Fakulta informačních technologií VUT v Brně

Statistika a pravděpodobnost

Projekt

Bc. Ondřej Ondryáš

10. prosince 2022

Úkol 1: Střídání času

> Český stát si objednal průzkum, jak lidé vnímají střídání zimního a letního času. Průzkum zahrnoval větší města (Praha, Brno), menší města (Znojmo, Tišnov) a obce (Paseky, Horní Lomná, Dolní Věstonice). V průzkumu zjišťovali, co lidem lépe vyhovuje — zda střídání letního a zimního času, pouze zimní čas, nebo pouze letní čas.

Cílem úkolu je především ověřit souvislosti mezi bydlištěm (nebo velikostí bydliště) osoby a jejím názorem na střídání času.

Analýza tedy vychází z kvalitativních nominálních dat, která tvoří kontingenční tabulku, kde sloupce představují diskrétní náhodné veličiny – jednotlivé třídy (obce), v řádcích jsou četnosti naměřených realizací. Protože jsou pro účely úkolů *d)* a *e)* seskupeny veličiny do větších tříd (velká města, menší města, jiné obce), v tabulce 1 jsou označeny pouze zkratkami VM, MM a OB.

	VM1	VM2	MM1	MM2	OB1	OB2	OB3	ST	Σ
zimní	510	324	302	257	147	66	87	13	1 706
letní	352	284	185	178	87	58	65	16	1 225
střídání	257	178	124	78	44	33	31	7	752
nemá názor	208	129	70	74	6	19	32	5	543
Σ	1 327	915	681	587	284	176	215	41	4 226

Tabulka 1: Naměřená data. Sloupec VM1 odpovídá datům z Prahy, VM2 Brnu, MM1 Znojmu, MM2 Tišnovu, OB1 Pasekám, OB2 Horní Lomné, OB3 Dolním Věstonicím a ST značí data z mého vlastního průzkumu.

K vypracování úkolu byl použit software Excel, soubor s patřičnými tabulkami `Ukol1.xlsx` je přiložen.

Části *a) – c)*

První tři části řeší otázku, zde je v jednotlivých obcích stejné relativní zastoupení obyvatel preferujících postupně zimní čas, letní čas nebo střídání času. V následujícím textu tedy uvažujme otázku části *a)* týkající se zimního času; řešení dalších částí je obdobné.

Pro účely otázky můžeme data pozměnit tak, jako bychom v každé třídě zjišťovali pouze, zda respondent preferuje, či nepreferuje zimní čas:

	VM1	VM2	MM1	MM2	OB1	OB2	OB3	ST	Σ
pref. zimní	510	324	302	257	147	66	87	13	1 706
nepref. zimní	817	591	379	330	137	110	128	28	2 520
Σ	1 327	915	681	587	284	176	215	41	4 226

Takto upravená data můžeme považovat za výběry z binomického rozdělení. Pokud bychom vycházeli z předpokladu, že se zastoupení „příznivých“ jevů mezi jednotlivými třídami rovnají, můžeme z původních dat provést bodový odhad parametru p jako podíl všech výskytů možnosti „zimní“ a celkového počtu měření:

$$\hat{p}_{\text{Zimni}} = \frac{\sum n_c}{n} = \frac{510 + 324 + \dots}{4\,226} \doteq 0,403\,69$$

Vynásobíme-li tímto odhadem počet respondentů v jednotlivých třídách, získáme očekávané četnosti v těchto třídách:

	VM1	VM2	MM1	MM2	OB1	OB2	OB3	ST	Σ
n_c	510	324	302	257	147	66	87	13	1 706
$\hat{n}_c$	535,70	369,38	274,91	236,97	114,65	71,05	86,79	16,55	1 706

