Přednáška 9 – Testy nezávislosti

Doposud jsme testovali buď shodu parametru/mediánu jednoho výběru s předpokládanou hodnotou, shodu dvou a více výběrů, nebo rozdělení hodnot výběru. Touto přednáškou začínáme nové téma a budeme si povídat o testech nezávislosti. To znamená, že budeme testovat, zda jsou dva výběry nezávislé – přesněji řečeno, zda soubory (pozorované veličiny), ze kterých výběry pochází, jsou nezávislé.

Testy nezávislosti budeme rozlišovat podle toho, zda pozorované veličiny, na kterých jsme naměřili výběry, jsou spojité, nebo diskrétní. V případě spojitých dat volíme mezi parametrickými testy s předpokladem normality a neparametrickými testy bez normality dat. V následující tabulce jsou testy nezávislosti, které budeme používat (je také dostupná na webu pod odkazem Jak zvolit test hypotéz). Pro všechny tyto testy obecně jsou nulová a alternativní hypotézy stejné:

$$H_0: \quad \text{veličiny jsou nezávislé},$$
$$H_A: \quad \text{veličiny nejsou nezávislé}.$$

Směr testu u testů nezávislosti nevolíme. Probereme každý z testů podrobněji.

Pro spojité veličiny	
Parametrické	Neparametrické
Pearsonův test korelačního koeficientu **pearson_test** Předpoklady: $N(\mu, \sigma^2)$, párové výběry H_0: jsou nezávislé	Spearmanův test **spearman_test** Předpoklady: bez $N(\mu, \sigma^2)$, párové výběry H_0: jsou nezávislé

Pro diskrétní veličiny
χ^2 test nezávislosti – **chisquare_test_i** Předpoklady: všechny četnosti > 2, alespoň 80% četností > 5, kontingenční tabulka H_0: jsou nezávislé
Fisherův exaktní test Předpoklady: nominální data H_0: jsou nezávislé
Gamma koeficient (Goodmanovo-Kruskalovo gamma) Předpoklady: ordinální data, kategorické rozdělení H_0: jsou nezávislé
Yule's Q koeficient Předpoklady: ordinální data, alternativní rozdělení H_0: jsou nezávislé

Nejdřív se podíváme na testy nezávislosti pro spojité hodnoty výběrů.

Pearsonův test korelačního koeficientu

Pearsonův test korelačního koeficientu (**pearson_test**) použijeme v případě, že potřebujeme zjistit na hladině významnosti α, zda jsou hodnoty dvou párových spojitých výběrů s předpokladem normality dat nezávislé.

O Pearsonově korelačním koeficientu jsme už mluvili v souvislosti s náhodnými vektory a regresní analýzou v přednášce 4, kde nás zajímala míra korelace (závislosti) mezi náhodnými veličinami x a y. Tady použijeme výběrový korelační koeficient a budeme pracovat s dvěma párovými výběry X a Y, které jsme pořídili při pozorování těchto náhodných veličin. Pearsonův výběrový korelační koeficient r_p spočítáme podobně jako souborový (viz pro porovnání přednáška 4 na webu) :

$$r_p = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}}, \qquad \text{kde} \quad s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) \quad \text{– vyběrová kovariance,} \quad s_x^2 s_y^2 \text{ – vyběrové rozptyly,}$$

tj., rozdíl je jenom v použití výběrové kovariance a výběrových směrodatných odchylek. Pořád platí, že korelační koeficient je definován na intervalu od -1 do 1, kde hodnoty blížící se k -1 znamenají nepřímou závislost a hodnoty blízko 1 přímou závislost:

- $0 < r_p \leq 1$ – přímá závislost: $X_i \Uparrow Y_i \Uparrow$ nebo $X_i \Downarrow Y_i \Downarrow$
- $-1 \leq r_p < 0$ – nepřímá závislost: $X_i \Uparrow Y_i \Downarrow$ nebo $X_i \Downarrow Y_i \Uparrow$
- pro nezávislé veličiny $r_p = 0$ (neplatí to obráceně).

