



The background is a blue-tinted financial candlestick chart. It features several white candlesticks with black outlines. Overlaid on the chart are several technical analysis lines: a thick white curved line (likely a moving average or trend line) that peaks and then descends, and a thinner white straight line labeled '61.6 %: 99.19' indicating a retracement level. Two specific price points are highlighted with white boxes: '104.19' at a peak and '86.72' at a subsequent low. The overall theme is financial data analysis.

Testy pro diskrétní data

Pavla Pecherková

29. 4. 2021

Diskrétní data

- Testy pro diskrétní data jsou speciální testy, které pracují s následujícími druhy diskrétních dat:

1. Nominální data – nezávisí na pořadí

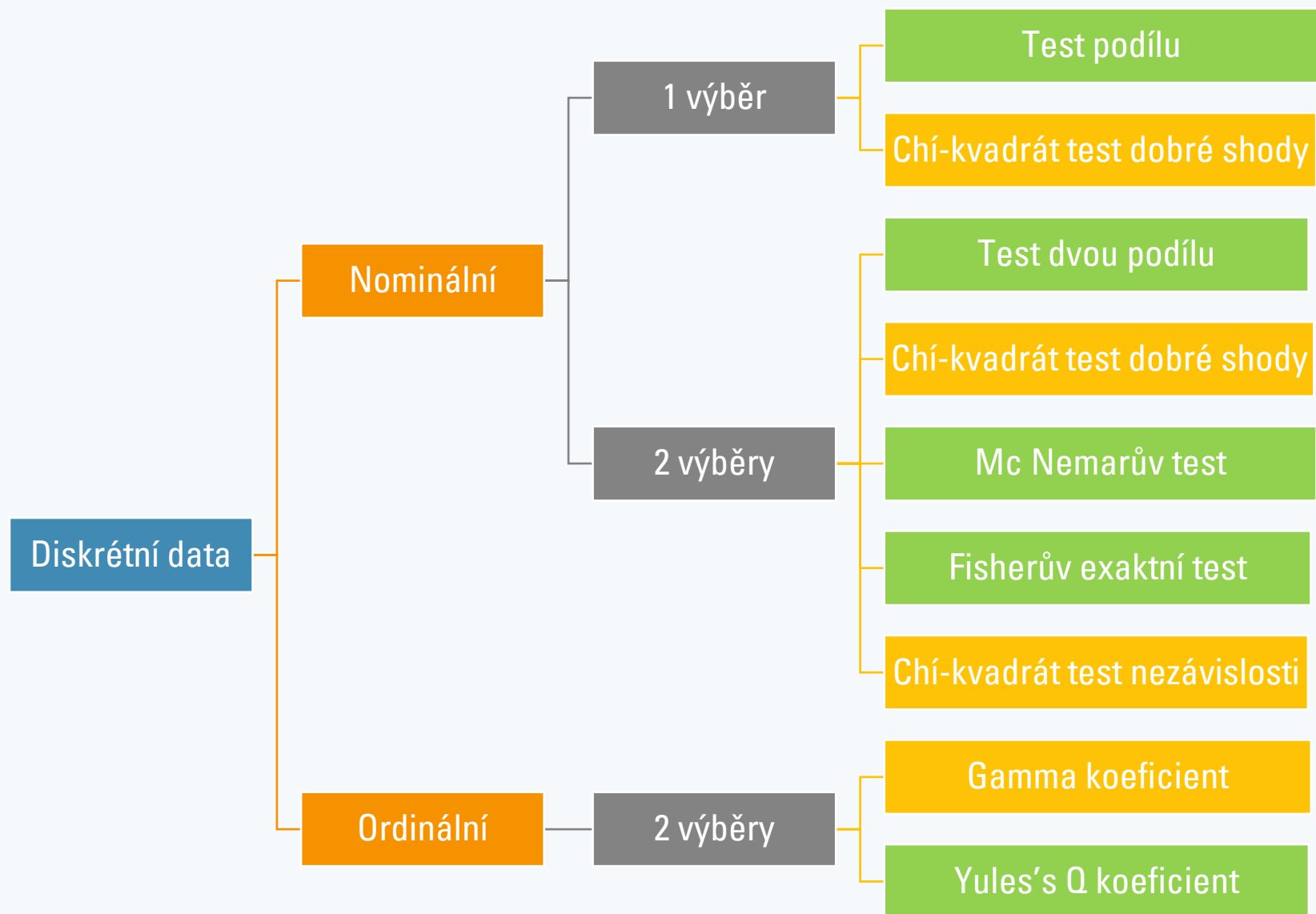
- a) Binární data $\rightarrow \{0,1\}$ – pneumatika: levá, pravá
– alternativní rozdělení
- b) Více jak 2 možnosti $\rightarrow \{0,1,2,\dots,n\}$ – barva: žlutá, červená, modrá, ..., černá
– binomické rozdělení, rovnoměrné rozdělení, kategorické rozdělení...