Nad těmito hodnotami můžeme použít χ^2 test dobré shody pro otestování hypotézy:

$$H_0 : p_{\text{VM1}} = p_{\text{VM2}} = p_{\text{MM1}} = \ldots = \hat{p}_{\text{Zimni}}$$

proti alternativní:

$$H_A : \exists c : p_c \neq \hat{p}_{\text{Zimni}}.$$

Uvedené očekávané četnosti jsou vypočítány z dat v prvním listu tabulky. V druhém listu jsou pak provedeny samotné testy. Testovacím kritériem je:

$$x = \sum_{c \in \{\text{VM1},\ldots\}} \frac{(n_c - \hat{n}_c)^2}{\hat{n}_c}$$

přičemž $x \sim \chi^2(k)$, kde k je počet stupňů volnosti, který je roven počtu tříd pokráceném o počet odhadnutých parametrů (zde pouze jeden – $\hat{p}_{\text{Zimni}}$) a o jedničku, tedy $x \sim \chi^2(8-1-1)$. Doplněk kritického oboru $\overline{W}_\alpha = \langle 0; \chi^2_{1-\alpha}(6)\rangle$, tedy pro $\alpha = 0,05$ je $\overline{W}_{0,05} = \langle 0; 12,591\,59\rangle$.

Konkrétně pro otázku *a)* ohledně zimního času je $x \doteq 21,42$, tedy $x \notin \overline{W}_{0,05}$, tedy na hladině významnosti 0,05 **zamítáme** hypotézu, že v městech, obcích a v okolí studenta je stejné procentuální zastoupení obyvatel, co preferují zimní čas.

Obdobný postup použijeme i pro části *b)* a *c)*, pouze použijeme bodové odhady pravděpodobnosti, četnosti a očekávané četnosti pro příslušné „řádky tabulky“.

b) Letní čas

$$\begin{aligned}\hat{p}_{\text{Letni}} &\doteq 0,289\,87\\ x &\doteq 8\\ \overline{W}_{0,05} &= \langle 0; 12,591\,59\rangle\end{aligned}$$

Doplněk kritického oboru zůstává stejný jako v *a)*. Pak $x \in \overline{W}_{0,05}$, tedy na hladině významnosti 0,05 **nezamítáme** hypotézu, že v městech, obcích a v okolí studenta je stejné procentuální zastoupení obyvatel, co preferují letní čas.

c) Střídání času

$$\begin{aligned}\hat{p}_{\text{Stridani}} &\doteq 0,177\,95\\ x &\doteq 12,349\,16\\ \overline{W}_{0,05} &= \langle 0; 12,591\,59\rangle\end{aligned}$$

$x \in \overline{W}_{0,05}$, tedy na hladině významnosti 0,05 **nezamítáme** hypotézu, že v městech, obcích a v okolí studenta je stejné procentuální zastoupení obyvatel, co preferují střídání čas. Povšimněme si však, že v tomto případě jsme hypotézu nezamítli jen velmi těsně (p-hodnota je zde 0,054 62).

Části *d) – e)*

V těchto částech je požadován obdobný test jako v těch předchozích, změnou však je, že je zde požadována agregace průzkumů z velkých měst, menších měst a z obcí. Ze zadání není úplně jasné, jestli mají být do jedné ze skupin zařazena i mnou naměřená data, testy jsem tedy provedl bez nich i s nimi, ačkoliv vzhledem k charakteru testu to nemělo šanci výsledek testu ovlivnit.

Data tentokrát do tří tříd rozdělíme sečtením hodnot ve velkých městech, menších městech a v obcích:

	VM	MM	OB	Σ
zimní	834	559	300	1 693
letní	636	363	210	1 209
střídání	435	202	108	745
nemá názor	337	144	57	538
Σ	2 242	1 268	675	4 185

Další postup je prakticky stejný jako v předchozích částech. Testujeme hypotézu:

$$H_0 : p_{\mathsf{VM}} = p_{\mathsf{MM}} = p_{\mathsf{OB}} = \hat{p}_{\mathsf{Zimni}}$$

proti alternativní:

$$H_A : \exists c : p_c \neq \hat{p}_{\mathsf{Zimni}}.$$

Očekávané četnosti jsou uvedeny v prvním listu tabulky, testy jsou provedeny ve třetím listu. Nutno podotknout, že snížením počtu tříd se sníží také stupně volnosti rozdělení χ^2 na pouhý jeden (3 třídy - 1 odhadnutý parametr - 1), což povede k extrémnímu rozšíření kritického oboru.

d) Zimní čas (3 průzkumy)

$$\begin{aligned} \hat{p}_{\mathsf{Zimni}} &\doteq 0,404\,54 \\ x &\doteq 12,661\,95 \\ \overline{W}_{0,05} &= \langle 0; 3,841\,46 \rangle \end{aligned}$$

$x \notin \overline{W}_{0,05}$, tedy na hladině významnosti 0,05 **zamítáme** hypotézu, že u větších měst, menších měst a obcí je stejné procentuální zastoupení obyvatel, co preferují zimní čas.