Jelikož se pro nezávislé veličiny korelační koeficient rovná 0, zní nulová hypotéza Pearsonova testu takto:

$$H_0 : r_p = 0, \quad \text{tj., veličiny jsou lineárně nezávislé.}$$

Alternativní hypotéza popírá nulovou:

$$H_A : r_p \neq 0, \quad \text{tj., veličiny nejsou lineárně nezávislé.}$$

Test používá statistiku

$$T = \frac{r_p}{\sqrt{\frac{1-r_p^2}{n-2}}} \sim \text{Studentovo rozdělení.}$$

Příklad Chystáme se kupovat byt. Zkoumáme 20 náhodně vybraných nabídek prodeje bytů v Praze (zdroj sreality.cz). V tabulce je uvedena rozloha v m^2 a cena bytu v Kč. Zajímá nás, zda je lineární závislost mezi rozlohou bytu a jeho cenou.

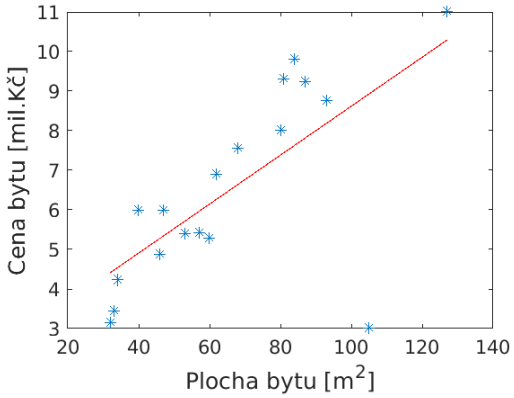
rozloha, m^2	cena, Kč
84	9 811 200
105	3 000 000
40	5 990 000
57	5 420 000
127	11 000 000
80	8 000 000
32	3 150 000
46	4 876 000
33	3 450 000
93	8 754 000
34	4 249 929
53	5 400 000
87	9 240 250
47	5 988 050
81	9 296 600
60	5 280 000
62	6 894 250
68	7 566 600

Řešení: Dá se předpokládat, že čím větší je byt, tím je dražší, takže očekáváme kladnou korelaci. Nicméně závislost mezi plochou a cenou bytu nemusí být nutně lineární – může být např., polynomiální nebo exponenciální. Proto otázka není tak jednoznačná. Zkusíme vyloučit možnost, že rozloha bytu a cena jsou lineárně nezávislé pomocí Pearsonova testu (**pearson_test**). Podle testu normality splňují oba párové výběry předpoklad normality, tedy ho můžeme použít. Jak jsme se zmínili, nulová hypotéza Pearsonova testu zní:

$$H_0 : \quad \text{rozloha bytu a cena jsou lineárně nezávislé,}$$
$$\text{alternativní hypotéza} \quad H_A : \quad \text{nejsou lineárně nezávislé.}$$

P-hodnota se rovná $0.0018 < 0.05$, což znamená, že zamítáme nulovou hypotézu, že rozloha bytu a cena jsou lineárně nezávislé. To znamená, že korelační koeficient (**correl**) $r_p \neq 0$, je roven 0.6824, tj., mezi veličinami je

vazba: čím větší je byt, tím je dražší. Taková data jsou vhodná pro analýzu pomocí lineární regrese, což můžeme ukázat na obrázku:



proto tento test nezávislosti využijeme nejvíce při testování dat na vhodnost k lineární regresi (což uvidíme příští týden). Pokud bychom nulovou hypotézu nezamítli, znamenalo by to, že data nejsou vhodná k lineární regresi, protože mohou být lineárně nezávislá. Může se ale projevit jiný typ závislosti (polynomiální, exponenciální, atd.).

