EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY

Analýza dat v programu SPSS pro začátečníky a mírně pokročilé I.

Patrik Vidlář

2019

Strategický projekt UTB ve Zlíně, reg. č. CZ.02.2.69/0.0/0.0/16_015/0002204

Informace o autorech:

Patrik Vidlář, Bc.

ACREA CR, spol. s r.o.

pvidlar@acrea.cz

OBSAH

ÚVOD 4
1 MATICE DAT 5
1.1 TYPY PROMĚNNÝCH (SLOUPCE V DATOVÉ MATICI) 5
1.2 POPIS PROMĚNNÝCH 5
2 ROZLOŽENÍ ČETNOSTÍ 7
2.1 TABULKY ČETNOSTÍ 9
2.1.1 Hodnoty rozložení - typy četností 9
2.1.2 Grafy 10
2.1.3 Příklady četnostních tabulek 11
3 BOXPLOT 12
3.1 KVANTILOVÉ MÍRY 12
4 TESTY HYPOTÉZ 14
4.1 POSTUP TESTOVÁNÍ HYPOTÉZ 16
4.2 NEPARAMETRICKÉ TESTY 16
4.2.1 Test dobré shody 16
4.2.2 Test pořadí 20
4.2.3 Kolmogorovův-Smirnovův test 21
SHRNUTÍ 23
SEZNAM POUŽITÉ LITERATURY 24
SEZNAM OBRÁZKŮ 25
SEZNAM TABULEK 26

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

ÚVOD

Text se zabývá vybranými tématy, která se týkají základů statistiky. Je pojat více jako statistický text, než text o softwaru SPSS a jednotlivých procedurách. Byl vytvořen tak, aby podpořil kurz jako celek. Není bezpodmínečně nutné, aby vše v tomto materiálu bylo zároveň probráno – danou látku lze vnímat jako rozšíření znalostí získaných na kurzu.

Čtenář se seznámí se základní maticí dat a typy proměnných. Následují tabulky četností včetně ukázek a grafů, které lze z tabulek četností tvořit. Blíže je popsán boxplot, neboli krabicový graf. Poslední část textu se věnuje obecnému popisu statistického testování hypotéz. Následují tři příklady neparametrických testů, a to test dobré shody, test pořadí a Kolmogorův-Smirnovův test. Poslední dva testy jsou již popsány bez statistických vzorců. Postup zadání jednotlivých testů bude probrán v rámci workshopu.

Veškeré uvedené informace spadají mezi obecné znalosti a dovednosti v oblasti kvantitativní analýzy dat (jsou pojednány v mnoha učebních textech jako je Hebák et al, 2007; Hron & Kunderová, 2013; Řehák & Brom, 2015), ale zároveň reflektují individuální přístup lektora k dané problematice a jsou podloženy příklady na cvičných datových souborech.

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY

1 MATICE DAT

Tabulka údajů pro statistické jednotky umístěné v řádcích tabulky a charakterizované sloupci tabulky se nazývá matice dat.

Data sloupce matice tvoří tzv. statistickou řadu. Jsou-li hodnoty číselné statistické řady uspořádány podle velikosti, tvoří uspořádanou statistickou řadu.

Datový soubor v počítači má své jméno, kterým je identifikován.

řádek = jednotka, objekt, vzorek, výrobek, případ

sloupec = proměnná, záznam informace o jedné vlastnosti jednotek

1.1 Typy proměnných (sloupce v datové matici)

a) číselné - spojité, počty, poměrové indexy
b) kategorizované - nominální, dichotomické, ordinální, kardinální
c) textové
d) datum a čas

1.2 Popis proměnných

a) název sloupce - pro práci programu a pro identifikaci
b) popis sloupce - charakteristika proměnné (sloupce)
c) popis kódů, resp. popis čísel (lze jimi zaměnit kódy v matici)
d) chybějící hodnoty -„missing values“ - určení kódů, které se vynechávají z výpočtů
e) formát zápisu (počet desetinných míst, text, apod.)

Termínům matice dat a datový soubor používaných v oblasti počítačového zpracování odpovídá v statistické teorii termín výběrový soubor či výběr.