Při započítání mnou naměřených dat do odpovídající třídy (velkých měst) se testovací kritérium ještě cca o 0,5 zvýšilo, hypotéza byla tedy stále zamítnuta.

e) Nerozhodnutí obyvatelé (3 průzkumy)

$$\begin{aligned} \hat{p}_{\mathsf{NN}} &\doteq 0,128\,55 \\ x &\doteq 20,688\,66 \\ \overline{W}_{0,05} &= \langle 0; 3,841\,46 \rangle \end{aligned}$$

$x \notin \overline{W}_{0,05}$, tedy na hladině významnosti 0,05 **zamítáme** hypotézu, že u větších měst, menších měst a obcí je stejné procentuální zastoupení obyvatel, co preferují zimní čas.

Při započítání mnou naměřených dat do odpovídající třídy (velkých měst) se testovací kritérium cca o 0,2 snížilo, což výsledek nijak neovlivní, hypotéza byla stále zamítnuta.

Část *f)*

V poslední části bylo cílem odhadnout z dat, v jaké ze tří tříd (velká m., menší m., obce) jsem prováděl výzkum, tedy jaké ze tříd mnou získaná data nejvíce odpovídají.

Zvolil jsem k tomu také χ^2 test dobré shody, zde se však bere v úvahu celá naměřená distribuce hodnot vlastního průzkumu. Ta se porovnává s „očekávanou distribucí“, kterou představují hodnoty naměřené v jednotlivých třídách. „Míra shody“ s očekávaným rozložením (z četností tříd) se pak určí podle p-hodnoty z provedeného testu: test s nejvyšší p-hodnotou vykazuje nejvyšší míru shody.

Test je proveden ve čtvrtém listu tabulky. Očekávané četnosti jsou odpovídají počtu respondentů ve třídě *ST* rozděleném v poměru daném původními daty postupně ze tříd *VM*, *MM* a *OB*. Pro každý z těchto sloupců P je pak spočítáno testovací kritérium podobně jako výše, zde však jednotlivé páry četností odpovídají četnostem jednotlivých možností odpovědi:

$$x_{ST\sim P} = \sum_{c\,\in\,\{\text{Zimni},\ldots\}} \frac{(n_c - \hat{n}_c^P)^2}{\hat{n}_c^P},$$

kde n_c je četnost odpovědi c ve třídě *ST*, $\hat{n}_c^P$ je očekávaná četnost odpovědi c podle třídy P spočítaná jako:

$$\hat{n}_c^P = \frac{n_c^P}{\sum_{b\,\in\,\{\text{Zimni},\ldots\}} n_b^P} \cdot \sum_{c\,\in\,\{\text{Zimni},\ldots\}} n_c.$$

Protože jde o pravostranný test, p-hodnota se z testovacího kritéria spočítá jako $1 - F_{\chi^2(k)}(x)$, kde $F_{\chi^2(k)}$ je distribuční funkce χ^2 rozdělení s k stupni volnosti. V tomto případě volíme $k = 3$ stupňů volnosti, protože test vychází ze čtyř tříd (kterými zde jsou, na rozdíl od předchozích částí, jednotlivé odpovědi) a nebyl zde využit bodový odhad žádného parametru.

P-hodnoty vyšly takto:

$$\begin{aligned} x_{ST\sim VM} &\doteq 0,511\,018 \\ x_{ST\sim MM} &\doteq 0,386\,729 \\ x_{ST\sim OB} &\doteq 0,386\,354 \end{aligned}$$

Je tedy možné vyvodit závěr, že výsledky mého šetření mají největší shodu s výsledky z velkých měst. Tento závěr odpovídá realitě – ačkoliv mí respondenti pocházejí z rozličných míst, aktuálně všichni dlouhodobě pobývají ve větších městech ČR (konkrétně 3 v Olomouci, 3 v Praze a 35 v Brně).

