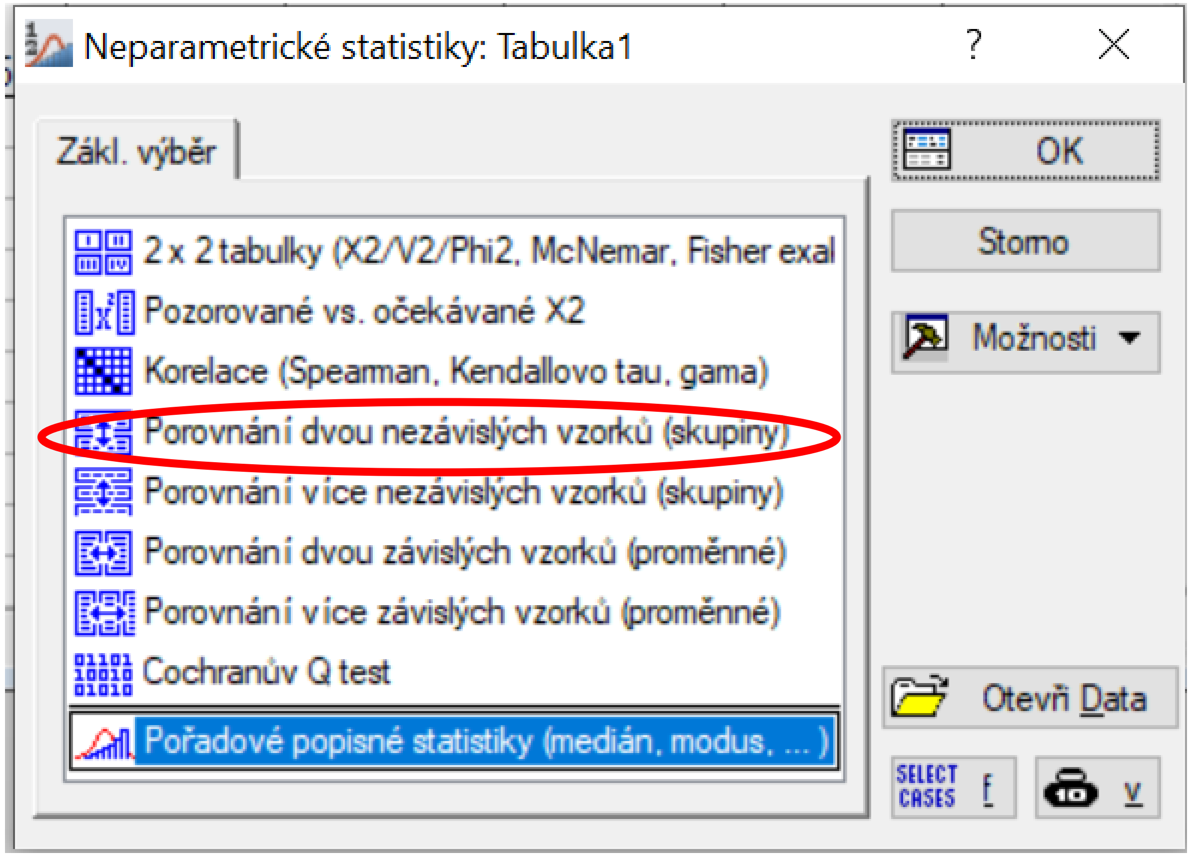
3. Testovací kritérium: $m = \min(13, 1) = 1$