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

Příklady:

- země - textová proměnná
- region - nominální kategorizovaná proměnná/ 1= západní a severní Evropa, 2 = jižní Evropa, 3 = střední Evropa, 4 = SNS a Balkán/
- vzdělání - ordinální kategorizovaná proměnná/ 1= základní a nedokončené základní, 2 = střední bez maturity, 3 = maturita, 4= vysokoškolské/

2 ROZLOŽENÍ ČETNOSTÍ

Vlastnosti rozložení dat v kategoriích (souboru četností) se hodnotí v závislosti na typu znaku (obdobně jako u číselných dat):

- poloha četností

Kde se soustřeďují jednotky? Ve které kategorii (ích)? Na které části škály?

- rozptýlení v kategoriích a podél škály

Jak se jednotky soustřeďují do jedné kategorie? Jak se polarizují na ordinální

škále? Je rozložení rovnoměrné v kategoriích nebo se soustřeďuje do (kolem) jedné kategorie.

- symetrie rozložení na preferenční nebo znaménkové škále

Převažují preference jedné strany škály proti druhé? Převažují kladné hodnoty proti záporným na znaménkové škále? Které obsahově protipolné kategorie porušují vyváženost rozložení?

Charakteristiky se liší podle typu proměnné:

a) míry polohy: modus, mediánová kategorie, ordinální medián, průměr

b) míry variability: nomvar (Giniho míra), dorvar, rozptyl

c) míry symetrie: koeficient asymetrie, šikmost

POLOHA:

modus = nejčetnější kategorie

mediánová kategorie = kategorie, v níž kumulativní četnost dosáhne 50%

ordinální medián = $\widetilde{X} = Me - .5 + \dfrac{.5 - F_{Me-1}}{f_{Me}}$

průměr = $\overline{X} = (1/N)\sum X_i$

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

VARIABILITA:

Variabilitu měříme těmito mírami:

<u>a) u nominálních proměnných</u>

koeficient variability v = $1 - f_{mod}$ (f_{mod} – relativní modální četnost)

nomvar = $1 - \Sigma_{i=1 \ldots K} \; fi^2$ (f_i – i-tá relativní četnost; K – počet kategorií)

normovaný nomvar = K * nomvar / (K – 1)

<u>b) u ordinálních proměnných</u>

dorvar = $2 * \Sigma_{i=1 \ldots K} \; F_i (1 - F_i)$ (F_i – i-tá kumulativní relativní četnost)

normovaný dorvar = 2 * dorvar / (K – 1)

<u>c) u kardinálních proměnných</u>

rozptyl var X = $s^2 = \Sigma_{i=1 \ldots N} (x_i - X)^2 / (N - 1)$ (N – počet měření; X – průměr x_i)

směrodatná odchylka s = $(\text{var } X)^{1/2}$

Čím je příslušná míra variability větší, tím více variability daná proměnná vykazuje

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání
MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY

2.1 Tabulky četností

U kategorizovaných dat je první informaci souboru rozložení případů (jednotek) v kategoriích - absolutní a relativní (resp. procentní). Toto rozložení získáváme ve formě tabulek a grafů. Tabulky i grafy mohou mít různé tvary.

Kategorizovaná data - kategorie tvoří úplný disjunktní systém: vzájemně se vylučují (disjunktnost) a jejich sjednocení pokrývá všechny možnosti (úplnost).

Analýza kategorizovaných dat závisí na **typu kategorizované proměnné**:

a) nominální typ - kategorie vyjadřují různé kvality
b) ordinální typ - kategorie vyjadřují uspořádané kvality
c) kardinální typ - kategorie vyjadřují kvantifikované kvality

Speciálním případem kategorizované proměnné je **dichotomie**, která má jen dvě hodnoty (např. ANO/NE, MÁ/NEMÁ VLASTNOST). Dichotomie lze tabelovat úsporně a lze vycházet i z toho, že při kódování 1=ANO, 0=NE je průměrný skór souboru roven relativní četnosti kategorie ANO, při kódování 100=ANO, 0=NE je průměr procentem. Proto lze použít pro tabelace těchto dat i postupy připravené pro číselné proměnné.