Závěrem

Pro zajímavost jsem provedl test dobré shody také nad celou kontingenční tabulkou, tedy pro ověření sdružené hypotézy, zda jsou obecně preference ohledně času nezávislé na místě původu. Test je proveden v posledním listu tabulky. Na hladině významnosti 0,05 přesvědčivě zamítá, preference jsou tedy zřejmě opravdu závislé na místě původu.

Úkol 2: Regresní analýza

Ve druhém úkolu je cílem pomocí regresní analýzy vhodně modelovat závislou proměnnou Z z náhodného vektoru (X, Y, Z). K dispozici máme 70 realizací. O datech nejsou dostupné žádné další informace, příčinnou závislost tedy předpokládáme.

Podle zadání tvoříme lineární regresní model s kvadratickými členy v proměnných:

$$Z = \beta_1 + \beta_2 X + \beta_3 Y + \beta_4 X^2 + \beta_5 Y^2 + \beta_6 XY$$

Součástí úkolu je zjednodušit tento model vynecháním některých členů, které jsou statisticky nevýznamné (na hladině významnosti 0,05 jsou jejich parametry rovny nule).

Úkol jsem realizoval pomocí Pythonu a knihovny `statsmodels` v prostředí Jupyter Notebook, notebook je přiložen v souboru `Ukol2.ipynb`. Dovolím si poznamenat, že zadání požaduje testování hypotéz na hladině významnosti „0,05 %“, což považuji za chybu – všechny testy jsem tedy prováděl na hl. významnosti 0,05 (tedy 5 %), jak je zvykem.

Část *a)*: Určení vhodného modelu

Celkový počet možných (sub)modelů je v tomto případě 32 (každý člen kromě konstanty je možné zahrnout, nebo vynechat, tedy 2^5 variací). Vzhledem k celkem nízkému počtu možných modelů a nízkému rozsahu dat jsem zde zvolil „brute-force“ způsob nalezení vhodného z nich: Vygeneroval jsem všechny možné modely, které jsem seřadil podle hodnoty R^2 a pro každý otestoval rovnost všech parametrů na nulu. Poté jsem vybral model s co největším R^2 a co nejmenším množstvím parametrů, pro které na hladině významnosti 0,05 nebyla zamítnuta hypotéza, že jsou nulové. Tento způsob mi umožnil lépe nahlédnout na rozdíly mezi jednotlivými modely.

Alternativním postupem mohlo být zvolení kompletního modelu (se všemi členy) a postupná eliminace členů s nejvyšší p-hodnotou testu na nulu. Po každé eliminaci by byly zbylé parametry otestovány znovu. Tímto způsobem bychom došli ke stejnému výsledku.

Funkce `try_model` sestaví z dat model se zvolenými členy pomocí knihovny `statsmodels`. Ta přitom automaticky vypočítá pro každý parametr modelu p-hodnotu *t-testu* daného parametru na nulu. Pro každý model je poté vypsán koeficient determinace R^2 a pomocí `zeroish_params` jsou vypsány parametry, pro které je p-hodnota *větší* než zadaná hladina významnosti – pro tyto parametry tedy nezamítáme hypotézu, že jsou nulové (což může znamenat, že jsou opravdu nevýznamné nebo že nemáme dostatek dat).

Níže je uvedeno prvních pět nejlepších modelů. U každého je uveden koef. determinace a seznam (potenciálně) nulových parametrů, v závorce p-hodnota testu pro každý z nich, přičemž nejvyšší hodnota (tj. parametr, který by bylo vhodné odstranit) je označena podtržením.