2. Ordinální data – lze je seřadit podle velikosti

- a) Binární data $\rightarrow \{0,1\}$ – **nemoc: má, nemá**
- b) Více jak 2 možnosti $\rightarrow \{0,1,2,\dots,n\}$ – **testy hypotéz: zcela chápu, spíše chápu, chápu tak napůl, spíše nechápu, vůbec nechápu**

	Test pro nominální data	Test pro ordinální data
Nominální data		
Ordinální data		

- Jak se liší diskrétní data i pro tato data je dobré používat správné testy.



Binární data

Binární/nebinární data

Test podílu pro 1 výběr

- Předpoklad:

- diskrétní náhodná veličina binární – nominální (ordinální) data,
- počet dat $n > 30$,
- $np \geq 5, n(1 - p) \geq 5$, kde p je pravděpodobnost úspěchu.

- $H_0: p = p_0$
- $H_A: p \neq p_0; p > p_0; p < p_0$

- Statistika:

$$T = \frac{p - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1)$$

- p_0 je předpokládaný podíl, tedy nulová hypotéza. Ač se jedná o diskrétní hodnoty aproximovat normálním rozdělením díky platnosti limitních vět.

Na hladině významnosti 0,05 otestujte, že podíl žen sekajících trávu je 0,2. Zeptali jsme se 100 lidí a hodnoty zapsali do tabulky:

Žena	Muž	M	M	Ž	Ž	...	M
------	-----	---	---	---	---	-----	---

- Diskrétní data: ano
- Počet dat > 30 : ano ($100 > 30$)
- $np = 100 * 0,2 = 20 \geq 5$: ano
- $n(1 - p) = 100 * 0,8 = 80 \geq 5$: ano

1. Spočteme statistiku T
2. Porovnáme s tabulkovými hodnotami pro normální rozdělení $N(0,1)$
3. Zjistíme, zda zamítáme či nezamítáme

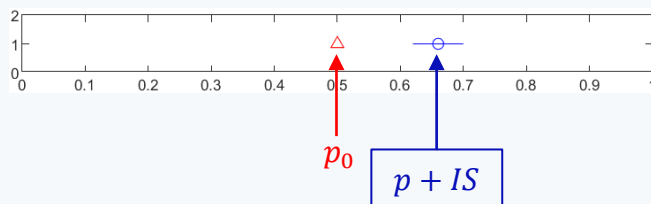
Zeptali jsme se rodičů 5. týden karantény, zda by poslali děti do školy. Očekává se, že toto číslo bude kolem hodnoty 0,5, tedy že 50 % rodičů by dítě poslalo do školy, zbytek by raději neriskoval onemocnění a nechají si ho doma. Zeptali jsme se 485 rodičů a z toho by 320 poslalo děti do školy. Na hladině významnosti 0,05 testujte, zda byl předpoklad správný.

- $H_0: p = 0,5$
- $H_A: p \neq 0,5$

- Spočítáme odhad $\hat{p} = \frac{320}{485} = 0,66$
 - Toto je pouze bodový odhad, interval spolehlivosti je $IS \in < 0,62; 0,70 >$

- Spočítáme statistiku

$$T = \frac{p - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0,66 - 0,5}{\sqrt{\frac{0,5(1-0,5)}{485}}} = 7,0382$$



- Zjistíme p-hodnotu (pomocí distribuční funkce $N(0,1)$)
 - $p_val = \text{normcdf}(-\text{abs}(T))$ – pravostranný test
 - $p_val = 1 - \text{normcdf}(-\text{abs}(T))$ – levostranný test
 - $p_val = \text{normcdf}(-\text{abs}(T)) * 2$ – oboustranný test

Máme oboustranný test, proto

```
p_oboustranny=normcdf(-abs(T))*2
p_oboustranny = 1.9476e-12
```

Zamítáme hypotézu, že 50 % rodičů by 5. týden poslalo děti do školy, protože $p < 0,05$.

Test o shodě dvou podílů

Testujeme, zda podíl jednoho souboru je roven podílu druhého souboru, jinak platí to co u testu shodě dvou podílů pro 1 výběr:

- $H_0: p_1 = p_2$
- $H_A: p_1 \neq p_2; p_1 > p_2; p_1 < p_2$

Statistika pro podíl dvou výběrů je

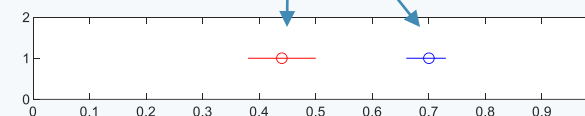
- $T = \frac{p_1 - p_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$
- $\hat{p} = \frac{Y_1 + Y_2}{n_1 + n_2}$ kde Y_1 resp. Y_2 je počet příznivých odpovědí

Zeptali jsme se 900 lidí, zda souhlasí s plošným zákazem kouření v restauracích. Získali jsme následující data. Z 250 kuřáků souhlasí s tímto zákazem 110, z 650 nekuřáků s tímto zákazem souhlasí 452. Na hladině významnosti 0,05 testujte tvrzení, že souhlasí v obou skupinách stejné množství lidí.