Spearmanův test korelačního koeficientu

Spearmanův test korelačního koeficientu (**spearman_test**) použijeme v případě, že potřebujeme zjistit na hladině významnosti α, zda jsou hodnoty dvou párových spojitých výběrů bez předpokladu normality dat nezávislé. Je neparametrickou alternativou Pearsonova testu. Spearmanův test používá tzv. Spearmanův koeficient pořadové korelace, který se počítá na základě Pearsonova koeficientu, ale s využitím pořadí hodnot výběrů místo dat:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \quad \text{kde } d_i \text{ jsou rozdíly pořadí z obou výběrů.}$$

Podobně jako v předchozím případě, je Spearmanův koeficient pořadové korelace r_s definován na intervalu od -1 do 1, kde hodnoty blížící se k -1 znamenají nepřímou závislost a hodnoty blízko 1 přímou závislost. Pro nezávislé veličiny $r_s = 0$. Nulová hypotéza Spearmanova testu proto zní:

$$H_0 : r_s = 0, \quad \text{tj., veličiny jsou nezávislé.}$$

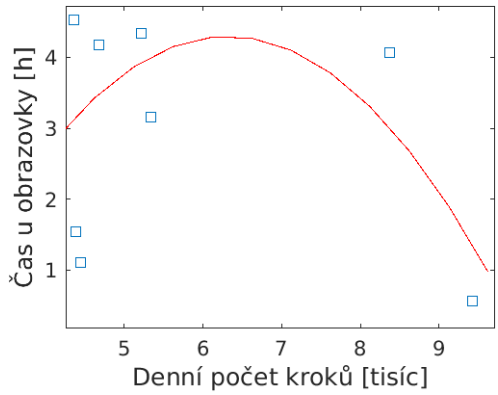
Alternativní hypotéza popírá nulovou:

$$H_A : r_s \neq 0, \quad \text{tj., veličiny nejsou nezávislé.}$$

Příklad Zjišťujeme, zda je závislost mezi časem stráveným u obrazovky mobilu a denním počtem kroků. Záznamy máme v tabulce:

denní počet kroků	čas u obrazovky, [h]
8374.6163	4.07
9419.2094	0.57
4364.4742	4.53
4682.0057	4.18
9885.3437	0.54
5214.719	4.34
4119.8086	4.57
4446.0034	1.11
5337.3764	3.16
4384.4456	1.54

Řešení: Ověříme předpoklad normality výběrů – v obou případech zamítáme nulovou hypotézu o normalitě. To znamená, že musíme použít neparametrický test nezávislosti, a to je Spearmanův test. P-hodnota je $0.03 < 0.05$, zamítáme tedy nulovou hypotézu Spearmanova testu, že čas u obrazovky a denní počet kroků jsou nezávislé. To znamená, že data jsou vhodná k polynomiální nebo exponenciální regresi.* Zkusíme např., polynomiální regresi 2.řádu, což vidíme na obrázku:



Pro zajímavost můžeme zkusit i Pearsonův test, kde p-hodnota je 0.1263, což znamená, že bychom nezamítli nulovou hypotézu, že data jsou lineárně nezávislá. Na obrázku je právě vidět, že lineární závislost by nebyla vhodná.

Testy nezávislosti pro diskrétní veličiny

Teď se podíváme na testy nezávislosti pro diskrétní data.

χ^2 test nezávislosti

χ^2 test nezávislosti použijeme v případě, že potřebujeme otestovat nezávislost dvou diskrétních náhodných veličin. Podobně jako χ^2 test dobré shody, se kterým jsme se seznamili při testování rozdělení (přednáška 7), používá i tento test stejnou statistiku

$$T = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i},$$

kde O_i – pozorované četnosti a E_i – očekávané četnosti. Pro použití tohoto testu by měly všechny četnosti naměřených hodnot být větší než 2 a alespoň 80% četností by mělo být větší než 5. Nulová a alternativní hypotézy jsou:

H_0 : veličiny jsou nezávislé,
H_A : veličiny nejsou nezávislé.

K použití testu potřebujeme data ve tvaru kontingenční tabulky četností obou veličin, nezávislost kterých testujeme a pozorované a očekávané četnosti O_i a E_i.

Příklad Ptali jsme se zákazníků různých věkových kategorií, kde objednávají potraviny během karantény. Mezi zákazníky ve věku od 20 do 30 let preferovalo 56 e-shop Rohlík, 42 Košík a 19 Tesco. Ve věku 31-45 let 46 zákazníků volilo Rohlík, 17 Košík a 37 využívalo služeb rozvozu potravin Tesco. Z věkové kategorie od 46 do 70 let 25 zákazníků by si objednalo potraviny v Rohlíku, 36 v Košíku a 44 v Tesco. Ha hladině významnosti 0.05 testujte tvrzení, že věk a volba e-shopu jsou nezávislé.