1. $Z = \beta_1 + \beta_2 X + \beta_3 Y + \beta_4 X^2 + \beta_5 Y^2 + \beta_6 XY$
   - $R^2 = 0{,}607\ 9$
   - β_1 (0,056 6), β_3 (0,453 6), β_4 (0,062 1), β_6 (<u>0,993 8</u>)
2. $Z = \beta_1 + \beta_2 X + \beta_3 Y + \beta_4 X^2 + \beta_5 Y^2$
   - $R^2 = 0{,}607\ 9$
   - β_3 (<u>0,408 5</u>), β_4 (0,060 1)
3. $Z = \beta_1 + \beta_2 X + \beta_4 X^2 + \beta_5 Y^2 + \beta_6 XY$

- $R^2 = 0{,}604\ 42$
- β_1 (0,060 3), β_4 (0,061 2), β_6 (0,737 1)

4. $Z = \beta_1 + \beta_2 X + \beta_4 X^2 + \beta_5 Y^2$
   - $R^2 = 0{,}603\ 72$
   - β_4 (0,059 4)
5. $Z = \beta_1 + \beta_2 X + \beta_3 Y + \beta_5 Y^2 + \beta_6 XY$
   - $R^2 = 0{,}585\ 81$
   - β_1 (0,217), β_2 (0,055 5), β_3 (0,462 3), β_6 (0,994)

Není překvapivé, že model s nejvyšším množstvím členů má nejvyšší koeficient determinace – ten má tendenci růst s každým přidaným členem (proměnnou), to však neznamená, že všechny členy přinášejí významné množství nové informace. Nahlédněme, že čtvrtý uvedený model má pouze jeden parametr, pro který jsme nezamítli hypotézu o nulovosti, a to navíc velmi těsně (např. na hladině významnosti 0,06 bychom ji zamítli). Jeho R^2 je přitom pouze nepatrně nižší než koeficienty předcházejících tří modelů, které mají větší množství nulových proměnných.

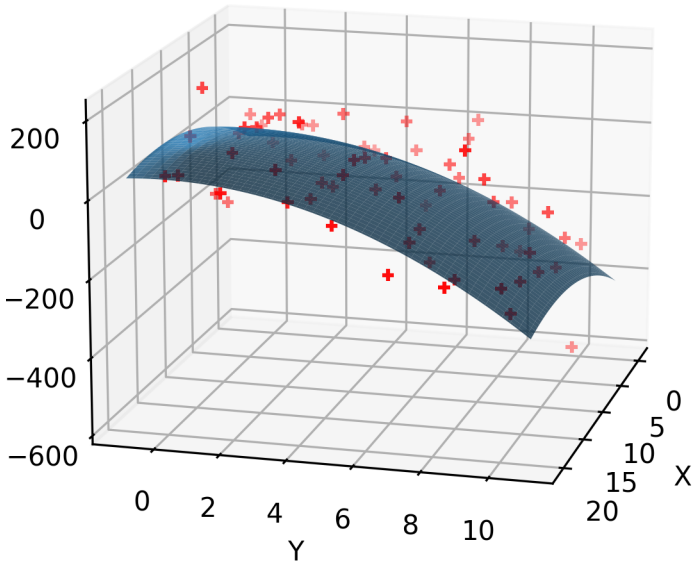
Za vhodný model jsem tedy prohlásil 4. model: $Z = \beta_1 + \beta_2 X + \beta_4 X^2 + \beta_5 Y^2$. Tento model tedy vysvětluje cca 60 % variability v datech. Můžeme také nahlédnout, že postupným odstraňováním „nejméně důležitého“ členu z úplného modelu také po dvou krocích dojdeme k tomuto modelu, další odstraňování už posune R^2 pod hodnotu 0,6.

Růst koeficientu determinace se zvyšujícím se množstvím proměnných se snaží kompenzovat různé „adjustované koeficienty determinace“. Knihovna automaticky kromě R^2 počítá také adjustovaný

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1},$$

kde n je počet měření a p je počet parametrů modelu. Ověřil jsem, že zvolený model má nejvyšší hodnotu tohoto koeficientu, což znovu potvrzuje, že jde o nejvhodnější submodel.

Výsledný polynomiální model je naznačen na grafu na obrázku 1.



Obrázek 1: Graf zvoleného regresního modelu a zdrojových dat.

Diagnostika modelu

Knihovna `statsmodels` pro model vyhodnotí také několik dalších parametrů, které mohou posloužit k ověření vhodnosti modelu.