- $H_0: p_1 = p_2$
- $H_A: p_1 \neq p_2$

- $p_1 = \frac{Y_1}{n_1} = \frac{110}{250} = 0,44$
- $p_2 = \frac{Y_2}{n_2} = \frac{452}{650} = 0,70$
- $T = -7,0862$

p_1 a p_2 včetně IS. p_1 má širší IS, protože je zde méně respondentů a tedy větší účinnost.



```
p_oboustranny=normcdf(-abs(T))*2
p_oboustranny = 1.3780e-12
```

Zamítáme hypotézu, že v obou skupinách je stejný názor na zákaz kouření v restauracích, protože $p < 0,05$.

Chí-kvadrát test dobré shody

Tento test už byl u testů rozdělení. Lze tedy porovnávat zda výběr pochází z nějakého diskrétního rozdělení (nebo diskretizovaného spojitého) – rovnoměrné, binomické, kategorické...

Lze také porovnávat, zda dva výběry pochází ze stejného rozdělení

Předpoklady:

- diskrétní data
- všechny četnosti > 2
- alespoň 80% četností > 5

H_0 : není statisticky významný rozdíl mezi pozorovanými a očekávanými četnostmi

H_A : je statisticky významný rozdíl mezi pozorovanými a očekávanými četnostmi

Statistika:

$$T = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2(n-1)$$

kde O_i jsou pozorované četnosti a E_i očekávané četnosti.

Pro každou pozorovanou četnost musí být i teoretická četnost

O_i	E_i
O_1	E_1
O_2	E_2
O_3	E_3
...	...
O_n	E_n

Řetězec cukráren nabízí 4 druhy zmrzliny ve svých provozovnách. Dosud se ve stávajících provozovnách zmrzlina prodávala takto: vanilková (62 %), čokoládová (18 %), jahodová (12 %), pistáciová (8 %). Při otevření nové provozovny se prodala za první den zmrzlina takto: vanilková (120 ks), čokoládová (40 ks), jahodová (18 ks), pistáciová (22 kusů).

Na hladině významnosti 0,05 testujte tvrzení, že prodej v nové lokalitě je stejný jako v ostatních provozovnách.

- Pozorované četnosti známe, musíme dopočíst očekávané. Jak by to vypadlo, kdyby to bylo přesně podle předpokladů?
- Počet prodaných zmrzlin: $\sum O_i = 120 + 40 + 18 + 22 = 200$
- Teoretická četnost tedy bude: $E_1 = 200 * 0,62 = 124, \dots$

```
h = 0
p = 0.2286
st = struct with fields:
  chi2stat: 4.3235
  df: 3
  edges: [0.5000 1.5000 2.5000 3.5000 4.5000]
  O: [120 40 18 22]
  E: [124 36 24 16]
```

O_i	E_i
120	124 (200*0,62)
40	36 (200*0,18)
18	24 (200*0,12)
22	16 (200*0,08)

Nezamítáme hypotézu, že prodej je v nové cukrárně stejný jako v předchozích provozovnách.

McNemarův test

McNemarův test je hodně podobný Chí-kvadrátu testu nezávislosti (Fisherovu exaktnímu testu). Ale na rozdíl od něj, sleduje jen poměr mezi dvěma hodnotami.

Tento test je pouze pro velikost 2x2 a sleduje se pouze změna stavu, tedy (b) a (c) v tabulce, nikoliv kolik lidí ve stavu zůstalo. Tento počet není pro výpočet vůbec důležitý na rozdíl právě od chí-kvadrát testu nezávislosti.

Předpoklady:

- diskrétní binární data
- všechny četnosti > 2
- alespoň 80% četností > 5

	ANO	NE
ANO	a	b
NE	c	d

- $H_0: p(b) = p(c)$
- $H_A: p(b) \neq p(c)$

Statistika:

$$T = \frac{(b-c)^2}{b+c} \sim \chi^2(1)$$

Ptali jsme se všech studentů v prvním ročníku, jakou mají náladu před zkouškou. Po zkoušce jsme položili stejný dotaz.

Studenti odpověděli takto:

Na hladině významnosti 0,05 otestujte, zda má na náladu vliv zkouška.

$H_0: p(b) = p(c)$ (zkouška nemá vliv na náladu)

$H_A: p(b) \neq p(c)$ (zkouška má vliv na náladu)

$$T = \frac{(b - c)^2}{b + c} = \frac{(99 - 126)^2}{99 + 126} = 3,0044$$

`p_McNemar=1-chi2cdf(T,1)`

`p_McNemar = 0.0830`

Nezamítám hypotézu o tom, že zkouška nemá vliv na náladu.

		Dobrá nálada po zkoušce	
		ANO	NE
Dobrá nálada před zkouškou	ANO	55	99
	NE	126	152

		Dobrá nálada po zkoušce	
		ANO	NE
Dobrá nálada před zkouškou	ANO	a	b
	NE	c	d