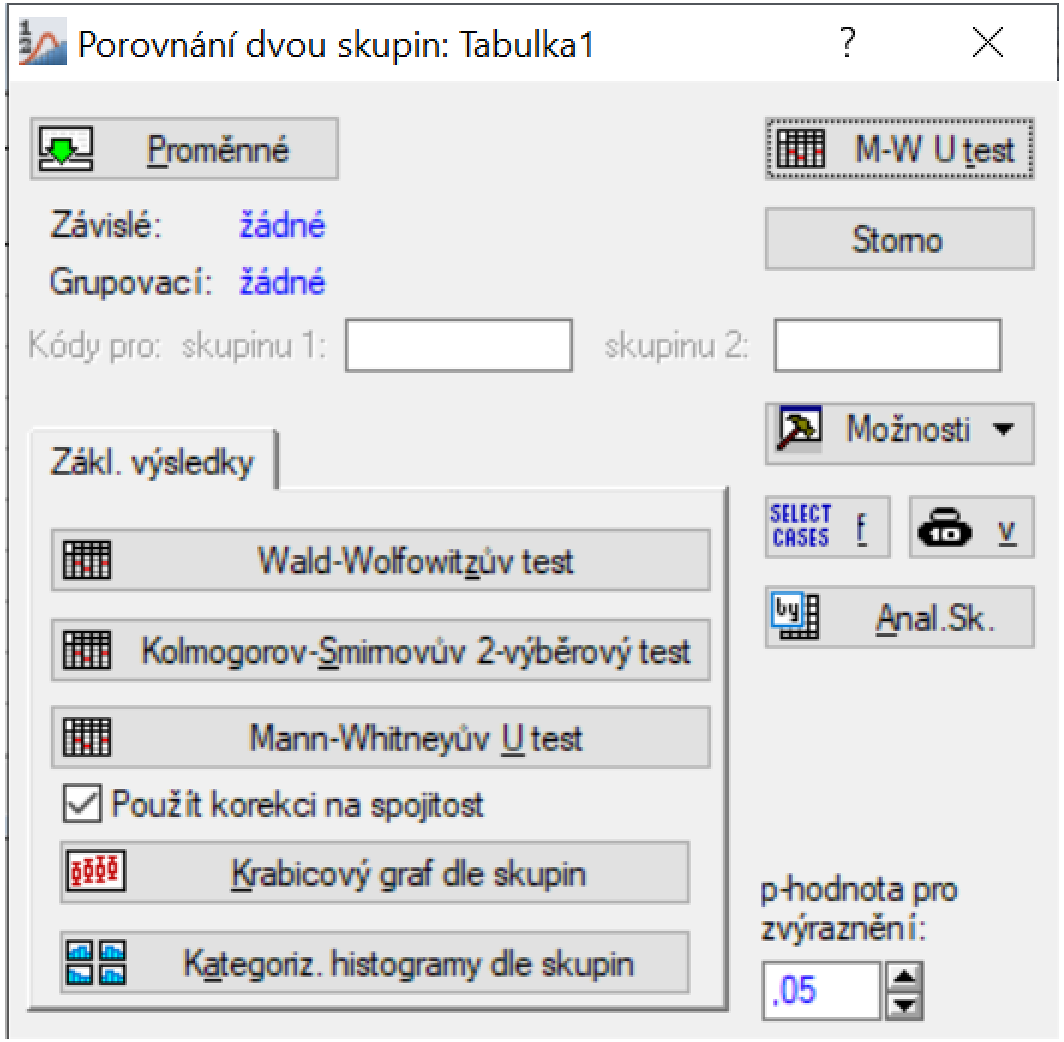
4. $m_{krit}(14; 0{,}05) = 2$ -> $m < m_{krit}$ -> zamítám $H_0$

5. Prokázán statisticky významný rozdíl ($p < 0{,}05$) v účinnosti Furantoinu a Penicilinu na růst bakterií ve vzorcích moči pacientů.


STATISTICA Cz - Tabulka1
Soubor
Domů
Statistiky
Data mining
Grafy
Feature Finder
Možnosti
Nový
Otevřít
Uložit
Projekt
Soubor
Přidat do sešitu
Přidat do protokolu
Přidat do Wordu
Přidat do Workspace
Výstup
Panel analýz
Makro
Zavřít vše
Uspořádat ikony
Přepnout okna
Okna
Data: Tabulka1 (10s krát 10ř)
1 Prom1
2 Prom2
3 Prom3
4 Prom4
5 Prom5
6 Prom6
10 Prom10
Vítá vás STATISTICA
Co byste chtěli udělat nejdříve?
Otevřít dat. soubor STATISTICA
Otevřít sešit Excelu
Vytvořit dotaz k externí databázi
Otevřít protokol
Otevřít pracovní sešit
Otevřít makro
Otevřít skript R
Otevřít projekt Data Mineru
Otevřít projekt STATISTICA
Otevřít STATISTICA Workspace
Otevřít elektronickou příručku
Spustit video
STATISTICA Data Miner
Naposledy otevřené soubory
Dohromady + statistika.xlsx
Dohromady.xlsx
Dohromady2.xlsx
Příště tento dialog nezobrazovat
OK
Zavřít
Připravena
CAP
NUM
REC