2.1.1 Hodnoty rozložení - typy četností

a) absolutní četnosti - počty jednotek v kategoriích
b) relativní četnosti - podíl kategorie na celém souboru
c) relativní četnosti z validních dat - podíl kategorie na souboru validních dat (tj. po redukci těch jednotek, jejichž údaj chybí, je chybně zapsán, respondent odmítl odpovědět, nebo i neuměl odpovědět, tedy po redukci o údaje, které deklarujeme jako „vynechávané“ (missing values)
d) procenta z celého souboru i z validních dat

e) kumulativní absolutní i relativní četnosti z validních dat - aplikujeme jen pro ordinální a kardinální data, tj. pro proměnné, které mají uspořádané kategorie nebo kategoriím jsou přiřazeny číselné skóry.

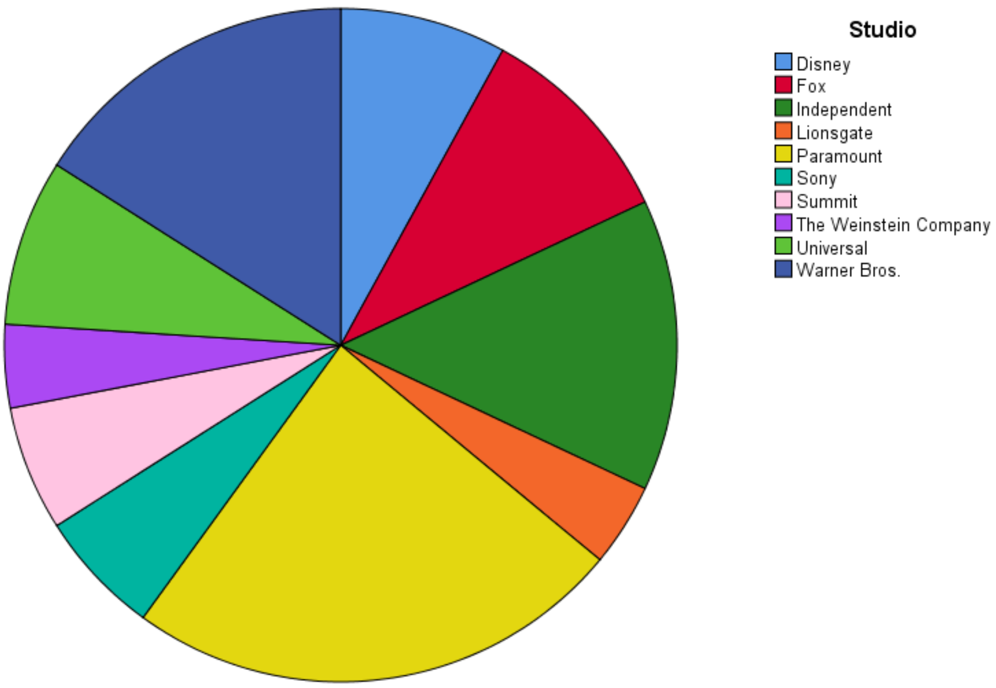
2.1.2 Grafy

a) histogram - vhodný pro číselné proměnné, jejichž hodnoty byly vytříděny do intervalů a kategorie číselně označeny; i pro kumulativní četnosti

b) sloupkový graf - bar chart vhodný pro jakýkoliv typ proměnné; i pro kumulativní četnosti

c) kruhový graf, koláčový graf - pie chart - vhodný jen pro nominální proměnné

Tyto údaje a grafy zobrazují tvar koncentraci dat do jednotlivých kategorií, posunutí dat na škále, tvar rozptýlení dat, rozložení jako celek a úplnou informaci o něm.



Obrázek 1 Koláčový graf

2.1.3 Příklady četnostních tabulek

Rozložení filmového žánru:

Žánr

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Akční	16	32,0	32,0	32,0
	Animovaný	6	12,0	12,0	44,0
	Dobrodružný	3	6,0	6,0	50,0
	Drama	7	14,0	14,0	64,0
	Horor	3	6,0	6,0	70,0
	Komedie	12	24,0	24,0	94,0
	Romantický	1	2,0	2,0	96,0
	Thriller	2	4,0	4,0	100,0
	Total	50	100,0	100,0	

Tabulka 1: Četnostní tabulka pro nominální proměnnou

Rozložení let, kdy film vyšel:

Rok, kdy film vyšel

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2007	6	12,0	12,0	12,0
	2008	12	24,0	24,0	36,0
	2009	12	24,0	24,0	60,0
	2010	14	28,0	28,0	88,0
	2011	6	12,0	12,0	100,0
	Total	50	100,0	100,0	

Tabulka 2: Četnostní tabulka pro ordinální proměnnou

3 BOXPLOT

Boxplot neboli krabicový graf je zobrazení dat, které využívá kvantilové míry polohy a charakteristiky rozptýlení.