*Příští týden uvidíme, jak ověřit, který druh regrese vyhovuje líp naměřeným datům.

Řešení: Tady máme diskrétní veličiny věk, která může nabývat tří možných hodnot {20-30 let, 31-45 let, 46-70 let} a e-shop ∈ {Rohlík, Košík, Tesco}. Pro χ^2 test nezávislosti potřebujeme z naměřených dat vytvořit kontingenční tabulku, což znamená, že zapíšeme data pro všechny kombinace hodnot obou veličin do tabulky takto:

věk \ e-shop	Rohlík	Košík	Tesco
20-30 let	56	42	19
31-45 let	46	17	37
46-70 let	25	36	44

kontingenční tabulka=O

V našem případě to je kontingenční tabulka 3×3. Představuje pozorované četnosti O_i – to jsou naše data obou veličin věk a e-shop. Potřebujeme ale ještě E_i – to jsou četnosti očekávané v případě nezávislosti veličin.

Uvědomíme si, že kdybychom tabulku vydělili počtem dat, dostaneme sdružené rozdělení, tj., pravděpodobnosti každé kombinace hodnot veličin věk a e-shop (viz přednáška 4). Připomeňme si, že pokud se sdružené rozdělení rovná součinu marginálních rozdělení, jsou náhodné veličiny nezávislé:

$$f(\text{věk}, \text{e-shop}) = f(\text{věk}) f(\text{e-shop}).$$

To znamená, že z našich dat musíme vyrobit takovou "nezávislou" tabulku a to budou očekávané četnosti E_i. Postup je následující:

1. Vydělíme četnosti v kontingenční tabulce počtem dat 332 – tímto dostaneme sdružené rozdělení f(věk,e-shop):

věk \ e-shop	Rohlík	Košík	Tesco
20-30 let	0.17	0.13	0.06
31-45 let	0.14	0.05	0.11
46-70 let	0.08	0.11	0.15

2. Spočítáme obě marginální rozdělení $f(\text{věk})$ a $f(\text{e-shop})$

věk \ e-shop	Rohlík	Košík	Tesco
20-30 let	0.17	0.13	0.06
31-45 let	0.14	0.05	0.11
46-70 let	0.08	0.11	0.15

$f(\text{věk})$
0.17+0.13+0.06=0.36
0.14+0.05+0.11=0.3
0.08+0.11+0.15= 0.34

$f(\text{e-shop})$	0.39	0.29	0.32

3. Dále vynásobíme vektory $f(\text{věk})$ a $f(\text{e-shop})$ – tímto dostaneme nové sdružené rozdělení pro dvě nezávislé veličiny věk a e-shop:

$f(\text{věk})f(\text{e-shop}) =$

věk \ e-shop	Rohlík	Košík	Tesco
20-30 let	0.14	0.1	0.12
31-45 let	0.12	0.09	0.09
46-70 let	0.13	0.1	0.11

Tak by vypadaly sdružené pravděpodobnosti, kdyby veličiny věk a e-shop byly nezávislé.

4. Teď pravděpodobnosti vynásobíme zase počtem dat 332 a dostaneme zpátky četnosti. Toto jsou četnosti E_i, které bychom očekávali v případě nezávislých veličin věk a e-shop.

věk \ e-shop	Rohlík	Košík	Tesco
20-30 let	0.14	0.1	0.12
31-45 let	0.12	0.09	0.09
46-70 let	0.13	0.1	0.11

sdružené rozdělení

$*332 =$

věk \ e-shop	Rohlík	Košík	Tesco
20-30 let	47	33	40
31-45 let	40	30	30
46-70 let	43	33	37

E

P-hodnota= $0.0000033 < 0.05$, proto zamítáme nulovou hypotézu, že věk zákazníků a výběr e-shopu jsou nezávislé.