- Durbin-Watsonův test: ověřuje autokorelaci prvního řádu (závislost hodnoty náhodné složky na hodnotách předcházejících). Hodnota kritéria d je zde 2,365. Můžeme tedy prohlásit, že $d \approx 2$, nezamítáme tak nezávislost.
- F-test vhodnosti regresního modelu: ověřuje, zda se model významně liší od aritmetického průměru Z, tedy jestli není stejně dobře možné vysvětlit závisle proměnnou pouze pomocí průměru. Hypotéza je $H_0 : (\beta_1, \beta_2, \beta_4, \beta_5) = (\overline{z}, 0, 0, 0)$. P-hodnota F-testu v tomto případě je $2{,}79 \cdot 10^{-13}$, čili můžeme na hladině významnosti 0,05 spolehlivě hypotézu zamítnout.

Dále je vhodné ověřit normalitu reziduí. K tomu jsem vyzkoušel použít dva testy, které poskytuje knihovna `scipy`: Kolmogorov-Smirnov test (`scipy.stats.kstest`) a „omnibus“ testu D'Agostina a Pearsona, který pro otestování normality vychází z šikmosti a špičatosti[1]. V obou případech je nulovou hypotézou, že data pochází z normálního rozdělení, alternativou je, že z něj nepochází. Pro první test vyšla p-hodnota 0,682, pro druhý 0,236, ani v jednom případě tedy na hl. spol. 0,95 nulovou hypotézu nezamítáme a můžeme předpokládat, že rezidua odpovídají normálnímu rozložení.

Část *b)*: Regresní parametry

Odhadnuté bodové parametry a 95% intervaly spolehlivosti jsou uvedeny v následující tabulce:

Parametr	Bodový odhad	95% int. spol.
β_1	-66,805 82	$\langle$-123,259; -10,353$\rangle$
β_2	17,187 378	$\langle$5,061; 29,314$\rangle$
β_4	-0,560 746	$\langle$-1,144; 0,023$\rangle$
β_5	-2,766 557	$\langle$-3,37; -2,163$\rangle$

Všechny hodnoty vypočítala metodou nejmenších čtverců knihovna `statsmodels` a je možné je zobrazit např. zavoláním funkce `summary()` nad objektem výsledků regrese.

Část *c)*: Odhad rozptylu

Bodový odhad rozptylu závisle proměnné Z odpovídá sumě reziduí vydělené $n - m$, kde n je počet realizací, m je počet odhadnutých parametrů. Knihovna tuto hodnotu uloží do atributu objektu výsledků regrese `mse_resid`, což jsem ověřil také ručním výpočtem z hodnot ve funkci `calc_variance_est()`.

Bodový odhad je **7 705,694**.

Části *d)* – *e)*: Testy nad parametry

V části *d)* je úkolem otestovat, zda jsou současně dva regresní parametry nulové. Nahlédnutí na 95% intervaly spolehlivosti pro jednotlivé parametry naznačuje, že takové parametry nebudou (nula leží pouze v int. spol. parametru β_4), otestovat to však můžeme. „Blízko“ k nule se pohybuje β_5, otestujeme tedy současnou nulovost a shodnost pro tyto dva parametry.

[1] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.normaltest.html

Testujeme sdruženou hypotézu:

$$H_0 : \beta_4 = 0 \wedge \beta_5 = 0$$

proti alternativní:

$$H_A : \beta_4 \neq 0 \vee \beta_5 \neq 0.$$

K otestování sdružené hypotézy je vhodný F-test, který nad modelem provede knihovna `statsmodels` (viz funkce `do_f_test` na konci notebooku). P-hodnota provedeného testu je $8{,}28 \cdot 10^{-13}$, nulovou hypotézu tedy zjevně na hl. významnosti 0,05 zamítáme, tedy alespoň jeden z koeficientů není nulový.

V části *e)* pak testujeme hypotézu:

$$H_0 : \beta_4 = \beta_5$$

proti alternativní:

$$H_A : \beta_4 \neq \beta_5.$$

Zde můžeme použít buď opět F-test, nebo t-test, v obou z nich však vyjde p-hodnota prakticky stejně: $1{,}78 \cdot 10^{-6}$. Nulovou hypotézu tedy opět zamítáme, tedy koeficienty β_4 a β_5 se na hl. významnosti 0,05 nerovnají.