3.1 Kvantilové míry

medián – míra polohy, hodnota ve středu uspořádané řady, která rozděluje řadu na dva stejně dlouhé úseky (= půlky)

- pro liché N … hodnota prostředního pozorování
- pro sudé N … průměr dvou středních pozorování

rozpětí R = max - min – míra rozptýlení

kvartilové rozpětí RK = KH - KD – míra rozptýlení

kvartilová odchylka dK = (KH - KD)/2 – míra rozptýlení

zóna Z = 1.5 x RK

vnitřní hradby = KD – Z, KH + Z

vnější hradby = KD – 2Z, KH + 2Z

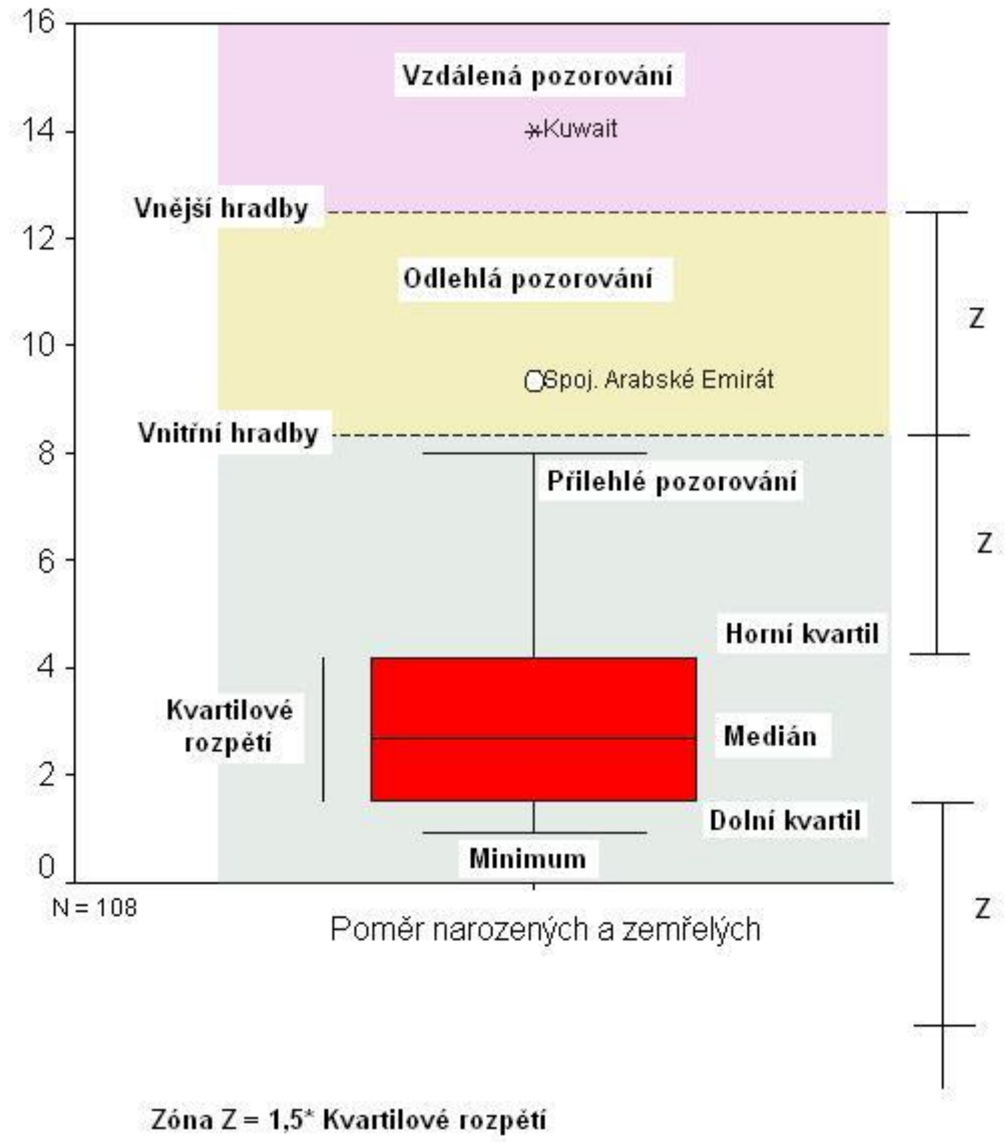
přilehlá pozorování: poslední body, které ještě leží uvnitř vnitřních hradeb (min a max té části řady, která leží mezi vnitřními hradbami