Fisherův exaktní test

Fisherův exaktní test použijeme v případě testování nezávislosti dvou diskrétních náhodných veličin, které mohou nabývat pouze dvou možných realizací. To znamená, že budeme pracovat s kontingenční tabulkou 2×2.[†] Fisherův exaktní test nevyžaduje podmínky minimálních četností jako χ^2 test nezávislosti. Na rozdíl od něj funguje pouze na základě pozorovaných četností. Nulová a alternativní hypotézy jsou:

$$\begin{aligned} H_0: &\quad \text{veličiny jsou nezávislé,} \\ H_A: &\quad \text{veličiny nejsou nezávislé.} \end{aligned}$$

X \ Y	$Y=1$	$Y=2$
$X=1$	a	b
$X=2$	c	d

Test používá referenční pravděpodobnost, která se počítá následovně:

$$p^* = \frac{\binom{a+c}{a}\binom{b+d}{b}}{\binom{n}{a+b}} = \frac{(a+b)!\,(a+c)!\,(c+d)!\,(b+d)!}{n!\,a!\,b!\,c!\,d!},$$

kde a, b, c, d jsou četnosti z kontingenční tabulky.

Dále se dopočítávají pravděpodobnosti pro všechny možné kombinace četností při zachování součtů z marginálního řádku a sloupce. P-hodnota je součet všech pravděpodobností, které jsou menší nebo se rovná p^*.

Příklad Testujeme nezávislost mezi volbou operačního systému na mobilním telefonu a pohlavím. Data jsou v tabulce:

pohlaví \ operační systém	iOS	Android
muži	19	26
ženy	31	22

Řešení: Máme dvě binární diskrétní náhodné veličiny: operační systém $\in$ {iOS, Android} a pohlaví $\in$ {muži,ženy}. Otestujeme jejich nezávislost pomocí Fisherova exaktního testu. Nulová a alternativní hypotézy jsou:

$$\begin{aligned} H_0: &\quad \text{volba operačního systému a pohlaví jsou nezávislé,} \\ H_A: &\quad \text{veličiny nejsou nezávislé.} \end{aligned}$$

P-hodnota= $0.1555 > 0.05$, proto nezamítáme nulovou hypotézu o nezávislosti mobilního operačního systému a pohlaví.

Důležitá poznámka: Všimněme si, že doposud jsme pracovali s nominálními diskrétními daty. Připomeňme si, že nominální jsou data, které nemůžeme uspořádat. Například, nemůžeme uspořádat hodnoty Rohlík, Košík a Tesco nebo iOS a Android, atd. V případě χ^2 testu nezávislosti a Fisherova exaktního testu mohou být použita nominální data. Teď se podíváme na testy nezávislosti, které jsou na rozdíl od nich určeny pro ordinální data, tj., data, které můžeme uspořádat. Například, známka u zkoušky $\in \{A, B, C, D, E\}$ je diskrétní náhodná veličina s ordinálními hodnotami. Můžeme říct, že A je lepší než B, B je lepší než C, a E je nejhorší.

[†] V poslední době se objevují studie o rozšíření testu pro vícepolní kontingenční tabulky.

Gamma koeficient (Goodmanovo-Kruskalovo gamma)

Gamma koeficient neboli Goodmanovo-Kruskalovo gamma je určen pro ordinální data s kategorickým rozdělením, což znamená, že každá z náhodných veličin může nabývat minimálně 3 možných realizací (viz přednáška 2).‡ Využívá se jako míra asociace mezi veličinami x a y.

Pro výpočet Gamma koeficientu

$$\gamma = \frac{N_c - N_d}{N_c + N_d},$$

kde N_c je celkový počet dat s přímou závislostí, N_d celkový počet dat s nepřímou závislostí, potřebujeme kontingenční tabulku uspořádaných hodnot. Podobně jako korelační koeficient,

$\gamma \in \langle -1, 1 \rangle$, kde

- $0 < \gamma \leq 1$ – přímá závislost
- $-1 \leq \gamma < 0$ – nepřímá závislost
- pro nezávislé veličiny $\gamma = 0$.

Nulová hypotéza testu hypotéz s Gamma koeficientem zní takto:

$$H_0 : \gamma = 0, \quad \text{tj., veličiny jsou nezávislé.}$$

Alternativní hypotéza popírá nulovou:

$$H_A : \gamma \neq 0 \quad \text{tj., veličiny nejsou nezávislé.}$$

Test používá statistiku

$$T = \gamma \sqrt{\frac{N_c + N_d}{n(1-\gamma^2)}}.$$

Příklad Zajímá nás, zda je vazba mezi vzděláním a dobou uplynulou od vydání řidičského průkazu. Máme data v tabulce:

doba od vydání řidičáku \ vzdělání	bez maturity	s maturitou	VŠ
do 5 let	25	38	11
6-10 let	56	47	37
více než 10 let	53	46	54

Řešení: Máme dvě diskrétní náhodné veličiny: doba od vydání řidičáku $\in$ {do 5 let, 6-10 let, víc než 10 let } a vzdělání $\in$ {bez maturity, s maturitou, VŠ}. Pro použití Gamma koeficientu musí být data v kontingenční tabulce uspořádana v souladu pro obě veličiny, například, od nejnižšího k nejvyššímu (tj., do 5 let a bez maturity jsou první hodnoty veličin). Hodnoty na hlavní diagonále vyjadřují přímou vazbu veličin ("vyšší vzdělání – déle vlastníme řidičák"). Hodnoty na vedlejší diagonále vyjadřují opačný směr – nepřímou vazbu.

Nejdříve spočítáme N_c – počet všech hodnot, které vyjadřují přímou vazbu. Proto vynásobíme každou hodnotu počínaje horním levým políčkem tabulky součtem hodnot z nižších řádků vpravo:

$$N_c = 25*(47+37+46+54)+38*(37+54)+11*0+56*(46+54)+47*54+37*0+53*0+46*0+54*0 = 16196.$$

Teď spočítáme N_d pro výpočet hodnot vyjadřujících nepřímou vazbu. Postup je obrácený, tj., začínáme od horního pravého políčka tabulky směrem vlevo dolů:

$$N_d = 11*(56+47+53+46)+38*(56+53)+25*0+37*(53+46)+47*53+56*0+54*0+46*0+53*0 = 12518.$$

‡Můžeme ho použít i pro spojité hodnoty.

Gamma koeficient je

$$\gamma = \frac{N_c - N_d}{N_c + N_d} = \frac{16196 - 12518}{16196 + 12518} = 0.128,$$

což znamená mírnou přímou závislost. P-hodnota$= 0.0592 > 0.05$, takže jen velmi těsně nezamítáme nulovou hypotézu, že vzdělání a doba od vydání řidičského průkazu jsou nezávislé.

Yule's Q koeficient

Yule's Q koeficient je speciální případ Gamma koeficientu pro binární diskrétní ordinální data, tj., pro kontingenční tabulku 2×2 s uspořádanými hodnotami. Nulová a alternativní hypotézy a statistika jsou stejné jako v případě Gamma koeficientu. Rozdíl je jenom ve výpočtu Yule's Q koeficientu:

$$Q = \frac{N_c - N_d}{N_c + N_d} = \frac{ad - bc}{ad + bc},$$

kde a, b, c, d jsou četnosti z kontingenční tabulky.

X \ Y	$Y = 1$	$Y = 2$
$X = 1$	a	b
$X = 2$	c	d

Příklad Zjišťujeme závislost mezi dobou učení a výsledkem u zkoušky. Data máme v tabulce:

doba učení \ prospěch	neprospěl	prospěl
málo se učil/a	13	5
hodně se učil/a	7	19

Řešení: Spočítáme Yule's Q koeficient:

$$Q = \frac{N_c - N_d}{N_c + N_d} = \frac{ad - bc}{ad + bc} = \frac{13 * 19 - 7 * 5}{13 * 19 + 7 * 5} = 0.752,$$

což je poměrně vysoká přímá závislost mezi dobou učení a prospěchem.

P-hodnota$= 4.8866e - 07 < 0.05$, takže zamítáme nulovou hypotézu, že doba učení a prospěch u zkoušky jsou nezávislé.