STATISTICA Cz - Tabulka1
Soubor
Domů
Upravit
Zobrazit
Formát
Statistiky
Data mining
Grafy
Nástroje
Data
Feature Find
Základní statistiky
Vícenásobná regrese
ANOVA
Neparametrické statistiky
Prokládání rozdělení
Rozdělení a simulace
Základ
Pokročilé modely
Vícer./průzkumné
Analýza síly testu
Neuron. sítě
PLS, PCA, ...
VEPAC
Pokročilé/Vícerozměrné
Diagramy řízení kvality
Multivariate
Predictive
Analýza procesu
DOE
Six Sigma
Průmyslová statistika
STATISTICA VB
Dávk. analza (dle skupin)
Kalkulátory
Statistiky bloku dat
Nástroje
Data: Tabulka1 (10s krát 10ř)
1 Prom1
2 Prom2
3 Prom3
4 Prom4
5 Prom5
6 Prom6
7 Prom7
8 Prom8
9 Prom9
10 Prom10
1
2
3
4
5
6
7
8
9
10
Spustí modul Navrhování experimentů (DOE)
Tabulka1
Ř1;S1
Filtr -
Váhy


Neparametrické statistiky: Tabulka1
Zákl. výběr
2 x 2 tabulky (X2/V2/Phi2, McNemar, Fisher exal
Pozorované vs. očekávané X2
Korelace (Spearman, Kendallovo tau, gama)
Porovnání dvou nezávislých vzorků (skupiny)
Porovnání více nezávislých vzorků (skupiny)
Porovnání dvou závislých vzorků (proměnné)
Porovnání více závislých vzorků (proměnné)
Cochranův Q test
Pořadové popisné statistiky (medián, modus, ... )
OK
Storno
Možnosti
Otevři Data
SELECT CASES


Porovnání dvou skupin: Tabulka1
Proměnné
Závislé: žádné
Grupovací: žádné
Kódy pro: skupinu 1:
skupinu 2:
M-W U test
Storno
Možnosti
Zákl. výsledky
Wald-Wolfowitzův test
Kolmogorov-Smirnovův 2-výběrový test
Mann-Whitneyův U test
Použít korekci na spojitost
Krabicový graf dle skupin
Kategoriz. histogramy dle skupin
SELECT CASES
Anal.Sk.
p-hodnota pro zvýraznění:
,05

# Zdroje:


- https://portal.matematickabiologie.cz/index.php?pg=aplikovana-analyza-klinickych-a-biologickych-dat--analyza-a-management-dat-pro-zdravotnicke-obory--testovani-hypotez-o-kvantitativnich-promennych
- http://ach.upol.cz/user-files/intranet/08-neparametricketesty-2012-1347562623.pdf?fbclid=IwAR1ddEa32MO4n5ge6LVMwGD6uWVedTpph9E9MLWrapvR1KhtQ8FsMG_JKIc
- https://math.feld.cvut.cz/ftp/prucha/ubmi/predn/u15.pdf?fbclid=IwAR2KOU6Rk6ddXL9KbszBQh_NRqxquHEMD0jMiiCGxbiT2qjUhiZy6nHPVjk
- http://www.biostatisticka.cz/wp-content/seminar/Motol-lekce5.pdf
- https://www.iba.muni.cz/res/file/ucebnice/pavlik-biostatistika-v2.pdf
- https://sms.nipax.cz/_media/planovani_experimentu:ppe_5.pdf
- https://wikisofia.cz/wiki/Znam%C3%A9nkov%C3%BD_test
- https://cit.vfu.cz/statpotr/POTR/Teorie/Predn4/znamenko.htm

Děkujeme za pozornost!