vnější pozorování (outliers): body, které jsou vně vnitřních hradeb

vzdálená pozorování (extremes): body které jsou vně vnějších hradeb

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



Obrázek 2 Boxplot

Graf vyjadřuje rozložení dat na svislé ose. Obdélník je shora a zdola ohraničen kvartily, uprostřed obdélníku je značka mediánu. Úsečky jdoucí od kvartilových hodnot končí u přilehlých pozorování. Body (a popisem) jsou označeny vnější (kroužky) a vzdálená (hvězdička) pozorování.

4 TESTY HYPOTÉZ

Testování statistických hypotéz je rozhodovací problém, v němž proti sobě stavíme dva výroky – dvě hypotézy: H_0 (**nulovou hypotézu**) a H_A (**alternativní hypotézu**). Neyman-Pearsonův princip testování je založen na ověřování modelu H_0 proti modelu H_A.

Výsledkem může být jedno ze dvou rozhodnutí:

- není důvod zamítnout H_0,
- data nulové hypotéze odporují, H_0 tedy neplatí, přijímáme H_A.

O tom, zda data nulové hypotéze odpovídají, či zda indikují HA, vypovídá vždy vhodně zvolená statistická funkce dat (**testová statistika**), která charakterizuje stupeň vzdálenosti dat od H_0 směrem k H_A, a tím stupeň platnosti H_0.

Test je statistické rozhodovací pravidlo, které stanoví, zda testová statistika nabývá takové hodnoty, aby nulová hypotéza, ze které vycházíme, byla odmítnuta.

Při testování hypotéz poznatky zjištěné na konkrétním výběrovém souboru zobecňujeme na základní – hypotetický soubor (někdy se označuje jako **populace**).

Možné chyby v procesu

Chyby se můžeme dopustit již při samotné formulaci statistické hypotézy (nulová a/nebo alternativní hypotéza neodpovídají řešenému problému).

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

Při samotném rozhodování se lze dopustit těchto chyb:

1. statistické chyby rozhodování:

		výsledek rozhodnutí	
		H_0	**H_A**
skutečnost	**H_0**	OK	chyba 1. druhu (α)
	H_A	chyba 2. druhu (β)	OK

Tabulka 3: Chyby při testování hypotéz

2. nesprávně zvolená testová statistika

3. nesprávně určené rozhodovací pravidlo

Pravděpodobnost chyby 1. druhu α je pravděpodobnost, že se rozhodneme pro H_A a ve skutečnosti platí H_0. Je menší nebo rovna předem dané hodnotě α (v praxi většinou volená jako 0,05 nebo 0,01). Mluvíme také o riziku zamítnutí nulové hypotézy, když tato platí.

Pravděpodobnost chyby 2. druhu β je pravděpodobnost, že se rozhodneme pro H_0 a ve skutečnosti platí H_A. Pravděpodobnost této chyby značíme β. Její doplněk do jedné se nazývá síla testu.

Rozhodovací pravidlo určujeme tak, abychom nepřekročili zvolené riziko neoprávněného zamítnutí nulové hypotézy a zároveň pokud možno minimalizovali její chybné přijetí. (Není možno minimalizovat obě rizika zároveň.)

Vzhledem k této asymetrii je třeba zformulovat nulovou hypotézu tak, abychom se jejím zamítnutím dostali k tomu, co chceme ukázat.

Při zvolené hodnotě α říkáme, že testujeme hypotézu na **hladině významnosti α** nebo na **hladině spolehlivosti** (1-α) x 100 (%).

4.1 Postup testování hypotéz

- zformulují se hypotézy H_0 a H^A
- zvolí se hladina významnosti α
- vybere se vhodný test a příslušná testová statistika – rozhodovací funkce dat
- do testové statistiky se dosadí hodnoty z dat
- provede se vlastní test hypotézy
- *a) manuální postup:* hodnota statistiky se porovná s kritickou hodnotou zjištěnou v tabulce příslušné danému testu (základní statistické tabulky jsou přílohou většiny učebnic, pro speciální testy je třeba použít samostatné publikace statistických tabulek). Při překročení kritické hodnoty se zamítne nulová hypotéza ve prospěch alternativní; nepřekročí-li testová statistika kritickou, je možno se domnívat, že odchylka od nulové hypotézy byla způsobena náhodnými vlivy a chybami
- *b) počítač:* zjistí se tzv. dosažená hladina významnosti, která znamená vypočtené empirické riziko odmítnutí nulové hypotézy za předpokladu, že H0 platí (je to odhad pravděpodobnosti chyby prvního druhu); je-li toto riziko menší než předem zvolená hranice α, rozhodujeme se pro alternativní hypotézu, je-li riziko větší než pro nás přijatelná hranice, nezamítáme nulovou hypotézu. (Na výstupech z počítače se označuje většinou jako P, tail probability nebo Sig = significance.)

4.2 Neparametrické testy

4.2.1 Test dobré shody

Úlohy:

a) Komparace rozložení s hypotetickým resp. normativním, stabilizovaným stavem.

b) Jsou výzkumná data reprezentativní?

c) Pokrývá trh výrobku proporcionálně jednotlivé sociální a demograficky definované skupiny?

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

Tyto úlohy mají společný rys: existuje buď objektivní, normativní nebo hypotetické rozložení četností, ke kterému se poměřuje rozložení dat. Jsou to tři typy úloh:

- **Kontrola reprezentativity výběrového šetření** - porovnáváme rozložení kategorizovaných proměnných s dostupnými statistickými daty, například věk, příjem, povolání ap. (ovšem jen takové proměnné, které se neúčastní výběru jako stratifikační, kvótní, řízené).
- **Odchylky od standardu a normy** - některé vlastnosti jsou stabilizované a po dlouhou dobu se opakují, jejich rozložení se stává standardem a vychází se z něj např. při plánování. To může být stabilní podíl značek na trhu, stabilně distribuovaný zájem o typy výrobků, stabilní poptávka po předmětech dlouhodobé spotřeby apod. Četnosti (i výběrové, při dostatečném opakování) se používají jako populační parametr. Testujeme v novém výzkumu, zda nedošlo ke změně.
- **Ověření modelu chování nebo vzniku dat** - naše představa o chování populace může být formulována jako váhy zastoupení v kategoriích. Testujeme platnost naší představy/hypotézy/modelu a odchylky od ní.

Úloha dobré shody:

Statistické hypotézy:

H_0: $p_k = \pi_k$ *pro všechna k;* H_A: $p_k \neq \pi_k$ alespoň pro jedno k

p_k jsou četnosti, které reprezentuje náš výběr

π_k jsou četnosti (proporce) které předpokládá hypotéza (stav, standard, model)

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

Testová statistika:

$$X^2 = \sum_k \frac{(n_k - n\pi_k)^2}{n\pi_k} \qquad df = K - 1$$

$$X^2 = \sum_k \frac{n_k^{\ 2}}{n\pi_k} - n \qquad df = K - 1$$

Podmínky pro aplikaci testu: n>=30

Hodnoty $n\pi_k$ se nazývají očekávané četnosti.

Pro rozhodnutí proti nulové hypotéze shody se použije tabulka kritických hodnot:

Tabulka kritických hodnot testu chí-kvadrát		
df = K - 1	**alfa = 0.05**	**alfa = 0.01**
1	3.84	6.63
2	5.99	9.21
3	7.81	11.34
4	9.49	13.28
5	11.07	15.09
6	12.59	16.81
7	14.07	18.48
8	15.51	20.09
9	16.92	21.67
10	18.31	23.21

Tabulka 4: Tabulka kritických hodnot testu chí-kvadrát

Po odmítnutí hypotézy si klademe další otázky, neboť výsledkem předchozího kroku je prosté negování shody, tedy závěrem je neshoda. Tu ale chceme specifikovat: ve které kategorii nastal významný rozdíl?

Úlohy:

a) Je některá skupina typická v zájmu o daný výrobek? Ve kterých skupinách není výrobek přijímán? Je značka přijímána rovnoměrně v populaci nebo se její akceptace diferencuje?

b) Jaké je uspořádání kategorií podle typičnosti, tj. četnostního nadhodnocení/podhodnocení oproti očekávání?

Úlohy se řeší testem pro odchylky v jednotlivých kategoriích.

Statistické hypotézy:

H_{0k}: $p_k = \pi_k$; H_{Ak}: $p_k \neq \pi_k$ postupně pro k = 1, ... K.

Každá z těchto hypotéz se testuje pomocí z-testu:

$$z_k = \frac{n_k - n\pi_k}{\sqrt{n\pi_k(1-\pi_k)}}$$

Podmínky pro použití: n je alespoň 30.

Očekávaná četnost $n\pi_k$ je alespoň 5.

Kritické hodnoty pro alfa = 0.05,0.01,0.001 jsou postupně: 1.96, 2.58, 3.29

Shodu zamítáme, je-li $|z_k|$ >= zvolená kritická hodnota

4.2.2 Test pořadí

Test pořadí je známý také jako Waldův-Wolfowitzův test. Testuje, zda je pořadí výskytu hodnot dichotomické proměnné náhodné. V případě, že proměnná není dichotomická, lze testovat náhodné střídání hodnot větších nebo menších než daná hodnota (medián, průměr apod.).

Příklady použití:

- *Výzkumný projekt mapuje chování dětí předškolního věku v určitých herních situacích. Šetření se účastní celkem 20 dětí z jedné mateřské školy. Protože se však psycholog může každý den věnovat maximálně dvěma dětem, existuje podezření, zda nejsou výsledky zkresleny tím, že si děti mezi sebou některé informace sdělí. Je tedy třeba prověřit, zda hodnocení nezávisí na pořadí, v jakém se děti výzkumu účastnily.*
- *Marketingový výzkum je zaměřen na hodnocení vzhledu nových webových stránek určité firmy. Všichni respondenti zahajují práci současně a odpověď každého z nich je zaznamenána okamžitě, jakmile dokončí prohlížení. Ukazuje se však, že názory se mohou lišit podle toho, kolik času respondenti věnovali prohlížení stránek. Tuto hypotézu je nutné ověřit.*

Princip testu je založen na uspořádání případů. Testová statistika vyjadřuje počet úseků se stejnou hodnotou (například pro kódy 0 a 1 počet posloupností typu 0,0,0 nebo 1,1,1,1). Příliš mnoho nebo málo takových úseků svědčí proti testované hypotéze. Testová statistika má asymptoticky normální rozložení.

Metoda nevychází ze žádných předpokladů o rozdělení dat, proto ji řadíme mezi tzv. neparametrické testy.

Je-li datový soubor vážený (Data, Weight Cases), dovoluje procedura využívat pouze celočíselné váhy, v ostatních případech jsou váhy zaokrouhleny na nejbližší přirozené číslo. Při realizaci výpočtu je potom každý případ zkopírován tolikrát, kolik udává příslušná hodnota váhy.

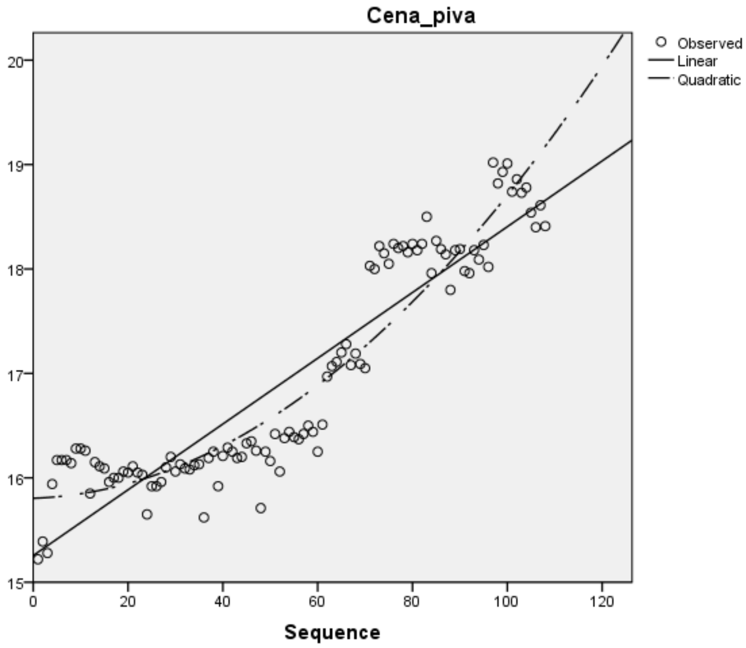
EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

Součástí výstupu může být rovněž přehled základních popisných statistik vstupních proměnných (průměr, směrodatná odchylka, minimum, maximum, počet platných hodnot a/nebo kvartily).

Volání procedury v IBM SPSS Statistics:

Analyze → Nonparametric Tests → Runs



Obrázek 3 Výstup testu pořadí

4.2.3 Kolmogorovův-Smirnovův test

Jednovýběrový Kolmogorov-Smirnovův test je určen k ověření shody mezi rozložením dané číselné proměnné a určitým teoretickým rozdělením. Nulová hypotéza je formulována tak, že data pocházejí z daného rozdělení (normální, rovnoměrné, exponenciální nebo Poissonovo rozdělení). Metoda nevychází ze žádných předpokladů o rozložení dat, proto ji řadíme mezi tzv. neparametrické testy.

Proceduru využijeme například k ověření, zda data pocházejí z normálního rozdělení, což je předpokladem celé řady statistických metod – například T-testů nebo analýzy rozptylu.

Test je založen na porovnání empirické distribuční funkce odhadnuté z dat a teoretické distribuční funkce. Testová statistika Z vychází z maximální odchylky těchto funkcí.

Jednovýběrový Kolmogorov-Smirnovův test předpokládá, že parametry teoretického rozdělení jsou předem dány. Tato procedura však vychází ze situace, kdy je třeba parametry odhadnout z dat a považuje získané odhady za známé parametry. Tímto způsobem však může dojít k výraznému snížení síly testu. Kolmogorov-Smirnovův test pro normální rozdělení je v IBM SPSS Statistics rovněž dostupný v proceduře Explore, kde je jeho součástí také Lillieforsova korekce zohledňující skutečnost, že hodnoty parametrů jsou odhadovány z dat.

Je-li datový soubor vážený (Data, Weight Cases), dovoluje procedura využívat pouze celočíselné váhy, v ostatních případech jsou váhy zaokrouhleny na nejbližší přirozené číslo. Při realizaci výpočtu je potom každý případ zkopírován tolikrát, kolik udává příslušná hodnota váhy.

Součástí výstupu může být rovněž přehled základních popisných statistik vstupních proměnných (průměr, směrodatná odchylka, minimum, maximum, počet platných případů, medián a kvartily).

Volání procedury v IBM SPSS Statistics

Analyze → Nonparametric Tests → 1-Sample K-S

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

SHRNUTÍ

Text se zabývá vybranými tématy, která se týkají základů statistiky. Není bezpodmínečně nutné, aby vše v tomto materiálu bylo zároveň probráno – danou látku lze vnímat jako rozšíření znalostí získaných na kurzu.

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

SEZNAM POUŽITÉ LITERATURY

Hebák P., Hustopecký J., Jarošová E. & I. Pecáková. (2007). Vícerozměrné statistické metody (1). Praha: Informatorium.

Hron, K., & Kunderová, P. (2015). Základy počtu pravděpodobnosti a metod matematické statistiky (2. dopl. vydání). Olomouc: Univerzita Palackého v Olomouci.

Řehák, J., & Brom, O. (2015). SPSS - Praktická analýza dat. Brno: Computer Press.

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

SEZNAM OBRÁZKŮ

Obrázek 1 Koláčový graf 10
Obrázek 2 Boxplot 13
Obrázek 3 Výstup testu pořadí 21

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

SEZNAM TABULEK

Tabulka 1: Četnostní tabulka pro nominální proměnnou 11
Tabulka 2: Četnostní tabulka pro ordinální proměnnou 11
Tabulka 3: Chyby při testování hypotéz 15
Tabulka 4: Tabulka kritických hodnot testu chí-kvadrát 18

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY