ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE

PROVOZNĚ EKONOMICKÁ FAKULTA

ZÁKLADY DATOVÉ ANALÝZY

Ing. Andrea Jindrová, Ph.D.

Provozně ekonomická fakulta

Česká zemědělská univerzita v Praze

Česká zemědělská univerzita v Praze
Provozně ekonomická fakulta
Katedra statistiky
Praha, 2026

Určeno pro posluchače oboru PAA

Tato publikace neprošla redakční ani jazykovou úpravou.

Lektoroval: Mgr. Roman Biskup, Ph.D.

ISBN 978-80-213-3530-1

PŘEDMLUVA

Vážení studenti,
předkládané studijní materiály jsou určeny právě Vám, posluchačům předmětu Statistika I bakalářského oboru Podnikání a administrativa na Provozně ekonomické fakultě ČZU v Praze. Cílem těchto skript je poskytnout Vám ucelený a přehledný souhrn teorie pravděpodobnosti a matematické statistiky, které tvoří základ pro jakoukoliv efektivní analýzu dat.

Celý text je koncipován tak, aby Vás podpořil v procesu studia – pomohl Vám lépe pochopit a prohloubit znalosti získané na přednáškách a seminářích a sloužil jako efektivní pomůcka pro přípravu na zkoušku. Je však důležité si uvědomit, že při své přípravě musíte věnovat pozornost nejen zde uvedeným tématům, ale současně též intenzivně pracovat formou samostudia z dalších doporučených zdrojů. Právě tímto komplexním přístupem budete rozvíjet své analytické schopnosti při práci s daty.

Materiály jsou přehledně rozděleny do tematických okruhů. V každém z nich se nejprve seznámíte se základní teorií a následně si osvojené principy prohloubíte prostřednictvím řešených příkladů. Skripta navíc obsahují i úvod do práce se statistickým softwarem IBM SPSS Statistics.

Na konci každého celku na Vás čekají příklady pro samostatnou práci, které Vám umožní ověřit a upevnit si nabyté dovednosti. Datové matice k řešeným i neřešeným příkladům jsou dostupné na platformě Moodle v rámci kurzu, případně je lze získat e-mailem na adrese jindrova@pef.czu.cz.

Přeji Vám mnoho studijních úspěchů a radosti z objevování světa statistiky.

Andrea Jindrová

OBSAH

OBSAH 4
1 Úvod 7
1.1 Základní pojmy 8
2 Náhodný jev a pravděpodobnost 21
2.1 Náhodný jev 21
2.2 Základní operace a vztahy mezi náhodnými jevy 22
2.3 Pravděpodobnost náhodného jevu 26
2.3.1 Axiomatická teorie pravděpodobnosti (teoretická) 26
2.3.2 Statistická definice pravděpodobnosti 30
2.3.3 Klasická definice pravděpodobnosti 32
3 Náhodné veličiny 38
3.1 Distribuční funkce 40
3.1.1 Funkce rozdělení diskrétní náhodné veličiny 41
3.1.2 Funkce rozdělení spojité náhodné veličiny 44
3.2 Kvantilová funkce 45
3.3 Číselné charakteristiky náhodných veličin 46
3.3.1 Momentové charakteristiky náhodných veličin 47
Obecný moment 47
Normovaný moment 48
3.3.2 Normovaná náhodná veličina 51
3.3.3 Kvantilové charakteristiky náhodných veličin 51
4 Pravděpodobnostní rozdělení 55
4.1 Pravděpodobnostní rozdělení diskrétních náhodných veličin 55
4.1.1 Alternativní rozdělení 56
4.1.2 Geometrické rozdělení 56
4.1.3 Binomické rozdělení 58
4.1.4 Hypergeometrické rozdělení 60
4.1.5 Poissonovo rozdělení 61
4.2 Pravděpodobnostní rozdělení spojitých náhodných veličin 63
4.2.1 Normální (Gausovo) rozdělení 64
4.2.2 Normované normální rozdělení 65
4.2.3 $Chí$-kvadrát rozdělení 68
4.2.4 Studentovo t-rozdělení 69
4.2.5 Fisher-Snedecorovo F-rozdělení 70
5 zpracování datového Souboru 76
5.1 Třídění, tabelování a grafické znázornění proměnných 77
5.1.1 Jednorozměrné rozdělení četností 77
5.1.2 Intervalové jednorozměrné rozdělení četností 79
5.1.3 Grafická analýza 80
5.2 Číselné charakteristiky 82
5.2.1 Číselné charakteristiky nominálních proměnných 83
5.2.2 Číselné charakteristiky ordinálních proměnných 84
5.2.3 Číselné charakteristiky kvantitativních proměnných 85
5.3 Průzkumová analýza datových souborů 97
5.3.1 Statistické zvláštnosti dat 97
6 Náhodný výběr 104
6.1 Výběrové charakteristiky 105
6.2 Statistická indukce 108

7 Výběrové zjišťování ... 111
7.1 Pravděpodobnostní výběrové zjišťování ... 112
7.1.1 Prostý pravděpodobnostní výběr ... 112
7.1.2 Systematický pravděpodobnostní výběr ... 113
7.1.3 Stratifikovaný pravděpodobnostní výběr ... 114
7.1.4 Vícestupňový pravděpodobnostní výběr ... 114
7.2 Nepravděpodobnostní výběrové zjišťování ... 115
7.2.1 Kvótní výběr ... 115
7.2.2 Anketa ... 115
7.2.3 Metoda základního masivu ... 116
7.2.4 Záměrný výběr ... 116
7.3 Rozsah výběru ... 117
7.3.1 Určení minimálního rozsahu výběru ... 119
7.4 Reprezentativnost výběrového souboru ... 121
7.4.1 Validita ... 122
7.4.2 Reliabilita ... 122
8 Teorie odhadu ... 125
8.1 Bodové odhady parametrů rozdělení ... 126
8.2 Intervalové odhady parametrů rozdělení ... 129
8.3 Bodový a intervalový odhad střední hodnoty ... 132
8.3.1 Intervalový odhad střední hodnoty při známém populačním rozptylu ... 132
8.3.2 Intervaly spolehlivosti pro μ při neznámém populačním rozptylu ... 133
8.4 Bodový a intervalový odhad variability ... 135
8.4.1 Bodový a intervalový odhad populačního rozptylu ... 135
8.4.2 Bodový odhad směrodatné odchylky ... 136
8.5 Bodový a intervalový odhad relativní četnosti ... 137
9 Testování statistických hypotéz ... 142
9.1 Statistická hypotéza ... 142
9.1.1 Parametrická hypotéza ... 143
9.1.2 Neparametrická hypotéza ... 143
9.1.3 Nulová a alternativní hypotéza ... 143
9.2 Statistický test ... 144
10 Parametrické testy ... 151
10.1 Jednovýběrové parametrické testy ... 151
10.1.1 Test o populačním rozptylu ... 151
10.1.2 Testy o populačním průměru ... 152
10.1.3 Test o populačním podílu ... 154
10.2 Dvouvýběrové parametrické testy ... 156
10.2.1 Testy shody dvou populačních rozptylu ... 156
10.2.2 Testy shody dvou nezávislých populačních průměru ... 157
10.2.3 Test o shodě dvou závislých populačních průměru ... 159
10.2.4 Test o populačních podílech ... 161
10.3 Vícevýběrové parametrické testy ... 162
10.3.1 Test o shodě více jak dvou populačních rozptylů ... 163
10.3.2 Test o shodě více jak dvou populačních průměrů ... 164
10.4 Vztah mezi testováním hypotéz a intervalovými odhady parametrů ... 170
11 Neparametrické testy ... 174
11.1 Testy shody rozdělení ... 175
11.1.1 $Chí$-kvadrát test dobré shody ... 175
11.1.2 Kolmogorovův-Smirnovův test ... 178

11.1.3 Shapiro-Wilkův test ... 178
11.2 Pořadové testy ... 180
11.2.1 Mannův-Whitneyho test ... 180
11.2.2 Dvouvýběrový Wilcoxonův test ... 181
11.2.3 Wilcoxonův test ... 182
11.2.4 Kruskal-Wallisův test ... 184
Seznam literatury ... 190

1 ÚVOD

Historie statistiky sahá do nejstarších dějin lidstva. Prvními velkými doloženými statistickými událostmi ve starověku bylo sčítání lidu ve 3. tisíciletí př. n. l. v Egyptě a Číně. Nejstarší statistikou je „*popis státu*“, státověda, státovědná statistika, v dnešní době nazývána **popisná statistika**.

V roce 1562 v tehdy hospodářsky vyspělých Benátkách vyšlo jedno z prvních státovědných děl Francesca Sansovina „*O vládě a správě v různých královstvích*“. Jeho základem byl popis území státu, utváření terénu, přírodního bohatství, počtu a struktury obyvatelstva, rozvoje výrobních a dalších odvětví, armády, organizace správy, soudů, škol a podobně. V této době se kromě popisu jen výjimečně objevovala číselná charakteristika, protože číselná data o sledovaných jevech většinou neexistovala.

S rychlým rozvojem hospodářství ke konci období feudalismu začala statistika získávat stále více číselných údajů. V 18. století pak statistici tato data začali systematicky řadit do tabulek a vzájemně je srovnávat.

Statistika je vědní obor zabývající se zkoumáním jevů, které mají hromadný charakter. Jevy, které můžeme opakovaně pozorovat, zkoumá **popisná statistika**. V reálném světě se však objevují i jevy, jejichž přesný výsledek není možno určit. Matematickou interpretací takovýchto jevů se zabývá **matematická statistika**, která vychází ze základů **teorie pravděpodobnosti**.

Teorie pravděpodobnosti je matematická disciplína, která vznikla v polovině 17. století. Podnětem pro její vznik byly úvahy o hazardních hrách (hod kostkou, karetní hry). **Teorie pravděpodobnosti se zabývá popisem chování náhodných veličin na základě teoretických informací (a priori) o povaze daného experimentu (náhodného pokusu).** Je to vědní obor, jehož logická struktura je budována axiomaticky. To znamená, že její základ je tvořen několika **tvrzeními** (tzv. **axiomy**), které vyjadřují základní vlastnosti axiomatizované veličiny a všechna další tvrzení jsou z nich odvozena deduktivně. Takto vytvořená abstraktní teorie již nemusí pracovat s konečnými množinami prvků. Obecně platí, že správnost axiomů je ověřena zkušeností, která se v rámci dané teorie nedokazuje.

Matematická statistika vychází z principu popisné statistiky a počtu pravděpodobnosti. Jedná se o vědní obor, který zkoumá charakter procesů vzniku hromadných dat, které jsou výsledkem náhodného pokusu. Opakování náhodného pokusu vede k nahodilosti výsledků a z toho vyplývá, že všechny závěry matematické statistiky mají náhodný charakter. **Vychází z popisu veličin (a posteriori) na základě měření či empirického (skutečného) poznání,**

které můžeme číselně změřit. Z výše uvedeného vyplývá, že matematická statistika a teorie pravděpodobnosti jsou úzce propojené, ale mají odlišné zaměření.

Popisná statistika (deskriptivní statistika) se zabývá elementárními metodami popisu stavu anebo vývoje hromadných jevů. Cílem popisné statistiky je popsat chování hromadných jevů a zkoumané poznatky o těchto jevech kvantifikovat. Číselné charakteristiky jevů mohou být předmětem statistického zkoumání anebo mohou být základem pro aplikaci složitějších statistických analýz.

Rozvoj informačních technologií v polovině 20. století dal možnost vzniknout *„nové statistice“*. Statistice, která je založena na analytickém úsudku a statistické inferenci (indukci), neboli zobecňování úsudků o vlastnostech celku (populace) na základě výběru a dílčích šetření. Počítače umožnily provádět rozsáhlé početní operace a daly možnost vzniku výpočetně náročným statistickým metodám. Další velký posun v této oblasti byl zaznamenám s rozšířením programového vybavení. Statistické programy umožňují i *„neznalým matematikům“* aplikovat složité statistické metody na datech jim blízkých.

1.1 Základní pojmy

Statistika umožňuje rozvíjet lidské znalosti na základě zjišťování, zpracování, analýzy a prezentace jevů, které jsou výsledkem působení mnoha náhodných i nenáhodných vlivů.

Statistiku je možné chápat jako konkrétní **hodnotu** (především číselnou, ale i slovní), která nám poskytuje informaci o sledovaných jevech (průměr, četnost) nebo jako **činnost**, která je založena na zjišťování dat o hromadných jevech (statistická evidence, výkazy, ročenky) a v neposlední řadě, jako **vědní disciplínu**, která zkoumá zákonitosti chování náhodných jevů a slouží k hlubší analýze chování těchto jevů (třídění, analýza dat, prezentace závěrů).

Jev je výsledkem experimentu, pozorování či jiné činnosti. Číselné či slovně popsaný jev však není dostatečným zdrojem informací k vyslovení závěru. Ve statistice pracujeme především s jevy, které se vyskytují za určitých podmínek opakovaně – tzv. hromadnými jevy.

Hromadné jevy můžeme definovat v souboru věcí, osob, událostí ve formě kvantitativní (číselné) anebo formě kvalitativní (slovní), kterou je možné převést na číslo.

Hromadné jevy mohou být povahy **deterministické** (jevy, jejichž výsledek je přesně určen předchozími podmínkami) anebo jevy **náhodné** (stochastické)**.**

Náhodné hromadné jevy mohou a nemusí nastat za předem stanovených podmínek. Tyto jevy jsou výsledkem působení velkého množství příčin a jejich vlastnosti se nemohou projevit

v jednotlivých jevech, ale jen v souboru jevů, a to prostřednictvím řady náhod (vznik náhodných jevů i přesto odpovídá určitým pravidlům a zákonitostem).

Příklad 1.1

Hromadné jevy z oblasti:

věcí → počet bytů, počet bazénů, počet PC;

osob → výška všech osob v české republice, dosažené vzdělání;

událostí → meteorologická měření, výskyt bouřek.

Zkoumání hromadných jevů je založeno na definování **množiny prvků** (jednotek, případů), které jsou nositelem celé řady vlastností → **znaků** (proměnných → slovně popsatelných anebo číselně kvantifikovatelných). Množina jednotek, která je nositelem stejných znaků, se nazývá **statistický soubor**.

Množinu prvků (**statistickou jednotku**) je potřeba přesně vymezit z hlediska **věcného** (co považujeme za statistickou jednotku), **prostorového** (vymezení území, ze kterého budou statistické jednotky zahrnuty do statistického souboru) a **časového** (určení časového okamžiku nebo období, ve kterém bude statistické šetření prováděno).

Příklad 1.2

Vymezení statistické jednotky

CO?	**KDY?**	**KDE?**
byty 3+1	rok 2024	Středočeský kraj
student VŠ	školní rok 2024/2025	Karlova univerzita
množství srážek v mm	13. 08. 2024	Stochov
výrobní podnik	první pololetí roku 2024	Jihočeský kraj

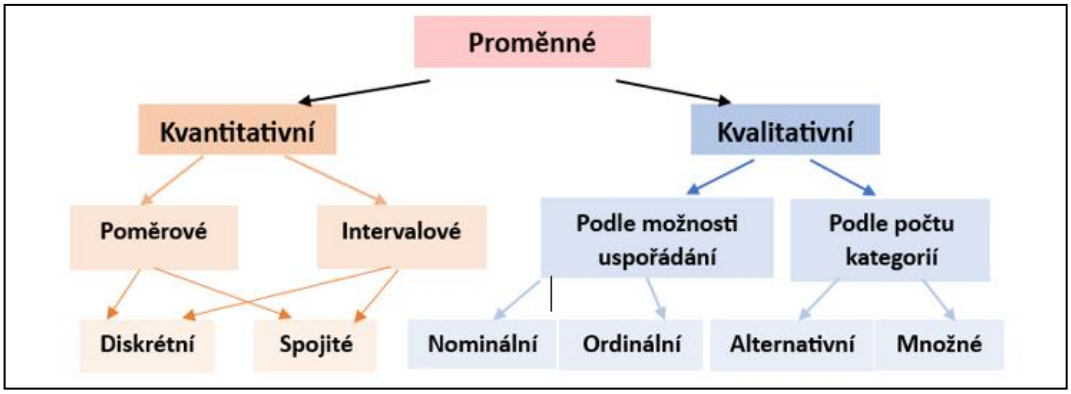
Statistické jednotky (prvky) jsou nositeli statistických znaků → **proměnných (ukazatelů)**.

Zpracování dat a **výběr konkrétních statistických metod** se vždy odvíjí od jejich charakteru. Klasifikaci proměnných je možné provádět podle různých hledisek. Nejčastěji využívané je dělení na kvantitativní a kvalitativní proměnné (obr. 1.1). Je-li sledovaná vlastnost prvku číselně měřitelná, označujeme příslušnou statistickou proměnnou jako **kvantitativní (kardinální)**, v opačném případě, kdy jsou jednotlivé obměny sledované vlastnosti popsány pouze slovně, hovoříme o proměnné **kvalitativní**.

Kvantitativní proměnné jsou číselně měřitelné údaje, u kterých můžeme přesně říct, o kolik je jedna hodnota vyšší než druhá (věk, hmotnost, počet osob, strojů atd). Jsou to takové

hodnoty, které vyjadřují skutečné množství sledovaných vlastností statistických jednotek (v žádném případě, se nejedná o číselné kódy, které přiřazujeme proměnným kvalitativním při převodu údajů do datové matice). Kvantitativní proměnné rozdělujeme na poměrové a intervalové. U **poměrových (podílových) proměnných** můžeme vyjádřit rozdíl a podíl dvou hodnot. Poměrová proměnná může nabývat pouze kladných číselných hodnot (např. Karel má 50 let, Jana 25 let → rozdíl 25 let; Karel je dvakrát starší než Jana). U tohoto typu proměnné hovoříme o existenci přirozené nuly tzv. *„smysluplné nuly"* (např. cílem statistického sledování je váha notebooku v případě, že notebook nic neváží, tak neexistuje). V případě **intervalové proměnné** můžeme číselně vyjádřit pouze velikost rozdílu mezi libovolnými dvěma hodnotami znaku (rozdíl teplot v místnostech, teplota vzduchu měřená v celých číslech). Intervalové proměnné mají nulu na stupnici určenou konvencí (dohodou), tzn. že nemají nulovou hodnotu ve smyslu neexistence dané vlastnosti. Kvantitativní proměnné rozdělujeme také na proměnné **diskrétní**, které nabývají celočíselných hodnot (počet studentů, počet domů) a **spojité**, které nabývají libovolných hodnot z určitého intervalu (výška, cena, příjem).

Obrázek 1.1: Základní rozdělení proměnných podle jejich charakteru



Kvalitativní proměnné jsou slovně vyjádřitelné vlastnosti a dělíme je podle **možnosti uspořádání** na nominální a ordinální. O hodnotách **nominálních proměnných** můžeme pouze říct, zda jsou stejné či různé. Kvantifikace těchto proměnných je možná na základě četnosti výskytu sledovaného znaku (škola, fakulta). U **ordinálních (pořadových) proměnných**, můžeme určit pořadí sledovaných znaků. Hodnoty proměnných se vyjadřují vzestupně nebo sestupně podle úrovně zkoumané vlastnosti (úroveň spokojenosti, vzdělání). Hodnoty nominální proměnné mohou být označeny čísly (zakódovány), ale nelze u nich provádět stejné statistické analýzy jako s proměnnými kvantitativními. U ordinálních proměnných závisí možnost provádění statistických analýz na předpokladu stejné vzdálenosti mezi kategoriemi.

Příklad 1.3

Vymezení statistické proměnné

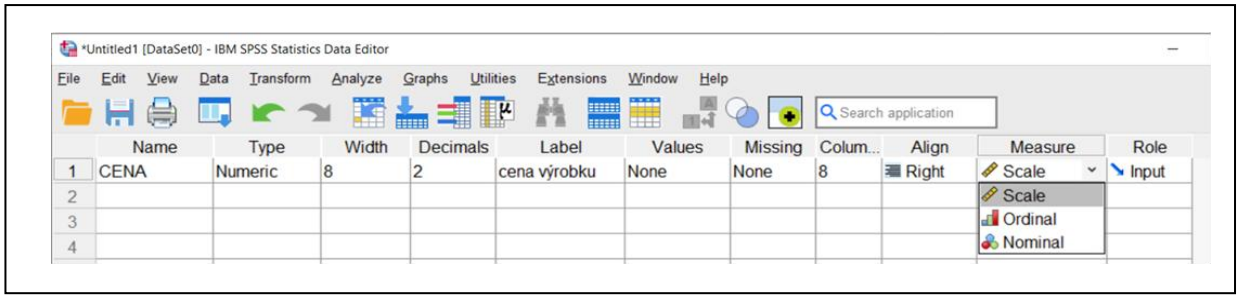
Proměnná	Znak	typ	podtyp
cena výrobku	0 neexistence, 1, …….	kvantitativní	poměrová; spojitá
počet sourozenců	0 neexistence, 1, …….	kvantitativní	poměrová; diskrétní
teplota	„- ∞ do +∞“	kvantitativní	intervalová; spojitá
dny v roce	den	kvantitativní	intervalová; diskrétní
barva očí	modrá, zelená, hnědá, …..	kvalitativní	nominální; množná
kouření	ano, ne	kvalitativní	nominální; alternativní
vzdělání	základní, středoškolské, …	kvalitativní	ordinální; množná

Podle **počtu kategorií** je dělíme na alternativní a množné. **Alternativní (dichotomické) proměnné** nabývají pouze dvou obměn sledované vlastnosti (kuřák a nekuřák). **Množné proměnné** nabývají více než dvou obměn sledované vlastnosti (obor studia, barva).

Uvedené způsoby klasifikace proměnných dle vlastností statistického znaku, jehož jsou nositelem, se v odborné literatuře a programových systémech mohou lišit. Často se setkáváme také s pojmem **kategoriální proměnné.** Obměny těchto proměnných nazýváme kategoriemi. Řadíme zde nejen kvantitativní proměnné, ale také proměnné, které mohou vzniknout odvozením z proměnné kvantitativní v případě, že ji rozdělíme do intervalů.

Proměnné můžeme vzájemně transformovat, např. hodnota BMI (Body Mass Index). Výpočtový vztah vychází ze srovnání dvou kvantitativních proměnných váhy a výška. Výsledná hodnota je tvořena jedním číslem, které můžeme následně roztřídit do kategorií. Volba kategorizace a tím i změna typu proměnné je založena na individuálním přístupu analytika, např: neobézní, obézní → proměnná nominální/alternativní; podváha <18,5; ideální váha 18,5–25; nadváha 25,1–30; obezita >30,1 → proměnná ordinální/množná.

Obrázek 1.2: Klasifikace proměnných v IBM SPSS



Data (hodnoty sledované proměnné) můžeme získat přímo (např. měřením, vážením, počítáním) → **primární data**, anebo nepřímo (např. dělením, sčítáním, násobením) → **sekundární data**.

Příklad 1.4

Vymezení statistické proměnné

Proměnná	**Statistická data**
věk	primární
počet zaměstnanců	primární
BMI	sekundární
produktivita práce	sekundární

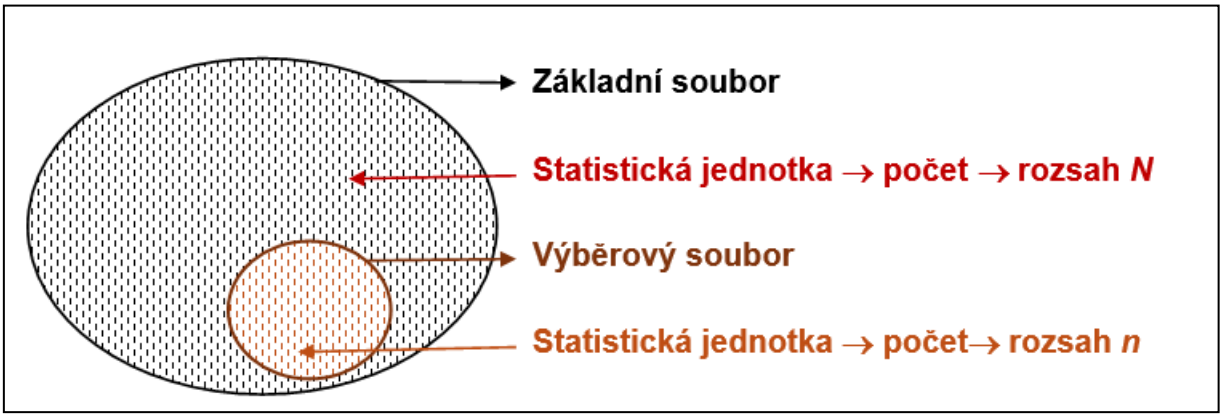
Rozdělení proměnných podle číselných hodnot nebo počtu vlastností, jejichž jsou nositelem, je vhodné ještě doplnit o rozdělení proměnných na nezávislé a závislé.

Nezávislé proměnné jsou takové hodnoty znaků, které ovlivňují **hodnoty závislých proměnných** (váha člověka závisí na jeho výšce; počet bodů z písemného testu je ovlivněn dobou přípravy na test).

Údaje o číselných nebo slovních hodnotách jednoho nebo více znaků ve statistickém souboru se nazývají **statistická data.**

Definování statistických jednotek a jejich znaků je důležitým krokem při sestavování **statistických souborů**.

Obrázek 1.3: Grafické znázornění statistického souboru



Statistický soubor může být **základní** nebo **výběrový** (Obr. 1.3). Definování základního a výběrového souboru se odvíjí od cílů statistického zkoumání a musí být rovněž jednoznačně věcně, prostorově a časově vymezené.

Základní soubor (populace) je množina všech prvků (jednotek), na kterých je prováděno statistické zkoumání. Počet prvků → **rozsah základního souboru** značíme $\boldsymbol{N}$.

V některých případech je statistické zkoumání na celé populaci nezastupitelné (výkaznictví, sčítání lidu, počty zvířat atd.). Rozsah základního souboru může být **konečný** – počet členů $\boldsymbol{N}$ je u takové populace přesně stanovitelný (počet strojů ve výrobním závodě, počet studentů daného ročníku atd.) anebo **„nekonečný“** – počet členů $\boldsymbol{N}$ je proměnlivý, nelze ho přesně zjistit (počet studentů vysokých škol v celé Evropě, počet jehličnatých stromů v kraji atd.). V případě, že rozsah základního souboru je konečný a *„rozumně velký“* můžeme, provádět úplné zjišťování o vlastnostech všech jeho jednotek. Zpracování a rozbor takto získaných údajů je pak přímo založen na metodách deskriptivní statistiky a poskytuje nám přesný obraz sledovaného chování celé populace.

Rozsah základního souboru je však často opravdu velký anebo nekonečný, proto není možné zjistit sledované vlastnosti u všech prvků. Vzhledem k tomu provádíme výběr určitého počtu prvků. V takovém případě pak hovoříme o **výběrovém souboru**. Výběrový soubor je podmnožinou souboru základního a počet jeho jednotek značíme $\boldsymbol{n}$**.** Vybrané jednotky souboru mohou být vybrány náhodně (náhodný výběr) anebo podle určitých pravidel (záměrný výběr). Obecně by měly vybrané jednotky vždy obsahovat podstatné a charakteristické rysy základního souboru a být jeho reprezentativním zástupcem → reprezentativním výběrem.

Tabulka 1.1: Výhody a nevýhody statistického zjišťování podle typu souboru

	Výhody	Nevýhody
Základní soubor	Poskytuje přesné výsledky o všech vlastnostech sledovaných jednotek.	Velké finanční náklady a pracnost. Nespolehlivost zjišťovaných dat. Těžko proveditelná kontrola správnosti údajů.
Výběrový soubor	Zjišťování je rychlejší. Vyžaduje nižší náklady.	Umožňuje pouze odhady o průběhu sledovaných jevů v základním souboru. Údaje nebo charakteristiky na jejich základě vypočtené jsou zatíženy chybou.

Na základě informací o výběrových statistických jednotkách a hodnotách jejich znaků (datech), usuzujeme na vlastnosti celé populace (viz kapitola Statistická indukce).

Zjišťujeme-li u každé jednotky statistického souboru hodnoty jedné proměnné (znaku), hovoříme o **jednorozměrném** statistickém souboru. Zjišťujeme-li u každé jednotky statistického souboru hodnoty dvou či více proměnných, hovoříme o **dvourozměrném**, respektive **vícerozměrném** statistickém souboru. V případě, že hodnoty proměnné pro

jednotlivé prvky souboru (statistická data) uspořádáme do datové matice, získáme **datový soubor**.

Příklad 1.5

Základní soubor představuje počet všech studentů nejmenované vysoké školy, jehož rozsah je N = 12 534. Výběrových souborů může být celá řada, vždy záleží na cíli statistického šetření. V našem případě se může jednat například o studenty Ekonomické fakulty n = 8 255; studenty oboru Provoz a ekonomika n = 1 053. Z uvedených výběrových souborů se dají vytvořit další výběrové soubory s ještě menším rozsahem a vymezením.

Při pořizování statistických dat je vždy důležité zaměřit se na to, aby data obsahovala potřebné informace o zkoumané problematice. Předmětem zkoumání jsou statistické proměnné, které mohou být povahy kvantitativní nebo kvalitativní. Proměnné jsou zkoumány u statistických jednotek, kterými jsou například regiony, obce, podniky či osoby. Z uvedeného vyplývá, že ve statistice pracujeme s různými datovými strukturami.

Příklad 1.6

Datový soubor je tvořen čtyřmi jednotkami a pěti proměnnými (ukazateli).

Jednotka /proměnná	Pohlaví	Věk	Výška	Sourozenci	Počet sourozenců
student Karel	muž	25	185	ano	2
studentka Jana	žena	23	162	ano	1
student Tomáš	muž	27	175	ne	0
student Petr	muž	22	180	ano	4

V **datovém souboru** jsou data uspořádána tak, že ve sloupcích jsou zaznamenány analyzované ukazatele – např. věk, pohlaví. Každý řádek tabulky (matice) se týká jedné statistické jednotky – např. student.

Příprava datové matice spočívá především v převedení různých formátů dat do formátu, který potřebujeme pro zpracování ve vybraném statistickém programu. Vybavení datové matice zahrnuje kódování ukazatelů, jejich popis i zařazení dle typu znaků, jejichž jsou nositeli. Hodnoty ukazatelů musí být srovnatelné mezi jednotkami v prostorovém a časovém srovnání.

Statistické zjišťování, organizace dat a jejich přenos do datového souboru s sebou přináší různé druhy chyb, kterými mohou být data zatížena. Proto je vždy nutné před samotným zpracováním provést formální kontrolu správnosti získaných údajů (převody jednotek,

matematické úpravy vstupních hodnot) a logické posouzení charakteru dat vzhledem k cílům daného šetření.

Jedná se například o **chyby při zpracování**, kterým je možné předejít pomocí zvýšené pečlivosti při záznamu dat, ruční či počítačovou kontrolou nebo duplicitním způsobem převodu dat do datové matice. Dále se jedná o **chyby metodické**, jichž je možné se dopustit zvolením nesprávného postupu zpracování dat. V neposlední řadě se může jednat o **chyby výběrové**, kterých se dopouštíme v případě zobecnění výsledků získaných pomocí výběrového souboru na soubor základní. Výběrová chyba se skládá z chyby systematické a pravděpodobnostní (viz kapitola Teorie odhadu).

Proces odstraňování chyb se nazývá čištění dat a úzce souvisí i s ověřováním kvality sledovaných jevů. Odstraňování chyb spočívá v rozpoznání nepřesných, nekompletních nebo nesmyslných údajů a jejich úpravě. Je zaměřeno na odhalování extrémních (odlehlých) či vybočujících pozorování, které mohou výrazně ovlivnit výsledky některých analýz. Nalezení odlehlých či vybočujících pozorování se v jednorozměrné analýze dat opírá zejména o grafickou analýzu (viz kapitola Průzkumová analýza datových souborů). Před samotným zpracování dat je důležité zaměřit se také na **chybějící hodnoty**, které jsou obsaženy téměř v každém datovém souboru.

Jednou z možností, která by měla být výjimečná, avšak bývá často používaná, je **vymazání záznamů** s chybějícími hodnotami. Tento krok by se měl provádět pouze v situacích, kdy chybějící údaje neovlivní výsledky zpracování, nebo v případě většího počtu chybějících hodnot u jednoho případu.

Mnohem vhodnější jsou metody imputace, neboli doplnění chybějících hodnot. Cílem **imputačních metod** je nahrazení chybějící hodnoty odhadem, který se provádí na základě známých hodnot pomocí různých algoritmů od nejjednodušších, které spočívají v prostém nahrazení chybějící hodnoty aritmetickým průměrem nebo mediánem, až po metody, které berou v úvahu vztahy mezi proměnnými napříč datovým souborem. Tyto metody jsou vhodné především pro kvantitativní proměnné. Pro práci s kvalitativními daty jsou vhodné **metody deduktivní** neboli subjektivní metody imputace, které vycházejí z pravidel založených na odvození logických vztahů mezi jednotlivými proměnnými.

POZOR! *Kvalita dat a kontrola datového souboru je klíčovým předpokladem a nezbytnou podmínkou pro úspěšnou aplikaci vybraných statistických metod.*

V neposlední řadě je žádoucí vstupní data vhodně popsat a zakódovat, neboli přidělit každé variantě znaku číselný kód, který bude použit při zpracování. Kódování může vycházet nejen z typu proměnných, ale i z požadavků zvoleného počítačového programu.

Pro zpracování datového souboru je možné využít různé typy statistických programů. Jednoduché procedury lze provádět již pomocí tabulkového procesoru (např. Microsoft Excel), ke složitějším analýzám je vhodné využít některý ze specializovaných statistických systémů (např. SPSS, SAS, STATISTICA), které představují komplexní nástroje pro zpracování rozsáhlých datových souborů. Programové systémy umožňují nejen komplexní statistickou analýzu dat, ale také následnou prezentaci výstupů ve formě tabulek a grafů.

Shrnutí kapitoly

Popisná statistika je popis dat, co „vidíme“. Popisná statistika se zabývá vyhodnocením a prezentací dat. Jejím cílem je popsat chování hromadných jevů a získané poznatky kvantifikovat. Používá k tomu především číselné charakteristiky a grafické metody. Netýká se primárně náhodných veličin, nýbrž konkrétních naměřených dat.

Teorie pravděpodobnosti je matematická disciplína, která se zabývá popisem chování náhodných jevů a veličin na základě teoretických informací.

Matematická statistika se zabývá zobecňování a vyvozování závěrů z dat o celé populaci, s využitím nástrojů teorie pravděpodobnosti, která poskytuje teoretický základ pro práci s náhodnými jevy.

Zkoumání hromadných jevů je založeno na definování množiny jednotek stejného druhu (srovnatelnost věcná, prostorová, časová) → předmět zkoumání je statistická jednotka.

Statistické jednotky jsou nositelem celé řady vlastností (znaků, proměnných).

Množina jednotek, která je nositelem stejných proměnných, se nazývá **statistický soubor**.

Základní soubor (populace) je množina všech statistických jednotek.

Výběrový soubor je podmnožinou základního souboru a je tvořen jednotkami, které byly ze základního souboru vybrány podle určitých předem zvolených hledisek.

Základní rozdělení **proměnných** je na **kvantitativní** a **kvalitativní**. Správná klasifikace proměnných je velice důležitá, neboť od typu proměnné se odvíjí následné postupy statistických analýz.

Při přenosu hodnot sledovaných proměnných do souboru může dojít k **chybám** při **zpracování** dat, nebo k **metodickým** a **výběrovým** chybám. Vstupní data a jejich prvotní kontrola je klíčovým parametrem, který určuje kvalitu prováděných analýz a následných výstupů.

Literatura

BUDÍKOVÁ, Marie; Maria KRÁLOVÁ a Bohumil MAROŠ. *Průvodce základními statistickými metodami*. Vydání první. Praha: Grada Publishing, a.s., 2010. ISBN 978-80-247-3243-5.

HEBÁK, Petr. *Statistické myšlení a nástroje analýzy dat*. 2. vydání. Praha: Informatorium, 2015. ISBN 978-80-7333-118-4.

JOHNSON, Richard A. a Dean W. WICHERN. *Applied multivariate statistical analysis*. Sixth edition. Harlow: Pearson. 2014. ISBN 978-1-292-02494-3.

1 Kontrolní otázky

1. Co je statistika a jak ji lze chápat?
2. Co jsou to hromadné jevy a jak se dělí? Uveďte příklady.
3. Vysvětlete pojem statistická jednotka.
4. Co je to statistický soubor a jaké jsou jeho hlavní typy?
5. Jaký je rozdíl mezi základním a výběrovým souborem?
6. Popište výhody a nevýhody úplného zjišťování (základního souboru) a výběrového zjišťování (výběrového souboru).
7. Co jsou to statistické znaky (proměnné, ukazatele) a jak se dělí?
8. Jaký je rozdíl mezi kvantitativními a kvalitativními proměnnými? Uveďte příklady.
9. Jak se dále dělí kvantitativní proměnné? Popište poměrové a intervalové proměnné.
10. Jaký je rozdíl mezi diskrétními a spojitými proměnnými? Uveďte příklady.
11. Jak se dále dělí kvalitativní proměnné?
12. Uveďte příklady nominálních, ordinálních, alternativních a množných proměnných.
13. Co jsou to závislé a nezávislé proměnné? Uveďte příklad.
14. Co jsou statistická data a jak se dělí podle způsobu získání?
15. Co je to datová matice a jak jsou v ní data uspořádána?
16. Jaké typy chyb mohou zatížit statistická data a jak se jim předchází?
17. Co je čištění dat a na co se zaměřuje?
18. Jaké jsou možnosti řešení chybějících hodnot v datovém souboru? Kdy je vhodné záznamy s chybějícími hodnotami vymazat a kdy se používají metody imputace?
19. K čemu slouží kódování ukazatelů?

1 Příklady k procvičení

Marketingová agentura provádí průzkum, jehož cílem je zjistit, jak spotřebitelé v České republice vnímají a uplatňují principy udržitelnosti při svých nákupech. Pro zjednodušení cvičení máte k dispozici následující data od 12 respondentů.

Věk	Vzdělání	Měsíční příjem (Kč)	Nakupuje ekologické produkty	Jste ochoten zaplatit více za udržitelné produkty	Informační zdroj o udržitelnosti	Počet recyklovaných druhů odpadu
35	VŠ	45000	často	Ano	Sociální sítě	4
22	SŠ	28000	někdy	Ne	Online zprávy	2
58	VŠ	55000	často	Ano	TV/rozhlas	5
41	SŠ	32000	někdy	Ano	Přátelé/rodina	3
29	VŠ	40000	často	Ano	Online zprávy	4
67	ZŠ	20000	nikdy	Ne	Žádný	1
38	VŠ	48000	často	Ano	Sociální sítě	4
25	SŠ	30000	někdy	Ne	Online zprávy	2
50	VŠ	60000	často	Ano	TV/rozhlas	5
33	SŠ	31000	někdy	Ano	Přátelé/rodina	3
46	VŠ	50000	často	Ano	Online zprávy	4
70	ZŠ	22000	nikdy	Ne	Žádný	1

1.1 **Vymezení statistické jednotky a souborů:**

a) Definujte statistickou jednotku v tomto průzkumu.

b) Jak byste vymezili základní soubor a výběrový soubor pro tento průzkum?

c) Jaký je rozsah výběrového souboru v tomto konkrétním případě?

1.2 **Klasifikace proměnných.** Pro každou proměnnou v tabulce určete její typ a podtyp podle klasifikace proměnných. Zdůvodněte své rozhodnutí.

a) Věk.

b) Vzdělání (ZŠ, SŠ, VŠ).

c) Měsíční příjem (Kč).

d) Nakupuje ekologické produkty (často/někdy/nikdy).

e) Je ochoten zaplatit více za udržitelné produkty (Ano/Ne).

1.3 **Možné chyby a jejich kontrola:**

a) Uveďte alespoň dva typy chyb, které by mohly nastat při sběru dat v takovém průzkumu (např. chyby při zpracování, metodické, výběrové).

b) Navrhněte, jak byste provedli prvotní kontrolu dat pro proměnnou "Měsíční příjem (Kč)" a "Informační zdroj o udržitelnosti".

1.4 **Definování závislé/nezávislé proměnné:**

a) Formulujte jednu otázku, kterou by bylo možné zkoumat na základě uvedených údajů (např. o souvislosti mezi proměnnými).
b) V rámci této otázky určete, která proměnná by byla závislá (vysvětlující) a která nezávislá (vysvětlovaná).

2 NÁHODNÝ JEV A PRAVDĚPODOBNOST

Pravděpodobnost se zabývá studiem zákonitostí náhodných jevů, jejich popisem a vytvořením pravidel pro její výpočet.

2.1 Náhodný jev

Výchozím pojmem teorie pravděpodobnosti je **náhodný pokus**.

Pokus je děj, činnost nebo pozorování, jehož výsledek je ovlivněn mnoha různými vlivy a podmínkami. Pokus je možné libovolněkrát nezávisle opakovat (teoreticky) při dodržení všech podmínek. Vzhledem k tomu, že některé podmínky se mohou měnit (náhoda), může každé opakování pokusu vést k jinému výsledku.

Náhoda je komplex drobných příčin, které se od jednoho provedení pokusu k druhému pokusu mění. Je příčinou toho, že výsledky některých pokusů není možné jednoznačně určit vzhledem k množství a neznalosti všech podmínek (náhod), při nichž probíhá.

Na základě výše uvedeného, definujeme pojem **náhodný pokus** jako pokus, jehož každé opakování vede k právě jednomu výsledku (hod klasickou hrací kostkou vede k právě jednomu ze šesti výsledků atd.), tzn., že žádné výsledky nemohou nastat současně. Množina všech výsledků musí být vyčerpávající, tzn., že při realizaci náhodného pokusu musí právě jeden výsledek z množiny výsledků nastat.

Příklad 2.1

Základní prostor Ω hod hrací kostkou.

Náhodný pokus	**Hod kostkou**	$\boldsymbol{\Omega = \{\omega_i\}}$
Elementární jevy (jednobodová množina) $\boldsymbol{i = 1; 2; \ldots; 6}$	„padne 1“	ω_1
	„padne 2“	ω_2
	„padne 3“	ω_3
	„padne 4“	ω_4
	„padne 5“	ω_5
	„padne 6“	ω_6

Řešení

Základní prostor Ω je vymezen množinou všech možných elementárních jevů $\omega_1, \omega_2, \ldots, \omega_6$. $\Omega = \{\omega_1, \omega_2, \ldots \omega_6\}$.

Náhodný jev je výsledek náhodného pokusu. Náhodný jev může a nemusí nastat při daném komplexu podmínek. Jev, který nastane vždy při realizaci daných podmínek, se nazývá **jev jistý**. Jev, který nikdy nemůže nastat při realizaci daných podmínek, se nazývá **jev nemožný**. Náhodný jev nelze s jistotou předpovědět, protože některé výsledky se stávají častěji než jiné. Při velkém počtu opakování však vykazují náhodné jevy určité zákonitosti a pravidelnosti.

Množinu všech možných výsledků náhodného pokusu nazýváme **základní prostor** a značíme ji symbolem Ω. Jednotlivé možné elementární výsledky pokusu $\omega \subseteq \Omega$ nazýváme **elementární jevy** $\Omega = \{\omega_1, \omega_2, \dots, \omega_i\}$.

Neprázdný systém všech podmnožin základního prostoru, který je uzavřen vzhledem k množinovým operacím nazýváme **jevovou σ-algebrou** tzv. **jevovým polem** $\mathcal{A} \subseteq \Omega$. Prvky jevového pole jsou sledované náhodné jevy, které obvykle označujeme velkými písmeny převážně z počátku latinské abecedy $(A, B, C, \dots)$. Definice jevového pole vychází ze dvou tvrzení (axiómů).

Axiom bezespornosti – jevové pole je možné sestrojit na každém základním prostoru.

Axiom neúplnosti – na dvouprvkovém či více prvkovém prostoru, lze vytvořit více jevových polí.

Příklad 2.2

Vypište níže uvedené podmnožiny náhodných jevů z množiny základního prostoru Ω hod hrací kostkou.

Náhodný pokus – hod kostkou	**Náhodný jev**	**Jevové pole**
„padne sudé číslo“	Náhodný jev A	$A = \{\omega_2, \omega_4, \omega_6\}$
„padne číslo ≥ 3“	Náhodný jev B	$B = \{\omega_3, \omega_4, \omega_5, \omega_6\}$
„padne číslo > 6“	Náhodný jev C	$C = \emptyset$ Jev nemožný
„padne číslo < 7“	Náhodný jev D	$D = \Omega$ Jev jistý

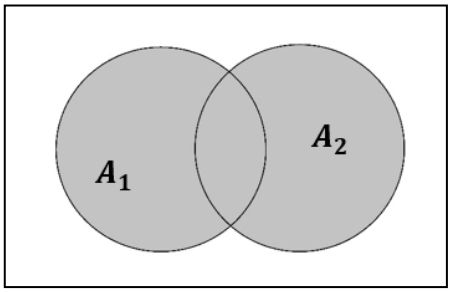
2.2 Základní operace a vztahy mezi náhodnými jevy

Základní prostor spolu s jevovým polem je měřitelný prostor. Jevy A, respektive $A_1, A_2, \dots, A_i$; jsou jevy na tomto prostoru – jedná se o podmnožinu množiny, a proto při základních operacích vycházíme ze vztahů, které odpovídají množinovým relacím (místo výrokových). Pro ilustraci je u základních operací uvedeno grafické znázornění uvedených vztahů tzv. Vennovy diagramy.

Sjednocení jevů (disjunkce) → $A_1 \cup A_2$ nastane alespoň jeden z jevů A_1, A_2. Může nastat pouze jeden z jevů anebo mohou nastat i oba jevy zároveň.

$$A_1 \cup A_2 = \{\omega \in \Omega; \omega \in A_1 \vee \omega \in A_2\}$$

Obrázek 2.1: Vennův diagram pro sjednocení dvou množin



Máme-li více náhodných jevů $A_1 \cup A_2 \cup A_3 \cup \ldots \cup A_n = \bigcup_{i=1}^{n} A_i$ značí uskutečnění alespoň jednoho z náhodných jevů $A_1, A_2, A_3, \ldots, A_n$ pro $1 \leq i \leq n$.

Příklad 2.3

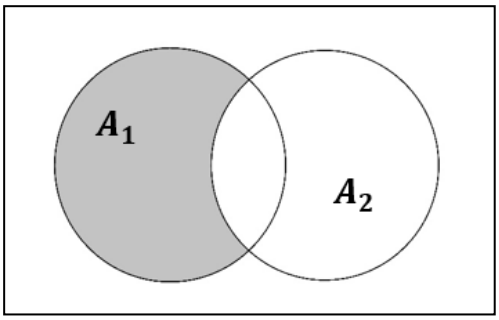
Jev A_1 na hrací kostce padne sudé číslo; jev A_2 na hrací kostce padne číslo $A_2 = \{2; 3; 4\}$.

Řešení $\boldsymbol{A_1 \cup A_2 = \{2; 3; 4; 6\}}$

Rozdíl jevů → $A_1 \setminus A_2$ při realizaci jevu A_1 současně nenastane jev A_2.

$$A_1 \setminus A_2 = \{\omega \in \Omega; \omega \in A_1 \wedge \omega \notin A_2\}$$

Obrázek 2.2: Vennův diagram pro rozdíl dvou množin



Příklad 2.4

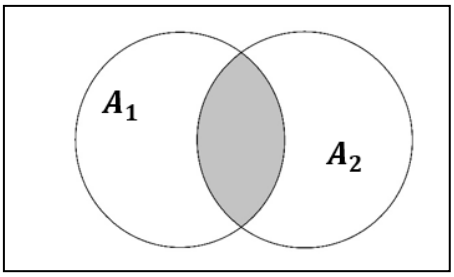
Jev A_1 na hrací kostce padne číslo větší než tři; jev A_2 na kostce padne liché číslo.

Řešení $\boldsymbol{A_1 \setminus A_2 = \{4; 6\}}$

Průnik jevů (konjunkce) → $A_1 \cap A_2$ nastane při současném výskytu obou jevů.

$$A_1 \cap A_2 = \{\omega \in \Omega; \omega \in A_1 \wedge \omega \in A_2\}$$

Obrázek 2.3: Vennův diagram pro průnik dvou množin



Máme-li více náhodných jevů $A_1, A_2, A_3, \ldots, A_n$, potom jejich průnik $A_1 \cap A_2 \cap A_3 \cap \ldots \cap A_n = \bigcap_{i=1}^{n} A_i$ spočívá v nastoupení všech jevů.

Příklad 2.5

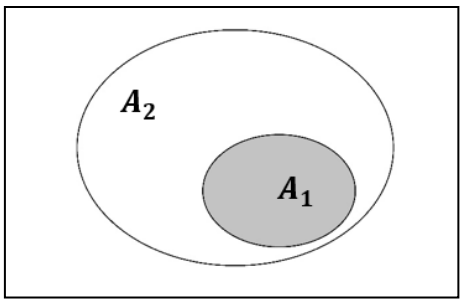
Jev A_1 na hrací kostce padne sudé číslo; jev A_2 na kostce padne číslo menší než pět.

Řešení $\boldsymbol{A_1 \cap A_2 = \{2; 4\}}$

Jev A_1 je podjevem jevu A_2 → $A_1 \subset A_2$ při realizaci jevu A_1 nastává i jev A_2. Jev A_1 má za následek jev A_2. Nenastane-li jev A_1 nenastane ani jev A_2.

$$A_1 \subset A_2 = \{\omega \in \Omega; \omega \in A_1 \Longrightarrow \omega \in A_2\}$$

Obrázek 2.4: Vennův diagram znázorňující podmnožinu



Příklad 2.6

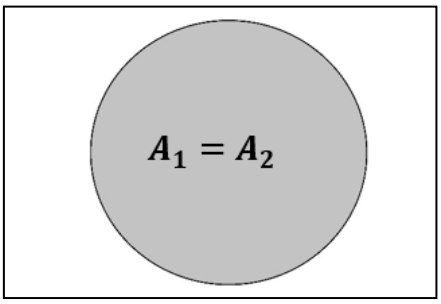
Jev A_1 na hrací kostce padne číslo dvě; jev A_2 na kostce padne sudé číslo. Jev A_1 je podjevem jevu A_2.

Řešení $\boldsymbol{A_1 \subset A_2 = \{2\}}$

Rovnocennost jevů $\rightarrow A_1 = A_2$ při realizaci jevu A_1 nastane také jev A_2 a naopak. Jev A_1 je ekvivalentní s jevem A_2.

$$A_1 \subset A_2 \wedge A_2 \subset A_1$$

Obrázek 2.5: Vennův diagram shodných množin

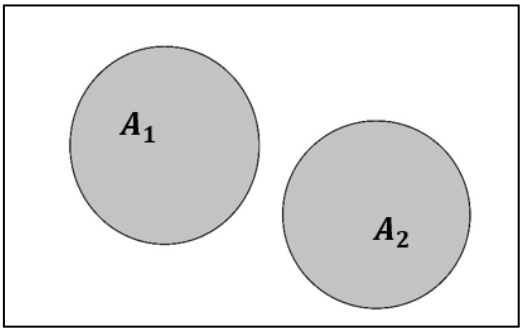


Příklad 2.7

Jev A_1 při hodu hrací kostkou padne sudé číslo; jev A_2 při hodu kostkou padne číslo dělitelné dvěma.

Neslučitelnost jevu $\rightarrow A_1 \cap A_2 = \emptyset$ výskyt jednoho jevu vylučuje možnost výskytu druhého jevu, tj. jejich průnik je jev nemožný. Dva jevy A_1, A_2 nemohou nastat současně, nemají-li spolu žádný možný společný výsledek.

Obrázek 2.6 Vennův diagram disjunktních množin



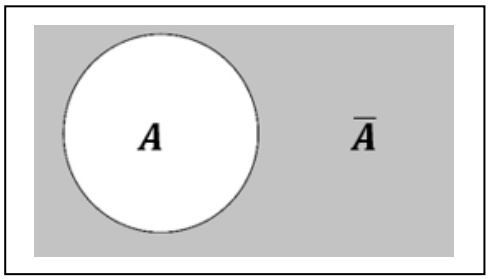
Příklad 2.8

Jev A_1 na hrací kostce padne číslo šest; jev A_2 na hrací kostce padne liché číslo.

Opačný jev $\rightarrow \bar{A}$ spočívá v nenastoupení jevu A. Nastanou všechny náhodné jevy neobsažené v jevu A.

$$\bar{A} = \Omega - A = \{\omega \subset \Omega; \forall\, \omega \notin A \Longrightarrow \omega \in \bar{A}\}$$

Obrázek 2.7: Vennův diagram pro doplněk



Příklad 2.9

Jev A při hodu kostkou padne číslo 6; jev opačný $\bar{A}$ při hodu kostkou padnou 1 nebo 2 nebo 3 nebo 4 nebo 5.

2.3 Pravděpodobnost náhodného jevu

V souvislosti s náhodnými jevy bylo uvedeno, že systém podmnožin základního prostoru, který **je uzavřen vzhledem k množinovým operacím, se nazývá jevové pole $\mathcal{A}$.**

Jevové pole spolu se **základním prostorem Ω** označujeme jako **měřitelný prostor**. Míru očekávání výskytu jevu vyjadřujeme pomocí pravděpodobnosti.

Pravděpodobnost náhodného jevu P je reálné číslo, které vyjadřuje míru možnosti nastoupení náhodného jevu v náhodném pokusu.

Existují různé definice pravděpodobnosti. Všechny však mají určité společné vlastnosti a axiomy (tvrzení), o jejichž platnosti nepochybujeme.

2.3.1 Axiomatická teorie pravděpodobnosti (teoretická)

Axiomatickou teorii formuloval A. N. Kolmogorov. Jedná se o teorii, která je základem celé současné pravděpodobnosti. Vychází ze skutečnosti, že šanci jevu na jeho uskutečnění lze vyjádřit pomocí pravděpodobnosti, což je funkce, která každému jevu přiřazuje hodnotu od 0 do 1 za předpokladu, že tato funkce splňuje určité axiomy. Axiomy jsou základní předpoklady (tvrzení), jejichž správnost byla ověřena zkušeností a v rámci dané teorie se nedokazují.

Axiom 1: **axiom nezápornosti** $\rightarrow P(A) \geq 0\ \forall\, A \in \mathcal{A}$; pravděpodobnosti náhodného jevu $P(A)$ je vždy přiřazeno nezáporné číslo.

Axiom 2: **axiom normovanosti** → $P(\Omega) = 1$; pravděpodobnost jistého jevu. Základní prostor obsahuje všechny možné výsledky měření, které mohou nastat (Ω je jev jistý).

Axiom 3: **axiom součtu pravděpodobnosti nekonečně mnoha neslučitelných náhodných jevů** → $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$; pravděpodobnost nastolení navzájem neslučitelných náhodných jevů se rovná pravděpodobnosti jejich součtu.

Systém axiómů, stejně jako, jevového pole (viz kapitola 2.1), vychází ze dvou tvrzení:

Bezespornost → pravděpodobnost, lze sestrojit na každém základním prostoru;

Neúplnost → na každém měřitelném prostoru lze sestrojit více pravděpodobností.

Tato teorie nedává návod na to, jak vypočítat pravděpodobnost náhodných jevů.

Příklad 2.10

Vypočítejte pravděpodobnost, že na hrací kostce padne číslo tři.

Řešení

Náhodný jev A „padne číslo tři“ $A = \{\omega_3\}$

Pravděpodobnost jevu A, který může nastat při hodu hrací kostkou vypočítáme pomocí axiomatické definice pravděpodobností. Vycházíme ze skutečnosti, že není potřeba experimentu, na základě teoretické znalosti, že základní prostor Ω je vymezen množinou všech možných výsledků elementárních jevů $\omega_1, \omega_2, \dots, \omega_6; \Omega = \{\omega_1, \omega_2, \dots, \omega_6\}$.

$$P(\omega_3) = \frac{1}{6} = 0{,}16\bar{6}$$

Vlastnosti pravděpodobnosti

1. Pravděpodobnost libovolného náhodného jevu A je číslo z intervalu $0 \leq P(A) \leq 1$.
2. Pravděpodobnost jistého jevu je rovna jedné $P(\Omega) = 1$.
3. Pravděpodobnost nemožného jevu je rovna nule $P(\emptyset) = 0$.
4. Pravděpodobnost opačného jevu $P(\bar{A}) = 1 - P(A)$.
5. Dva rovnocenné jevy jsou i stejně pravděpodobné $P(A) = P(B)$.
6. Jestliže jev A je částí jevu B, pak $P(A) \leq P(B)$.

7. Pro dva libovolné **slučitelné jevy** A, B platí, že pravděpodobnost jejich sjednocení se rovná součtu **(věta o sčítání pravděpodobností**) pravděpodobností jednotlivých jevů zmenšenému o pravděpodobnost jejich průniku. Zjednodušeně při sčítání pravděpodobnosti jevů musíme odečíst pravděpodobnost výsledků, které jsou příznivé oběma jevům.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \qquad (2.1)$$

Příklad 2.11

Student Jan a studentka Jana počítají současně příklad z matematiky. Každý ze studentů vypočítá správně příklad s pravděpodobností 0,5. Jaká je pravděpodobnost, že příklad bude vyřešen alespoň jedním studentem?

Řešení

Jev A … Jan vypočítá příklad … $P(A) = 0{,}5$
Jev B … Jana vypočítá příklad … $P(B) = 0{,}5$
V tomto případě se jedná o jevy slučitelné. K řešení použijeme vztah 2.1.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0{,}5 + 0{,}5 - 0{,}25 = 0{,}75$$

Pravděpodobnost, že příklad bude vyřešen alespoň jedním studentem, je 0,75.

8. Pro dva **neslučitelné jevy** A a B platí, že průnik jevů je jev nemožný. Prosté sčítání pravděpodobnosti je možné pouze v případě jevů, které se navzájem vylučují.

$$P(A \cup B) = P(A) + P(B) \tag{2.2}$$

Příklad 2.12

Jaká je pravděpodobnost, že při hodu kostkou padne číslo větší než 3 a padne číslo 2.

Řešení

V tomto případě se jedná o jevy neslučitelné. K řešení použijeme vztah 2.2.

$P(A) = P(4) + P(5) + P(6) = 1/6 + 1/6 + 1/6 = 3/6$

$P(B) = P(2) = 1/6$

$$P(A \cup B) = \frac{3}{6} + \frac{1}{6} = \frac{4}{6} = 0{,}66\overline{6} \cong 67\ \%$$

9. Pro pravděpodobnost průniku dvou **závislých** jevů platí

$$P(A \cap B) = P(A) \cdot P(B|A) = P(B) \cdot P(A|B)^{1} \tag{2.3}$$

Pravděpodobnost průniku jevů nám umožňuje vyjádřit **věta o násobení pravděpodobností**. Abychom mohli toto pravidlo definovat, musíme nejdříve vysvětlit pojem **podmíněná pravděpodobnost**. Tyto pravděpodobnosti vyjadřují **závislost jevů**.

[1] podmíněná pravděpodobnost

Pravděpodobnosti $P(A|B)$ a je pravděpodobnost náhodného jevu A, která je ovlivněna podmínkou, že nastal nějaký náhodný jev B, který má nenulovou pravděpodobnost (vztah 2.4). Obdobně vyjadřujeme i pravděpodobnost $P(B|A)$ (vztah 2.5).

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad P(B) \neq 0 \qquad (2.4)$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \qquad P(A) \neq 0 \qquad (2.5)$$

Z uvedených vztahů vychází tzv. **pravidlo o násobení pravděpodobností** a lze jej rozšířit i na průnik více jevů. Pravděpodobnost průniku dvou libovolných jevů A a B je rovna součinu pravděpodobnosti jevu A a podmíněné pravděpodobnosti jevu B vzhledem k jevu A nebo součinu nepodmíněné pravděpodobnosti jevu B a podmíněné pravděpodobnosti jevu A vzhledem k jevu B vztah 2.3.

Příklad 2.13

Jaká je pravděpodobnost výběru hlíz při třídění určitého druhu brambor v případě, kdy brambor má více než dvě očka a jeho hmotnost je menší než 70 g. Bylo zjištěno, že pravděpodobnost, že brambor má více než dvě očka, je 0,7. Dalším tříděním brambor s více než dvěma očky podle hmotnosti bylo zjištěno, že pravděpodobnost výběru brambor o hmotnosti menší než 70 g je 0,5.

Řešení

Jev A … brambor má více než dvě očka … $P(A) = 0{,}7$

Jev B … brambory s 2 očky a hmotností menší než 70 g

Podmíněná pravděpodobnost $P(B|A) = 0{,}5$

V tomto případě se jedná o jevy závislé. K řešení použijeme vztah 2.3.

$$P(A \cap B) = P(A) \cdot P(B|A) = 0{,}7 \cdot 0{,}5 = 0{,}35 = 35\ \%$$

Pravděpodobnost výběru hlíz brambor, které mají více než dvě očka a jsou současně lehčí než 70 g, je 35 %.

10. Pro pravděpodobnost průniku dvou **nezávislých** jevů platí

$$P(A \cap B) = P(A) \cdot P(B) \qquad (2.6)$$

Pokud nastoupení jevu A neovlivňuje pravděpodobnost nastoupení jevu B, tj. $P(B|A) = P(B)$, neovlivňuje ani nastoupení jevu B pravděpodobnost jevu A, tj. $P(A|B) = P(A)$ znamená

to, že jevy A a B jsou **jevy nezávislé**. Pro nezávislé jevy A a B se pravidlo pro výpočet průniku zjednoduší a průnik jevů A, B je přímo jen součinem pravděpodobností těchto jevů.

Příklad 2.14

Jaká je pravděpodobnost, že student uspěje v akademickém roce 2024/25 u zkoušek ze Statistiky I (zimní semestr) a Statistiky II (letní semestr)? Víme, že v zimním semestru je pravděpodobnost úspěchu 0,75. V letním semestru je pravděpodobnost, že student zkoušku úspěšné vykoná, 0,8.

Řešení

Jev A … student uspěje u zkoušky ze Statistiky I … $P(A) = 0{,}75$

Jev B … student uspěje u zkoušky ze Statistiky II … $P(B) = 0{,}8$

Jedná se o jevy nezávislé. K řešení použijeme vztah 2.6.

$$P(A \cap B) = P(A) \cdot P(B) = 0{,}75 \cdot 0{,}8 = 0{,}6$$

2.3.2 Statistická definice pravděpodobnosti

Statistická definice pravděpodobnosti (empirická, četnostní) je založena na předpokladu n-krát opakování nezávislého náhodného pokusu, při kterém nastoupí sledovaný jev m-krát. Při rostoucím počtu pokusů kolísá pozorovaná relativní četnost (m/n) jevu A stále v užších mezích (ustaluje se na určité hodnotě) kolem určitého čísla, toto číslo můžeme považovat za pravděpodobnost jevu A nebo číslo blízké této pravděpodobnosti. Hodnotu pravděpodobnosti nastolení jevu nemůžeme přesně experimentálně zjistit, ale můžeme se jí přiblížit prodloužením posloupnosti prováděných pokusů.

Statistická definice pravděpodobnosti jevu A je definována vztahem:

$$P(A) = \lim_{n \to +\infty} \frac{m}{n} \tag{2.7}$$

m …. počet pokusů, ve kterých nastal jev A,

n …. počet všech pokusů.

Výhodou statistické definice pravděpodobnosti je skutečnost, že v případě, kdy jsme schopni zajistit, aby měření náhodných pokusů probíhalo za stejných vstupních podmínek, máme předpověď nastolení náhodného jevu podloženou reálným měřením. Určitou nevýhodou je pak skutečnost, že nikdy nejsme schopni experimentálně zjistit přesnou hodnotu limity, počet

pokusů je vždy limitně konečný a tím se vylučuje předpoklad, že každý elementární jev má stejnou možnost nastolení.

Příklad 2.15

Vypočítejte, jaká je pravděpodobnost, že při opakovaných nezávislých pokusech hodem kostkou padne číslo větší než 3. Při statistickém výpočtu pravděpodobnosti vycházíme z opakování experimentu.

Řešení

Náhodný jev A … „padne číslo> 3“ je $A = \{\omega_4, \omega_5, \omega_6\}$

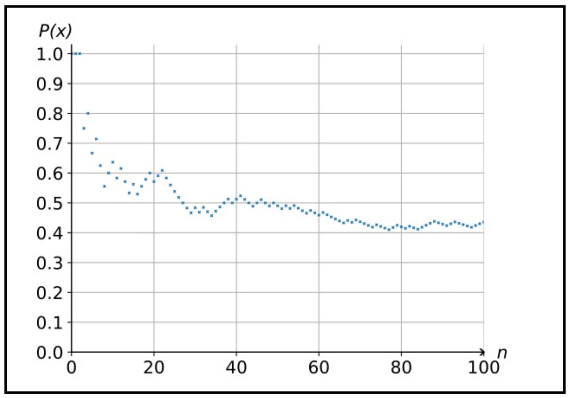
$$\boldsymbol{n = 10} \qquad P(A) = \frac{6}{10} = 0{,}6$$

$$\boldsymbol{n = 100} \qquad P(A) = \frac{43}{100} = 0{,}43$$

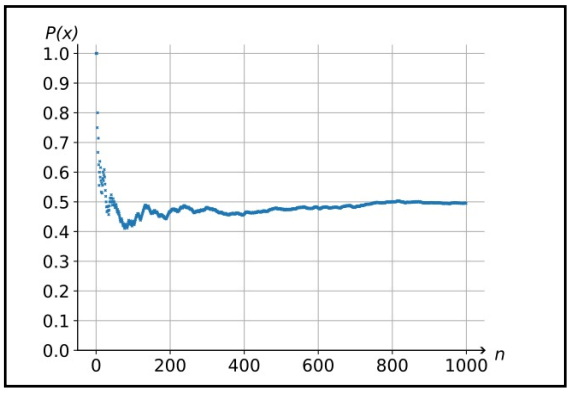
$$\boldsymbol{n = 1000} \qquad P(A) = \frac{495}{1000} = 0{,}495$$

Náhodný pokus není možno donekonečna opakovat, avšak při dostatečně velkém množství náhodných pokusů lze pravděpodobnost nastolení jevu s velkou přesností odhadnout. Toto pravidlo vychází ze ***slabého zákona velkých čísel***[2]**.** Zákon popisuje skutečnost, že s rostoucím počtem opakovaných nezávislých náhodných pokusů se empirické charakteristiky, které popisují výsledky těchto pokusů, blíží k teoretickým charakteristikám.

Obrázek 2.8: Počet pokusů *n* =100



Obrázek 2.9: Počet pokusů *n* =1000



Z grafického znázornění příkladu 2.15, které vychází z 1000 opakování vyplývá, že u prvních 100 hodů kostkou (obrázek 2.8) je patrné, že relativní četnosti jevu A se stabilizuje a blíží se hodnotě 0,4. Z obrázku 2.9, ve kterém jsou zaznamenány výsledky u více než 1000 náhodných pokusů, je zřejmé, že hodnota relativní četnosti se ustaluje kolem hodnoty 0,5.

[2] ANDĚL, Jiří. *Základy matematické statistiky*. Vyd. 3. Praha: Matfyzpress, 2011. ISBN 978-80-7378-162-0.

2.3.3 Klasická definice pravděpodobnosti

Klasická definice pravděpodobnosti je historicky nejstarším způsobem zavedení pravděpodobnosti. Definice je založena na pojmu stejného množství výskytu všech jevů a je použitelná pouze v případě, že množina všech elementárních jevů je konečná (můžeme vyjádřit množinu všech čísel) a žádné dva výsledky nemohou nastat současně. I přes tuto skutečnost, je tato definice často využívána k výpočtu pravděpodobnosti.

Může-li určitý náhodný pokus vykázat konečný počet n různých výsledků, které jsou stejně možné, a jestliže m těchto výsledků má za následek nastoupení jevu A, kdežto zbývajících $n - m$ výsledků realizaci jevu A vylučuje, pak pravděpodobnost jevu A je definována vztahem:

$$P(A) = \frac{m}{n} \tag{2.8}$$

m … $\|A\|$ počet prvků množiny A počet pokusů příznivých pro opakování jevu A,

n …. $\|\Omega\|$ počet prvků množiny všech jevů (počet všech pokusů).

Příklad 2.16

Vycházíme z příkladu 1.3. Jaká je pravděpodobnost, že při jednom hodu kostkou:

a) Náhodný jev A „padne sudé číslo“.

b) Náhodný jev B „padne číslo ≥ 3“

Řešení

a) $n = 6, m = 3 \quad P(A) = \frac{3}{6} = 0{,}5 = 50\ \%$

b) $n = 6, m = 4 \quad P(A) = \frac{4}{6} = 0{,}66\overline{6} \cong 67\ \%$

Shrnutí kapitoly

Axiomatická definice pravděpodobnosti → teoretická, vychází z teoretických předpokladů, jak se bude náhodný jev za určitých předpokladů chovat. Nemáme možnost provést měření a nemáme ani naměřené hodnoty.

Konkrétní výpočet pravděpodobnosti je možný např. na základě definice pravděpodobnosti statistické nebo klasické.

Statistická definice pravděpodobnosti → empirická, je založena na pozorování. Popisuje budoucí náhodnost nastolení jevu na základě již uskutečněných náhodných pokusů, kterých je limitně mnoho → Ω limitně *„nekonečně konečná“*.

Klasická definice pravděpodobnosti → empirická, je založena na pozorování. Popisuje budoucí náhodnost nastolení před provedením náhodných pokusů. Všechny elementární výsledky mají stejnou šanci nastat a množina všech elementárních jevů je konečná → Ω konečná množina prvků.

Literatura

ANDĚL, Jiří. *Základy matematické statistiky*. Vyd. 3. Praha: Matfyzpress, 2011. ISBN 978-80-7378-162-0.

BUDÍKOVÁ, Marie; Maria KRÁLOVÁ a Bohumil MAROŠ. *Průvodce základními statistickými metodami*. Vydání první. Praha: Grada Publishing, a.s., 2010. ISBN 978-80-247-3243-5.

HEBÁK, Petr a Jana KAHOUNOVÁ. 2014. *Počet pravděpodobnosti v příkladech*. Praha: Informatorium. ISBN 978-80-7333-109-2.

2 Kontrolní otázky

1. Co je náhodný jev?
2. Jaký je rozdíl mezi jevem jistým a jevem nemožným?
3. Co je základní prostor a jak se značí?
4. Uveďte příklad základního prostoru pro hod hrací kostkou.
5. Popište sjednocení jevů a uveďte příklad.
6. Popište rozdíl jevů a uveďte příklad.
7. Popište průnik jevů a uveďte příklad.
8. Co znamená, když je jev A_1 podjevem jevu A_2?
9. Kdy jsou jevy rovnocenné?
10. Kdy jsou dva jevy neslučitelné?
11. Co je opačný jev?
12. Co vyjadřuje pravděpodobnost náhodného jevu?
13. Kdo formuloval axiomatickou teorii pravděpodobnosti?
14. Jaké jsou vlastnosti pravděpodobnosti?
15. Kdy jsou jevy A a B nezávislé?
16. Na čem je založena statistická definice pravděpodobnosti?
17. Na čem je založena klasická definice pravděpodobnosti?
18. Kdy je klasická definice pravděpodobnosti použitelná?

2 Příklady k procvičení

2.1 Náhodný jev a základní prostor:

2.1.1 Z pytlíku plného modrých, červených a zelených kuliček náhodně vytáhnete jednu kuličku.

a) Co je v tomto případě náhodný pokus?
b) Jaký je základní prostor Ω pro tento pokus?
c) Uveďte příklad jednoho elementárního jevu.
d) Uveďte příklad jistého jevu.
e) Uveďte příklad nemožného jevu.

2.2 Základní operace a vztahy mezi náhodnými jevy

2.2.1 Ptáme se zákazníka, jak je spokojený s novým smartphonem.

Jev A: "*Zákazník je spokojen s výdrží baterie"*

Jev B: "*Zákazník je spokojen s kvalitou fotoaparátu*"

Definujte sjednocení jevu A a B.

2.2.2 Sledujeme zápis jednoho studenta do předmětů Statistika a Informatika. Volba předmětů.

Jev A: *"Student si zapsal kurz Statistika"*

Jev B: *"Student si zapsal kurz Informatika"*

Definujte průnik jevu A a B.

2.2.3 Student si vybírá obor na univerzitě.

Jev A: *"Student si vybral obor Elektrotechnika "*

Jev B: *"Student si vybral obor Strojírenství "*

Definujte neslučitelnost jevu A a B.

2.2.4 Sledujeme stav semaforu.

Jev A: *"Semafor svítí červeně"*

Definujte opačný jev jevu A.

2.3 Pravděpodobnost náhodného jevu

2.3.1 V obchodě 70 % zákazníků kupuje chléb (jev A) a 40 % zákazníků kupuje mléko (jev B). 30 % zákazníků kupuje chléb i mléko. Jaká je pravděpodobnost, že náhodně vybraný zákazník koupí chléb nebo mléko (nebo obojí)?

[0,80]

2.3.2 Při výběru jednoho studenta z kurzu je pravděpodobnost, že student je z prvního ročníku 0.6, a pravděpodobnost, že student je z druhého ročníku je 0.4. (Student nemůže být současně z prvního i druhého ročníku). Jaká je pravděpodobnost, že náhodně vybraný student je z prvního nebo druhého ročníku?

[1]

2.3.3 Ve třídě je 60 % dívek. Z dívek 30 % nosí brýle. Jaká je pravděpodobnost, že náhodně vybraný student je dívka a nosí brýle?

[0,18]

2.3.4 Dva stroje pracují nezávisle na sobě. Pravděpodobnost, že stroj A pracuje bez poruchy je 0.9. Pravděpodobnost, že stroj B pracuje bez poruchy je 0.8. Jaká je pravděpodobnost, že oba stroje budou pracovat bez poruchy současně?

[0,72]

2.3.5 Provedli jste 50 hodů symetrickou mincí a panna padla 28krát.

a) Jaká je statistická pravděpodobnost, že padne panna na základě tohoto experimentu?

[0,56]

b) Jak by se změnila tato pravděpodobnost, kdybyste provedli 1000 hodů místo 50?

[cca 0,50]

2.3.6 V balíčku 32 mariášových karet jsou 4 esa. Jaká je pravděpodobnost, že náhodně vytažená karta bude eso?

[0,125 %]

2.3.7 Určete pravděpodobnost, že při hodu třemi stejnými mincemi padne:

a) dvakrát rub a jednou líc; [0,375]

b) třikrát líc; [0,125]

c) na všech mincích stejná strana? [0,25]

2.3.8 V nákupní tašce máme osm banánů. Šest banánů je zralých a dva banány jsou nahnilé. Jaká je pravděpodobnost, že z 8 banánů vytáhneme náhodně:

a) jeden nahnilý banán; [0,25]

b) jeden zralý banán? [0,75]

2.3.9 Jan a Zuzka jsou na procházce v parku. Potkají babičku a ta se jich zeptá, jestli mají napsané úkoly. Předpokládáme, že Jan zalže s pravděpodobností 0,2 a Zuzka s pravděpodobností 0,3.

a) Zeptáme-li se obou nezávisle na sobě, zda mají napsané úkoly, jaká je pravděpodobnost, že budou lhát? [0,06 %]

b) Zeptáme-li se obou nezávisle na sobě, zda mají napsané úkoly, jaká je pravděpodobnost, že řeknou pravdu? [0,56 %]

2.3.10 Studen jde do studovny, kde je učebna s 25 počítači. Jaká je pravděpodobnost, že ve studovně bude volný alespoň jeden počítač za předpokladu, že pravděpodobnost obsazení PC ve studovně je 0,90. [0,928 %]

3 NÁHODNÉ VELIČINY

V této kapitole se budeme věnovat číselnému vyjádření výsledků náhodného pokusu.

Příklad 3.1

V náhodně vybrané porodnici se během jednoho dne narodily tři děti. Definujte základní prostor elementárních jevů, které se vztahují k možným variantám pohlaví narozených dětí (děvče D; kluk K) a podmnožinu náhodného jevu narození alespoň jedné dívky.

Řešení

Množina všech možných výsledků se skládá z osmi elementárních jevů:

$\Omega = \{\omega_{KKK}, \omega_{KKD}, \omega_{KDK}, \omega_{DKK}, \omega_{DDK}, \omega_{DKD}, \omega_{KDD}, \omega_{DDD}\}$

Z této množiny můžeme definovat podmnožinu náhodných jevů narození alespoň jedné dívky: $\Omega_D = \{\omega_{KKD}, \omega_{KDK}, \omega_{DKK}, \omega_{DDK}, \omega_{DKD}, \omega_{KDD}, \omega_{DDD}\}$

Tento výčet výskytu všech elementárních jevů plynoucích z náhodného pokusu neposkytuje informaci o pravděpodobnosti daného jevu, ale jen o možných kombinacích jeho výskytu. Takový kvalitativní výrok je často nepostačující. Výsledky náhodných pokusů je třeba kvantifikovat neboli číselně vyjádřit (počet dívek, počet bodů z testů, počet bodů na hrací kostce).

Příklad 3.2

Vycházíme z příkladu 3.1. Přiřaďte každému z možných výsledků náhodného pokusu hodnotu, která odpovídá počtu narozených dívek.

Řešení

Náhodná veličina X počet narozených dívek je realizována pouze konečně mnoha jevy $KKK, KKD, KDK, DKK, DDK, DKD, KDD, DDD$.

Počet dívek	Možné výsledky náhodného pokusu
0	KKK
1	$KKD; KDK; DKK$
2	$DDK; DKD; KDD$
3	DDD

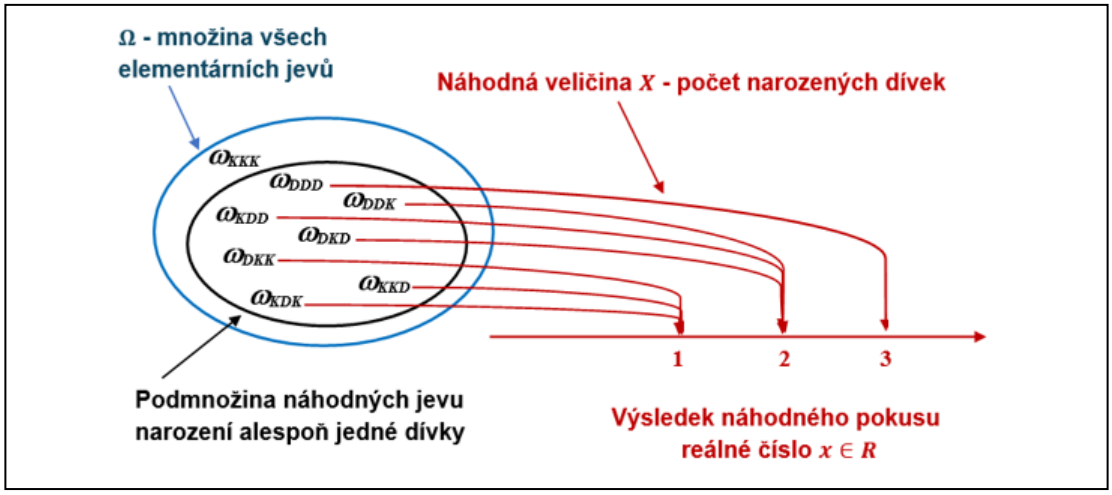
Kvantitativní charakteristiku náhodného pokusu nazýváme náhodná veličina. Náhodnou veličinu definujeme jako proměnnou, která nabývá různých hodnot v závislosti na náhodě. Náhodná veličina je reálné číslo a je jednoznačně určena výsledkem náhodného pokusu.

Náhodnou veličinu obvykle označujeme velkými písmeny z konce abecedy (X, Y). Konkrétní hodnoty náhodné veličiny značíme odpovídajícími malými písmeny $(x: x_1, x_2, \dots ; y: y_1, y_2, \dots)$.

Náhodná veličina *X* je reálná funkce $X(\omega)$ definovaná na množině všech elementárních jevů $\omega \in \Omega$, která každému možnému výsledku náhodného pokusu přiřadí reálné číslo $(x \in R)$ z množiny možných reálných hodnot.

Množina všech hodnot $\{x = X(\omega), \omega \in \Omega\}$ se nazývá **základní soubor**. Celý základní prostor Ω často není znám (množina Ω může být i nekonečná). Výhodou náhodné veličiny X je, že převádí základní prostor možných výsledků náhodných jevů na čísla. Tyto hodnoty „čísla" nám následně slouží k popisu vlastností základního souboru (obr. 3.1).

Obrázek 3.1: Grafické znázornění příkladu 3.1 a 3.2



Z hlediska pravidel matematických operací a analýz rozlišujeme dva typy náhodných veličin diskrétní a spojité.

Spojitá náhodná veličina X nabývá všech hodnot z konečného či nekonečného intervalu. Příkladem může být výška rostliny, hmotnost zvířete, životnost žárovky, atmosférický tlak.

Diskrétní (nespojitá) náhodná veličina X nabývá od sebe vzájemně oddělené hodnoty. Je to taková veličina, jejíž obor hodnot neboli množina všech čísel, kterým se může rovnat je konečná, nebo nanejvýš spočtená (počet možných hodnot je nekonečný, ale lze je uspořádat do posloupnosti). Jako příklad můžeme uvést počet zkoušek za semestr, počet pokusů u zkoušek – diskrétní náhodná veličina s konečným oborem hodnot. Počet přihrávek fotbalistů, než padne gól – diskrétní náhodná veličina s nekonečným oborem hodnot.

Diskrétní náhodná veličina se netýká pouze kvantitativních dat, neboť číselné vyjádření výsledku náhodného pokusu může popisovat i kvalitativní data, jako např.: stupeň vzdělání, známky ve škole atd.

K úplnému charakterizování náhodné veličiny je potřeba znát informace o množině možných hodnot, ale také je důležité znát **pravděpodobnost, s jakou náhodná veličina nabude určité hodnoty nebo hodnoty z určitého intervalu.** Náhodná veličina je tedy z pravděpodobnostního hlediska zcela popsána, jestliže známe její hodnoty či intervaly hodnot a pravděpodobnosti těchto hodnot nebo intervalů. Výskyt hodnot náhodné veličiny podléhá určitým zákonitostem. Zákonitost výskytu hodnot náhodné veličiny vyplývá z rozdělení hodnot náhodné veličiny.

Každý předpis, který určuje vztah mezi možnými hodnotami náhodné veličiny a pravděpodobností jejich výskytu, se nazývá **zákon rozdělení náhodné veličiny.** Jinými slovy rozdělení náhodné veličiny je pravidlo, které každé hodnotě nebo každému intervalu hodnot přiřazuje pravděpodobnost, že náhodná veličina nabude této hodnoty nebo hodnoty z tohoto intervalu – hovoříme o **teoretickém rozdělení pravděpodobnosti.**

Zákon rozdělení náhodné veličiny neboli teoretické rozdělení pravděpodobnosti můžeme ve statistice charakterizovat pomocí:

- **průběhu funkce** – průběhem funkce v tomto případě rozumíme určení vlastností funkce a její následné grafické vyjádření.
- **parametrů rozdělení** – jedná se o číselné charakteristiky náhodných veličin, které nám poskytují souhrnnou informaci o určitém chování náhodných veličin (viz kapitola Číselné charakteristiky náhodných veličin).

3.1 Distribuční funkce

Distribuční funkce kumulativním způsobem popisuje rozdělení diskrétních i spojitých náhodných veličin. Distribuční funkcí náhodné veličiny X nazýváme funkci $F(x)$, která je pro všechny reálné hodnoty x definována vztahem (3.1).

Distribuční funkce přiřazuje každému reálnému číslu pravděpodobnost, že náhodná veličina X nabude hodnoty menší nebo rovno x

$$F(x) = P(X \leq x) = P(\{\omega: X(\omega) \leq x\}) \tag{3.1}$$

a

$$P\ (a < X \leq b) = \ F_x(b) - F_x(a), \text{ pro libovolná reálná čísla } a, b, \text{ kde } a \leq b \tag{3.2}$$

Vlastnosti distribuční funkce

Obor funkčních hodnot distribuční funkce leží mezi 0 a 1, tj.

$0 \leq F(x) \leq 1.$

Distribuční funkce libovolné náhodné veličiny je neklesající, tj. pro libovolné $a, b \in R$, $a \leq b$, platí: $F_x(a) \leq F_x(b)$.

Pro každou distribuční funkci platí:

$F(-\infty) = \lim_{x \to -\infty} F(x) = 0; F(+\infty) = \lim_{x \to +\infty} F(x) = 1.$

Distribuční funkce je zprava spojitá v libovolném bodě $x \in R$. Lze se setkat i s definicí, která vychází z funkce zleva spojité, ale vzhledem k tomu, že budeme využívat při výpočtu statistické programy, budeme vycházet ze stejného předpisu, jako je v softwaru a vždy budeme používat funkci spojitou zprava.

U distribuční funkce je teoretický předpis, který definuje pravděpodobnost pro náhodnou veličinu X. V případě, že neznáme přesné vyjádření teoretického předpisu funkce, můžeme ji aproximovat **výběrovou distribuční funkci $F_n(x)$**, která kumulativním způsobem popisuje pravděpodobnostní chování pozorovaných hodnot za splnění předpokladu reprezentativnosti experimentálního vzorku pozorování. Z hodnot výběrové distribuční funkce a jejího grafického znázornění můžeme usuzovat na vlastnosti teoretické distribuční funkce (viz kapitola Statistická indukce).

3.1.1 Funkce rozdělení diskrétní náhodné veličiny

Distribuční funkce $F(x)$ diskrétní náhodné veličiny je funkce schodovitého tvaru (obr. 3.2). Jedná se o funkci, která je mezi jednotlivými body x_i konstantní, a právě v každé hodnotě $x_1, x_2, \ldots, x_n$ dochází ke skoku. Výška skoku je rovna hodnotě pravděpodobnosti $p(x_i) \rightarrow p_i$. Body vyznačené prázdným kolečkem znamenají, že distribuční funkce není určena v bodě skoku, ale až výše na úrovni dalšího schodu, kde je vyznačena plným kolečkem, tedy tak, že funkce $F(x)$ je zprava spojitá – uzavřený interval je ten levý a otevřený ten pravý.

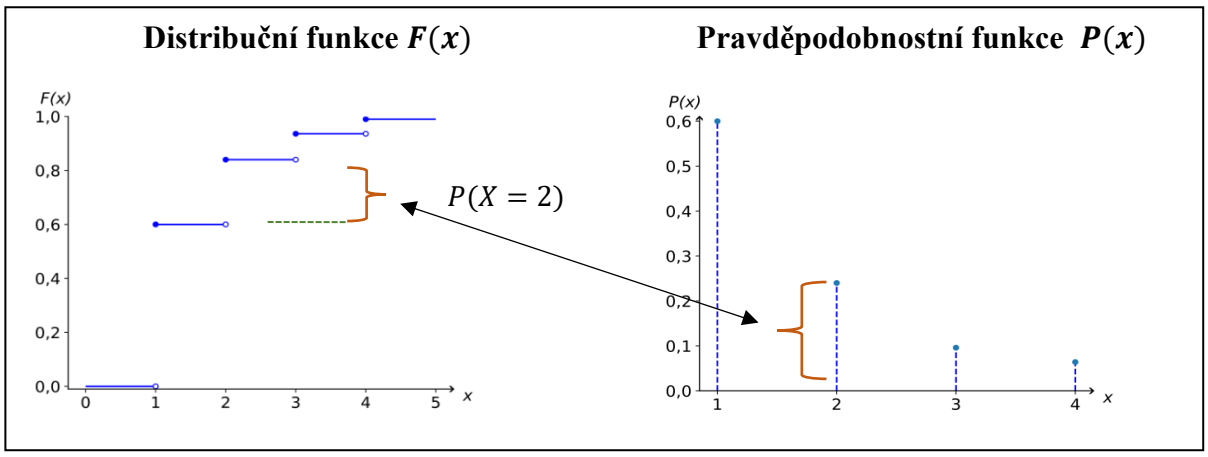
Diskrétní náhodnou veličinu X s distribuční funkcí $F(x)$ charakterizuje pravděpodobnostní funkce, která vychází ze vztahu 3.1.

$$F(x) = P(X \leq x) \tag{3.3}$$

Pravděpodobnostní funkce $P(x)$ je nejjednodušší formou vyjádření zákona rozdělení diskrétní náhodné veličiny (obr. 3.3). Pravděpodobnostní funkce udává pravděpodobnost, že náhodná veličina X nabude právě hodnoty x.

$$P(x) = P(X = x) \tag{3.4}$$

Obrázek 3.2 a 3.3: Funkce rozdělení diskrétníá náhodné veličiny z příkladu 3.3



Pravděpodobnostní funkce nabývá hodnot od nuly do jedné, což plyne z axiomu o pravděpodobnosti náhodného jevu: $0 \leq P(x) \leq 1$; součet všech hodnot pravděpodobnostní funkce je roven jedné: $\sum P(x) = 1$.

Roztřídění hodnot a uspořádání distribuční funkce a pravděpodobnostní funkce lze provést různými způsoby. V příkladu 3.2 a 3.3 si budeme ilustrovat **matematický zápis**, **tabulku rozdělení** a **graf**.

Příklad 3.3

Zákazník si jde koupit zboží, které bývá na trhu k dostání s pravděpodobností 0,6. Je rozhodnutý, že navštíví maximálně čtyři prodejny. Náhodný jev můžeme označit A_i $(i = 1; 2; 3; 4)$. Tento jev nastane v případě, když v *i*-té prodejně je požadované zboží $p(A_i) = 0{,}6$. Pro jednoduchost budeme předpokládat, že jevy $A_1, \ldots, A_4$ jsou skupinově nezávislé.

a) Jaká je pravděpodobnost, že zákazník koupí zboží už v první, nebo až v druhé, třetí či čtvrté prodejně.

b) Jaká je pravděpodobnost, že zákazník si koupí zboží v první až čtvrté prodejně.

Pokračování příkladu 3.3

Náhodná veličina X představuje počet navštívených prodejen. Výsledky realizace náhodné veličiny mohou nabývat v tomto případě konkrétních hodnot $x = 1; x = 2;\ x = 3; x = 4$.

Počet navštívených prodejen – **diskrétní náhodnou veličinu**.

ad a) Hodnoty pravděpodobnosti vyjádříme pomocí **pravděpodobnostní funkce $P(x)$**.

Matematický zápis

Zákazník navštíví pouze jednu prodejnu, kde mají požadované zboží.

$$p_1 = P(X = 1) = P(A_1) = \mathbf{0,6}$$

Zákazník navštíví dvě prodejny; v první zboží nemají, ve druhé ano.

$$p_2 = P(X = 2) = P(\bar{A}_1 \cap A_2) = P(\bar{A}_1) \cdot P(A_2) = 0{,}4 \cdot 0{,}6 = \mathbf{0,24}$$

Zákazník navštíví tři prodejny; v prvních dvou neuspěje, ve třetí ano.

$$p_3 = P(X = 3) = P(\bar{A}_1) \cdot P(\bar{A}_2) \cdot P(A_3) = 0{,}4^2 \cdot 0{,}6 = \mathbf{0,096}$$

Zákazník navštíví čtyři prodejny; v prvních třech neuspěje (bez ohledu na výsledek návštěvy čtvrté prodejny). $p_4 = P(X = 4) = P(\bar{A}_1) \cdot P(\bar{A}_2) \cdot P(\bar{A}_3) = 0{,}4^3 = \mathbf{0,064}$

$$\sum_{i=1}^{n} p_i = p_1 + p_2 + p_3 + p_4 = 0{,}6 + 0{,}24 + 0{,}096 + 0{,}064 = 1$$

Tabulka rozdělení pravděpodobností

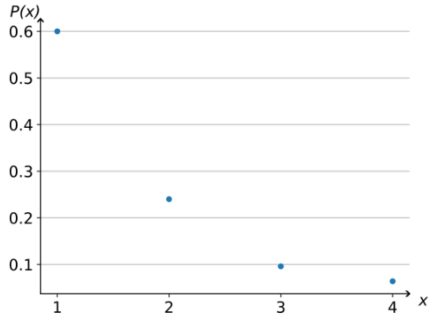
Tato tabulka se také nazývá **řada rozdělení** a znázorňuje nám množinu možných variant realizací (variační řadu x_i) a konkrétní hodnoty výsledků uvedených realizací (odpovídající pravděpodobnost náhodné veličiny p_i). Výsledky realizace náhodné veličiny X počet navštívených prodejen jsou vyjádřeny v následující tabulce.

x_i	1	2	3	4	$\sum$
p_i	0,6	0,24	0,096	0,064	1

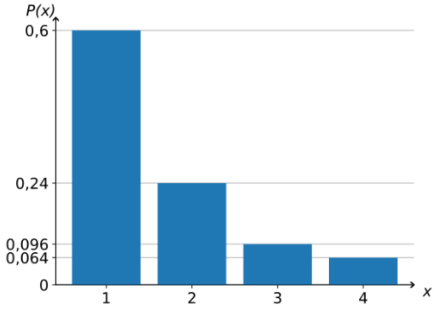
Graf

Znázornění pravděpodobnostní funkce $P(x)$ pomocí bodového a sloupcového grafu.

Bodový graf



Sloupcový graf



Pokračování příkladu 3.3

ad b) Kumulaci hodnot pravděpodobnosti náhodné veličiny budeme definovat pomocí **distribuční funkce $F(x)$**.

Zákazník si koupí zboží v první prodejně.

$$F(1) = P(X \leq 1) = P(X = 1) = P(1) = 0{,}6$$

Zákazník si koupí zboží v první prodejně nanejvýše ve druhé prodejně.

$$F(2) = P(X \leq 2) = P(1) + P(2) = 0{,}6 + 0{,}24 = 0{,}84$$

Zákazník si koupí zboží v první prodejně, ve druhé prodejně a nanejvýš ve třetí prodejně.

$$F(3) = P(X \leq 3) = P(1) + P(2) + P(3) = 0{,}6 + 0{,}24 + 0{,}096 = 0{,}936$$

Zákazník si koupí zboží v první, ve druhé, ve třetí a nanejvýš ve čtvrté prodejně.

$$F(3) = P(X \leq 4) = P(1) + P(2) + P(3) + P(4) = 0{,}6 + 0{,}24 + 0{,}096 + 0{,}064 = 1$$

Distribuční funkci zapisujeme následovně

$F(x)$	0	pro $x \in (-\infty; 1)$
	0,6	pro $x \in \langle 1; 2)$
	0,84	pro $x \in \langle 2; 3)$
	0,936	pro $x \in \langle 3; 4)$
	1	pro $x \in \langle 4; \infty)$

Grafické znázornění distribuční funkce $F(x)$ a pravděpodobnostní funkce je na obrázcích 3.2 a 3.3.

3.1.2 Funkce rozdělení spojité náhodné veličiny

Distribuční funkci $F(x)$ **spojité náhodné veličiny** nelze popsat pravděpodobnostní funkcí v určitém bodě. Spojité náhodné veličiny mohou nabývat všech hodnot z určitého intervalu, a proto grafem distribuční funkce spojité náhodné veličiny je spojitá křivka (příklad 3.3). U spojité náhodné veličiny pravděpodobnost roste spojitě s obsahem plochy.

Rozdělení pravděpodobnosti spojité náhodné veličiny se určuje prostřednictvím funkce, kterou označujeme **hustota rozdělení pravděpodobnosti**.

$$F(x) = P(X \leq x) = P(X \in (-\infty, x\rangle) = \int_{-\infty}^{x} f(t)dt \tag{3.5}$$

Hustotu rozdělení pravděpodobnosti $f(x)$ získáme derivováním distribuční funkce.

$$f(x) = F'(x) = \frac{d}{dx}F(x) \tag{3.6}$$

Funkční hodnota $F(x)$ spojité náhodné veličiny vyjadřuje obsah plochy z hustoty $f(t)$ od mínus nekonečna do hodnoty x. Hustota pravděpodobnosti nabývá nezáporných hodnot

$f(x) \geq 0$; integrál hustoty pravděpodobnosti přes všechny hodnoty, kterých může náhodná veličina nabýt (celková plocha pod funkcí), je roven jedné: $\int_{-\infty}^{\infty} f(x)dx = 1$.

Příklad 3.4

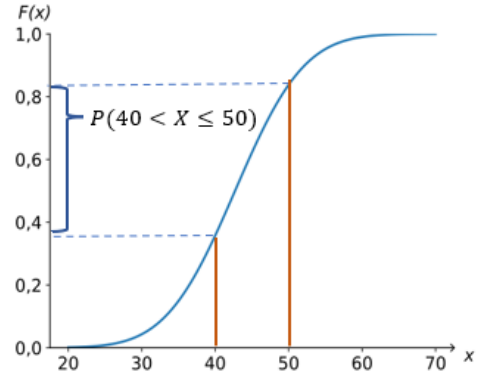
Průměrná čistá mzda zaměstnanců nejmenovaného národního koncernu je 42,7 tisíc Kč, se směrodatnou odchylkou 7,4 tisíce Kč. Kolik procent zaměstnanců má čistou mzdu v intervalu od 40 do 50 tisíc Kč?

Mzda zaměstnanců – spojitá náhodná veličina, která se řídí normálním rozdělením.

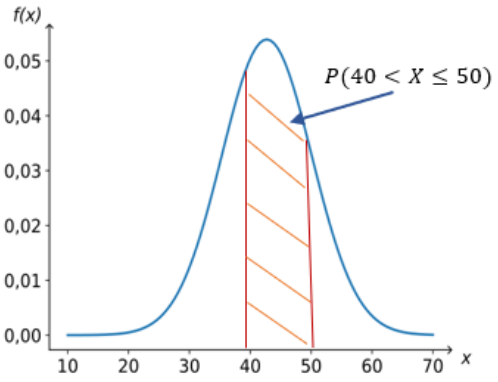
Řešení

Pro výpočet pravděpodobnosti spojité náhodné veličiny X použijeme vzorec odvozený ze vztahu 3.2: $P(a < X \leq b) = = P(40 < X \leq 50) = F(50) - F(40) = 0{,}83805278 - 0{,}35760614 = 0{,}48044664 =$ **48,04 %**

Distribuční funkce $F(x)$



Hustota rozdělení $f(x)$



Hodnoty distribuční a kvantilové funkce jsou tabelovány z důvodů matematické náročnosti výpočtu.

3.2 Kvantilová funkce

Z distribuční funkce lze jednoznačně určit funkci kvantilovou. Kvantilová funkce náhodné veličiny je teoretická funkce, která je charakteristikou rozdělení náhodné veličiny X.

U **spojitých náhodných veličin**, kde je distribuční funkce $F(x)$ oboustranně spojitá, je možné p-kvantil jednoznačně určit (vztah 3.7). Výsledkem není pravděpodobnost, ale číslo na reálné ose, které odpovídá určité pravděpodobnosti.

$$P[X \leq Q_x(p)] = F[F^{-1}(p)] = \alpha, P[X \geq Q_x(p)] = 1 - p \qquad (3.7)$$

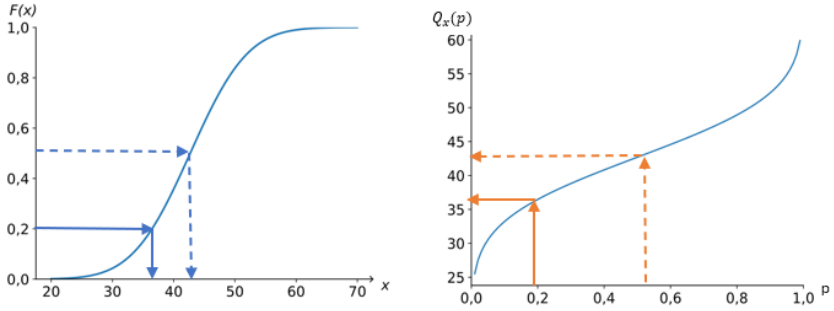
Příklad 3.5

Vycházíme z příkladu 3.3. Průměrná čistá mzda zaměstnanců nejmenovaného národního koncernu je 42,7 tisíc Kč se směrodatnou odchylkou 7,4 tisíce Kč. Jaká je čistá mzda prvních 20 % zaměstnanců[3]?

Řešení

Pro výpočet pravděpodobnosti spojité náhodné veličiny X (mzda zaměstnanců) použijeme vztah 3.6: $P[X \leq Q_x(0{,}2)] = F[F^{-1}(0{,}2)]$ =36,472 tisíc Kč.

Distribuční funkce $F(x)$ **Kvantilová funkce $Q_x(p)$**



U **diskrétních náhodných veličin**, kde je distribuční funkce $F(x)$ schodovitá (po částech konstantní), není možné p-kvantil jednoznačně určit.

$$P[X < Q_x(p)] = 1 - P[X \geq Q_x(p)] \leq p, \tag{3.8}$$

kde limita zleva distribuční funkce v bodě $Q_x(p)$ je $\leq p$.

Kvantilová funkce úzce souvisí s pojmem kvantil (viz kapitola Číselné charakteristiky náhodných veličin).

3.3 Číselné charakteristiky náhodných veličin

Distribuční funkce, pravděpodobnostní funkce a hustota pravděpodobnosti[4] nám poskytují úplnou, ale ne na první pohled, přehlednou informaci o charakteru rozdělení náhodné veličiny. Z tohoto důvodu se k popisu tvaru rozdělení používají **číselné charakteristiky**, které nám umožňují shrnout informace o náhodné veličině do několika hodnot (čísel). Jedná se o charakteristiky (parametry), které nám poskytují souhrnnou informaci o chování náhodných veličin.

Jejich konstrukce vychází ze dvou základních principů: **momentového** a **kvantilového**.

[3] Kvintil dělí statistický soubor na pět stejných dílů. 20 % prvků souboru má hodnoty menší (nebo rovné) hodnotě prvního kvintilu.
[4] Funkcionální charakteristiky

3.3.1 Momentové charakteristiky náhodných veličin

Zavedení pojmu momenty umožnuje definovat nejdůležitějších statistické charakteristiky, které označujeme jako momentové charakteristiky.

Jedná se o charakteristiky náhodné veličiny vystupující jako číselné proměnné, které jsou založené na mocninách a součinech hodnot pozorování.

Obecný moment $\Rightarrow$ moment okolo počátku náhodné veličiny *X.*

Obecný moment *k*-tého řádu, který značíme $\mu_k'(X) = E(X)^k$, je definován vztahem:

pro diskrétní NV $$\mu_k'(X) = \sum_{i=1}^{\infty} x_i^k \cdot p_i \tag{3.9}$$

pro spojitou NV $$\mu_k'(X) = \int_{-\infty}^{\infty} x^k \cdot f(x)dx \tag{3.10}$$

Kde: k ……. stupeň momentu $k = 1, 2, 3\ldots$,

x_i ……. hodnoty náhodné veličiny *X*, pro $i = 1, 2, \ldots..$,

p_i ……. pravděpodobnost, že *X* nabývá hodnoty x_i,

$f(x)$ ….. hustota pravděpodobnosti *X*.

V teorii pravděpodobnosti je nejvíce používaný **první obecný moment náhodné veličiny *X*** (vztah 3.11; 3.12), který vyjadřuje nejčastěji používanou **charakteristiku polohy** neboli úrovně a který se nazývá **střední hodnota** náhodné veličiny $\Rightarrow \mu_1'(X) = E(X) = \mu$. První obecný moment udává střední hodnotu náhodné veličiny $E(X)$.

Diskrétní NV $k = 1$ $\mu_1'(X) = \sum_{i=1}^{\infty} x_i^k \cdot p_i = \sum_{i=1}^{\infty} x_i^1 \cdot p_i = \sum_{i=1}^{\infty} x_i \cdot p_i = E(X)$ (3.11)

Spojitá NV $k = 1$ $$\mu_1'(X) = \int_{-\infty}^{\infty} x^1 \cdot f(x)dx = E\,(X) \tag{3.12}$$

Definice obecného momentu je analogická s definicí centrálního momentu. Obecné momenty jsou ovšem vztaženy k počátku souřadného systému, zatímco centrální momenty jsou vycentrovány ke střední hodnotě (těžišti všech možných hodnot) náhodné veličiny.

Centrální moment $\Rightarrow$ moment okolo konstanty, kterou je první obecný moment.

Centrální moment *k*-tého řádu, který značíme $\mu_k(X) = E[X - E(X)]^k$, je definován vztahem:

pro diskrétní NV
$$\mu_k(X) = \sum_{i=1}^{\infty} [x_i - E(X)]^k \cdot p_i \qquad (3.13)$$

pro spojitou NV
$$\mu_k(X) = \int_{-\infty}^{\infty} [x - E(X)]^k \cdot f(x)dx \qquad (3.14)$$

Kde: $E(X)$ ….. střední hodnota.

Nejdůležitější charakteristikou z centrálních momentů je **druhý centrální moment** (vztah 3.15; 3.16) náhodné veličiny, který se nazývá **rozptyl** (disperze) $\Rightarrow \mu_2(X) = D(X) = \sigma^2$.

Diskrétní NV

k = 2
$$\mu_2(X) = \sum_{i=1}^{\infty} [x_i - E(X)]^2 \cdot p_i = \sum_{i=1}^{\infty} x_i^2 \cdot p_i - E(X)^2 = D(X) \qquad (3.15)$$

Spojitá NV

k = 2
$$\mu_2(X) = \int_{-\infty}^{\infty} [x - E(X)]^2 \cdot f(x)dx = D(X) \qquad (3.16)$$

Normovaný moment $\Rightarrow$ moment, který je bezrozměrný a neměnný, vůči aditivní i multiplikativní konstantě. Při normování vycházíme z číselných charakteristik polohy a variability náhodné veličiny (*X*).

Normovaný moment *k*-tého stupně $\bar{\mu}_k$ náhodné veličiny *X* je určen vztahem:
$$\bar{\mu}_k(X) = \frac{E[X - E(X)]^k}{\sqrt{[D(X)]}} \qquad (3.17)$$

Kde: $E(X)$ ….. střední hodnota náhodné veličiny.
$D(X)$ …. rozptyl náhodné veličiny.

Normovaný moment je výchozím počtem pro charakteristiky šikmosti a špičatosti, které se využívají k porovnání průběhu posuzovaného rozdělení pravděpodobností s průběhem normovaného normálního rozdělení pravděpodobností *N*(0,1) více v kapitole Rozdělení spojitých náhodných veličin. Normování se také využívá při standardizaci hodnot náhodných veličin.

Mezi základní vlastnosti níže uvedených číselných charakteristik náhodných veličin, které při svém výpočtu vycházejí z momentů, patří skutečnost, že všechny tyto charakteristiky jsou citlivé na extrémní hodnoty.

Střední hodnota náhodné veličiny

Střední hodnota náhodné veličiny určuje střed rozdělení náhodné veličiny, kolem kterého kolísají pozorované hodnoty náhodné veličiny.

Střední hodnota diskrétní náhodné veličiny (vztah 3.18) představuje těžiště soustavy všech hodnot náhodné veličiny, jejichž četnost je popsána pravděpodobnostní funkcí. Střední hodnota spojité náhodné veličiny (vztah 3.19) je pak těžištěm hmotné přímky, na níž je rozprostření všech možných hodnot náhodné veličiny popsáno hustotou pravděpodobností.

Střední hodnota *E*(*X*)

Diskrétní NV $$E(X) = \sum_{i=1}^{\infty} x_i \cdot p_i \qquad (3.18)$$

Spojitá NV $$E(X) = \int_{-\infty}^{+\infty} x \cdot f\,(x)dx \qquad (3.19)$$

Při mnohačetném opakování nezávislých náhodných pokusů, které jsou realizovány za stejných podmínek, je možné s velkou pravděpodobností očekávat, že mezi střední hodnotou náhodné veličiny *X* (teoretickou hodnotou) a aritmetickým průměrem (empirická hodnota) konkrétních hodnot *x*, které vycházejí z realizace pokusu, bude jen nepatrný rozdíl.

Pro střední hodnotu mimo jiné platí:

a) střední hodnota konstanty *c* je rovna této konstantě $E(c) = c$;
b) $E(c \cdot X) = c \cdot E(X)$;
c) střední hodnota součtu nebo rozdílu dvou náhodných veličin je rovna:
   $E(X \pm Y) = E(X) \pm E(Y)$;
d) střední hodnota součinu dvou nezávislých náhodných veličin *X* a *Y* je rovna:
   $E(X \cdot Y) = E(X) \cdot E(Y)$.

Pro ucelený popis pravděpodobnostního rozdělení náhodné veličiny *X* potřebujeme znát nejen střední hodnotu $E(X) = \mu$, ale také hodnotu variability $D(X) = \sigma^2$ hodnot kolem střední hodnoty.

Variabilita náhodné veličiny

Variabilita neboli proměnlivost hodnot náhodné veličiny je charakterizována **rozptylem**, který představuje **druhý centrální moment** a je definován jako hodnota kvadrátů odchylek od střední hodnoty a značí se $\mu_2(X) = D(X) = \sigma^2$.

Rozptyl $D(X)$

Diskrétní NV $D(X) = E[X - E(X)]^2 = E(X^2) - [E(X)]^2 = \sum_{i=1}^{\infty} x_i^2 \cdot p_i - [E(X)]^2$ (3.20)

Spojitá NV $D(X) = \int_{-\infty}^{+\infty} x^2 \cdot f(x)dx - [E(X)]^2$ (3.21)

Pro rozptyl mimo jiné platí:

a) rozptyl konstanty c je roven nule, $D(c) = c$;

b) pro libovolnou konstantu c platí $D(c \cdot X) = c^2 \cdot D(X)$;

c) rozptyl součtu dvou nezávislých náhodných veličin je roven: $D(X + Y) = D(X) + D(Y)$;

d) $\sqrt{D(X)} = \sqrt{\sigma^2} = \sigma(X)$… směrodatná odchylka vyjadřuje variabilitu v původních jednotkách náhodné veličiny.

Charakteristika šikmosti

Šikmost (asymetrie) nám udává nesouměrnost rozložení četností náhodné veličiny X. Objektivní míra k určení asymetrie vychází z třetího normovaného centrálního momentu.

Šikmost $\bar{\mu}_3(X)$

$$\bar{\mu}_3(X) = \frac{E[X - E(X)]^3}{\sqrt{[D(X)]^3}} \quad (3.22)$$

Charakteristika špičatosti

Špičatost je číselná charakteristika náhodné veličiny X, která souvisí s mírou koncentrace rozložení četností hodnot náhodné veličiny. Míra špičatosti vychází ze čtvrtého normovaného centrálního momentu.

Špičatost $\bar{\mu}_4(X)$

$$\bar{\mu}_4(X) = \frac{E[X - E(X)]^4}{\sqrt{[D(X)]^4}} \quad (3.23)$$

Příklad 3.6

Dlouhodobým pozorováním bylo zjištěno, že pravděpodobnost, že studenti půjčí v univerzitní knihovně jednu knihu, je 0,22, že si půjčí dvě knihy je 0,25, tři knihy si půjčí 0,30, čtyři knihy si odnese 0,13 studentů. Maximální počet knih, které si studenti mohou vypůjčit, je 5 kusů.

a) Vypočítejte, jaká je pravděpodobnost, že student si půjčí právě 5 knih.

b) Stanovte pravděpodobnostní funkci $P(x)$ a distribuční funkci $F(x)$.

c) Vypočítejte: střední hodnotu, rozptyl a směrodatnou odchylku.

Řešení příkladu 3.6

Náhodná veličina X …počet vypůjčených knih je nezávislá diskrétní náhodná veličina.

a) Hodnoty realizace náhodné veličiny X pravděpodobnostní a distribuční funkce

x_i	1	2	3	4	5
$P(x)$	0,22	0,25	0,30	0,13	**0,10**
$F(x)$	0,22	0,47	0,77	0,90	1

b) $E(X) = \sum_{i=1}^{n} x_i \cdot p_i = 1 \cdot 0{,}22 + 2 \cdot 0{,}25 + 3 \cdot 0{,}30 + 4 \cdot 0{,}13 + 5 \cdot 0{,}10 =$ **2,64**

$D(X) = \sum_{i=1}^{n} x_i^2 \cdot p_i - [E(X)]^2 = 1^2 \cdot 0{,}12 + 2^2 \cdot 0{,}25 + 3^2 \cdot 0{,}30 + 4^2 \cdot 0{,}13 + 5^2 \cdot 0{,}10 -$ $(2{,}64)^2 = (0{,}12 + 1 + 2{,}7 + 2{,}08 + 2{,}5) - 6{,}9696 =$ **1,1304**

3.3.2 Normovaná náhodná veličina

Normováním náhodné veličiny nazýváme takovou operaci, která spočívá ve zmenšení hodnot náhodné veličiny o první obecný moment (střední hodnotu) a dělením druhou odmocninou druhého centrálního momentu (směrodatnou odchylkou).

Normovanou (standardizovanou) náhodnou veličinu značíme U a definujeme ji vztahem

$$U = \frac{X - E(X)}{\sqrt{D(X)}} \tag{3.24}$$

Pro normovanou náhodnou veličinu U platí, že střední hodnota je rovna nule $E(U) = 0$ a rozptyl jedné $D(U) = 1$. Z uvedeného vyplývá, že u normovaných náhodných veličin nemá smysl rozlišovat obecné a centrální momenty. Obecně tedy nazýváme momenty normované náhodné veličiny **normované momenty**. Střední hodnota normovaného znaku je vždy rovna nule a směrodatná odchylka hodnotě jedna.

Všechny uvedené momentové charakteristiky vykazují menší citlivost vůči extrémním hodnotám, proto dále uvádíme charakteristiky, které označujeme jako robustní a jejich výpočet je založen na kvantilech.

3.3.3 Kvantilové charakteristiky náhodných veličin

Kvantily Q_p dělí hodnoty náhodné veličiny na části. Kvantily jsou inverzní funkcí k distribuční funkci a jsou určeny pořadím ve vzestupně uspořádaném souboru hodnot.

Je to taková hodnota náhodné veličiny X, která rozděluje uspořádaný soubor hodnot určité statistické proměnné na dvě části – jedna obsahuje ty hodnoty, které jsou menší nebo rovny p-

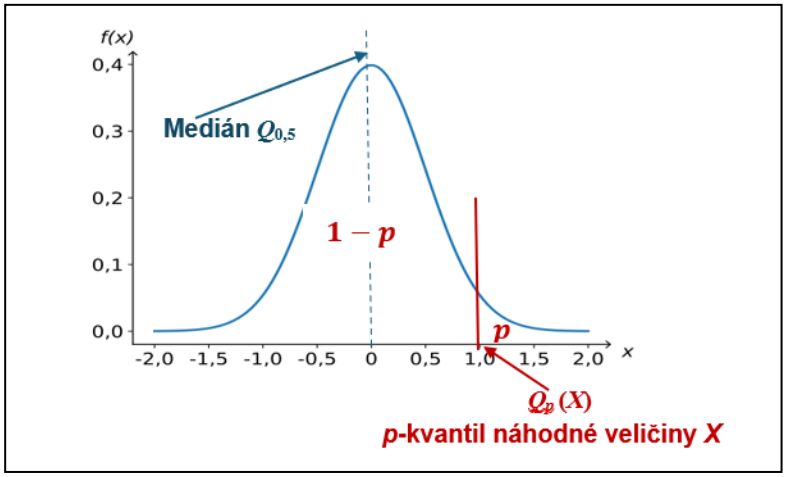
procentnímu kvantilu, druhá část obsahuje hodnoty, které jsou větší nebo rovny p-procentnímu kvantilu. Jestliže $p \in (0, 1)$, pak je číslo $Q_p(X)$, které splňuje následující vlastnosti:

$$P[X \geq Q_p] \geq 1 - p \tag{3.25}$$

$$P[X \leq Q_p] \geq p \tag{3.26}$$

p-kvantil dělí plochu pod grafem hustoty pravděpodobnosti v poměru p: $(1 - p)$.

Obrázek 3.4: Grafické znázornění kvantilů



Kvantily pro některé význačné hodnoty jsou označovány jmény a pro nejdůležitější rozdělení jsou hodnoty základních kvantilů tabelovány. Kvantil je bod (obr. 3.4), který rozděluje prostor hodnot náhodné veličiny v určitém pravděpodobnostním poměru. Volba kvantilů závisí na tom, jak podrobné informace o rozdělení pravděpodobností jsou požadovány.

Nejčastěji používané kvantily

Medián $Q_{0,5}$ rozdělující statistický soubor na dvě stejně početné množiny. U symetrického rozdělení je medián roven střední hodnotě, pokud střední hodnota existuje. Jeho hodnota je přirozeně interpretovatelná. **Kvartil** $Q_{0,25}$ (dolní kvartil), $Q_{0,5}$ (medián), $Q_{0,75}$ (horní kvartil) rozdělující hodnoty znaku na čtvrtiny. **Kvintil** dělí statistický soubor na pět stejných dílů. 20 % prvků souboru má hodnoty menší (nebo rovné) hodnotě prvního kvintilu, 80 % hodnoty větší (nebo rovné). **Decil** $Q_{p/10}$ dělí soubor na desetiny. **Percentil** $Q_{p/100}$ dělí soubor na setiny.

Kvantily jsou považovány za důležitý prostředek popisu celého pravděpodobnostního rozdělení a jejich znalost umožňuje konstruovat intervaly, do nichž hodnota náhodné veličiny padne s předem zvolenou pravděpodobností více viz kapitola Rozdělení spojitých náhodných veličin.

Shrnutí kapitoly

Náhodná veličina je proměnná, která nabývá různých hodnot v závislosti na náhodě. Náhodná veličina je reálné číslo, které je jednoznačně určeno výsledkem náhodného pokusu.

Diskrétní náhodná veličina X nabývá od sebe vzájemně oddělené hodnoty.

Spojitá náhodná veličina X nabývá všech hodnot z konečného či nekonečného intervalu.

Zákon rozdělení náhodné veličiny je pravidlo, které každé hodnotě nebo každému intervalu hodnot přiřazuje pravděpodobnost, že náhodná veličina nabude této hodnoty nebo hodnoty z tohoto intervalu.

Diskrétní náhodná veličina	**Spojitá náhodná veličina**
Distribuční funkce $\boldsymbol{F(x)}$	**Distribuční funkce** $\boldsymbol{F(x)}$
Pravděpodobnostní funkce $\boldsymbol{P(x)}$	**Hustota rozdělení pravděpodobnosti** $\boldsymbol{f(x)}$

Číselné charakteristiky poskytují souhrnnou informaci o chování náhodných veličin.

Jejich konstrukce vychází ze dvou základních principů: **momentového** a **kvantilového**.

Střední hodnota náhodné veličiny $\Rightarrow \mu_1'(X) = E(X) = \mu$.

Rozptyl neboli variabilita náhodné veličiny $\Rightarrow \mu_2(X) = D(X) = \sigma^2$.

Šikmost udává nesouměrnost rozložení četností náhodné veličiny $\Rightarrow \bar{\mu}_3(X)$.

Špičatost je charakteristika míry koncentrace rozložení četností náhodné veličiny $\Rightarrow \bar{\mu}_4(X)$.

Normovaná náhodná veličina $\Rightarrow$ **s**třední hodnota normovaného znaku je vždy rovna nule a směrodatná odchylka hodnotě jedna $N\ (0; 1)$.

Kvantily Q_p dělí hodnoty náhodné veličiny na části. Kvantilové charakteristiky jsou odolnější vůči extrémním hodnotám (robustnost).

Literatura

ANDĚL, Jiří. *Základy matematické statistiky*. Vyd. 3. Praha: Matfyzpress, 2011. ISBN 978-80-7378-162-0.

HEBÁK, Petr a Jana KAHOUNOVÁ. *Počet pravděpodobnosti v příkladech*. Praha: Informatorium, 2014. ISBN 978-80-7333-109-2.

3 Kontrolní otázky

1. Co je to náhodná veličina a jaké jsou její základní vlastnosti?
2. Jakým způsobem se obvykle označuje náhodná veličina a její konkrétní hodnoty?
3. Vysvětlete rozdíl mezi diskrétní a spojitou náhodnou veličinou. Uveďte příklady pro každý typ.
4. Proč je potřeba výsledky náhodných pokusů kvantifikovat neboli číselně vyjádřit?
5. Jakými způsoby můžeme ve statistice charakterizovat zákon rozdělení náhodné veličiny?
6. Jaké jsou hlavní vlastnosti distribuční funkce?
7. Jaký má tvar distribuční funkce u diskrétní náhodné veličiny? Co znamenají skoky na grafu?
8. Jaký je tvar distribuční funkce u spojité náhodné veličiny a v čem se liší od diskrétní?
9. Co je to pravděpodobnostní funkce $P(x)$ a pro jaký typ náhodné veličiny se používá?
10. Co je to hustota rozdělení pravděpodobnosti $f(x)$ a pro jaký typ náhodné veličiny se používá? Jaký je její vztah k distribuční funkci?
11. Co je to kvantilová funkce a jaký je její vztah k distribuční funkci?
12. Jaký je účel číselných charakteristik náhodných veličin?
13. Definujte střední hodnotu náhodné veličiny a rozptyl náhodné veličiny. Uveďte jejich značení.
14. Jaké jsou základní vlastnosti střední hodnoty náhodné veličiny?
15. Jsou momentové charakteristiky citlivé na extrémní hodnoty?
16. Jaké jsou nejčastěji používané typy kvantilů a jak dělí statistický soubor?
17. Vysvětlete pojem robustnost kvantilových charakteristik. Proč se označují za robustní?

4 PRAVDĚPODOBNOSTNÍ ROZDĚLENÍ

Rozdělení náhodné veličiny je pravděpodobnostní model chování náhodné veličiny v cílové populaci. Pravděpodobností rozdělení vychází ze zákona rozdělení náhodné veličiny (viz kapitola Náhodná veličina).

V programu IBM SPSS Statistics se pro výpočet pravděpodobnosti a kritických hodnot rozdělení (kvantilů) využívají následující funkce:

CDF (Cumulative Distribution Function) – Kumulativní distribuční funkce

Pro diskrétní rozdělení: vypočítá pravděpodobnost, že náhodná veličina s daným rozložením a určenými parametry nabude hodnoty, která se rovná anebo je menší než zadaná hodnota argumentu $\Rightarrow P(X \leq x)$.

Pro spojité rozdělení: vypočítá pravděpodobnost, že náhodná veličina s daným rozložením a určenými parametry nabude hodnoty menší než zadaná hodnota argumentu $\Rightarrow P(X < x)$.

PDF (Probability Density Function) – Funkce hustoty pravděpodobnosti

Pro diskrétní rozdělení: Vypočítá pravděpodobnost, že náhodná veličina X nabude přesně hodnoty $x \Rightarrow P(X = x)$.

Pro spojité rozdělení: *pravděpodobnost libovolné konkrétní hodnoty u spojitého rozdělení je rovna 0! Více v kapitole 3.1.2 – Paradox nulové pravděpodobnosti.*

IDF (Inverse Distribution Function) – Kvantilová funkce

Pro spojité rozdělení: vypočítá hodnotu kvantilu, který odděluje P procent nejnižších hodnot od zbývajících hodnot.

4.1 Pravděpodobnostní rozdělení diskrétních náhodných veličin

Zákon rozdělení náhodné veličiny neboli teoretické rozdělení budeme charakterizovat pomocí:

Matematické funkce → distribuční funkce $F(x)$ a pravděpodobnostní funkce $P(x)$.

Číselných charakteristik → střední hodnoty $E(X)$ a rozptylu $D(X)$.

Diskrétní rozdělení má konečný nebo spočetný počet realizací. Existuje mnoho typů rozdělení diskrétních náhodných veličin. V následujícím textu budou uvedeny základní poznatky o nejběžnějších rozděleních využívaných ve statistice.

4.1.1 Alternativní rozdělení

Náhodná veličina X, která představuje počet nastoupení sledovaného jevu A při realizaci jednoho pokusu, který se neopakuje $(\boldsymbol{n = 1})$ a který má jen dva výsledky $\{\mathbf{0}; \mathbf{1}\}$. Jestliže jev A nastane (úspěch) přiřadíme náhodné veličině hodnotu p, jestliže jev A nenastane (neúspěch) přiřadíme ji $1 - p$.

Označení

$X \sim A(p)$

Pravděpodobnostní funkce

$$P(x) = p^x(1-p)^{1-x}$$

Distribuční funkce

$$F(x) = \begin{cases} 0 & x < 0; \\ 1 - p & 0 \leq x < 1; \\ 1 & x \geq 1. \end{cases} \qquad (4.1; 4.2)$$

Kde: $x = 0, 1;$

p ….. pravděpodobnost úspěchu pro $p \in (0; 1)$.

Číselné charakteristiky

Střední hodnota $E(X) = p$ (4.3)

Rozptyl $D(X) = p \cdot (1-p)$ (4.4)

Příklad 4.1

Dlouhodobým zjišťováním se odhaduje pravděpodobnost úspěchu obchodní akce 0,7 a pravděpodobnost neúspěchu této akce je 0,3. Vypočítejte střední hodnotu $E(X)$ a rozptyl $D(X)$.

Řešení

Možnými hodnotami veličiny *X* jsou 0 (neúspěch) a 1 (úspěch), přičemž pravděpodobnosti úspěchu obchodní akce je 0,7.

Náhodná veličina $X \sim A(0{,}7)$

Střední hodnota $E(X) = p = 0{,}7$

Rozptyl $D(X) = p \cdot (1-p) = 0{,}7 \cdot 0{,}3 = 0{,}21$

4.1.2 Geometrické rozdělení

Náhodná veličina X udává celkový počet neúspěchů, které v nekonečné posloupnosti **opakovaných (*n*-krát)** nezávislých pokusů předcházejí prvnímu úspěšnému pokusu nastolení sledovaného jevu A. Pravděpodobnost úspěchu při každém opakování pokusu je p a pravděpodobnost neúspěchu je při každém z pokusu $1 - p$. Náhodná veličina může mít v jednom náhodném pokusu jen dva výsledky $\boldsymbol{X} = \{\mathbf{0}, \mathbf{1}\}$

Označení

$X \sim Ge(p)$

Pravděpodobnostní funkce **Distribuční funkce**

$$P(x) = p(1-p)^x \qquad F(x) = \begin{cases} 0 & x < 0; \\ \sum_{0 \le i \le x} p(1-p)^i & 0 \le x. \end{cases} \qquad (4.5; 4.6)$$

Kde: $x = 0, 1, 2, \ldots\ldots, n$ ….. počet neúspěšných výsledků pokusů;

p ….. pravděpodobnost úspěchu.

Číselné charakteristiky

Střední hodnota $$E(X) = \frac{1-p}{p} \qquad (4.7)$$

Rozptyl $$D(X) = \frac{1-p}{p^2} \qquad (4.8)$$

Geometrické rozdělení vychází ze stejného předpokladu jako rozdělení alternativní. Náhodný pokusu má vždy jen dva výsledky úspěch/neúspěch. Geometrické rozdělení počítá **pravděpodobnost počtu neúspěšných výsledků** náhodného pokusu, které předcházejí prvnímu úspěšnému pokusu o nastoupení jevu A.

Příklad 4.2

Studenti v rámci předmětu Finanční a pojistná matematika píšou 5 průběžných testů, které jsou na sobě nezávislé. Hodnocení testů je označeno znaménky „+“, když student test napsal a „–“, když student test nenapsal. Pravděpodobnost, že student test úspěšně napíše, je 0,6. Vypočítejte, jaká je pravděpodobnost, že student úspěšně napíše až třetí test? Následně střední hodnotu $E(X)$ a rozptyl $D(X)$.

Řešení

Náhodná veličina $X \sim Ge(0{,}6)$

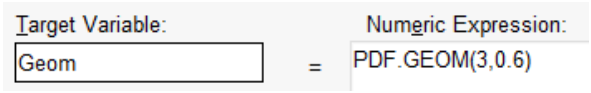
Pravděpodobnost $P(x) = p(1-p)^x = P(X = 2) = 0{,}6(1-0{,}6)^2 = 0{,}09$

Střední hodnota $$E(X) = \frac{1-p}{p} = \frac{1-0{,}6}{0{,}6} = 25{,}0416\overline{6} \cong 0{,}66\overline{6}$$

Rozptyl $$D(X) = \frac{1-p}{p^2} = \frac{1-0{,}6}{0{,}6^2} = 1{,}11\overline{1}$$

Pokračování příkladu 4.2

SPSS Transform → Compute Variable → Function group **PDF & Noncentral PDF** → Functions and Special Variable **Pdf. Geom**



POZOR! *Datová matice při výpočtu v SPSS musí obsahovat proměnnou, jinak je kalkulačka nedostupná! Při zápisu hodnot určených k výpočtu pravděpodobnosti se používá desetinná tečka.*

4.1.3 Binomické rozdělení

Binomickým rozdělením se řídí rozdělení diskrétní náhodné veličiny v případech, kdy nezávislý náhodný pokus **opakujeme** vícekrát **(*n*-krát)**, přičemž pravděpodobnost p výskytu náhodného jevu A je ve všech pokusech stejná (konstantní). Jedná se o modely založené na *náhodný výběr s vracením prvků*. Náhodná veličina může mít v jednom náhodném pokusu jen dva možné výsledky úspěch/neúspěch $\{\mathbf{0}, \mathbf{1}\}$.

Z uvedeného vyplývá, že binomické rozdělení počítá **pravděpodobnost počtu úspěšných výsledků** nastoupení sledovaného jevu A při opakovaných nezávislých pokusech.

Označení

$X \sim Bi(n, p)$

Pravděpodobnostní funkce

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Distribuční funkce

$$F(x) = \begin{cases} 0 & x < 0; \\ \sum_{0 \le i \le x} \binom{n}{i} p^i (1-p)^{n-i} & 0 \le x < n; \\ 1 & x \ge n. \end{cases} \quad (4.9; 4.10)$$

Kde: $x = 0, 1, \ldots, n$ ….. počet úspěšných výsledků pokusu;

p ….. pravděpodobnost úspěchu v jednom pokusu;

n ….. celkový počet pokusů.

Číselné charakteristiky

Střední hodnota $E(X) = np$ (4.11)

Rozptyl $D(X) = np(1-p)$ (4.12)

Příklad 4.3

Student uspěje u testu ze statistiky, jestliže správně odpoví nejméně na 8 otázek z 10. Každá otázka má 4 možné odpovědi, z nichž jediná je správná. Vypočítejte:

a) S jakou pravděpodobností student zodpoví právě 6 z 10 otázek správně, když zvolí odpovědi náhodně?

b) S jakou pravděpodobností student uspěje u testu, je-li zcela nepřipraven (odpovědi volí náhodně)?

c) Kolik otázek v průměru student zodpoví správně, bude-li odpovídat na otázky náhodně?

Řešení

Náhodná veličina X představuje počet otázek, u kterých bude zaškrtnuta správná odpověď z celkového počtu 10 otázek. Náhodná veličina X má binomické rozdělení a parametry $n = 10; p = 0{,}25$ (1 otázka 4 možnosti odpovědi ¼)

Náhodná veličina $X \sim Bi(10; 0{,}25)$

a) $$P(X = 6) = \binom{10}{6} 0{,}25^6 (1 - 0{,}25)^{10-6} = 210 \cdot 0{,}25^6 (0{,}75)^4 = 0{,}016$$

$$K(n,k) = \binom{n}{k} = \frac{n!}{k!\,(n-k)!} = \binom{10}{6} = \frac{10!}{6!\,4!} = 210$$

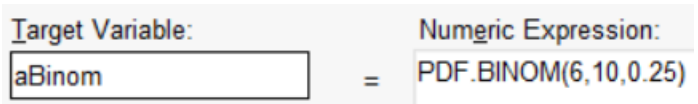
b) $$P(X \geq 8) = 1 - P(X < 7) = 1 - [P(X = 0) + P(X = 1) + \cdots + P(X = 7] =$$

$$= 0{,}0004158$$

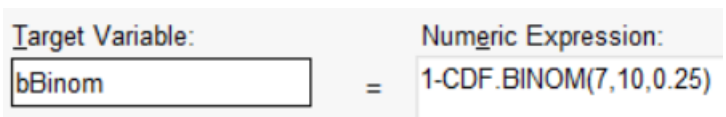
c) $$E(X) = np = 10 \cdot 0{,}25 = \mathbf{2,5}$$

SPSS Transform → Compute Variable →

a) Function group **PDF & Noncentral PDF** → Functions and Special Variable **Pdf. Binom**



b) Function group **CDF & Noncentral CDF** →Functions and Special Variable **Cdf. Binom**



4.1.4 Hypergeometrické rozdělení

Hypergeometrické rozdělení vychází z ***n*-krát** opakování náhodného pokusu, kde pravděpodobnost nastoupení sledovaného jevu je **závislá** na výsledcích předcházejících pokusů (opakované náhodné pokusy jsou závislé). Náhodná veličina může mít v jednom pokusu jen dva výsledky $\{\mathbf{0}, \mathbf{1}\}$. Hypergeometrické rozdělení je základním pravděpodobnostním rozdělením náhodné veličiny, které je založeno na *náhodném výběru bez vracení*.

Definice pravděpodobnostní funkce hypergeometrického rozdělení vychází z klasické definice pravděpodobnosti: počet příznivých možností ku počtu všech možností.

Označení

$X \sim Hg(N, M, n)$

Pravděpodobnostní funkce **Distribuční funkce**

$$P(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \qquad F(x) = \sum_{max \le x \le min} \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \qquad (4.13; 4.14)$$

Kde: $x = max\{0, n-(N-M)\}, \ldots, min\{n, M\}$ ….. počet prvků se sledovanou vlastností,

N ….. celkový počet prvků,

M ….. celkový počet prvků se sledovanou vlastností,

n ….. počet náhodně vybraných prvků z celkového počtu *N*.

Číselné charakteristiky

Střední hodnota $$E(X) = n\frac{M}{N} \qquad (4.15)$$

Rozptyl $$D(X) = n\frac{M}{N}\left(1 - \frac{M}{N}\right)\frac{N-n}{N-1} \qquad (4.16)$$

Limitním případem hypergeometrického rozdělení náhodných veličin je rozdělení binomické, pro $x \to \infty$ a $\frac{M}{N} \to 0$. Pro velká *N* můžeme zanedbat rozdíl mezi *náhodným výběrem bez vracení prvků* a *výběrem s vracením prvků* (v praxi je možné rozhodnout se podle hodnoty tzv. výběrového poměru (n/N). Je-li tento poměr menší než 0,05, lze hypergeometrické rozdělení aproximovat rozdělením binomickým.

Příklad 4.4

Vysoká škola „odAdoZ“ si objednala 180 zářivkových trubic k osvětlení nových počítačových učeben. Při přejímce výrobků bylo z dodávky náhodně vybráno bez vracení 25 trubic, u kterých byla provedena kontrola funkčnosti. Z dohody mezi prodejcem a vysokou školou plyne, že dodávka zářivkových trubic bude přijata a zaplacena jen v případě, že mezi kontrolovanými výrobky budou nanejvýš tři nefunkční.

Vypočítejte, jaká je pravděpodobnost, že dodávka bude přijata, jestliže prodejce předpokládá, že 15 % zářivkových trubic může vykazovat známky poškození a nebýt tedy plně funkčních.

Řešení

Náhodná veličina $\boldsymbol{X \sim Hg(180, M, n)}$

$x = 3$

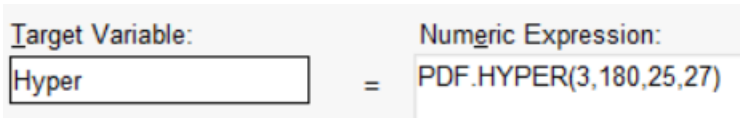
$M \rightarrow 27 \rightarrow$ 15 % poškozených

$N \rightarrow 180 \rightarrow$ celkový počet

$n \rightarrow 25$ počet náhodně vybraných ke kontrole

$$P(X = 3) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} = \frac{\binom{27}{3}\binom{180-27}{25-3}}{\binom{180}{25}} = 0{,}22789$$

SPSS Transform → Compute Variable → Function group **PDF & Noncentral PDF** → Functions and Special Variable **Pdf. Hyper**



4.1.5 Poissonovo rozdělení

Poissonovo rozdělení patří k jednomu z nejdůležitějších rozdělení náhodných veličin v teorii pravděpodobnosti. Popisuje výskyt ***n*-náhodných událostí** (počet událostí – diskrétní NV) v jednotce času, objemu nebo plochy, za předpokladu, že události jsou na sobě vzájemně **nezávislé** a dochází k nim náhodně a jednotlivě (počet volání na telefonní ústřednu v určitém časovém intervalu, počet poruch stroje za dobu t apod.).

Náhodná veličina X může nabývat pouze nezáporných celočíselných hodnot. Náhodná veličina může mít v jednom náhodném pokusu jen dva výsledky $\{\mathbf{0}, \mathbf{1}\}$.

Označení

$X \sim Po(\lambda)$

Pravděpodobnostní funkce **Distribuční funkce**

$$P(x) = \frac{\lambda^x}{x!} e^{-\lambda} \qquad F(x) = \sum_{o \le x} \frac{\lambda^x}{x!} e^{-\lambda} \qquad (4.17; 4.18)$$

Kde: $x = 0, 1, \ldots$. počet výskytu událostí,

λ ……. průměrný počet událostí na jednotku času, objemu nebo prostoru.

Číselné charakteristiky

Střední hodnota $E(X) = \lambda$ (4.19)

Rozptyl $D(X) = \lambda$ (4.20)

Poissonovo rozdělení je limitním případem binomického rozdělení. Používáme ho v případě, že máme větší počet událostí $n \to \infty$ a malou pravděpodobnost výskytu události v jednotce času, objemu nebo plochy $p \to 0$. Poissonovo rozdělení dobře aproximuje binomické za podmínek $n > 30$ a $p < 0{,}1 \Rightarrow Po\,(\lambda) \approx Bi\,(n, p) \to \lambda = np$.

Příklad 4.5

Na revize zápočtových testů náhodně chodí „v průměru“ 6 studentů za hodinu. Vypočítejte, s jakou pravděpodobností přijdou na konzultaci během prvních 30 minut alespoň 2 studenti.

Řešení

Náhodná veličina $X \to$ počet studentů, kteří přijdou během prvních 30 minut. V „průměru“ přijdou na konzultaci 3 studenti za půl hodiny.

$X \sim Po(3)$

$$P(X \ge 2) = 1 - [P(X = 0) + P(X = 1)] = 1 - \left[\frac{3^0}{0!} e^{-3} + \frac{3^1}{1!} e^{-3}\right] = 0{,}8008$$

SPSS Transform → Compute Variable → Function group **CDF & Noncentral CDF** → Functions and Special Variable **Cdf. Poisson**

Příklad 4.6

Na pult centrální ochrany je napojeno 50 subjektů. Pravděpodobnost, že během hodiny zavolá všech 50 subjektů, je 0,01. Ze zkušenosti víme, že střední počet subjektů, kteří během 1 hodiny zavolají na pult centrální ochrany, je $E(X) = 0{,}5$. Poissonovo rozdělení v tomto případě aproximuje binomické rozdělení parametr, λ vypočteme ze vztahu $\lambda = np$.

Vypočítejte:

a) Jaká je pravděpodobnost, že během hodiny zavolají dva klienti?

b) Jaká je pravděpodobnost, že během hodiny nezavolá žádný klient?

Řešení

Náhodná veličina $X \rightarrow$ počet klientů, kteří během hodiny zavolají na pult centrální ochrany.

Náhodná veličina $X \sim Po(0{,}5)$

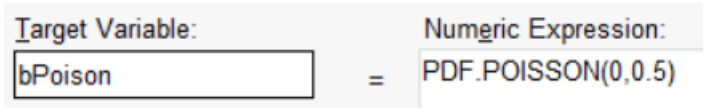
a) $P(X = 2) = \frac{0{,}5^2}{2!} e^{-0{,}5} = 0{,}076$

b) $P(X = 0) = \frac{0{,}5^0}{0!} e^{-0{,}5} = 0{,}607$

SPSS Transform → Compute Variable →

a) Function group **PDF & Noncentral PDF** → Functions and Special Variable **Pdf. Poisson**

b) Function group **PDF & Noncentral PDF** →Functions and Special Variable **Pdf. Poisson**



4.2 Pravděpodobnostní rozdělení spojitých náhodných veličin

Rozdělení pravděpodobnosti spojité náhodné veličiny, které představuje model chování náhodné veličiny v cílové populaci, charakterizujeme pomocí:

a) Matematické funkce → v tomto případě pomocí distribuční funkce $F(x)$ a hustoty pravděpodobnosti $f(x)$.

b) Číselných charakteristik → střední hodnoty $E(X)$ a rozptylu $D(X)$.

Pro popis spojitých náhodných veličin, které jsou předmětem statistických analýz, nejčastěji používáme níže uvedené typy rozdělení náhodné veličiny.

4.2.1 Normální (Gausovo) rozdělení

Normální rozdělení je nejčastěji se vyskytující pravděpodobnostní model rozdělení spojité náhodné veličiny X. Toto rozdělení je velmi důležité, neboť je klíčovým předpokladem pro použití celé řady základních statistických testů a modelů (modelování náhodných chyb – chyby měření, které jsou způsobeny velkým množstvím neznámých a vzájemně nezávislých vlivů).

Normální rozdělení má dva parametry střední hodnotu μ, charakterizující polohu rozdělení, a rozptyl σ^2, který charakterizuje proměnlivost hodnot náhodné veličiny kolem střední hodnoty.

Označení

$X \sim N(\mu,\ \sigma^2)$

Pravděpodobnostní funkce hustoty

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

Distribuční funkce

$$F(x) = \int_{-\infty}^{x} f(x)\,dx \qquad (4.21;\ 4.22)$$

Kde: $x \in (-\infty;\ \infty)$,

π ….. matematická konstanta (Ludolfovo číslo),

μ ….. střední hodnota,

σ^2…. rozptyl.

Číselné charakteristiky

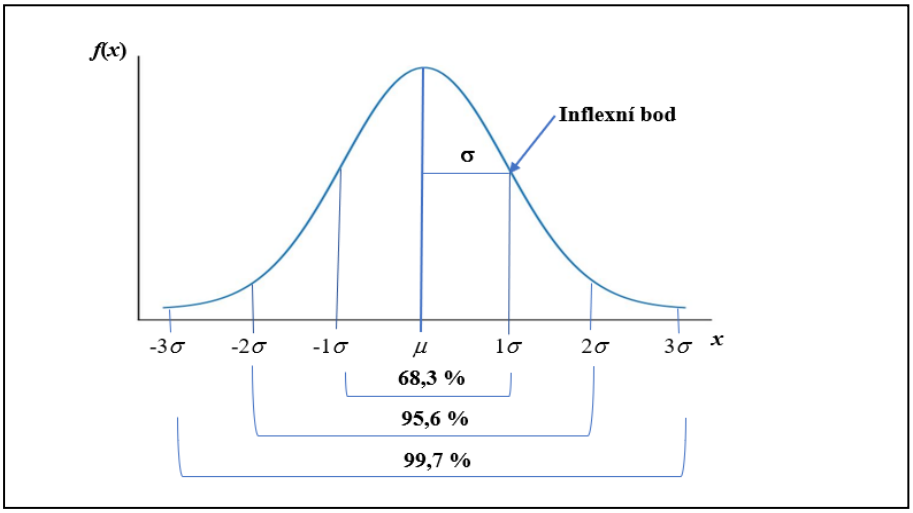
Střední hodnota $E(X) = \mu$ (4.23)

Rozptyl $D(X) = \sigma^2$ (4.24)

Hodnoty náhodné veličiny, které pocházejí z normálního rozdělení se **symetricky** rozprostírají kolem střední hodnoty. Střední hodnota $E(X) \rightarrow$ parametr μ je zároveň mediánem (50% kvantilem) i modem normálního rozdělení. Normální rozdělení je vhodným pravděpodobnostním modelem tehdy, jestliže při opakovaném měření téže náhodné veličiny za stejných podmínek způsobuje množství nezávislých náhodných vlivů kolísání náhodné veličiny kolem skutečné hodnoty. Kolísání náhodných vlivů je vyjádřeno směrodatnou odchylkou $\sigma > 0$, která nám udává šířku křivky v inflexním bodě $(\mu \pm \sigma)$.

Normální rozdělení má tvar zvonovité křivky (obr. 4.1), která se nazývá Gaussova křivka. Tvar hustoty pravděpodobnosti nabývá svého maxima v bodě $x = \mu$ a při $x \rightarrow \pm\infty$ se limitně přibližuje k ose x.

Obrázek 4.1: Hustota rozdělení pravděpodobnosti normálního rozdělení NV



V případě, že náhodná veličina X má normálního rozdělení $N(\mu, \sigma^2)$, můžeme vyčíslit procento pozorování, která by se měla realizovat v rozmezí ± x-násobku směrodatné odchylky σ od střední hodnoty μ.

Pravděpodobnostní intervaly náhodné veličiny kolem střední hodnoty:

- $(\mu - \sigma, \mu + \sigma)$ s pravděpodobností 68,27 %,
- $(\mu - 2\sigma, \mu + 2\sigma)$ s pravděpodobností 95,45 %,
- $(\mu - 3\sigma, \mu + 3\sigma)$ s pravděpodobností 99,73 %.

Pravděpodobnost realizace náhodné veličiny s normálním rozdělením uvnitř intervalu $\mu \pm 3\sigma$ je 99,73 %. Tato skutečnost je označována jako **pravidlo ± 3 sigma.** Pravidlo lze použít pro vizuální ověření normality rozložení sledované náhodné veličiny a pro identifikaci odlehlých hodnot. Význam normálního rozdělení spočívá v možnosti aproximovat (nahradit) řadu jiných rozdělení náhodných veličin. Toto je možné při dodržení určitých podmínek, nejdůležitější je dostatečně velký počet sledovaných náhodných veličin.

4.2.2 Normované normální rozdělení

Speciálním typem normálního rozdělení je normované (standardizované) normální rozdělení s nulovou střední hodnotou $\mu = 0$ a jednotkových rozptylem $\sigma^2 = 1$. Vztah pro přepočet (standardizaci) normální náhodné veličiny na veličinu normovanou je:

$$U = \frac{X - \mu}{\sigma} \tag{4.25}$$

Standardizace nám umožní převést hodnoty x náhodné veličiny X, které jsou vyjádřeny v jednotkách měřené veličiny, na normované normální veličiny U, kde jednotlivé hodnoty u jsou vyjádřeny ve směrodatných odchylkách.

Označení

$X \sim U \sim N(0; 1)$

Pravděpodobnostní funkce hustoty

$$f(x; 0{,}1) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}}$$

Distribuční funkce

$$F(x) = \int_{-\infty}^{x} f(x)\, dx \qquad (4.26; 4.27)$$

Kde: $x \in (-\infty; \infty)$

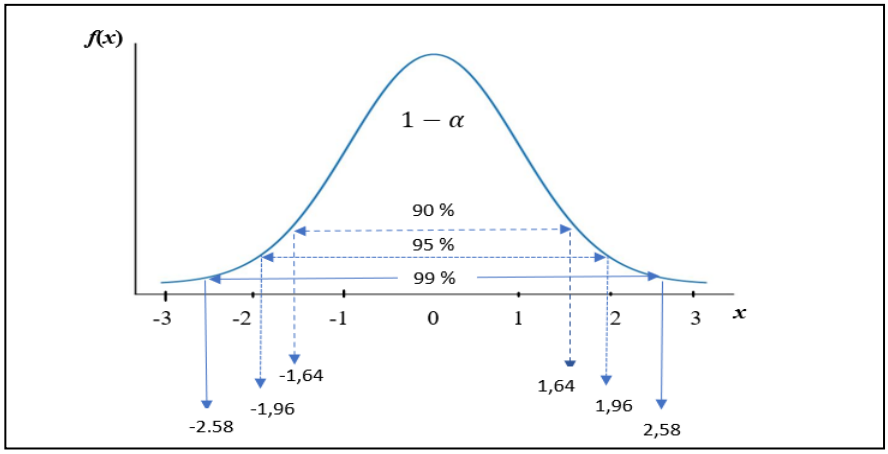
π ….. matematická konstanta (Ludolfovo číslo),

0 …. střední hodnota,

1 …. rozptyl.

Distribuční funkce nám udává plochu, ve které se náhodná veličina vyskytuje s pravděpodobností $1 - \alpha$. Vymezení této plochy je realizováno pomocí kvantilů. Nejčastěji používané kvantily normovaného normálního rozdělení jsou graficky znázorněny v obr. 4.2.

Obrázek 4.2: Kvantily normovaného normálního rozdělení pravděpodobnosti NV



V případě normovaného normálního rozdělení se:

- s pravděpodobností 90 % náhodná veličina vyskytuje v intervalu $\mu \pm 1{,}64\sigma$,
- s pravděpodobností 95 % náhodná veličina vyskytuje v intervalu $\mu \pm 1{,}96\sigma$,
- s pravděpodobností 99 % náhodná veličina vyskytuje v intervalu $\mu \pm 2{,}58\sigma$.

Hodnoty distribuční a kvantilové funkce, stejně jako hustota pravděpodobností, jsou z důvodů matematické náročnosti výpočtu tabelovány právě pro normované normální rozdělení. Tyto hodnoty jsou obsaženy také ve všech statistických softwarech.

Příklad 4.7

Předpokládáme, že doba čekání na obsloužení v restauraci má normální rozdělení s parametry $N(8;\ 0{,}49)$. Jaké procento zákazníků bude obslouženo:

a) za méně než 7 minut,

b) za více jak 9,5 minuty,

c) v rozmezí 6–9 minut?

Řešení

Náhodná veličina $X \rightarrow$ doba čekání na obsloužení v restauraci má normální rozdělení. Hodnota pravděpodobnosti vychází z výpočtu distribuční funkce normálního rozdělení $N(8;\ 0{,}49)$ podle vztahu (4.22). V případě ručního výpočtu a za předpokladu, že máme k dispozici tabulky **distribuční funkce normálního rozdělení**, přistoupíme ke standardizaci náhodné veličiny podle vztahu 4.25. Tím převedeme původní hodnotu náhodné veličiny $X \sim N(\mu,\ \sigma^2)$ na standardizovanou náhodnou veličinu $U \sim N(0;1)$, která je vzdálená 0,7 směrodatné odchylky (σ) od střední hodnoty (μ), tedy na relativní vyjádření nezávislé na původních jednotkách.

a) $P(X < 7) = P\left(U < \frac{7-8}{0{,}7}\right) = P(U < -1{,}43) = 1 - F(1{,}43) = 1 - 0{,}924 \cong 0{,}076.$

Do 7 sekund bude obslouženo 7,6 % zákazníků.

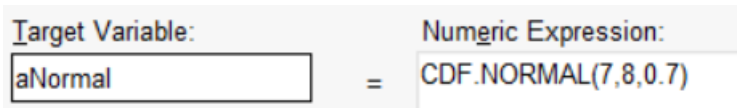
Distribuční funkce normálního rozdělení $F(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-\frac{u^2}{2}}$

u	0,00	0,01	0,02	0,03	0,04
1,2	0,885	0,887	0,889	0,891	0,893
1,3	0,903	0,905	0,907	0,908	0,910
1,4	0,919	0,921	0,922	0,924	0,925
1,5	0,933	0,934	0,936	0,937	0,938

POZOR! *Řešení daného příkladu bylo pouze nastíněno. Rozdíl ve výsledcích je způsoben nepřesností při zaokrouhlování. Pro veškeré další výpočty budeme využívat program IBM SPSS, který nabízí přesnější a efektivnější možnosti výpočtu.*

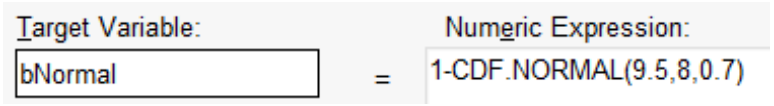
SPSS Transform → Compute Variable → Function group **CDF & Noncentral CDF** → Functions and Special Variable **Cdf. Normal**

a) $P(X < 7) = 0{,}076564$

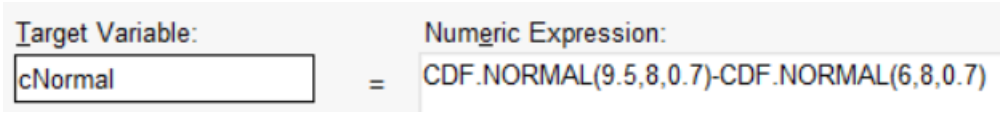


Pokračování příkladu 4.7

b) $P(X > 9{,}5) = 0{,}016062$



c) $P(6 < X < 9{,}5) = 0{,}9818$



	aNormal	bNormal	cNormal
1	,076564	,016062	,981800

V případě, že provádíme statistické šetření na celé populaci, jsme schopni přesně pomocí popisných statistik zjistit parametry jejího chování. V následujícím textu se budeme věnovat pravděpodobnostním rozdělením náhodných veličin, které jsou odvozeny od normovaného normálního rozdělení a využívají se ve statistických postupech a metodách, jejichž součástí jsou induktivní úsudky z výběru populace.

V odborné literatuře jsou níže uvedená rozdělení souhrnně nazývána rozděleními výběrovými. **Výběrová pravděpodobnostní rozdělení** využíváme především při odhadech populačních parametrů a při testování statistických hypotéz. Jestliže máme k dispozici výběr z populace, na jehož základě odhadujeme populační parametry, předpokládáme, že v případě dalšího výběru z téže populace se námi odhadované populační parametry budou lišit → jedná se tedy o náhodnou veličinu. Hodnoty uvedených výběrových rozdělení jsou, stejně jako hodnoty distribuční a kvantilové funkce normovaného normálního rozdělení, tabelovány z důvodů matematické náročnosti výpočtu.

4.2.3 *Chí*-kvadrát rozdělení

Chí-kvadrát rozdělení (Pearsonovo χ^2-rozdělení) je rozdělení spojitých náhodných veličin, které se v matematické statistice využívá především k testování rozdělení pravděpodobnosti nebo k odhadu a testování rozptylu náhodné veličiny.

Označení

$X \sim \chi^2(\nu)$

Chí-kvadrát rozdělení vzniká jako součet druhých mocnin nezávislých náhodných veličin $U_1, U_2, \dots, U_n$, které mají normované normální rozdělení $N(0,1)$.

$$\chi^2(\nu) = \sum_{i=1}^{n} U_i^2 \tag{4.28}$$

Kde: U ….. je normovaná normální náhodná veličina pro $i = 1, \dots, n$,
n ….. počet vzájemně nezávislých normovaných veličin,
ν ….. počet stupňů volnosti.

Průběh této funkce (tvar rozdělení pravděpodobnosti) je dán jedním parametrem, který se nazývá **počet stupňů volnosti** ν (řecké *ný*). Pojem „počet stupňů volnosti“ udává počet vzájemně nezávislých veličin. V případě testování statistických hypotéz (viz kapitola Testování hypotéz) slouží počet stupňů volnosti jako parametr pro porovnání testového kritéria s odpovídajícím rozdělením. V takovém případě je většinou hodnota parametru určena počtem pozorovaných náhodných veličin, sníženým o počet odhadovaných charakteristik.

Příklad 4.8

Vypočítejte 95% kvantil pro náhodnou veličinu X, která pochází z *chí*-kvadrát rozdělení o pěti stupních volnosti.

SPSS Transform → Compute Variable → Function group **Inverse DF** → Functions and Special Variable **Idf. Chisq**

$X \sim \chi^2_{0,95}(5) = 11{,}07$

Hustota pravděpodobnosti $f(x)$ *chí*-kvadrát rozdělení výběrové náhodné veličiny je **asymetrická** funkce. S rostoucím počtem vzájemně nezávislých normovaných veličin $n > 30$ je možné *chí*-kvadrát rozdělení aproximovat normovaným normálním rozdělením.

4.2.4 Studentovo *t*-rozdělení

Studentovo *t*-rozdělení vzniká jako podíl dvou nezávislých náhodných veličin: náhodné veličiny s normovaným normálním rozdělením $N(0,1)$ a náhodné veličiny s *chí*-kvadrát rozdělením $\chi^2(\nu)$.

Označení

$X \sim t(\nu)$

Studentovo t-rozdělení o ν stupních volnosti je definováno vztahem:

$$t(\nu) = \frac{N(0;1)}{\sqrt{\chi^2(\nu)/\nu}} \tag{4.29}$$

Kde: $N(0;1)$ ….. náhodná veličina s normovaným normálním rozdělením.

$\chi^2(\nu)$ …… náhodná veličina s *chí*-kvadrát rozdělením.

ν ……. počet stupňů volnosti.

Veličina Studentovo t-rozdělení je teoreticky odvozená hodnota, která je v matematické statistice základem pro odhady a testování populačních průměrů v případě, kdy neznáme skutečnou variabilitu (populační rozptyl σ^2 náhodné veličiny). V takovém případě vycházíme při výpočtech z odhadovaného výběrového rozptylu (s^2).

Příklad 4.9

Vypočítejte 90% kvantil pro náhodnou veličinu X, která pochází ze Studentova t-rozdělení o patnácti stupních volnosti.

SPSS Transform → Compute Variable → Function group **Inverse DF** → Functions and Special Variable **Idf. T**

$X \sim t_{0,90}(15) = 1{,}34$

Hustota pravděpodobnosti $f(x)$ Studentova t-rozdělení výběrové náhodné veličiny je **symetrická** funkce a její zvonovitý tvar závisí na počtu stupňů volnosti výběrového souboru ν. S rostoucím počtem stupňů volnosti se tvar Studentova pravděpodobnostního rozdělení aproximativně přibližuje hustotě normovaného normálního rozdělení.

4.2.5 Fisher-Snedecorovo *F*-rozdělení

Fisher-Snedecorovo *F*-rozdělení pravděpodobnosti vzniká jako podíl dvou náhodných veličin s *chí*-kvadrát rozdělením.

Označení

$X \sim F(\nu_1; \nu_2)$

Fisher-Snedecorovo *F*-rozdělení je definováno vztahem:

$$F\left(\nu_{1;}\,\nu_2\right) = \frac{\chi_1^2(\nu_1)/\nu_1}{\chi_2^2(\nu_2)/\nu_2} \tag{4.30}$$

Kde: $\chi_1^2(\nu_1),\ \chi_2^2(\nu_2)$….. náhodné veličiny s *chí*-kvadrát rozdělením,

$\nu_{1;}\,\nu_2$ ….. počty stupňů volnosti.

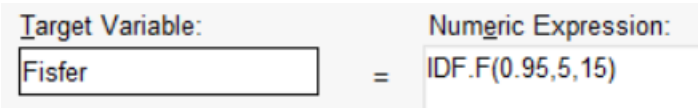
V matematické statistice se *F*-rozdělení využívá především při testování dvou výběrových rozptylů a dále například při testování hypotéz o rovnosti středních hodnot náhodné veličiny *X* u více než dvou výběrových souborů. Hustota pravděpodobnosti *F*-rozdělení je **asymetrická** a její tvar je dán parametry dvou nezávislých náhodných veličin.

Příklad 4.9

Vypočítejte 95% kvantil pro náhodnou veličinu, která má pět a patnáct stupňů volnosti.

SPSS Transform → Compute Variable → Function group **Inverse DF** → Functions and Special Variable **Idf. F**

$X \sim F_{0,95}(5;\ 15) = 2{,}90$



Shrnutí kapitoly

Vybraná rozdělení **diskrétních** náhodných veličin.

Rozdělení	Parametr	Opakování	Výpočet SPSS
Alternativní	$X \sim A(p)$	$n = 1$	$Bernpoulli(x, p, 1)$
Geometrické	$X \sim Ge(p)$	n-krát/nezávislá	$Geom(x, p)$
Binomické	$X \sim Bi(n, p)$	n-krát/nezávislá	$Binom\ (x, n, p)$
Hypergeometrické	$X \sim Hg(N, M, n)$	n-krát/závislá	$Hyper\ (x, N, n, M)$
Poissonovo	$X \sim Po(\lambda)$	n-krát/nezávislá	$Poisson\ (p, \lambda)$

Vybraná rozdělení **spojitých** náhodných veličin.

Rozdělení	Parametr	Výpočet SPSS
Normální (Gausovo)	$X \sim N(\mu,\ \sigma^2)$	$Normal(x, \mu,\ \sigma^2)$
Normované normální	$X \sim U \sim N(0; 1)$	$Normal(x, 0,1)$
***Chí*-kvadrát**	$X \sim N(0; 1) \sim \chi^2(v)$	$Chisq(x, nu)$
Studentovo *t*	$X \sim N(0; 1),\ \chi^2(v) \sim t(v)$	$T(x, v)$
Fisher-Snedecorovo *F*	$X \sim \chi_1^2(v_1),\ \chi_2^2(v_2) \sim F(v_1, v_2)$	$F(x, v_1, v_2)$

Literatura

ANDĚL, Jiří. *Základy matematické statistiky*. Vyd. 3. Praha: Matfyzpress, 2011. ISBN 978-80-7378-162-0.

BUDÍKOVÁ, Marie; Maria KRÁLOVÁ a Bohumil MAROŠ. *Průvodce základními statistickými metodami*. Vydání první. Praha: Grada Publishing, a.s., 2010. ISBN 978-80-247-3243-5.

HEBÁK, Petr a Jana KAHOUNOVÁ. *Počet pravděpodobnosti v příkladech*. Praha: Informatorium, 2014. ISBN 978-80-7333-109-2.

NEUBAUER, Jiří; Marek SEDLAČÍK a Oldřich KŘÍŽ. *Základy statistiky: aplikace v technických a ekonomických oborech*. 2., rozšířené vydání. Praha: Grada, 2016. ISBN 978-80-247-5786-5.

ŘEHÁK, Jan a Ondřej BROM. *SPSS – Praktická analýza dat.* 1. vydání. Brno: Computer Press, 2015. ISBN 978-80-251-4609-5.

4 Kontrolní otázky

1. Jakými způsoby se charakterizuje zákon rozdělení diskrétní náhodné veličiny (matematickými funkcemi a číselnými charakteristikami)?
2. Vysvětlete, co představuje náhodná veličina X u alternativního rozdělení a jaké má parametry.
3. Co udává náhodná veličina X u geometrického rozdělení?
4. V jakých případech se řídí diskrétní náhodná veličina binomickým rozdělením?
5. Jaké parametry má binomické rozdělení a co znamenají?
6. Jaký je klíčový rozdíl mezi hypergeometrickým a binomickým rozdělením z hlediska opakování pokusů?
7. Co popisuje Poissonovo rozdělení a jaké typy událostí modeluje?
8. Co znázorňuje Gaussova křivka?
9. Jaké dva parametry charakterizují normální rozdělení a co každý z nich vyjadřuje?
10. Popište tvar hustoty pravděpodobnosti normálního rozdělení.
11. Uveďte pravděpodobnostní intervaly náhodné veličiny kolem střední hodnoty pro 1σ, 2σ a 3σ a vysvětlete pravidlo ± 3 sigma.
12. Kdy se provádí standardizace náhodné veličiny, co to znamená?
13. Jaké jsou nejčastěji používané kvantily normovaného normálního rozdělení s pravděpodobností 90 %, 95 % a 99 %?
14. K čemu se využívají výběrová pravděpodobnostní rozdělení?
15. Co znamená aproximace rozdělení.

4 Příklady k procvičení

4.1 Rozdělení diskrétních náhodných veličin

4.1.1 Z dlouhodobého sledování víme, že měřící přístroj se dopouští závažné chyby pro 7% měření, přičemž jednotlivá měření jsou nezávislá.

a) Jaká je pravděpodobnost, že při dvaceti měřeních nedojde k žádné chybě? [0,2342]

b) Jaká je pravděpodobnost, že při dvaceti měřeních dojde maximálně ke dvěma chybám? [0,8389]

4.1.2 Ve třídě 1.A je 18 chlapců a 10 dívek.

a) Jaká je pravděpodobnost, že v náhodně sestavené pětici žáků budou alespoň dva chlapci? [0,9589]

b) Jaká je pravděpodobnost, že v náhodně sestavené pětici žáků budou maximální dvě dívky? [0,7722]

c) Jaká je pravděpodobnost, že při náhodném rozdělení třídy na dvě poloviny bude v jedné polovině právě 10 chlapců? [0,2291]

d) Tři chlapci odešli na záchod. Jaká je pravděpodobnost, že v náhodně sestavené pětici žáků budou alespoň dva chlapci? [0,9359]

4.1.3 V rámci výzkumu mezi kuřáky uvedlo 80 % dospělých, že začalo kouřit před dovršením 18 let. Uvažujte skupinu deseti náhodně vybraných dospělých kuřáků a vypočítejte:

a) Pravděpodobnost, že právě osm z nich začalo kouřit před osmnáctým rokem? [0,3019]

b) Pravděpodobnost, že více než osm z nich začalo kouřit před osmnáctým rokem? [0,3758]

c) Pravděpodobnost, že alespoň čtyři z nich nezačali kouřit před osmnáctým rokem? [0,1209]

4.1.4 Žáci fotbalového klubu soutěží, kolik gólů dají v případě, že mají 10 pokusů. Předpokládaná pravděpodobnost úspěšného vstřelení gólu je 0,75. Jaká je pravděpodobnost, že jeden z členů klubu dá:

a) právě 4 góly, [0,0162]

b) méně než 3 gólů, [0,0197]

c) více než 6 gólů. [0,7758]

4.1.5 Na telefonní ústřednu přijde během 8 hodin v průměru 360 žádostí o spojení. Jaká je pravděpodobnost, že během příštích 10 minut přijdou:

a) 4 žádosti o spojení, [0,0729]

b) nejvýše 4 žádosti o spojení? [0,132]

4.2 Rozdělení spojitých náhodných veličin

4.2.1 Výška dospělých mužů v určité populaci má normální rozdělení s průměrem μ =180 cm a směrodatnou odchylkou σ=8 cm. Jaká je pravděpodobnost, že náhodně vybraný bude vyšší než 190 cm? [0,106]

4.2.2 Bylo zjištěno, že průměrný počet odpracovaných hodin je u zdravotnického personálu 81,7 hod./týden se směrodatnou odchylkou 6,9 hodiny. Předpokládejme, že se jedná o náhodnou veličinu s normálním rozdělením.

a) Jaké procento pracovníků vykazuje pracovní dobu v rozmezí 70 až 90 hodin za týden? [0,8405]

b) Jaká je pravděpodobnost, že náhodně vybraný pracovník bude pracovat déle než 95 hodin? [0,0269]

c) Jaké procento zdravotníků pracuje týdně méně než 70 hodin? [0,0446]

4.2.3 Náhodná veličina F má Fisherovo rozdělení $F(12; 7)$

a) Určete 5% a 95% kvantily náhodné veličiny F. [0,3432; 3,5746]

b) Určete pravděpodobnost $P(F < 4{,}666)$ [0,975]

4.2.4 Náhodná veličina χ^2 má Pearsonovo rozdělení $\chi^2(15)$.

a) Určete dolní a horní kvartil náhodné veličiny χ^2 . [11,036; 18,245]

b) Určete pravděpodobnost $P(\chi^2 < 7{,}26)$ [0,0499]

4.2.5 Náhodná veličina t má Studentovo rozdělení $t(\nu)$.

a) Určete 99% kvantily pro $\nu = 4$. [3,7469]

b) Určete 99% kvantily pro $\nu = 23$. [2,4998]

5 ZPRACOVÁNÍ DATOVÉHO SOUBORU

Je-li stanoven jasný cíl statistického zkoumání a je-li k dispozici příslušný soubor statistických jednotek, začíná vlastní statistická práce, kterou lze rozdělit do několika na sebe navazujících etap.

První etapou je **statistické zjišťování**, které je založeno na získání údajů o hodnotách znaků sledovaných u statistických jednotek (viz kapitola Teorie odhadu). Jedná se o jednu z velmi důležitých činností, neboť chybně získané nebo neúplné údaje nelze ve většině případů již nijak získat nebo opravit. Výsledkem statistického zjišťování jsou ve většině případů neuspořádané kvalitativní a kvantitativní proměnné, které postrádají řád a je velmi obtížné se v nich orientovat. Před aplikací výpočetních metod je proto potřeba provést uspořádání (sumarizaci) a kontrolu těchto údajů, a to z hlediska formálního, logického i početního.

Získání uceleného přehledu o datech a jejich základní uspořádání umožňuje další etapa statistické práce, a tou je **statistické zpracování**, které je založeno na technikách třídění, tabelování a grafickém znázornění. Cílem grafického znázornění je podat rychlou a srozumitelnou informaci o studovaném jevu či o vzájemném vztahu více jevů.

Hlavním úkolem celého statistického zkoumání a také jednou z etap statistické práce je vlastní **statistická analýza** neboli rozbor, který vychází z výstižného popisu zkoumaných jevů na základě výpočtu elementárních statistických charakteristik. Statistickými charakteristikami nazýváme číselné hodnoty, které nám podávají základní informaci o vlastnostech datového souboru. Tyto statistiky mohou být nejen předmětem vlastního statistického zkoumání, ale také základem pro aplikaci složitějších statistických analýz, jejichž cílem je určení statistických zákonitostí a jejich vzájemných souvislostí (např. korelační analýza, regresní analýza atd.).

Poslední etapou statistické práce je slovní vyhodnocení a porovnání získaných výsledků o chování sledovaných jevů s teoretickými předpoklady a jejich **prezentace** pomocí vhodně zvolených grafů a tabulek. Při formulování závěrů a rozsahu platnosti dosažených výsledků je důležité neopomenout skutečnost, že vzhledem k pravděpodobnostnímu charakteru zkoumaných jevů není matematicky možné přesně vystihnout všechny sledované vlastnosti statistických jednotek.

Výběr zpracování statistických dat a početních postupů se vždy odvíjí od typu proměnných a jejich počtu. Přesná klasifikace proměnných, které statisticky zpracováváme, je důležitá. Pro kvantitativní proměnné používáme jiné metody než pro kvalitativní proměnné. Nesprávné

definování typu proměnných vede ke zkreslení výsledků a často také ke ztrátě informací obsažených v datech.

V následujícím výkladu se podrobněji zaměříme na **popisnou statistiku** (deskriptivní statistiku). Popisná statistika je předmětem statistického zpracování a zabývá se elementárními metodami popisu vlastností (údajů, dat) statistických jednotek získaných z populačních nebo výběrových zjišťování. Elementární metody popisu dat v tomto pojetí zahrnují **třídění, tabelování, grafické znázornění a metody výpočtu číselných charakteristik**.

Číselný a grafický popis dat budeme dále uvádět s ohledem na jednotlivé typy proměnných, ale bez ohledu na to, zda se jedná o základní nebo výběrový soubor (popisujeme jen získaná data, neprovádíme statistickou indukci – zobecňování).

5.1 Třídění, tabelování a grafické znázornění proměnných

Prvotní zpracování získaných údajů spočívá v roztřídění hodnot jednotlivých proměnných do tabulek rozdělení četností a grafického znázornění těchto rozdělení.

Tříděním se rozumí rozdělení statistického souboru do skupin podle hodnot jednoho nebo více třídicích znaků. Je-li třídicí znak jeden, hovoří se o **jednorozměrném rozdělení četností**, při třídění podle dvou znaků pak o **dvourozměrném rozdělení četností**.

5.1.1 Jednorozměrné rozdělení četností

Absolutní četnost n_i představuje počet opakování hodnoty znaku v původní řadě dat.

$$n_1 + n_2 + \ldots + n_k = \sum_{i=1}^{k} n_i = n \tag{5.1}$$

Kde: n_i …. absolutní četnost výskytu dané varianty znaku, pro i = 1, 2, 3, ..., k,
n ….. rozsah souboru.

Statistický soubor má n statistických jednotek, u kterých sledujeme vlastnosti (znaky) proměnné X. Číselné charakteristiky znaků tvoří neuspořádanou řadu hodnot $x_1, x_2, x_3, \ldots, x_n$. V této řadě je však pouze k různých variant (kategorií) hodnot, ostatní se opakují. Těchto k různých variant seřadíme vzestupně do tabulky a každé z variant přiřadíme číslo, které představuje počet opakování hodnoty (četnost výskytu dané varianty znaku) v původní řadě dat. Součet četností musí být vždy roven rozsahu statistického souboru n.

Relativní četnost p_i udává, jaká část vyšetřovaného souboru má hodnotu znaku x_i. Relativní četnosti se při interpretaci výsledků vyjadřují většinou v procentech. Pro součet relativních četností platí:

$$p_1 + p_2 + p_3 + \ldots\ldots\ldots\ldots\ldots + p_k = \sum_{i=1}^{k} p_i = 1 \tag{5.2}$$

$$p_i = \frac{n_i}{n} \tag{5.3}$$

Kde: p_i …. relativní četnost, pro i = 1, 2, 3, .., k,
n_i … absolutní četnost,
n ….. rozsah souboru.

Kumulativní četnosti představují počet statistických jednotek s hodnotou znaku menší nebo rovnou x_i. Kumulativní četnost může být **absolutní N_i**, tak **relativní P_i**.

$$N_k = n_1 + n_2 + n_3 + \ldots\ldots\ldots\ldots\ldots + n_k = \sum_{i=1}^{k} n_i \tag{5.4}$$

$$P_k = p_1 + p_2 + p_3 + \ldots\ldots\ldots\ldots\ldots + p_k = \sum_{i=1}^{k} p_i \tag{5.5}$$

Kde: p_i ….. relativní četnost, pro i = 1, 2, 3, …., k,
n_i …. absolutní četnost,
n ….. rozsah souboru.

Výsledky třídění se zapisují do tabulek (obr. 5.1), které se k tomu účelu sestavují.

Obrázek 5.1: Tabulka rozdělení četností

Varianta znaku x_i	Absolutní četnost n_i	Relativní četnost p_i	Kumulativní absolutní četnost N_i	Kumulativní relativní četnost P_i
x_1	n_1	p_1	$N_1 = n_1$	$P_1 = p_1$
x_2	n_2	p_2	$N_2 = n_1 + n_2$	$P_2 = p_1 + p_2$
…..	…..	…..		
x_k	n_k	p_k	$N_k = n_1 + n_2 + \ldots + n_k$	$P_k = p_1 + p_2 + .. + p_k$

Příklad 5.1

Univerzitní knihovnu během jedné hodiny navštívilo 26 studentů. V následující tabulce jsou uvedeny počty knih, které si studenti zapůjčili.

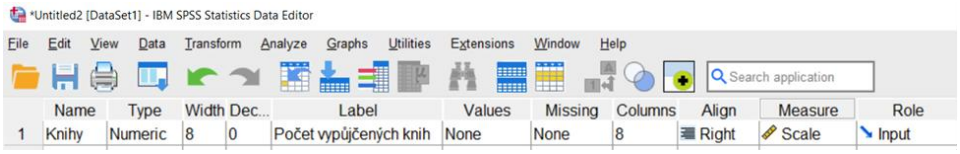
4	2	3	4	1	2	5	4	5	4	2	3	4
1	1	1	1	1	4	2	3	4	1	2	5	4

Na základě uvedených údajů vytvoříme tabulku rozdělení četností.

SPSS

Analyze → Descriptive Statistics → Frequencies → proměnná *Knihy* do okna Variable(s) → OK

	Knihy
1	4
2	2
3	3
4	4
5	1
6	2
7	5
8	4
9	5
10	4
11	2
12	3
13	4
14	1
15	1
16	1
17	1
18	1
19	4
20	2
21	3
22	4
23	1
24	2
25	5
26	4



Statistics

Počet vypůjčených knih

N	Valid	26
	Missing	0

→ Počet všech platných jednotek = rozsah souboru n

Počet vypůjčených knih

		n_i Frequency	p_i Percent	Valid Percent	P_i Cumulative Percent
Valid	1	7	26,9	26,9	26,9
	2	5	19,2	19,2	46,2
	3	3	11,5	11,5	57,7
	4	8	30,8	30,8	88,5
	5	3	11,5	11,5	100,0
	Total	26	100,0	100,0	

5.1.2 Intervalové jednorozměrné rozdělení četností

Při velkém množství údajů (např. třídění hodnot spojitých nebo diskrétních znaků s mnoha navzájem různými hodnotami) je použití tabulky rozdělení četností nepřehledné a její vypovídací schopnost je velmi malá. V tomto případě je vhodnější použít **intervalové rozdělení četností**.

Intervalové rozdělení četností má své výhody i nevýhody. Při konstrukci intervalového rozdělení četností dochází k určitému zjednodušení v tom smyslu, že se zachycuje četnost intervalu místo zjištěných hodnot. Hlavní nevýhodou je však ztráta informace, která plyne z agregace hodnot znaků u sledované proměnné. V současné době, kdy dochází ke zpracování dat pomocí statistických programů a není třeba se obávat velkého množství údajů, se toto

rozdělení využívá především pro zjednodušení dané úlohy či k vytvoření kategorií pro další potřeby zkoumání a rozboru dat.

Pro stanovení **počtu intervalů *k*** neexistuje jednotný předpis. Ve statistické literatuře je sice možné nalézt doporučení pro určení počtu intervalů, ale vždy záleží na rozhodnutí analytika a cílech prováděného statistického šetření.

Doporučená pravidla pro určení počtu intervalů:

Sturgessovo pravidlo

$$k \cong 1 + 3{,}3 \log n \tag{5.6}$$

Yuleho pravidlo

$$k \cong \sqrt{n} \tag{5.7}$$

Kde: k ….. počet intervalů,

n ….. rozsah souboru,

log .. dekadický logaritmus.

Tvar intervalového rozdělení lze podstatně ovlivnit **délkou intervalů *h***. Obecně platí, že při volbě příliš malé intervalové šíře nebudou podstatné vlastnosti rozdělení dostatečně zřetelné, protože budou zastřeny přítomností náhodných faktorů (náhodným kolísáním). Bude-li naopak šíře intervalů příliš velká, bude tvar vyšetřovaného rozdělení patrný jen ve velmi hrubých rysech a nevyniknou tak zákonitosti charakteristické pro daný soubor.

$$h = \frac{R}{k} \tag{5.8}$$

Kde: k ….. počet intervalů,

R ….. variační rozpětí → $(x_{max} - x_{min})$,

x_{min}… minimální hodnota znaku,

x_{max}… maximální hodnota znaku.

POZOR! *Vždy musí být jednoznačně určeno, do kterého intervalu jednotku zařadit!*

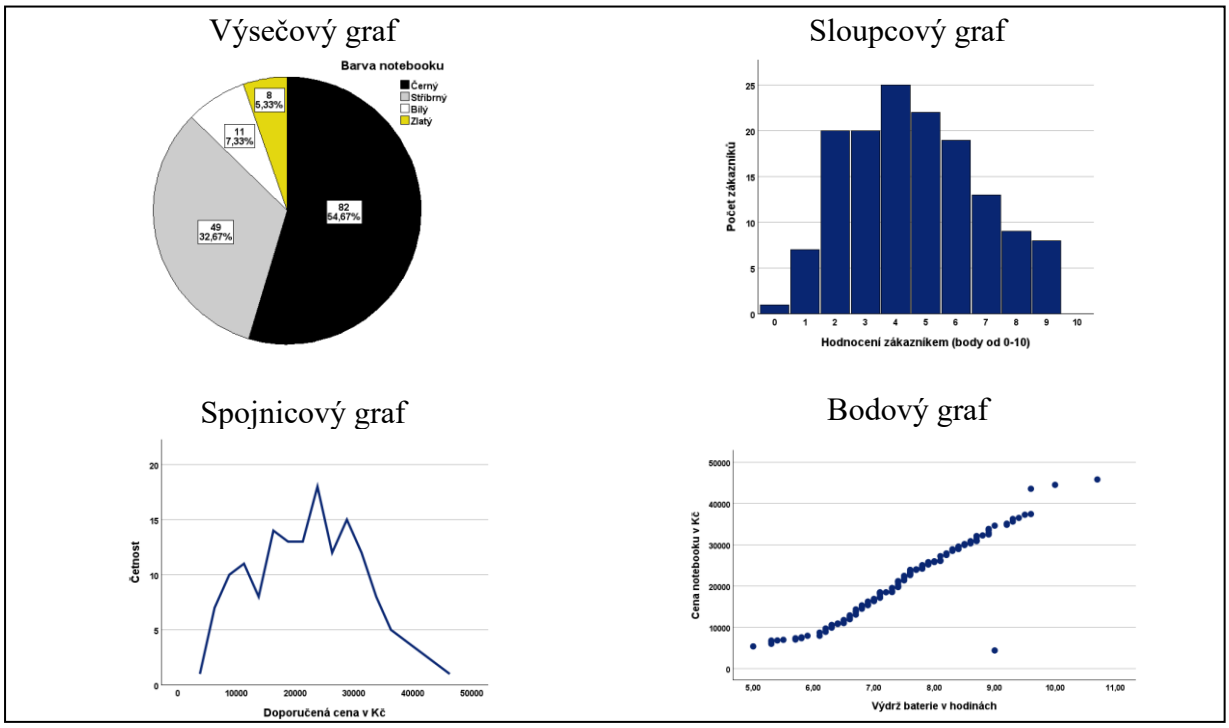
5.1.3 Grafická analýza

Názornou představu o studovaném souboru poskytuje grafické znázornění příslušného rozdělení četností. Grafické zobrazení umožňuje rychlou a přehlednou představu o tendencích a charakteristických rysech analyzovaných proměnných, a také rychlou kontrolu sledovaných údajů. Grafy (obr. 5.2) jsou rovněž vhodným prostředkem pro prezentaci statistických výsledků. Volba vhodného typu grafu musí zohledňovat typ zobrazované proměnné.

Spojnicový graf (polygon) je jednoduchý a všestranný nástroj pro grafické shrnutí hodnot proměnných. Na vodorovnou osu (osa x) se vynáší jednotlivé varianty proměnné v pořadí od „nejmenší“ do „největší“ a na svislou osu (osa y) příslušné hodnoty četností. Nejčastěji se používá pro zobrazení hodnot proměnných v závislosti na čase t. Tento graf není vhodný pro zobrazení diskrétních proměnných.

Bodový graf zobrazuje datové body bez propojovacích čar. Je vhodný pro vizualizaci kvantitativní diskrétní proměnné, anebo pro grafické znázornění vztahů mezi dvěma proměnnými (viz níže Vícerozměrné rozdělení četností).

Obrázek 5.2: Vybrané typy grafického znázorní dat



Výsečový graf (koláčový) je graf, kde plochy kruhových výsečí prezentují četnost výskytu jedné varianty proměnné v poměru k součtu četností všech variant. Jednotlivé výseče se označují nejen relativní četností, ale i četností absolutní, aby se předešlo chybné prezentaci výsledků šetření.

Sloupcový graf může nabývat nejrůznějších tvarů. Obvykle zobrazuje na ose x kategorie proměnné a na ose y hodnoty četnosti jejich výskytu. Tento typ grafu se používá především k zobrazování a srovnávání nespojitých, nominálních a ordinálních proměnných.

Na rozhraní mezi tabulkami a grafy, a mezi zobrazením zdrojových dat a souhrnných údajů (např. četností), se nachází grafy, které nám poskytují informace o rozdělení četností a také o základních číselných charakteristikách proměnných (např. minimum, maximum, medián atd.). Patří sem například box-plot a histogram. Jedná se o grafy, které řadíme do části popisné

statistiky, nazývané **průzkumová (explorační) analýza** dat, které se budeme podrobněji věnovat v další kapitole.

5.2 Číselné charakteristiky

Tabulka rozdělení četností a její grafické znázornění je sice přehlednější než původní neutříděná řada čísel, ale aby byly některé vlastnosti proměnné lépe patrné, musíme informace o nich koncentrovat do jediného (vhodného) čísla, tzv. **číselné charakteristiky**.

Statistickými charakteristikami nazýváme číselné hodnoty, které nám podávají základní informaci o vlastnostech statistického souboru. Podle toho, jaké vlastnosti rozdělení tyto charakteristiky popisují, je dělíme na charakteristiky **polohy**, **variability** a **tvaru (šikmost** a **špičatost)**.

Charakteristika polohy (úrovně) poskytuje základní informaci o daném rozdělení. Míra polohy je taková hodnota náhodné veličiny, kolem které se soustřeďují všechny ostatní hodnoty náhodné veličiny. Volba nejvhodnější charakteristiky úrovně závisí na vlastnostech sledovaného rozdělení a na účelu, k němuž chceme tuto charakteristiku použít. Často popisujeme polohu rozdělení několika charakteristikami polohy současně.

Charakteristiky variability (proměnlivosti, rozptýlení) hodnot znaků jsou čísla, která popisují z různých hledisek proměnlivost sledovaného kvantitativního znaku. Pokud je variabilita hodnot znaku nízká, potom příčiny, které ji způsobují, lze pokládat za nepodstatné či náhodné. Vysoká variabilita naopak znamená nevyrovnanost statistických jednotek z hlediska zkoumaného znaku. Vlivy, které toto kolísání způsobují, musíme pokládat za podstatné a provést jejich identifikaci.

Pokud uvažujeme o variabilitě jako o absolutních rozdílech hodnot znaků od střední hodnoty nebo od sebe navzájem, nazýváme ji **absolutní variabilitou** a měříme ji mírami absolutní variace (např. variační rozpětí, rozptyl). V některých případech se vyskytnou souvislosti, kdy absolutní vyjádření a absolutní míry nejsou vhodné ke srovnání kolísání znaku. Jde většinou o případy, kdy zkoumáme variabilitu dvou nebo více znaků, které se liší úrovní, nebo znaků měřených v nestejných měrných jednotkách. V těchto případech používáme **relativní charakteristiky variability**, které měří variabilitu pomocí vyjádření některé absolutní míry v poměru ke střední hodnotě sledovaného znaku (např. variační koeficient).

Charakteristiky tvaru se obvykle používají společně s charakteristikami polohy a variability a slouží k popisu specifických vlastností dat (míry nesouměrnosti). Přestože

statistické soubory mají stejnou úroveň a variabilitu znaku, mohou se přesto od sebe odlišovat tvarem rozdělení četností.

Podle konstrukce se číselné charakteristiky dělí na **momentové** a **kvantilové**. Informace uvedené v této části kapitoly jsou úzce propojené s kapitolou „Číselné charakteristiky náhodných veličin“.

Pro účely definování základních číselných charakteristik (ne všechny číselné charakteristiky jsou vhodné pro všechny typy proměnných) je potřeba určit jejich typ (viz kapitola Základní pojmy). V následujícím textu budeme vycházet z rozdělení proměnných podle **vztahů mezi jejich hodnotami**. Podle tohoto hlediska rozlišujeme proměnné na **nominální**, **ordinální** a **kvantitativní**.

5.2.1 Číselné charakteristiky nominálních proměnných

Nominální proměnná nabývá různých, avšak rovnocenných variant hodnot znaků. Nelze ji smysluplně porovnávat ani seřadit (např. pohlaví, barva, značka auta). Hodnotami nominální proměnné mohou být slovní varianty (text) nebo číselné kódy.

MÍRY POLOHY

Nejjednodušším a jedinou charakteristikou polohy u nominálních proměnných je **modus** $\hat{x}$. Modus určujeme v případech, kdy potřebujeme vystihnout nejtypičtější hodnotu znaku x_i daného souboru.

Jeho stanovení je založeno na určení **modální kategorie** k, tedy kategorie, která má nejvyšší četnost. V případě asymetrických rozdělení je jeho nevýhodou, že necharakterizuje polohu rozdělení příliš přesně. V praktických příkladech, kde je výsledná hodnota znaku ovlivněna různými poměrně silnými faktory, se stává, že dané rozdělení má více než jeden modus. Pak hovoříme o vícemodálním rozdělení četností.

MÍRY VARIABILITY

Při charakterizování variability vycházíme z koncentrace hodnot (hromadění hodnot kolem některé z variant proměnné) v jednotlivých kategoriích. V případě, že popisovaná nominální proměnná má celkem k kategorií, jejichž četnosti jsou $n_1, n_2, n_3, \ldots, n_k$, pak je možné k výpočtu míry variability použít:

Variační poměr

$$V = 1 - p_{mo} \text{ pro}, \quad V \in \langle 0, (k-1)/k \rangle \tag{5.9}$$

Kde: p_{mo} …. relativní četnost modální kategorie.

$$p_{mo} = \frac{n_{mo}}{n} \tag{5.10}$$

Kde: n_{mo} … četnost modální kategorie.

n ……. rozsah souboru.

Výpočet míry variability u nominálních proměnných má význam pouze tehdy, pokud je cílem výzkumu porovnání více nominálních proměnných mezi sebou, anebo pokud provádíme porovnání u jedné proměnné, ale mezi více skupinami (výběry).

5.2.2 Číselné charakteristiky ordinálních proměnných

Ordinální (pořadová) proměnná je typická tím, že u jejích hodnot vždy existuje určité uspořádání. Jednotlivé varianty znaků je možné seřadit a vzájemně porovnávat, ale není možné určit velikost rozdílů mezi nimi (např. známka ve škole, velikost oděvu S, M, L).

MÍRY POLOHY

Polohu rozdělení u tohoto typu proměnných můžeme popsat pomocí výše zmíněného **modu** nebo další charakteristikou, kterou je **medián $\tilde{x}$**.

Medián je hodnota, která dělí soubor na dvě poloviny. Padesát procent jednotek má hodnotu nižší než medián a padesát procent jednotek ji má vyšší než medián. Medián je hodnota znaku prostřední jednotky souboru, pokud jsou jednotky uspořádány podle velikosti a rozsah souboru je lichý. V případě sudého počtu hodnot je medián vypočítán jako aritmetický průměr dvou středních hodnot.

Medián

Lichý rozsah souboru

$$\tilde{x} = x_{\frac{n+1}{2}} \tag{5.11}$$

Sudý rozsah souboru

$$\tilde{x} = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} \tag{5.12}$$

Kde: x_i … hodnota znaku, pro *i* = 1, 2, 3, …., *n*.

Medián řadíme mezi kvantilové charakteristiky, a pro určení míry polohy u ordinálních proměnných lze využít i jiné kvantily rozdělení, nejčastěji horní ($\tilde{x}_{75}$) a dolní kvartil ($\tilde{x}_{25}$) (více viz kapitola Průzkumová analýza dat).

MÍRY VARIABILITY

Pro určení míry variability ordinálních proměnných můžeme použít:

Interkvartilové rozpětí

$$IQR = \tilde{x}_{75} - \tilde{x}_{25} \tag{5.13}$$

Ordinální rozptyl

$$dorvar = 2\sum_{i=1}^{k} P_i(1 - P_i), \text{ pro } Dorvar \in \langle 0, (k-1)/2 \rangle \tag{5.14}$$

Kde: P_i……kumulativní relativní četnosti, pro $i = 1, 2, \ldots, k$.

Normalizovaný ordinální rozptyl

$$norm.dorvar = 2 - \frac{dorvar}{(k-1)}, \text{ pro } nom.dorvar \in \langle 0; 1 \rangle \tag{5.15}$$

Kde: p_i……relativní četnosti jednotlivých kategorii pro $i = 1, 2, \ldots, k$.

5.2.3 Číselné charakteristiky kvantitativních proměnných

Kvantitativní proměnné, nebolí číselné proměnné, jsou vždy měřitelné. Lze s nimi provádět běžné matematické operace a vyhodnocovat je pomocí většiny metod popisné statistiky. Dělíme je na **diskrétní** (například počet studentů) a **spojité** (například příjem).

MÍRY POLOHY

Pro popis **diskrétní kvantitativní proměnné** můžeme využít již výše uvedený **modus** a **medián**. U **spojitých kvantitativních proměnných** je pro určení polohy rozdělení vhodný pouze medián. Výpočet modu z nesetříděných spojitých dat neposkytuje smysluplnou interpretaci.

Číselné charakteristiky, které používáme u obou typů kvantitativních proměnných, jsou **průměry**. V popisné statistice se nejčastěji používá **průměr aritmetický**. Další dva níže uvedené průměry, **harmonický** a **geometrický**, se využívají především v pokročilejších metodách statistických analýz (analýza časových řad, indexní analýza).

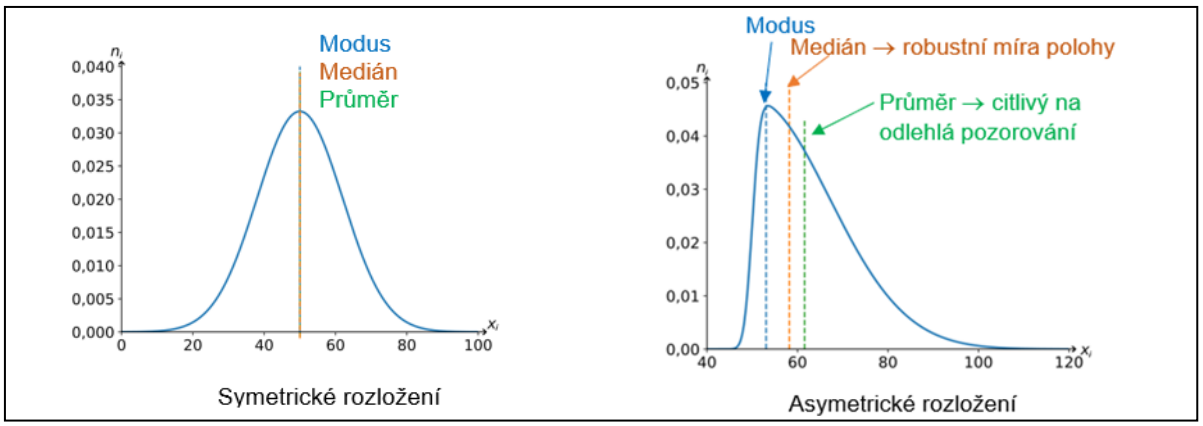
Pro všechny tyto charakteristiky je společné, že jsou určovány na základě všech naměřených hodnot znaku x. Tato vlastnost způsobuje, že průměry jsou citlivé na hodnoty znaků, které se od ostatních hodnot výrazně liší (**odlehlá pozorování**). U souborů s malým rozsahem může odlehlé pozorování značně zkreslit výslednou hodnotu.

V případech, kdy soubor dat obsahuje hodnoty, které se výrazně liší, je u kvantitativních charakteristik vhodnější k určení míry polohy zvolit **medián**, jehož velkou výhodou je **robustnost**. Robustnost číselných charakteristik obecně spočívá v tom, že vypočtené hodnoty

nejsou ovlivněny odlehlými hodnotami. Medián se tak často používá v případech, kdy kvantitativní proměnná má silné zešikmení či extrémní hodnoty.

Uvedené vztahy ilustruje obrázek 5.3. První obrázek představuje symetrické rozložení údajů, ve kterém se všechny základní charakteristiky polohy sobě rovnají. Druhý obrázek znázorňuje případ pravostranné nesouměrnosti dat, kdy převažují hodnoty nižší než aritmetický průměr.

Obrázek 5.3: Grafické znázornění charakteristik polohy



Aritmetický průměr $\bar{x}$ charakterizuje střed polohy rozdělení. Je to nejčastěji používaná míra polohy, která vychází z definice prvního obecného momentu náhodné veličiny X (viz kapitola Číselné charakteristiky náhodných veličin). Aritmetický průměr je definován jako součet všech naměřených hodnot $x_1, x_2, x_3, \ldots, x_n$ dělený celkovým rozsahem souboru n.

Prostá forma

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{5.16}$$

Kde: x_i … hodnota znaku, pro i = 1, 2, 3, …., n,

n … celkový rozsah souboru.

Vážená forma

$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + \cdots + x_k n_k}{n_1 + n_2 + \cdots + n_k} = \frac{\sum_{i=1}^{k} x_i n_i}{n} = \sum_{i=1}^{k} x_i p_i \tag{5.17}$$

$$\sum_{i=1}^{k} n_i = n \tag{5.18}$$

Kde: x_i … hodnota znaku nebo středy intervalů hodnot (aritmetický průměr mezních hodnot znaků), pro i = 1, 2, 3, …., n,

n_i … absolutní četnost znaku v jednotlivých kategoriích, pro i = 1, 2, 3, …., k,

p_i … relativní četnost znaku v jednotlivých kategoriích, pro i = 1, 2, 3, …., k,

k … počet kategorií.

V praxi se můžeme dostat do situace, kdy je třeba určit aritmetický průměr z intervalového rozdělení (setříděné údaje), aniž bychom znali původní řadu hodnot nebo aritmetické průměry intervalů. V takovém případě intervalové průměry „*odhadujeme*“. K tomu je však zapotřebí splnit určitý předpoklad o rozdělení hodnot znaku uvnitř intervalů. Nejčastěji předpokládáme, že je toto rozdělení symetrické, a pro výpočet aritmetického průměru odhadujeme středy intervalů. Celkový aritmetický průměr pak vypočítáme jako vážený aritmetický průměr středů intervalů. Musíme si však uvědomit, že takto získaná hodnota představuje přesnou hodnotu aritmetického průměru pouze v případě, že skutečné aritmetické průměry intervalů jsou totožné se středy intervalů, nebo že se chyby vzniklé asymetrií rozdělení kompenzují.

Obecně se odhad aritmetického průměru vypočítaný pomocí středů intervalů může od skutečné hodnoty aritmetického průměru lišit až o polovinu průměrné délky intervalů (mají-li všechny intervaly stejnou šíři).

Vlastnosti aritmetického průměru

Součet odchylek jednotlivých hodnot od průměru je roven nule.

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0 \tag{5.19}$$

Součet čtverců odchylek jednotlivých hodnot znaku od jejich průměru je vždy menší než součet čtverců odchylek těchto hodnot od libovolné jiné hodnoty $A = x$.

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 < \sum_{i=1}^{n}(x_i - A)^2; A \neq x \tag{5.20}$$

Přičteme-li ke všem hodnotám znaku libovolnou kladnou nebo zápornou konstantu, zvýší se nebo sníží průměr o tuto konstantu.

Vážený aritmetický průměr se nezmění, pokud násobíme váhy (tj. četnosti) libovolnou konstantou různou od nuly.

Transformujeme-li hodnoty původního znaku X na hodnoty znaku Y tak, že $y_i = ax_i + b$, $i = 1,..,n$, kde a, b jsou libovolné konstanty $(a \neq 0)$, pak průměry obou znaků splňují relaci.

$$\bar{y} = a\bar{x} + b \tag{5.21}$$

Aritmetický průměr součtu znaků X a Y se rovná součtu aritmetických průměrů těchto znaků při stejných rozsazích souborů.

$$\overline{x + y} = \bar{x} + \bar{y} \tag{5.22}$$

Nevýhodou aritmetického průměru je jeho citlivost na odlehlá pozorování.

Harmonický průměr $\bar{x}_H$ je definován jako podíl rozsahu souboru a součtu převrácených hodnot znaku. Používá se v situacích, kdy je potřeba zohlednit váhu hodnoty sledované proměnné. Využití harmonického průměru ve statistice je především v indexní analýze, kdy se počítá průměr z veličin vyjádřených podílem (například cenové indexy).

$$\bar{x}_H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \tag{5.23}$$

Harmonický průměr nepoužíváme v případě, že se v souboru dat nacházejí nulové hodnoty. Interpretace jeho výsledků je komplikovanější vzhledem k povaze vstupních proměnných (podíl hodnot).

Geometrický průměr $\bar{x}_G$ je statistická veličina, která je definována jako n-tá odmocnina součinu nezáporných reálných čísel. Jeho hodnota není (na rozdíl od aritmetického průměru) výrazně ovlivněna odlehlými hodnotami ze souboru dat.

$$\bar{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdot \ldots \cdot x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i} \tag{5.24}$$

V analýze časových řad se geometrický průměr využívá pro výpočet průměrného tempa růstu.

MÍRY VARIABILITY

Míry variability vyjadřují rozmístění hodnot dané proměnné okolo střední hodnoty. K popisu míry rozptýlenosti používáme dvě skupiny charakteristik – míry absolutní variability a míry relativní variability.

ABSOLUTNÍ míry variability

Absolutní míry variability popisují proměnlivost statistického souboru ve stejných měrných jednotkách, v jakých je zaznamenána statistická proměnná.

Variační rozpětí *R* je nejjednodušší mírou absolutní variability. Variační rozpětí vyjadřuje šířku intervalu, v němž se pohybují jednotlivé hodnoty sledovaného znaku.

$$R = x_{max} - x_{min} \tag{5.25}$$

Kde: x_{min}… minimální hodnota znaku,
x_{max}… maximální hodnota znaku.

Výhodu variačního rozpětí je velmi jednoduchý výpočet. Není to však charakteristika příliš vhodné pro nehomogenní soubory, protože odlehlé krajní hodnoty, vzhledem k ostatním

hodnotám, zvyšují hodnotu variačního rozpětí. Její nevýhodou je také to, že poskytuje jen hrubý a předběžný odhad variability.

Rozptyl S^2 je základní mírou variability. Měří současně variabilitu hodnot kolem aritmetického průměru a variabilitu ve smyslu vzájemných odchylek jednotlivých hodnot znaku.

Prostá forma

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} \tag{5.26}$$

Kde: x_i … hodnota znaku, pro *i* = 1, 2, 3, …., *n,*
$\bar{x}$ …. aritmetický průměr,
n … celkový rozsah souboru.

Vážená forma

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot n_i}{\sum_{i=1}^{k} n_i} = \sum_{i=1}^{k}(x_i - \bar{x})^2 p_i \tag{5.27}$$

Kde: x_i … hodnota znaku nebo středy intervalů hodnot (aritmetický průměr mezních hodnot znaků), pro *i* = 1, 2, 3, …., *n,*
n_i … absolutní četnost znaku v jednotlivých kategoriích, pro *i* = 1, 2, 3, …., *k,*
p_i … relativní četnost znaku v jednotlivých kategoriích, pro *i* = 1, 2, 3, …., *k,*
k … počet kategorií.

Vlastnosti rozptylu

Jestliže se ke všem individuálním hodnotám znaku přičte nebo odečte konstanta A, rozptyl se nezmění.

Jestliže se všechny individuální hodnoty znaku násobí nebo vydělí konstantou A, je rozptyl počítaný z upravených hodnot A^2- krát větší (nebo menší) než rozptyl počítaný z původních hodnot.

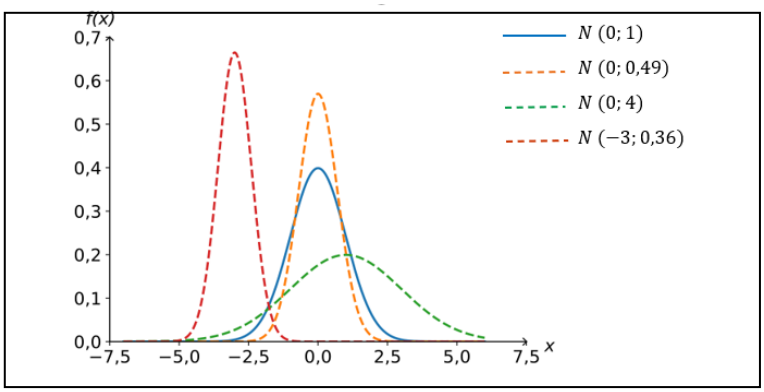
Rozptyl je minimální průměrnou čtvercovou odchylkou, což znamená, že rozptyl kolem libovolné hodnoty a ($A \neq \bar{x}$) znaku X je vždy větší, než rozptyl kolem aritmetického průměru.

Je-li znak konstantní, je rozptyl roven nule. Rozptyl proměnlivého znaku je vždy kladný.

Výše uvedené vzorce pro výpočet rozptylu (5.28; 5.29) se používají jen v popisné statistice obecně k určení proměnlivosti statistického souboru. Pokud však provádíme statistické usuzování a odhadujeme rozptyl neznámé náhodné veličiny z výběrového souboru, je nutné použít upravený vzorec pro zobecnění, tzv. **výběrový rozptyl.** Takto vypočtený výběrový

rozptyl je nestranným odhadem skutečné hodnoty rozptylu neznámé náhodné veličiny (viz kapitola Teorie odhadu).

Obrázek 5.4: Grafické znázornění četností proměnných odlišných ve variabilitě N(0,1)



Rozptyl se obvykle nepoužívá k přímému měření variability, ale jako charakteristika, která umožňuje aplikaci složitějších statistických metod. Rozptyl se uvádí ve čtvercích měrných jednotek hodnot sledovaných proměnných. Při praktickém rozboru variability používáme směrodatnou odchylku.

Směrodatná odchylka S je nejpoužívanější charakteristikou variability, která je definována jako kladná druhá odmocnina rozptylu. Rozměr (jednotky) směrodatné odchylky je stejný jako rozměr veličiny, což je její hlavní výhodou oproti rozptylu.

U souboru s normálním rozdělením četností platí, že v určitých intervalech, které jsou dány násobky směrodatné odchylky kolem aritmetického průměru, je určitá část hodnot sledované náhodné veličiny (viz kapitola Číselné charakteristiky náhodných veličin).

$$S = \sqrt{S^2} \tag{5.28}$$

RELATIVNÍ míry variability

Míry relativní variability jsou bezrozměrná čísla, obvykle se udávají v procentech. Umožňují srovnávání variability proměnných, které jsou vyjádřené v různých měrných jednotkách.

Variační koeficient V je definován jako poměr směrodatné odchylky a průměru. Není ovlivněn absolutními hodnotami sledovaného statistického znaku. Je vhodný pro porovnávání variability dvou či více souborů s odlišnou úrovní hodnot (např. variabilita příjmů v Kč s variabilitou objemu produkce v kg). V takovýchto případech míra variability odstraňuje vliv obecné úrovně daných hodnot.

$$V = \frac{S}{\bar{x}} \cdot 100 \tag{5.29}$$

MÍRY TVARU

Mezi charakteristiky tvaru rozdělení, které jsou založené na momentových charakteristikách, patří **koeficienty šikmosti a špičatosti**, které byly podrobně popsány v kapitole Číselné charakteristiky náhodných veličin. Zde si pro připomenutí uvedeme jejich definici a mezní hodnoty.

Koeficient šikmosti m_3 popisuje soubor hodnot proměnné z hlediska koncentrace nízkých a vysokých hodnot v porovnání se symetrickým rozdělením četností. Vyjadřuje asymetrii rozložení hodnot proměnné kolem jejího průměru (obr. 5.5).

$$m_3 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S} \right)^3 \qquad (5.30)$$

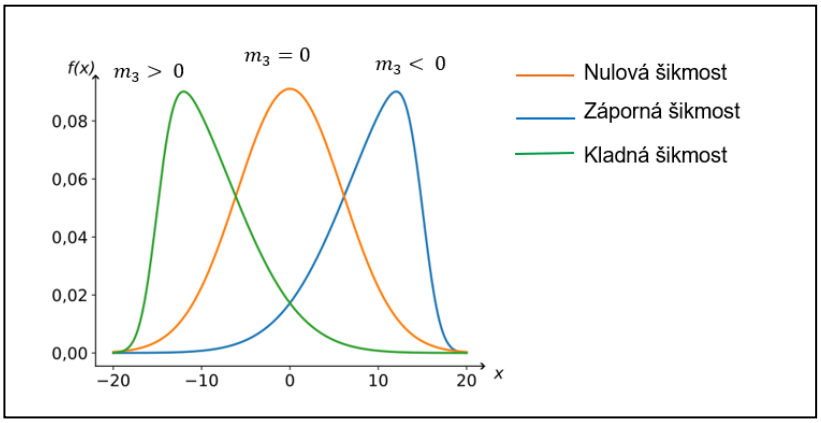
Kde: x_i … hodnota znaku, pro i = 1, 2, 3, …., n,

$\bar{x}$ …. aritmetický průměr,

S …. směrodatná odchylka,

n …. celkový rozsah souboru.

Obrázek 5.5: Grafické znázornění šikmosti



Šikmost rozdělení posuzujeme následovně (obr. 5.5):

$m_3 = 0 \Rightarrow$ nulová šikmost (znamená symetrické rozdělení hodnot nalevo a napravo od střední hodnoty),

$m_3 > 0 \Rightarrow$ kladná hodnota signalizuje levostrannou (odtud kladnou) asymetrii. To znamená vyšší koncentraci podprůměrných hodnot v porovnání s koncentrací hodnot nadprůměrných, převažují zde hodnoty menší než průměr,

$m_3 < 0 \Rightarrow$ záporná hodnota vypovídá o pravostranné (a tedy záporné) asymetrii. To znamená vyšší koncentraci nadprůměrných hodnot v porovnání s koncentrací hodnot podprůměrných, převažují zde hodnoty větší než průměr.

Koeficient špičatosti m_4 popisuje soubor hodnot sledované proměnné z hlediska koncentrace hodnot v souboru kolem střední hodnoty.

$$m_4 = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S}\right)^4 \qquad (5.31)$$

Kde: x_i … hodnota znaku, pro i = 1, 2, 3, …., n,

$\bar{x}$ …. aritmetický průměr,

S …. směrodatná odchylka,

n …. celkový rozsah souboru.

Korigovaný koeficient špičatosti m_4 umožňuje porovnat rozdělení sledované proměnné s normálním pravděpodobnostním rozdělením (obr. 5.6).

$$kurt = m_4 = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S}\right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)} \qquad (5.32)$$

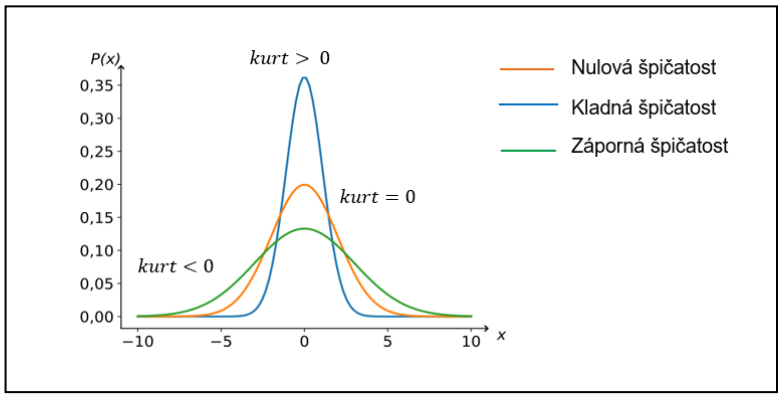
Kde: x_i … hodnota znaku, pro i = 1, 2, 3, …., n,

$\bar{x}$ …. aritmetický průměr,

S …. směrodatná odchylka,

n …. celkový rozsah souboru.

Obrázek 5.6: Grafické znázornění špičatosti



Špičatost rozdělení posuzujeme následovně (obr. 5.6):

$kurt = 0$ $\Rightarrow$ nulová špičatost. Nulová hodnota koeficientu špičatosti znamená, že rozdělení proměnné má tvar Gaussovy křivky – odpovídá normálnímu rozdělení,

$kurt > 0 \Rightarrow$ kladná špičatost znamená, že většina hodnot náhodné veličiny leží blízko její střední hodnotě. Křivka rozdělení náhodné veličiny je špičatější a hlavní vliv na rozptyl mají málo pravděpodobné odlehlé hodnoty,

$kurt < 0 \Rightarrow$ záporná špičatost znamená, že rozdělení je rovnoměrnější a jeho křivka je plošší.

Příklad 5.2

Na základě datové matice *„Notebooky“* o rozsahu 150 jednotek, kterou naleznete v kurzu v Moodle, budeme prezentovat základní zpracování dat pomocí metod z popisné a průzkumové analýzy dat.

Datový soubor obsahuje celkem **16 proměnných**, které můžeme rozdělit následovně:

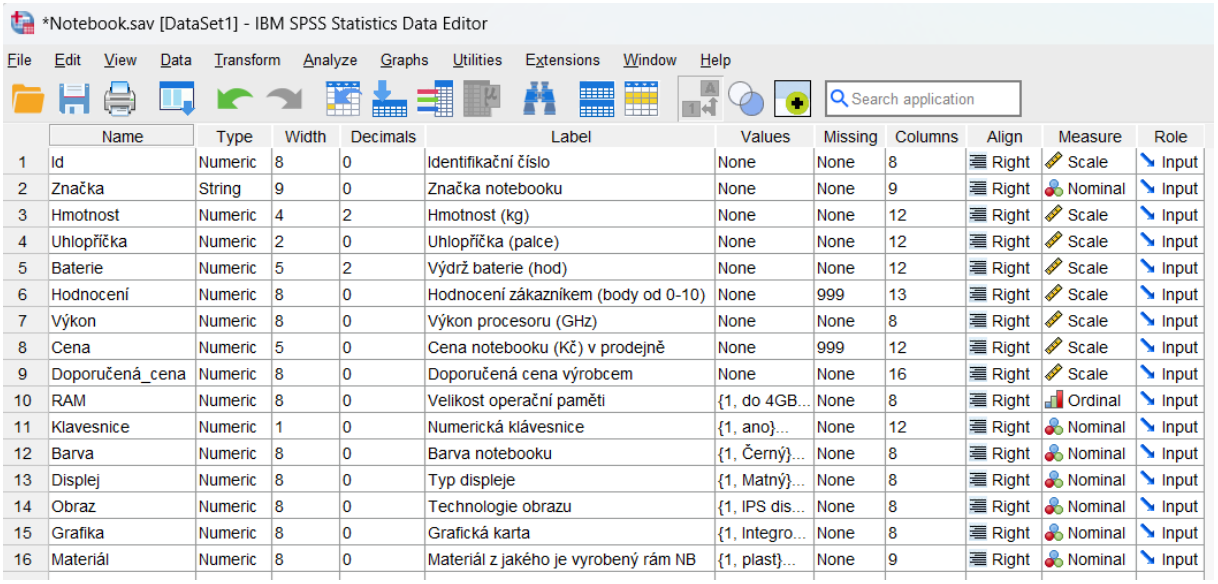
1 slovní proměnná: značka notebooku.

1 identifikační proměnná: ID notebooku.

7 kvantitativních proměnných: cena, úhlopříčka apod.

7 kvalitativních proměnných: barva, grafická karta apod.

Tento vzorový příklad nám pomůže nejen lépe porozumět vlastnostem datového souboru, ale zároveň poslouží jako základ pro procvičování a demonstraci dalších metod, **se kterými se budeme v průběhu kurzu seznamovat**.


*Notebook.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help

Search application

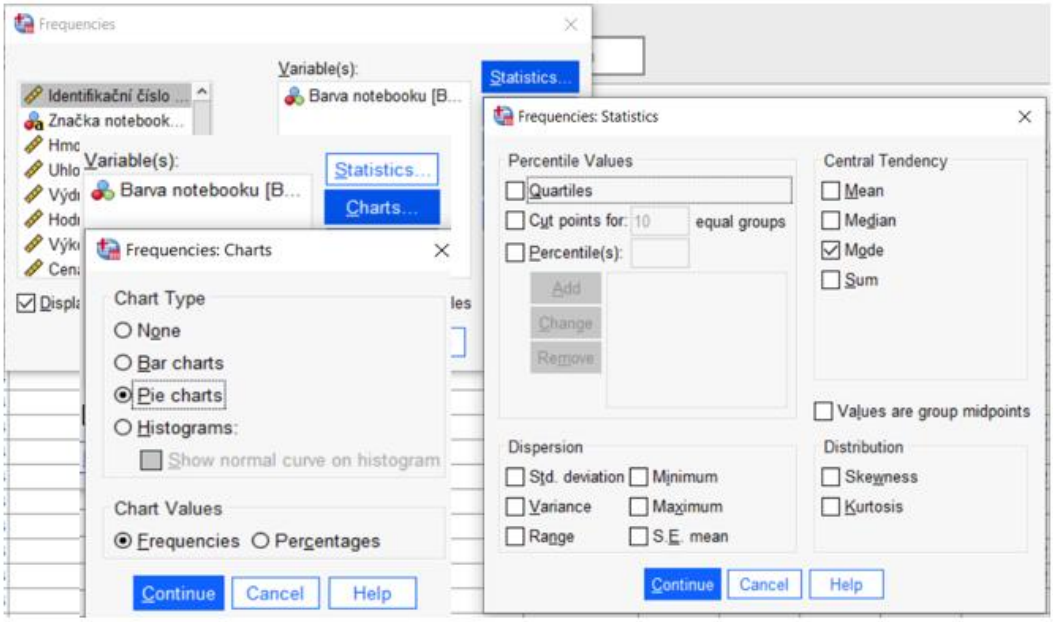
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Id	Numeric	8	0	Identifikační číslo	None	None	8	Right	Scale	Input
2	Značka	String	9	0	Značka notebooku	None	None	9	Right	Nominal	Input
3	Hmotnost	Numeric	4	2	Hmotnost (kg)	None	None	12	Right	Scale	Input
4	Uhlopříčka	Numeric	2	0	Uhlopříčka (palce)	None	None	12	Right	Scale	Input
5	Baterie	Numeric	5	2	Výdrž baterie (hod)	None	None	12	Right	Scale	Input
6	Hodnocení	Numeric	8	0	Hodnocení zákazníkem (body od 0-10)	None	999	13	Right	Scale	Input
7	Výkon	Numeric	8	0	Výkon procesoru (GHz)	None	None	8	Right	Scale	Input
8	Cena	Numeric	5	0	Cena notebooku (Kč) v prodejně	None	999	12	Right	Scale	Input
9	Doporučená_cena	Numeric	8	0	Doporučená cena výrobcem	None	None	16	Right	Scale	Input
10	RAM	Numeric	8	0	Velikost operační paměti	{1, do 4GB...	None	8	Right	Ordinal	Input
11	Klavesnice	Numeric	1	0	Numerická klávesnice	{1, ano}...	None	12	Right	Nominal	Input
12	Barva	Numeric	8	0	Barva notebooku	{1, Černý}...	None	8	Right	Nominal	Input
13	Displej	Numeric	8	0	Typ displeje	{1, Matný}...	None	8	Right	Nominal	Input
14	Obraz	Numeric	8	0	Technologie obrazu	{1, IPS dis...	None	8	Right	Nominal	Input
15	Grafika	Numeric	8	0	Grafická karta	{1, Integro...	None	8	Right	Nominal	Input
16	Materiál	Numeric	8	0	Materiál z jakého je vyrobený rám NB	{1, plast}...	None	9	Right	Nominal	Input

Pokračování příklad 5.2

Nyní se budeme podrobněji věnovat zpracování jednotlivých typů proměnných.

a) V prvním kroku se zaměříme na zpracování **nominální proměnné** „*Barva*“.

SPSS Analyze → Descriptive Statistics → Frequencies → proměnná *Barva* do okna Variable(s) → Statistics → Continue → Charts → Continue → OK



Řešení

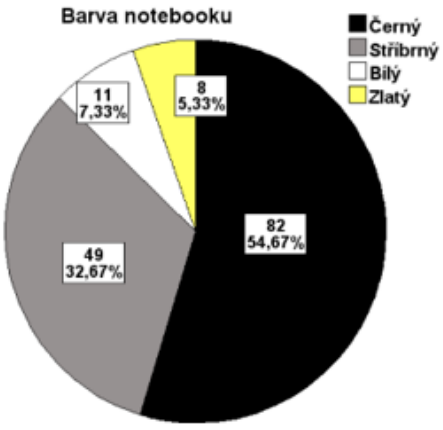
Statistics

Barva notebooku

N	Valid	150
	Missing	0
Mode		1

Barva notebooku

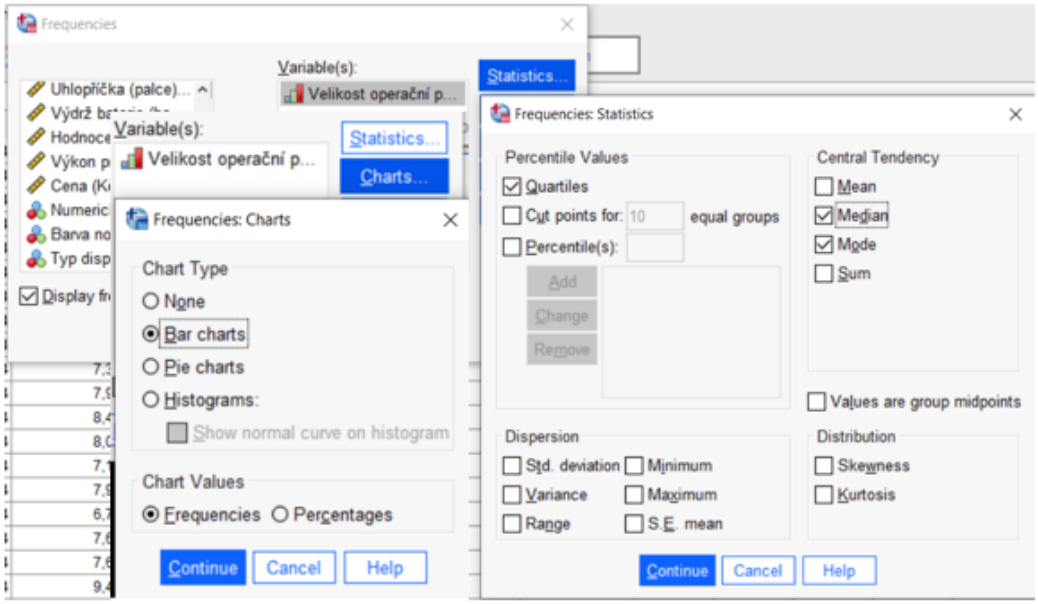
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Černý	82	54,7	54,7	54,7
	Stříbrný	49	32,7	32,7	87,3
	Bílý	11	7,3	7,3	94,7
	Zlatý	8	5,3	5,3	100,0
	Total	150	100,0	100,0	



Pokračování příklad 5.2

b) V druhém kroku se zaměříme na zpracování **ordinální proměnné** „*Ram*“.

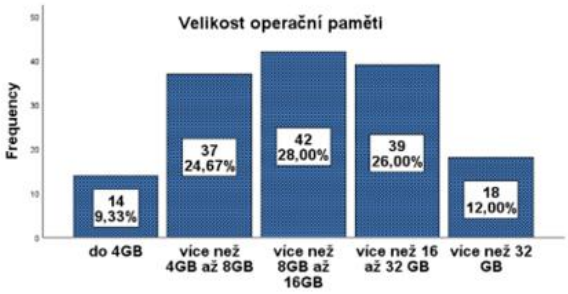
SPSS Analyze → Descriptive Statistics → Frequencies → proměnná *Ram* do okna Variable(s) → Statistics → Continue → Charts → Continue → OK



Řešení

Velikost operační paměti

N	ROZSAH SOUBORU	n	Valid	150
			Missing	0
Median	MEDIÁN	$\tilde{x}$		3,00
Mode	MODUS	$\hat{x}$		3
Percentiles	KVARTIL		25 $\tilde{x}_{25}$	2,00
			50 $\tilde{x}_{50}$	3,00
			75 $\tilde{x}_{75}$	4,00

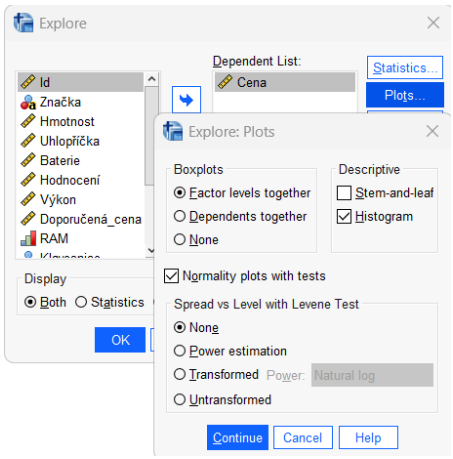


Velikost operační paměti		Frequency Absolutní četnost n_i	Percent Relativní četnost p_i	Valid Percent Relativní četnost z platných jednotek p_i	Cumulative Percent Kumulativní četnost P_k
Valid	do 4 GB	14	9,3	9,3	9,3
	více než 4 GB až 8 GB	37	24,7	24,7	34,0
	více než 8 GB až 16 GB	42	28,0	28,0	62,0
	více než 16 GB až 32 GB	39	26,0	26,0	88,0
	více než 32 GB	18	12,0	12,0	100,0
	Total	150	100,0	100,0	

Pokračování příklad 5.2

c) Ve třetím kroku se zaměříme na zpracování **kvantitativní proměnné** „*Cena*“.

SPSS Analyze → Descriptive Statistics → Explore → do okna Dependent List: proměnná *Cena* → Plots → Normality plots with tests → Continue → OK

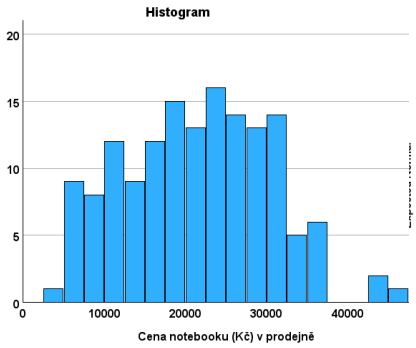


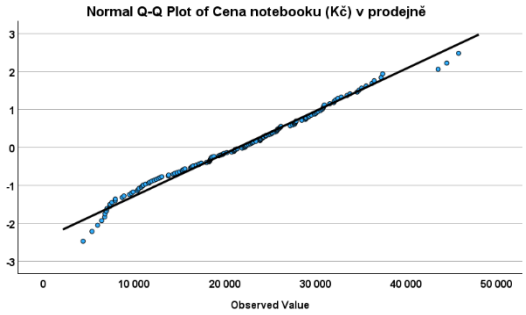
Řešení

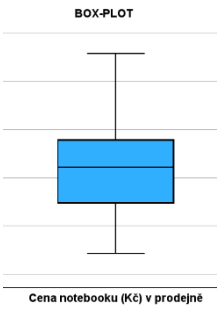
			Statistic	Značení
Cena	**Mean**	**PRŮMĚR**	**21492,44**	$\bar{x}$
	95% Confidence Interval for Mean	Lower Bound	20052,37	
		Upper Bound	22932,51	
	5% Trimmed Mean		21325,44	
	Median	**MEDIÁN**	**22179,50**	$\tilde{x}$
	Variance	**ROZPTYL**	**79666921,3**	S^2
	Std. Deviation	**SMĚRODATNÁ ODCHYLKA**	**8925,633**	S
	Minimum		4365	
	Maximum		45776	
	Range	**VARIAČNÍ ROZPĚTÍ**	**41411**	R
	Interquartile Range	**INTERKVANTILOVÉ ROZPĚTÍ**	**13154**	IRQ
	Skewness	**ŠIKMOST**	**,144**	m_3
	Kurtosis	**ŠPIČATOST**	**-,453**	$kurt$

Tests of Normality

	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Cena	,048	150	,200	,983	150	,069







5.3 Průzkumová analýza datových souborů

Průzkumová analýza (EDA – Exploratory Data Analysis) umožňuje popsat důležité vlastnosti datového souboru stejně jako statistika popisná. Rozdíl mezi průzkumovou a popisnou analýzou není tolik v použitých metodách, jako ve formě jejich zpracování.

Cílem průzkumové analýzy je popsat statistické zvláštnosti dat a ověřit předpoklady o výběru, ze kterého pocházejí. Většina postupů průzkumové analýzy dat je založena na základních popisných charakteristikách a na grafických metodách, které lze efektivně provádět pouze s použitím statistických programů.

K popisu dat se využívají především robustní kvantilové charakteristiky, jejichž základem je pořádková statistika, která pracuje s vzestupně seřazenými hodnotami datové proměnné.

Robustnost metody lze vyjádřit jako odolnost metody vůči malým změnám či odchylkám v postupu nebo v parametru modelu. U robustních metod není vliv vybočujících hodnot na výsledky analýzy tak výrazný. Robustnost poskytuje informaci o spolehlivosti modelu.

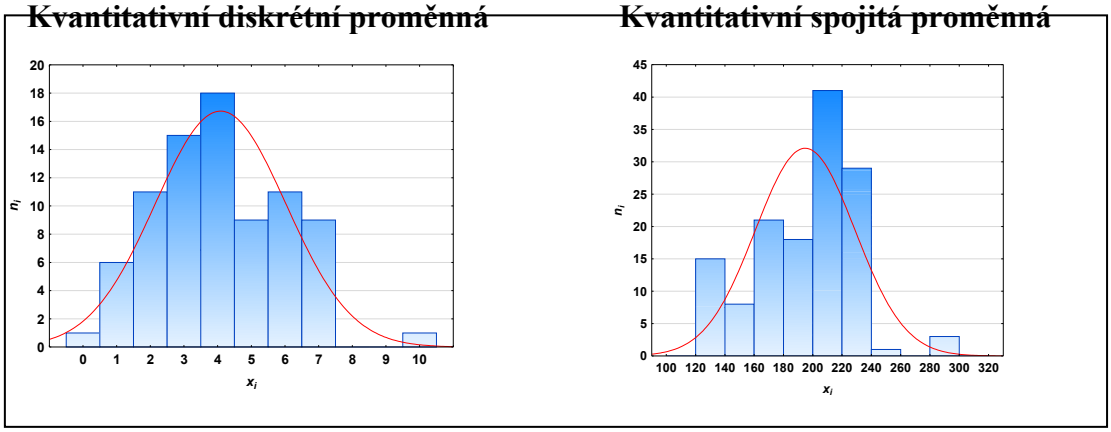
Čím je metoda odolnější proti vlivu chyb, tím je robustnější.

5.3.1 Statistické zvláštnosti dat

Nejprve se zaměříme na grafické metody průzkumové analýzy dat, které jsou zaměřeny na identifikaci vybočujících pozorování a stupně rozložení šikmosti a špičatosti dat.

Histogram je jeden z nejpoužívanější nástrojů prezentace kvantitativní proměnné (obr. 5.7). Každý sloupec a výška tohoto sloupce v histogramu znázorňuje četnost výskytu daného znaku. Na rozdíl od sloupcového grafu, v němž jsou při zobrazování četností kategorií pro jednu proměnnou sloupce oddělené, jsou v histogramu sloupce umístěny těsně vedle sebe.

Obrázek 5.7: Histogramy

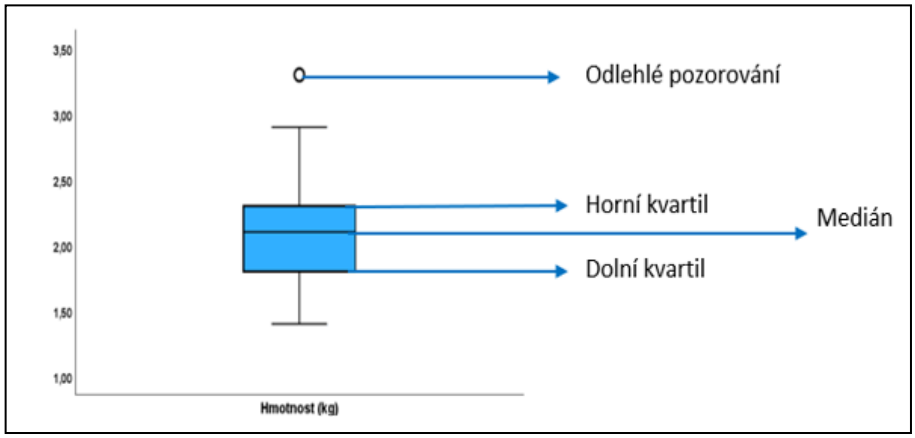


U diskrétní proměnné je na vodorovné ose znázorněna četnost výskytu vztažená k variantě znaku sledované proměnné. U spojitých proměnných jsou na vodorovné ose znázorněny stejně široké, vzájemně navázané (disjunktivní) intervaly sledované proměnné (pravidla pro stanovení počtu intervalů vztahy 5.5 a 5.6). Nevýhodou tohoto grafu je, že v případě nevhodného stanovení počtu intervalů (sloupců) jsou jeho interpretační možnosti značně omezené.

Výhodou histogramu je jeho jednoduchost, někdy nám však chybí informace o konkrétních hodnotách proměnné.

Grafem, který je doplněn o základní číselné charakteristiky je **box-plot** neboli **krabicový graf**. Tento graf zobrazuje data ve tvaru obdélníkové krabice, která je doplněna tzv. vousy (obr. 5.8).

Obrázek 5.8: Box-plot



Jednotlivé části krabicového grafu odpovídají pětičíselnému souhrnu kvantilových charakteristik (minimum, maximum, medián, dolní a horní kvartil), které podávají rychlou a přehlednou informaci o poloze, variabilitě a případném asymetrickém rozložení hodnot zkoumaného statistického souboru.

Horní hranice krabice je ohraničena **75% kvantilem** a **dolní hranici tvoří 25% kvantil.** Tyto dva kvartily odpovídají **kvartilovému rozpětí**, které ohraničuje 50 % pozorovaných údajů. Z hodnot kvartilů lze usuzovat na **velikost variability** zobrazeného souboru, početně vyjádřenou **interkvartilovým rozpětím** (vztah 5.15). Uprostřed krabice je **bodem nebo svislou příčkou** označen **medián**, tedy **50% kvantil**.

V případě symetricky rozdělených dat je medián přesně uprostřed krabice a v takovém případě data odpovídají normálnímu rozdělení.

Vousy, které vybíhají z krabice, signalizují **míru asymetrie rozložení** dat. Data jsou asymetrická v případě, kdy je jeden vous (úsečka vycházející z krabice) větší než druhý.

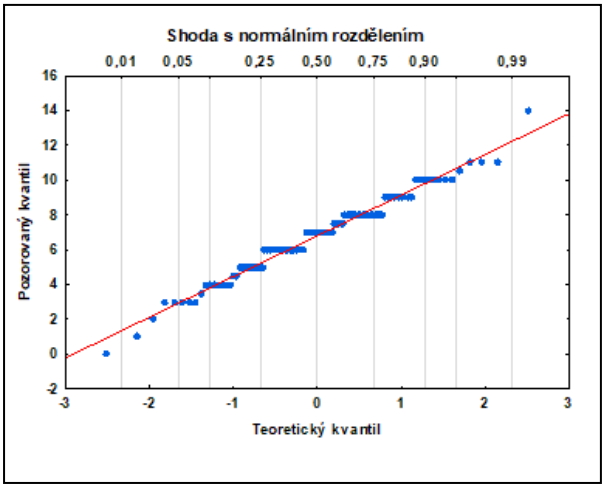
Koncové úsečky na vousech označují maximální a minimální „nevybočující“ hodnoty proměnné.

Hodnoty, které se nacházejí mimo grafické znázornění krabice a v grafu jsou vyobrazeny jako izolované body, jsou považovány za **odlehlá (vybočující) pozorování**.

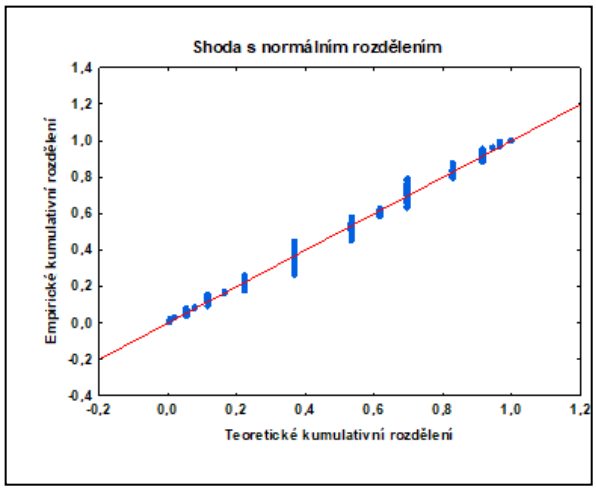
Shodu **kvantilů výběrového** (empirického) rozdělení s vybraným **teoretickým** rozdělením nám umožňuje posoudit **Kvantil-kvantilový graf ($Q-Q$ graf).** Princip jeho sestrojení spočívá ve **vzestupném** uspořádání naměřených hodnot v porovnání s hodnotami stanovenými pomocí pravděpodobnostní funkce zvoleného rozdělení (obr. 5.9).

V případě, že výběrové rozdělení plně odpovídá teoretickému, je grafem přímka. Jakékoli odchylky naměřených hodnot od přímky znamenají odchylky od předpokládaného teoretického rozdělení. Graf $Q-Q$ lze sestrojit pro různá rozdělení, pouze se jinak stanovují příslušné hodnoty na ose x (hodnoty teoretického kvantilu Q_T) a ose y (hodnoty empirického kvantilu Q_E).

Obrázek 5.9: $Q-Q$ graf



Obrázek 5.10: $P-P$ graf



Výstupem podobný předcházejícímu $Q-Q$ grafu, jen graf **Graf $P-P$** (obr. 5.10) ve kterém jsou na osách x a y vykresleny empirické a teoretické kumulativní četnosti.

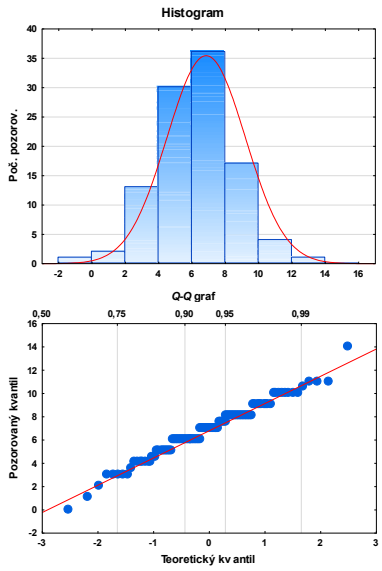
Grafy $P-P$ a $Q-Q$ se vzájemně doplňují. Grafy $P-P$ jsou citlivé na odchylky od teoretického rozdělení ve střední části – v okolí modu. Grafy $Q-Q$ jsou citlivé na odchylky od teoretického rozdělení v oblasti konců.

Příklad 5.3

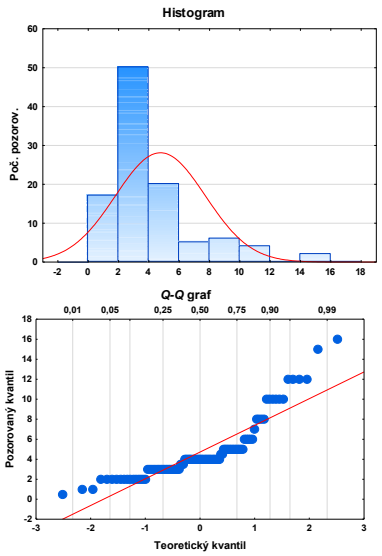
Pro ilustraci provedení a interpretace uvedených metod průzkumové analýzy dat byly vybrány tři proměnné s různým rozložením dat. Na základě níže uvedených pravděpodobnostních grafů provedeme vizuální kontrolu předpokladů o datech a jejich rozdělení – **průzkumovou analýzu**. Pomocí základních číselných charakteristik posoudíme kvalitu dat – **popisnou statistiku**.

Průzkumová analýza

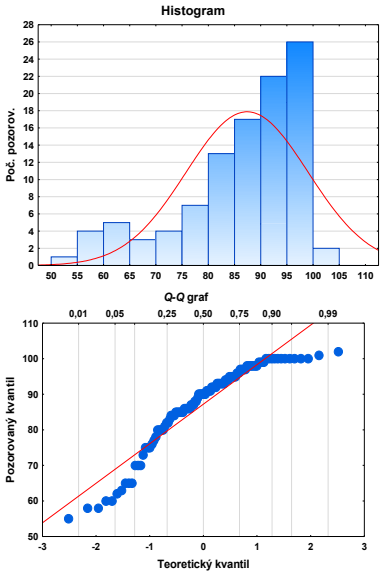
Počet hodin ve škole za den[5]



Počet hodin na PC/den[6]



Tepová frekvence po sportu[7]



Popisná statistika

Průměr	6,803
Medián	7
Minimální	0
Maximál.	14
Rozpětí	14
Kvartil	3
Sm.odch.	2,340
Var.koef.	34,404
Šikmost	-0,108
Špičatost	0,493

Průměr	4,721
Medián	4
Minimální	0,5
Maximál.	16
Rozpětí	15,5
Kvartil	2
Sm.odch.	2,950
Var.koef.	62,487
Šikmost	1,747
Špičatost	3,078

Průměr	87,221
Medián	90
Minimální	55
Maximál.	102
Rozpětí	47
Kvartil	15
Sm.odch.	11,604
Var.koef.	13,304
Šikmost	-1,042
Špičatost	0,392

[5] Data vykazují dobrou shodu, která je například v grafu *Q-Q* indikována tím, že jednotlivé body (kvantily) leží těsně kolem přímky. Odchylku vykazuje pouze jedna nejvyšší a jedna nejnižší hodnota. Můžeme tedy předpokládat, že vzhledem k velikosti výběru (n=104) nebudou mít tyto body větší vliv na normalitu rozložení dat. Oba uvedené grafy potvrzují typické vlastnosti normálního rozdělení. Srovnání mediánu a aritmetického průměru indikuje velmi dobrou shodu, což je typické právě pro normální rozdělení nebo symetrická rozdělení blízká normálnímu.

[6] Z grafů vyplývá, že rozdělení dat je špičatější a neodpovídá průběhu normálního rozdělení. Předpokládáme, že hlavní vliv na rozptýlenost mají spíše méně odlehlé hodnoty.

[7] Jedná se o předpoklad špičatosti, a především levostranné nesouměrnosti (šikmosti).

Shrnutí kapitoly

Popisná statistika společně s průzkumovou analýzou v sobě zahrnují vzájemně provázanou skupinu základních metod, které slouží k prvotnímu posouzení kvality dat.

Popisná statistika se zabývá uspořádáním souborů, jejich popisem a početními postupy, které poskytují souhrnné informace o vlastnostech analyzovaného datového souboru. Úkolem popisné statistiky je tabelování a číselný popis datového soubor.

Proměnné / **Míry**	**Nominální**	**Ordinální**	**Kvantitativní**
Polohy	Modus	Modus	Modus
		Medián	Medián
			Průměry
Variability absolutní	Variační poměr	Interqvartilové rozpětí	Variační rozpětí
	Modální rozptyl	Ordinální rozptyl	Rozptyl
			Směrodatná odchylka
Variability relativní			Variační koeficient
Tvaru			Šikmost a špičatost

Úkolem **průzkumové analýzy** je odhalit zvláštnosti v chování dat a ověřit předpoklady, které data musí splňovat (tj. nezávislost, homogenitu a normalitu). Pro průzkumovou analýzu je stěžejní grafické zobrazení dat. Při posuzování dat je vhodné kombinovat obě výše uvedené metody. Věcná znalost řešené problematiky představuje nenahraditelnou podmínku pro získání kvalitních výsledků a jejich následnou interpretaci.

Literatura

BUDÍKOVÁ, Marie; Maria KRÁLOVÁ a Bohumil MAROŠ. *Průvodce základními statistickými metodami. vydání první*. Praha: Grada Publishing, a.s., 2010. ISBN 978-80-247-3243-5.

CAMM, Jeffrey; James COCHRAN, David ANDERSON, Dennis SWEENEY, Thomas WILLIAMS et al. *Statistics for Business and Economics*. 15th Edition. Boston: Cengage Learning, 2023. ISBN 978-0357715857.

HINDLS, Richard; Stanislava HRONOVÁ, Jan SEGER a Jakub FISCHER. *Statistika pro ekonomy*. 8. vyd. Praha: Professional Publishing, 2007. ISBN 978-80-86946-43-6.

MELOUN, Milan, Jiří MILITKÝ a Martin HIL. *Statistická analýza vícerozměrných dat v příkladech*. Vyd. 2. Praha: Academia, 2012. ISBN 978-80-2002-071-0.

ŘEZANKOVÁ, Hana a Tomáš LÖSTER. *Základy statistiky*. Vyd. 1. Praha: Oeconomica, 2013. ISBN 978-80-245-1957-9.

5 Kontrolní otázky

1. Jaké jsou hlavní etapy statistické práce a co je jejich cílem?
2. Co je statistická jednotka? Uveďte příklad.
3. Co je statistický znak?
4. Co je cílem grafického znázornění?
5. Jaký je rozdíl mezi absolutní a relativní četností?
6. Definujte nominální a ordinální proměnnou. Uveďte příklady.
7. Jaký je rozdíl mezi diskrétními a spojitými kvantitativními proměnnými?
8. Jak dělíme číselné charakteristiky proměnných?
9. Co jsou charakteristiky polohy a jakou informaci nám poskytují?
10. Jaký je rozdíl mezi absolutními a relativními charakteristikami variability?
11. Co je modus a medián. Vysvětlete pojem robustnost.
12. Které grafy jsou vhodné pro zobrazení kvantitativních diskrétních proměnných a které pro nespojité, nominální a ordinální proměnné?
13. Jak se dělí charakteristiky variability a jaký je mezi nimi rozdíl?
14. Jaké znáte absolutní charakteristiky míry variability?
15. Co udává variační koeficient a proč je vhodný pro srovnání variability různých souborů?
16. K čemu slouží krabicový graf (box-plot) a jaké informace z něj můžeme vyčíst?
17. Proč je důležité správně klasifikovat proměnné před statistickým zpracováním?

5 Příklady k procvičení

Použijte datovou matici **„***Byty***“** kterou naleznete v kurzu Moodle, a odpovězte na následující otázky.

5.1 Nominální proměnné

5.1.1 Pracujte s proměnnými *Město* a *Výtah.*

a) Vytvořte tabulku rozdělení četností.
b) Vytvořte vhodný graf, který vizuálně ukáže podíl bytů v jednotlivých městech.
c) Určete modus u obou proměnných.
d) Vytvořte graf, který bude prezentovat průměrnou plochu bytu v závislosti na městě ve kterém se byt nachází.

5.2 Ordinální proměnné

5.2.1 Pracujte s proměnnými *Dispozice* a *Patro.*

a) Vytvořte tabulku rozdělení četností.
b) Vytvořte sloupcový graf pro proměnnou *Dispozice*.
c) Jaký závěr lze z grafu vyvodit o nabídce bytů na trhu?
d) Jaké charakteristiky polohy jsou vhodné pro proměnné *Dispozice* a *Patro*?
e) Vytvořte graf, který bude prezentovat průměrnou cenu nájmu v závislosti na dispozici bytu.

5.3 Kvantitativní proměnné

5.1.1 Pracujte s proměnnými *Cena* a *Plocha.*

a) Vypočítejte a interpretujte vhodné charakteristiky polohy a variability pro proměnnou *Cena*.
b) Vytvořte histogram pro proměnnou *Cena*. Posuďte, zda data vykazují symetrické rozdělení.
c) Vypočítejte, jaká je průměrná *Cena* nájmu v Praze pro bytu 2+kk?
d) Vypočítejte vhodné charakteristiky polohy pro proměnnou *Plocha*. Co vám tyto hodnoty říkají o velikosti bytů?
e) Vytvořte krabicový graf (box-plot) pro proměnnou *Plocha*. Identifikujte a interpretujte případná odlehlá pozorování.

6 NÁHODNÝ VÝBĚR

Dosud jsme se věnovali popisné statistice, která nám umožňuje shrnout a vizualizovat data, a teorii pravděpodobnosti, která nám dává nástroje pro pochopení chování náhodných jevů. Tyto teoretické základy jsou nezbytným předpokladem pro přechod k matematické statistice, která se zabývá zevšeobecňováním.

Abychom mohli efektivně usuzovat na vlastnosti celého základního souboru (populace), musíme nejprve pochopit, jak z něj získáváme data. Slouží k tomu náhodný výběr, který je tvořen náhodnými veličinami, které jsou spojeny s libovolným náhodným pokusem.

Náhodný výběr z cílové populace $X = (X_1, X_2, \dots, X_k)$, je vektor nezávislých náhodných veličin (k-náhodných výběrů) stejného typu rozdělení. Realizací náhodného výběru X získáme hodnoty $x = x_1, x_2, \dots, x_n$, kde x jsou nazývány „pozorování“ (reálná čísla) a n je rozsah výběru. Tyto konkrétní hodnoty tvoří datový soubor (neboli výběrový soubor).

Příklad 6.1

Chceme získat informace o průměrné spotřebě paliva u všech vozidel určité značky (např. Škoda) a modelu (např. Octavia), které byly vyrobeny v roce 2023.

Řešení

Základní soubor (populace): Všechny vozy Škoda Octavia vyrobené v roce 2023. Tento soubor je tak rozsáhlý, že není možné změřit spotřebu u každého jednotlivého vozu.

Náhodný pokus: Vybereme jeden vůz z populace a změříme jeho spotřebu paliva na standardní trase 100 km. Výsledek tohoto měření je náhodná veličina, protože se může mírně lišit v závislosti na řidiči, stavu vozovky, počasí a dalších faktorech.

Náhodný výběr: Protože nemůžeme změřit všechny vozy, provedeme náhodný výběr. Vybereme například 500 vozů. Těchto 500 měření spotřeby představuje náš **náhodný výběr**. Každé jednotlivé měření $(x_1, x_2, \dots, x_{500})$ je nezávislá náhodná veličina, která pochází ze stejného rozdělení jako spotřeba celé populace vozů.

Realizace náhodného výběru: Když skutečně provedeme měření, získáme konkrétní číselné hodnoty například 5,2l/100 km; 5,5l/100 km; 5,0l/100 km atd. Tento soubor 500 hodnot je **realizací náhodného výběru** (nebo datovým souborem). Právě s těmito konkrétními hodnotami budeme pracovat v praktické části statistiky.

6.1 Výběrové charakteristiky

Vlastnosti populace definujeme pomocí **populačních (teoretických) charakteristik** neboli **parametrů** (např. střední hodnota, rozptyl), které popisují určitou vlastnost základního souboru. Jsou to **konstantní hodnoty**, které v případě, že neznáme rozdělení pozorované náhodné veličiny, nedokážeme většinou přesně určit.

Tabulka 6.1: Značení populačních a výběrových charakteristik

	Populační parametr (Teorie pravděpodobnosti)	**Populační parametr** (Matematická statistika)	**Výběrová charakteristika** (Popisná statistika)
Průměr	$E(X)$	μ	$\bar{x}$
Medián	$Q_{0,5}(X)$	$\tilde{\mu}$	$\tilde{x}$
Rozptyl	$D(X)$	σ^2	S^2
Směrodatná odchylka	$\sqrt{D(X)}$	σ	S
Relativní četnost	P	π	p

Z výběrových dat však můžeme populační parametry odhadnout. Na základě výběrových pozorování vypočítáme **empirickou charakteristiku** neboli **statistiku**, kterou obecně značíme $T(X)$ a konkrétně ji značíme písmeny z latinky.

Statistika $T \rightarrow T(X = x)$ je funkcí náhodného výběru a v určitém pravděpodobnostním smyslu ji lze chápat, jako číselnou hodnotu, která nám umožňuje přiblížit se skutečné hodnotě populačního parametru anebo jako hodnotu, která nám charakterizuje stupeň nesouladu mezi předpokladem o datech (charakteristikách náhodných veličin) a hodnotami, které získáme na základě výběrových šetření.

Kritéria a metody výpočtu výběrové statistiky, se odvozují z charakteru rozdělení náhodné veličiny a vychází z principů Centrální limitní věty (CLV). Ta říká, že náhodná veličina, která vznikne jako součet dostatečně **velkého počtu**[8] vzájemně nezávislých náhodných veličin, má za velmi obecných podmínek přibližně normální rozdělení, a to bez ohledu na to, jaké bylo původní rozdělení těchto veličin.

[8] Neexistuje žádná konkrétní hodnota pro dostatečný počet. Běžně se však setkáváme s pravidlem, že výběry o rozsahu 30 a větším jsou považovány za dostatečně velké pro platnost centrální limitní věty, a to i když je toto pravidlo spíše zjednodušením, které se ustálilo pro potřeby ručních výpočtů v minulosti.

Tabulka 6.2: Výběrové rozdělení náhodných veličin

Rozdělení	Parametr	Kvantil
Normální (Gausovo)	$X \sim N(\mu, \sigma^2)$	
Normované normální rozdělení	$X \sim N(0; 1)$	u_α
Alternativní rozdělení	$X \sim A(\pi)$	
Aproximace na normované normální	$X \sim A(\pi) \sim N(0; 1)$	u_α
***Chí*-kvadrát rozdělení**	$X \sim \chi^2(\nu)$	$\chi^2_\alpha(\nu)$
Studentovo *t*-rozdělení	$X \sim \ \sim t(\nu)$	$t_\alpha(\nu)$
Fisher-Snedecorovo-rozdělení	$X \sim F(\nu_1; \nu_2)$	$F_\alpha(\nu_1; \nu_2)$

Uvedené koncepty nám umožňují přejít k praktickému modelování nejdůležitějších charakteristik náhodného výběru (statistik), které budeme využívat k odhadu neznámých populačních parametrů a k testování hypotéz o populaci s použitím reálných dat.

Přehled nejpoužívanějších výběrových charakteristik a jejich rozdělení

JEDEN NÁHODNÝ VÝBĚR

Modelování PRŮMĚRU

Známe populační rozptyl $\boldsymbol{\sigma^2}$

$\boldsymbol{n > 30}$

$$\boldsymbol{U = \frac{\bar{X} - \mu}{\sigma}\sqrt{n} \sim u_\alpha} \tag{6.1}$$

Neznáme populační rozptyl $\boldsymbol{\sigma^2}$

$\boldsymbol{n > 30}$

$$\boldsymbol{U = \frac{\bar{X} - \mu}{S}\sqrt{n} \sim u_\alpha} \tag{6.2}$$

Neznáme populační rozptyl $\boldsymbol{\sigma^2}$

$\boldsymbol{n < 30}$

$$\boldsymbol{T = \frac{\bar{X} - \mu}{S}\sqrt{n} \sim t_{\alpha(n-1)}} \tag{6.3}$$

Modelování ROZPTYLU

$$\boldsymbol{K = \frac{n-1}{\sigma^2}s^2 \sim \chi^2_{\alpha(n-1)}} \tag{6.4}$$

Modelování RELATIVNÍ ČETNOSTI

$\boldsymbol{n > \frac{9}{p(1-p)}}$

$$\boldsymbol{P = \frac{p - \pi}{\sqrt{\pi(1-\pi)}}\sqrt{n} \sim u_\alpha} \tag{6.5}$$

DVA NEZÁVISLÉ NÁHODNÉ VÝBĚRY

Zkoumáme rozdílnost hodnot náhodné veličiny pozorované v různých podmínkách.

Modelování PRŮMĚRŮ

Známe populační rozptyl $\boldsymbol{\sigma^2}$

$$\boldsymbol{U_{(\bar{X}-\bar{Y})} = \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim u_\alpha} \tag{6.6}$$

Neznáme rozptyl, ale $\boldsymbol{\sigma_1^2 = \sigma_1^2}$

$$\boldsymbol{T_{(\bar{X}-\bar{Y})} = \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{S_1^2(n_1-1)+S_2^2(n_2-1)}}\sqrt{\frac{n_1 n_2(n_1+n_2-2)}{n_1+n_2}} \sim t_{\alpha(n_1+n_2-2)}} \tag{6.7}$$

Neznáme rozptyl, ale $\boldsymbol{\sigma_1^2 \neq \sigma_1^2}$

$$\boldsymbol{T_{(\bar{X}-\bar{Y})} = \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \sim t_{\alpha(\nu)}}, \text{ kde } \boldsymbol{\nu \cong \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2\frac{1}{n_1-1}+\left(\frac{s_2^2}{n_2}\right)^2\frac{1}{n_2-1}}} \tag{6.8}$$

Modelování ROZPTYLŮ

$$\boldsymbol{F = \frac{S_1^2}{\sigma_1^2} / \frac{S_2^2}{\sigma_2^2} \sim F_{n_1-1;n_2-1}} \tag{6.9}$$

Modelování RELATIVNÍCH ČETNOSTÍ

$$\boldsymbol{\left(n_1 > \frac{9}{p_1(1-p_1)}\right)} \quad \boldsymbol{\left(n_2 > \frac{9}{p_2(1-p_2)}\right)}$$

$$\boldsymbol{P_{P_1-P_2} = \frac{(P_1-P_2)-(\pi_1-\pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}} \sim u_\alpha} \tag{6.10}$$

DVA ZÁVISLÉ NÁHODNÉ VÝBĚRU

Párové pozorování je charakterizováno jedním náhodným výběrem z dvourozměrného rozložení. Párem se rozumí dvojice rozdílných pozorování u jedné statistické jednotky.

Modelování PRŮMĚRŮ

$$\boldsymbol{T_{(\bar{X}-\bar{Y})} = \frac{\bar{X}_d-(\mu_1-\mu_2)}{s_d}\sqrt{n} \sim u_\alpha}, \tag{6.11}$$

kde: $\bar{X}_d = \frac{1}{n}\sum_{i=1}^{n} d_i$ …. průměr diferencí, kde: $d_i = x_{1i} - x_{2i}$ pro $i = 1, \dots, n$.

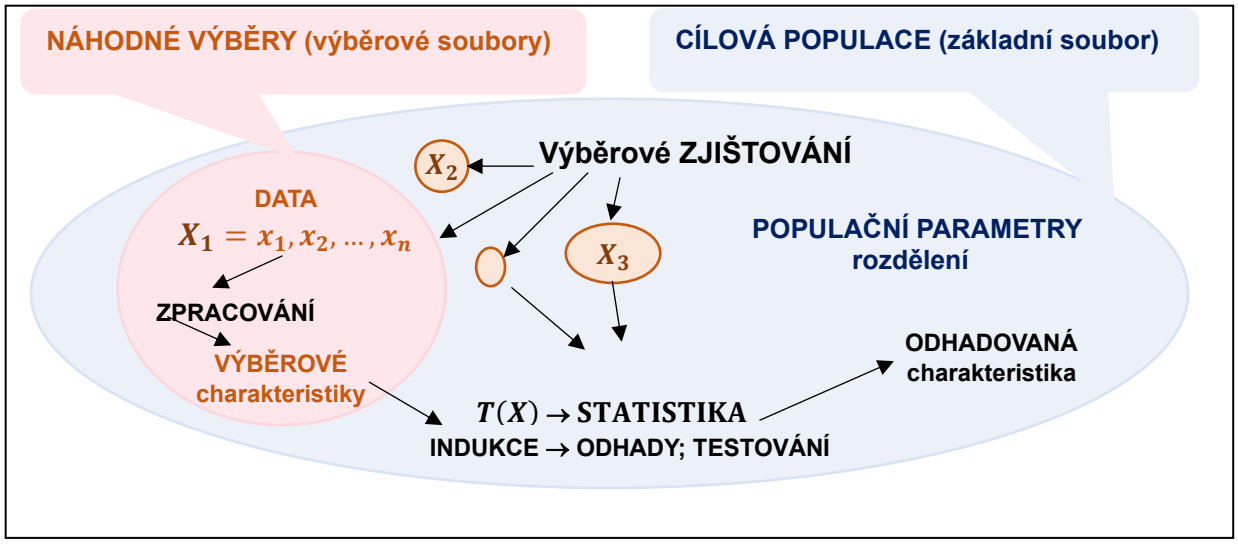
$S_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - d)^2$

6.2 Statistická indukce

Postup, kterým na základě výběrových údajů činíme úsudky o charakteru dat v základním souboru, se nazývá statistická indukce.

Jejím hlavním úkolem je zevšeobecnění, které nám umožňuje na **základě zkoumání dat z reprezentativního výběru** učinit závěry o vlastnostech a chování celého základního souboru (populace). Tento induktivní způsob myšlení znamená, že na základě znalostí konkrétních vlastností dat odvozujeme a usuzujeme na jejich obecné vlastnosti. Jak už bylo mnohokrát řečeno, ve většině případů je základní soubor tvořen velkým počtem jedinců, a není proto možné získat všechny informace o jejich vlastnostech. Na vlastnosti populace usuzujeme na základě údajů získaných z podmnožiny jedinců, tedy z výběrového souboru.

Obrázek 6.1: Postup statistického usuzování



Obor statistické indukce je zaměřen na řešení dvou základních úloh:

- **Odhad populačních rozdělení** a jejich parametrů.
- **Testování statistických hypotéz** o populačních parametrech a rozdělení populace.

V případě, že úsudky o datech jsou založeny na určitých předpokladech o rozdělení náhodných veličin (např. že data jsou normálně rozdělená a hledáme parametry $N(\mu, \sigma^2)$, hovoříme o **parametrických metodách** zpracování. Jestliže neznáme typ rozdělení a jeho parametry, pak popisujeme vlastnosti dat získaných na základě výběru pomocí **neparametrických metod**.

Shrnutí kapitoly

Matematická statistika navazuje na popisnou statistiku a teorii pravděpodobnosti s cílem zobecnit poznatky z dat. Umožňuje přejít od hypotetických modelů k analýze reálných dat.

Náhodný výběr je tvořen nezávislými náhodnými veličinami, které mají stejný typ rozdělení. Jeho realizací jsou konkrétní číselné hodnoty, nazývané pozorování, které tvoří datový soubor.

Populační (teoretické) charakteristiky neboli **parametry** popisují celou populaci a jsou to konstantní hodnoty.

Výběrové charakteristiky neboli **statistiky** se vypočítávají z datového souboru. Jsou to náhodné veličiny a slouží k odhadu populačních parametrů

Výpočet výběrových statistik vychází z principů centrální limitní větu (CLV). Ta říká, že průměr z dostatečně velkého počtu nezávislých náhodných veličin má přibližně normální rozdělení, a to bez ohledu na to, jaké bylo původní rozdělení těchto veličin.

Statistická indukce je postup, kterým na základě údajů získaných z výběrového souboru činíme úsudky (odhadujeme, předpovídáme) o charakteru dat v celém základním souboru (populaci). Obor statistické indukce řeší dvě základní úlohy:

Odhad populačních rozdělení a jejich parametrů.

Testování statistických hypotéz o populačních parametrech a rozděleních populace.

Literatura

BUDÍKOVÁ, Marie; Maria KRÁLOVÁ a Bohumil MAROŠ. *Průvodce základními statistickými metodami*. vydání první. Praha: Grada Publishing, a.s., 2010. ISBN 978-80-247-3243-5.

JAMES, Gareth; Daniela WITTEN, Trevor HASTIE a Robert TIBSHIRANI. *An Introduction to Statistical Learning: with Applications in R.* 2. vydání. Springer, 2021. ISBN 978-10-71618-19-6.

NEUBAUER, Jiří; Marek SEDLAČÍK a Oldřich KŘÍŽ. *Základy statistiky: Aplikace v technických a ekonomických oborech.* 3. vydání. Praha: Grada Publishing, 2021. ISBN 978-80-271-3421-2.

6 Kontrolní otázky

1. Vysvětlete rozdíl mezi základním souborem (populací) a náhodným výběrem.
2. Jaký je rozdíl mezi populačními parametry a výběrovými charakteristikami?
3. Co je to centrální limitní věta a k čemu slouží?
4. Vysvětlete rozdíl mezi parametrickými a neparametrickými metodami zpracování dat.
5. Vysvětlete rozdíl mezi dvouvýběrovým pozorováním (dvěma nezávislými náhodnými výběry) a párovým pozorováním (dvěma závislými náhodnými výběry).
6. Co je charakteristické pro párové pozorování?
7. Co je to statistická indukce a jaký je její hlavní úkol?
8. Jaké dvě základní úlohy řeší obor statistické indukce?
9. Jaký je rozdíl mezi parametrickými a neparametrickými metodami zpracování dat?

7 VÝBĚROVÉ ZJIŠŤOVÁNÍ

Výběrové zjišťování představuje jednu z etap statistické činnosti, která je zaměřena na sběr dat. Problémem statistického zjišťování je reprezentativnost sledovaných údajů. V praxi nemáme většinou k dispozici všechny údaje vztahující se k analyzovanému problému.

Základní soubory jsou často velmi rozsáhlé a v mnoha případech se setkáme jen s průzkumy, které se opírají o relativně malý počet zkoumaných jednotek. V takovém případě hovoříme o výběrových zjišťováních. Úkolem výběrových zjišťování je **vytvoření reprezentativních souborů, které zaznamenávají charakteristické rysy populace**[9] – základního souboru dat.

Číselné charakteristiky vypočítané z výběrových dat jsou podkladem pro odhady parametrů a testování hypotéz o tvaru rozdělení náhodných veličin – statistické usuzování (viz kapitola Teorie odhadu a Testování statistických hypotéz).

Kvalita a integrita shromážděných údajů přímo ovlivňuje platnost a spolehlivost výsledků analýzy. Aby byl výběrový soubor typickým (věrným) zástupcem základního souboru, musí být splněny určité požadavky.

- Přesné **definování základního souboru** z hlediska znalosti základních rysů a homogenity (heterogenity) sledované populace. Jednoznačné vymezení toho, které případy do zkoumané populace spadají a které již nikoliv. Zvážení míry homogenity souboru.
- **Vhodná metoda výběru**. Předpokladem kvalitních dat je pravděpodobnostní výběr. Všechny jednotky základního souboru by měly mít stejnou pravděpodobnost, že budou zařazeny do výběru.

Znalost, a především splnění uvedených požadavků umožňuje provést odhad velikosti výběrového souboru, a následně lze s využitím statistických metod přistoupit k posouzení výběru z hlediska jeho reprezentativity.

V kapitole Základní pojmy, byl uveden rozdíl mezi **úplným (vyčerpávajícím)** zjišťováním ve kterém je hodnota příslušného znaku zjišťována u všech statistických jednotek – **základní soubor (populace) a neúplným (výběrovým)** zjišťování, které je omezeno jen na část statistického souboru a na zkoumání jen některých statistických jednotek – **výběrový soubor**.

Základní druhy výběrových zjišťování můžeme rozdělit do dvou skupin. První skupina metod v sobě zahrnuje **nepravděpodobnostní/nenáhodné výběrové zjišťování**. Druhá

[9] Populace v takovém to pojetí neznamená pouze množinu osob. Zkoumanými jednotkami mohou být stroje, podniky, města, školy, organizace, umělecká díla apod. Definování pojmu populace se odvíjí od tématu výzkumného zaměření.

skupina, která je základem celé teorie statistického usuzování, je založena na metodách **pravděpodobnostního/náhodného výběrového zjišťování.**

7.1 Pravděpodobnostní výběrové zjišťování

Pravděpodobnostní[10] výběr je obecně považován za objektivní způsob výběru jednotek. Základní princip výběru spočívá v tom, že každá jednotka populace má nenulovou pravděpodobnost, že bude vybrána do výsledného vzorku. Tato skutečnost je důležitým předpokladem nestranné analýzy dat. Postup náhodných výběrových zjišťování vychází z principů matematické statistiky a teorie pravděpodobnosti.

7.1.1 Prostý pravděpodobnostní výběr

Základním typem pravděpodobnostního výběrů je prostý (jednoduchý) výběr. Jeho význam spočívá nejen ve snadném použití, ale také ve schopnosti vytvářet reprezentativní vzorky, které odrážejí skutečné vlastnosti populace.

Prostý pravděpodobnostní výběr se obvykle provádí tak, že se celý soubor se rozdělí na tzv. výběrové jednotky, které jsou zpravidla totožné se statistickými jednotkami, a každé se přiřadí určitá pravděpodobnost jejího zahrnutí do výběrového souboru. O nezahrnutí či zahrnutí určité výběrové jednotky do výběrového souboru rozhoduje jen náhoda. Výhodou pravděpodobnostního výběru je, že vybrané jednotky, potažmo jejich charakteristiky, které jsou předmětem zkoumání, mají obdobné rozložení dat jako má výchozí populace.

Mezi nejjednodušší **techniky prostého výběru** patří losování. Správné losování je spojeno s důkladným předchozím promícháním všech jednotek nebo jejich zástupců. V případě, kdy je technicky nemožné použití losování (např. při výběrech z velmi rozsáhlých základních souborů) využívá se pro výběr statistických jednotek tabulka náhodných čísel tzv. pseudonáhodných čísel generovaných počítačem (generátor náhodných čísel). Jednotky základního souboru **mohou**, ale zároveň **nemusí** mít možnost **být vybrány opakovaně**.

Podle tohoto hlediska dělíme prostý náhodný výběr na **výběr s vracením/opakováním** a na **výběr bez vracení/opakování.**

Při **výběru s vracením/opakováním**, je každá vybraná jednotka před dalším vybíráním vrácena zpátky do souboru, ze kterého vybíráme (rozsah i složení souboru zůstává neměnné).

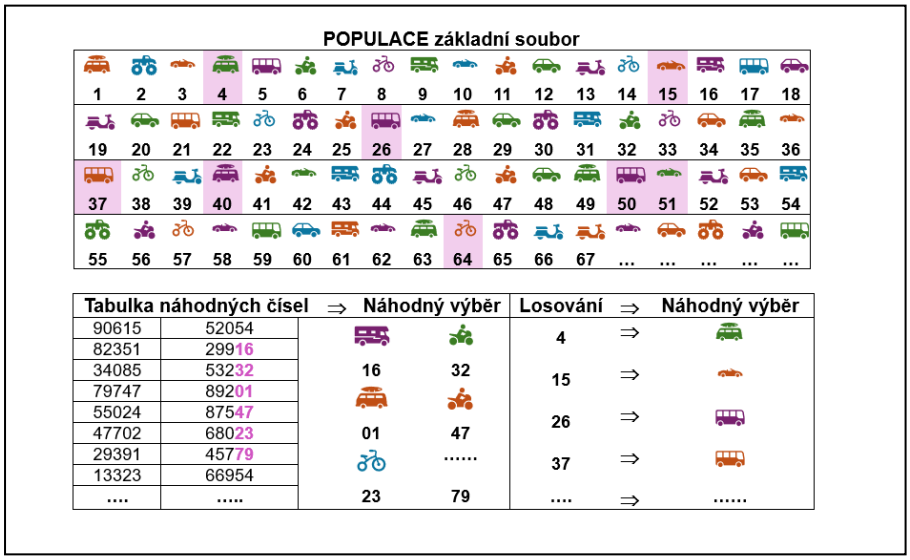
[10] Pravděpodobnostní hledisko náhodného výběru je natolik významné, že v současnosti název "pravděpodobnostní výběr" převažuje nad dosud užívaným názvem "náhodný výběr". Určitý význam v názvu vyplývá ze skutečnosti, že pravděpodobnosti vybrání jednotek ze základního souboru nemusí u všech jednotek stejné, ale může se lišit.

Z uvedeného vyplývá, že každá jednotka může být vybrána několikrát a že se nemění pravděpodobnost s jakou je vybírána (výběry s vracením mají z pravděpodobnostního hlediska charakter nezávislých pokusů a řídí se binomickým rozdělením).

Při **výběru bez vracení/opakování** nejsou již vybrané jednotky vráceny zpět do základního souboru. Každá jednotka základního souboru může být vybrána jen jednou. Složení základního souboru se při každém výběru mění (rozsah souboru se snižuje). V důsledku toho se zvyšuje pravděpodobnost výběru zbylých jednotek ze základního souboru. Výběr bez vracení má charakter pokusů závislých a řídí se tedy pravděpodobnostními pravidly hypergeometrického rozdělení.

Rozdíl mezi prostým pravděpodobnostním výběrem bez vracení a s vracením je třeba respektovat pouze při výběrech ze základního souboru menšího rozsahu. Jestliže rozsah základního souboru je dostatečně veliký (přesahuje několikrát svojí velikostí rozsah výběrového souboru) a výběr představuje pouze malou část základního souboru, jsou důsledky rozdílu mezi výběrem s opakováním a výběrem bez opakování zanedbatelné – rozdíly klesají s relativní četností neboli podílem rozsahu výběrového souboru na rozsahu základního souboru.

Obrázek 7.1: Grafická interpretace prostého pravděpodobnostního výběru



7.1.2 Systematický pravděpodobnostní výběr

V případě, že je zaručeno náhodné pořadí statistických jednotek v základním souboru (populaci), je vhodnou alternativou k prostému náhodnému výběru výběr systematický. Tento výběr vyžaduje, aby jednotky byly seřazeny do posloupnosti, jejíž pořadí nesouvisí se zjišťovanou skutečností. První jednotka do výběru se volí náhodně a teprve od tohoto výchozího bodu se dále vybírá každá n-tá jednotka.

7.1.3 Stratifikovaný pravděpodobnostní výběr

Tento typ výběru se využívá tehdy, když je základní soubor přirozeně rozdělen do několika menších či větších podskupin.

Stratifikovaný výběr je realizován tak, že základní soubor je rozdělen do nepřekrývajících se (nezastupitelných) podskupin/stratu, které obsahují jednotky stejných vlastností (jsou homogenní vzhledem k určitému kritériu). Podskupiny se mezi sebou liší, ale jednotky uvnitř se vzhledem k vybrané vlastnosti neliší. Statistické jednotky jsou z podskupin následně vybírány metodou prostého náhodného výběru. **Při stratifikovaném výběru si musíme dát pozor na stanovení správného podílu jednotlivých strat na základním souboru** (např. názory mužů a žen v jednotlivých věkových kategorií na otázky globalizace). K tomu, abychom mohli stanovit podíl jednotlivých strat (pohlaví, věk) ve výběru, musíme znát podíly obyvatelstva těchto skupin v základním souboru.

7.1.4 Vícestupňový pravděpodobnostní výběr

Vícestupňový výběr spočívá v tom, že statistické jednotky nevybíráme přímo, ale v několika stupních. Využívá se při výběru jednotek ze základního souboru, ve kterém jsou jednotlivé podskupiny přirozeně hierarchicky uspořádány. Na rozdíl od stratifikovaného výběru, jsou však tyto podskupiny/klastry navzájem zastupitelné (město, podnik, provoz, zaměstnanec).

Tabulka 7.1: Výhody a nevýhody pravděpodobnostního výběru

Metody výběrů	Výhody	Nevýhody
Prostý	Snadno se používá a poskytuje reprezentativní vzorek populace. Vyjadřuje všechny známé i neznámé vlastnosti populace; úplně eliminuje možnosti ovlivnit podobu vzorku ze strany výzkumníka.	
Systematický	Méně časově náročný než prostý náhodný výběr; poskytuje reprezentativní vzorek populace.	
Stratifikovaný	Zvyšuje reprezentativnost vzorku zahrnutím důležitých podskupin do výběru.	V porovnání s přecházejícími typy výběrů je náročnější na organizaci a zpracování výsledků.
Vícestupňový	Užitečný pro velké geograficky rozptýlených skupiny obyvatel.	Může snížit reprezentativnost vzorku, pokud shluky nejsou reprezentativní pro populaci.

7.2 Nepravděpodobnostní výběrové zjišťování

Nepravděpodobnostní výběr (nenáhodný) je takový výběr, u kterého neznáme pravděpodobnost zařazení jednotek do výběru. Výběr jednotek je založen na jiných než pravděpodobnostních faktorech, nejčastěji na základě zkušeností a poznatků analytika. Existuje několik typů nepravděpodobnostního výběru, v následujícím textu jsou uvedeny nejčastěji používané.

7.2.1 Kvótní výběr

Jedná se o nepravděpodobnostní metodu zajišťující **výběr reprezentativního souboru.**

Reprezentativita takového výběru **je omezena pouze na reprezentativitu podle definovaných kvótních znaků**, které nejsou ovlivněny cíli statistické analýzy. Při kvótním výběru nelze stanovit výběrovou chybu. Na základě známých charakteristik populace (věk, pohlaví, bydliště, národnost atd.), získaných např. ze Sčítání lidu, domů a bytů, stanovujeme kvóty a s jejich pomocí do vzorku zařazujeme vybrané jednotky tak, aby struktura vzorku odpovídala struktuře populace.

Jako příklad kvótního výběru můžeme uvést výběr jednotek na základě bydliště a věkové kategorie. Pokud je v populaci podle výsledků sčítání lidu, které bylo provedeno v roce 2021, v Praze v produktivním věku (15-64 let) 66,3 % osob, na základě správně provedeného kvótního výzkum bude zajištěno, že tyto skupiny jsou v šetření zastoupeny odpovídajícím způsobem. Reprezentativita takového výběru je omezena pouze na reprezentativitu podle definovaných kvótních znaků.

7.2.2 Anketa

Anketa (samovýběr) oslovuje zpravidla jen určitou vybranou část statistických jednotek (osob, podniků, institucí apod.).

Zpravidla se provádí formou rozeslání dotazníků, jejichž distribuce v dnešní době probíhá především **elektronicky** nebo **telefonicky** a v neposlední řadě na základě **osobního kontaktu** s dotazovanými.

Každý z uvedených způsobů má své výhody a nevýhody, ale jedno mají společné. Cílová skupina takto oslovených respondentů **není ve většině případů reprezentativním vzorkem** populace. Dotazník vyplní zpravidla jen menší část dotázaných (v průměru asi jedna třetina), často jen na základě finanční odměny anebo jiných benefitů, které z dané činnosti plynou. Takto získaná data nelze ve většině případů považovat za obecně platné informace o zkoumaných

jevech a nejsou tedy ani vhodné ke statistickému zobecňování[11]. Reprezentativita není narušena jen počtem vyšetřovaných jednotek, ale i způsobem jejich výběru.

7.2.3 Metoda základního masivu

Metoda základního masivu se používá tehdy, skládá-li se soubor z několika velkých objektů (jednotek) a z velkého počtu malých objektů.

Metoda základního masivu **nedovoluje zobecňovat získané výsledky na celý základní soubor**, neboť zpravidla nevystihuje specifika malých objektů, které mají jiné zákonitosti a tendence chování než velké objekty (např. může dojít k situaci, kdy šetření v oblasti potravinářství, bude vycházet z informací od několika opravdu velkých společnosti a specifika drobných živnostníků zabývající se výrobou lokálních potravin budou úplně vynechána).

7.2.4 Záměrný výběr

Záměrný výběr je druh výběru, u kterého rozhodují o zahrnutí jednotek do výběrového souboru různá **logická hlediska a subjektivní názor** osoby, která výběr realizuje.

V některých případech mohou přinést záměrné výběry užitečné informace, v jiných případech naopak bezcenné a zkreslené. Jeho nevýhodou je, že vyžadují určité předběžné znalosti o základním souboru a je založen na subjektivním přístupu osob, které výběr provádějí.

Jednou z nejčastěji používaných technik záměrného výběru je technika sněhové koule, která je zaměřena na výběr jednotek osob, které jsou nositelem specifických vlastností**.** Metoda spočívá v tom, že nové respondenty nabíráme na základě doporučení respondentů, kteří již byli vybráni. Tato technika výběru může být užitečná také při výběru z populace, kterou je obtížné identifikovat nebo k níž je obtížné získat přímý přístup (např. výběr osob ze stejných zájmových skupin a následné předávání kontaktů – nelegální migranti, uživatelé omamných látek atd.).

Nepravděpodobnostních výběrových zjišťování je v odborné literatuře uváděna celá řada (dostupný, příležitostní, typický či kriteriální výběr atd.). Jednotná klasifikace nenáhodných zjišťování neexistuje, vždy záleží na autorovi vědeckého pojednání o dané problematice. Oproti tomu v případě definování **pravděpodobnostních výběrových zjišťování** a jejich metod, panuje v odborné komunitě téměř terminologická shoda.

[11] Nejčastější problematika v závěrečných pracích studentů. Vzorek není dostatečně reprezentativní, aby pokryl všechna specifika zkoumané populace.

Tabulka 7.2: Výhody a nevýhody nepravděpodobnostního výběru

Metody výběrů	Výhody	Nevýhody
Kvótní výběr	Relativní pružnost, rychlost, nahrazuje známé vlastnosti ve struktuře populace; vytvoření model populace.	Obtížná kontrola; subjektivní přístup zpochybňuje možnost zobecnění; je reprezentativní pouze z hlediska znaků použitých v kvótách; použití jen pro dobře zmapované populace, u kterých známe podíly zastoupení kvót; vytváření kvót znesnadňuje výběr.
Anketa	Nejjednodušší forma	Oslovuje nesystematicky vybranou část populace; malá návratnost; výběr založený na rozhodnutí respondenta; informace získané anketním šetřením nelze zobecňovat.
Metoda základ. Masívu	Menší pracnost, časová a finanční nenáročnost.	Zobecnění poznatků má menší platnost; nevystihuje specifika menších jednotek.
Záměrný výběr	Užitečný pro těžko dostupné či důležité podskupiny obyvatel.	Může dojít ke zkreslení a nemusí být reprezentativní pro populaci.

7.3 Rozsah výběru

Splnění požadavků, které jsou kladeny na výše uvedené metody výběrového zjišťování nezaručuje, že budeme pracovat s daty, které jsou vhodným zástupcem základního souboru.

V následujícím textu se budeme věnovat nejčastějšímu problému, který ovlivňuje kvalitu výběrových zjišťování a tou je nedostatečný rozsah výběrového souboru. Velikost (rozsah) výběru je závislá na zaměření a cílech daného šetření. Jako příklad můžeme uvést rozdíl ve stanovení rozsahu výběru a následném sběru dat mezi kvantitativním a kvalitativním výzkumem.

Tabulka 7.3: Stanovení rozsahu výběrového souboru

Velikost základního souboru (populace[12])	Velikost výběrového vzorku v % (pravděpodobnostní výběr)
do 30 jednotek	100 %
do 100 jednotek	80 %
do 1 000 jednotek	40 %
do 10 000 jednotek	7,5 %
do 100 000 jednotek	1,5 %
do 1 000 000 jednotek	0,25 %
do 10 000 000 jednotek	0,06 %

[12] Základním souborem není vždy celá populace ČR, základním souborem jsou myšleny např. všichni zaměstnanci jednoho podniku, všechny stoje v jednom výrobním závodě, všichni studenti druhého ročníku VŠ atd.)

Kvantitativní výzkum probíhá podle předem daných pravidel a je založen na ověřování předem stanovených hypotéz. Předmětem takového šetření je **početná skupina jedinců** – reprezentativních zástupců celé populace (snažíme se získat vybrané informace o velkém vzorku populace). V odborné literatuře se uvádí následující poměr (tab. 7.3) v rozsahu mezi základním a výběrových souborem (tab. 7.3).

Kvalitativní výzkum, není přesně definován, zaměřuje se na nové pohledy (indukce) a předmětem šetření **může být i jedinec** (zaměřujeme se na získání mnoha informací o jednotlivci). Počet jednotek se odvíjí od „teoretického nasycení". Rozsah souboru je menší, avšak flexibilnější (je možné v průběhu měnit jeho velikost). Takto získané informace jsou validní, ale nejsou zobecnitelné na větší populaci.

Mezi další faktory, které mají vliv na rozsah výběrového souboru řadíme:

Velikost výběrové chyby

Výběrová chyba nám určuje, nakolik se může výsledek zjištěný na základě z výběrového souboru odchylovat od skutečnosti v souboru základním (spolehlivost výsledků).

Chyba se udává v procentech na zvolené hladině významnosti (více viz Teorie odhadu). Stanovíme-li si přípustnou výběrovou chybu, můžeme určit velikost výběrového souboru. Snižování výběrové chyby znamená zvyšování rozsahu výběrový soubor. Je třeba dobře zvážit, zda je to možné z praktických důvodů (např. podle počtu lidí v síti tazatelů, podle počtu rozhovorů, které lze provést atd.).

Homogenita populace

Variabilita základního souboru. Pokud je základní soubor nehomogenní (skládá se z více podskupin), je potřeba zvýšit rozsah souboru.

Typ proměnných a velikost výchozích podmnožin

Rozsah souboru záleží na počtu proměnných a jejich typu. Čím je více proměnných a čím je nižší úroveň jejich typu (nominální → ordinální → kvantitativní), tím větší rozsah souboru je potřeba. Rozsah výběrového soubor záleží také na počtu podmnožin, které budeme analyzovat. Čím více podmnožin tím je potřeba mít větší počet jednotek.

Finanční, časová a organizační hlediska

Skutečná velikost vzorku bývá kompromisem mezi metodologickými požadavky a dostupnými finančními, časovými a organizačními zdroji.

7.3.1 Určení minimálního rozsahu výběru

Stanovení dostatečné velikosti výběru je možné provést několika způsoby. Nejznámější postup, avšak v praxi málo využívaný je založený na šíři intervalů spolehlivosti dle typu sledovaných proměnných.

Kvantitativní proměnné – postup na základě odhadu intervalu pro průměr

V případě, že data mají normální rozdělení, určí se minimální velikost výběru tak, aby s pravděpodobností (nejčastěji stanovujeme 95 %) platilo:

$$P\left(\bar{x} - u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{s^2}{n}} \leq \mu \leq \bar{x} + u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{s^2}{n}}\right) = 1 - \alpha \qquad (7.1)$$

Kde: P…… …. pravděpodobnost $\rightarrow P = 1 - \alpha \rightarrow$ požadovaná spolehlivost,
α ……. hladina významnosti,
$\bar{x}$ ……. výběrový aritmetický průměr,
s^2…… výběrový rozptyl,
u_α …….. kvantil normálního rozdělení,
μ ……. populační průměr.

Minimální rozsah souboru se pak stanovuje na základě vztahu:

$$n \geq \frac{u_{1-\frac{\alpha}{2}}^2 \cdot s^2}{\Delta^2} \qquad (7.2)$$

Kde: Δ …. požadovaná polovina šíře intervalu spolehlivosti $\rightarrow$ přípustná chyba odhadu,
n … celkový rozsah souboru,
u_α …kvantil normálního rozdělení,
s^2 …. výběrový rozptyl.

Příklad 7.1

Jak velký soubor dat z populace s normálním rozdělením je potřeba vybrat, abychom mohli provést odhad střední hodnoty (průměru) proměnné počet hodin ve škole během dne. Předpokládáme, že variabilita vyjádřena směrodatnou odchylkou je 2,34 hodiny. Požadujeme, aby výsledný 95% interval spolehlivosti měl šířku:

a) $\pm$ 1 hodinu;

b) $\pm$ 3 hodiny.

Řešení příkladu 7.1

K výpočtu využijeme vztah 7.2

a) $\Delta \rightarrow \pm$ 1 hodina
$S^2 \rightarrow 2{,}34^2$
$P = 1 - \alpha \rightarrow 0{,}95 \; = \; 1 \; - \; 0{,}05$
u_α kvantil normálního rozdělení je tabelovaná hodnota $\rightarrow$ pro $u_{1-\frac{\alpha}{2}} = u_{0{,}975}$ = 1,96

$$\boldsymbol{n \geq \frac{1{,}96^2 \cdot 2{,}34}{1^2} \geq 22}$$

b) $\Delta \rightarrow \pm$ 3 hodina

$$\boldsymbol{n \geq \frac{1{,}96^2 \cdot 2{,}34}{3^2} \geq 3}$$

Kvalitativní proměnné – postup na základě odhadu relativní četnosti

V případě, že nás zajímá relativní četnost výskytu jediné kategorie proměnné, je rozsah souboru stanoven následovně:

$$P\left(p - u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}} \leq \pi \leq p + u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}}\right) = 1 - \alpha \qquad (7.3)$$

Kde: p ……. výběrová relativní četnost,
u_α …….. kvantil normálního rozdělení,
π ……. populační relativní četnost.

Minimální rozsah souboru se pak stanovuje na základě vztahu:

$$n \geq \frac{u_{1-\frac{\alpha}{2}}^2 \cdot p(1-p)}{\Delta^2} \qquad (7.4)$$

Ostatní symboly mají stejný význam jako ve vzorci 7.1 a 7.2.

Příklad 7.2

Jak velký soubor bude potřebný pro zjištění podílů žen, které chodí do menzy na obědy. Předpokládáme, že v menze se stravuje 35 % žen. Požadujeme, aby výsledný 95% interval spolehlivosti měl šířku:

a) $\pm$ 2 %;
b) $\pm$ 15 %.

Řešení příkladu 7.1

K výpočtu využijeme vztah 7.4

a) $\Delta \rightarrow \pm 2\ \%$
$p \rightarrow 0{,}35$
u_α kvantil normálního rozdělení je tabelovaná hodnota → pro $u_{1-\frac{\alpha}{2}} == u_{0{,}975}$ 1,96

$$\boldsymbol{n \geq \frac{1{,}96^2 \cdot 0{,}35(1-0{,}35)}{0{,}02^2} \geq 2185}$$

b) $\Delta \rightarrow \pm 15\ \%$

$$n \geq \frac{1{,}96^2 \cdot 0{,}35(1-0{,}35)}{0{,}15^2} \geq 39$$

Uvedené postupy pro stanovení rozsahu souboru jsou vhodné tehdy, když máme o chování dat předběžné informace.

Z výše uvedeného vyplývá, že čím je rozsah výběrového souboru větší, tím přesnější jsou i závěry z prováděného šetření. Velikost výběrového souboru je sice významná, ale není rozhodující. S rostoucí velikostí výběrového souboru se shoda mezi strukturou populace a výběrovými daty zvyšuje, ALE rozhodující je **reprezentativnost výběrového souboru**.

Závěry analýz prováděné na menším reprezentativním výběrovém souboru jsou často mnohem kvalitnějším zdrojem informací než závěry plynoucí z analýzy velkého souboru dat, které jsou nereprezentativním zástupcem populace.

7.4 Reprezentativnost výběrového souboru

Obecně můžeme uvést, že reprezentativnost výběrového souboru je dána tím, nakolik je tento soubor ve zmenšeném měřítku odrazem populace a může ji tedy zastupovat. **Jedná se tedy o to, jak věrně výběr reprezentuje známé parametry cílové populace** (např. národností složení, věkovou strukturu, strukturu podniků atd.), které jsou předmětem našeho šetření.

Reprezentativnost výběrového zjišťování, znamená také možnost zobecnění informaci, které jsou obsažené ve výběrovém souboru na soubor základní. Stoprocentní reprezentativnost neexistuje, vždy je nějaká možnost vzniku chyby – náhodná chyba výběru (viz kapitola Teorie odhadu).

Nakolik je soubor reprezentativní (spolehlivý a přesný) závisí na dvou faktorech; **validitě** (platnost) a **reliabilitě** (spolehlivost), které přestavují charakteristiky kvality měření. Validita a reliabilita jsou jednou ze základních charakteristik měření v kvantitativním výzkumu, a především důležitým předpokladem pro zabezpečení jeho kvality.

7.4.1 Validita

Validita znamená platnost toho, že měříme to, co předpokládáme, že měříme (pravítko je validním nástrojem pro stanovená rozměrů uměleckého díla – obrazu, nikoli však pro stanovení jeho hodnoty). Neexistuje exaktní metoda, jak ji zjišťovat. Validitu rozlišujeme kriteriální, konstruktovou a obsahovou.

Kriteriální validita posuzuje shodu výsledků, zaměřenou na výsledky zaváděné procedury v porovnaní s nějakou kriteriální proměnou nebo s jiným měřením, které je již ověřené. Největším problémem prokazování kriteriální validity je právě nalezení vhodného kritéria. Tato ověřená proměnná bývá někdy nazývána „kriteriální standard“ nebo „zlatý standard“.

Konstruktová[13] **validita** umožňuje srovnání a zabývá se teoretickými aspekty měřeného konstruktu (proměnné). Je objektivně měřitelná, neboť pracuje s více ukazateli sledovaného jevu a vzájemné vztahy těchto ukazatelů číselně vyjadřuje.

Obsahová validita ověřuje, do jaké míry je měření skutečně zaměřeno na všechny sledované aspekty (např. přijímací testy z matematiky – obsahují všechna potřebná témata). Na rozdíl od konstruktové či kriteriální validity ji nedokládáme analýzou shody více ukazatelů. Její ověření je založeno na věcném rozboru struktury testu, jeho výstupů a opírá se např. o znalost určité teorie nebo o odbornou literaturu. Stanovuje se expertním posouzením.

7.4.2 Reliabilita

Znamená spolehlivost a přesnost, která je vyjádřena stupněm shody výsledků měření provedeného za stejných podmínek (např. stanovení přesné teploty – kapalinový, plynový, termoelektrický teploměr).

Reliabilita se zaměřuje na náhodné chyby měření, kdy za reliabilní považujeme takové měření, kde je chyba co nejmenší. Chybovost mimo jiné závisí na vlastnostech a dovednostech hodnotitele. Na počtu a kvalitě pozorování. Technice pozorování, jeho účelu, typu použitého nástroje, na subjektu výzkumu a v neposlední řadě na okolnostech pozorování neboli vlivu vnějších faktorů (více viz GWET, K. 2014)

POZOR! *Bez reliability nemůžeme dosáhnout validity a naopak. Dostatečně vysoká reliabilita je nutnou podmínkou dobré validity měření, vysoká reliabilita však ještě nezaručuje dobrou validitu!*

[13] Konstrukt (jev) je teoretický pojem, který se označuje přímo nepozorovatelný faktor. Měřitelně vyjádřený konstrukt je proměnná.

Shrnutí kapitoly

Problémem statistického zjišťování je **reprezentativnost výběrového souboru**, která je dána tím, nakolik je tento soubor ve zmenšeném měřítku odrazem populace a může ji tedy zastupovat.

Požadavek reprezentativnosti výběru jedinců z populace se odvíjí také od metody výběru.

Pravděpodobnostní výběr vycházejí z předpokladu, že jednotky do výběrového souboru zahrnujeme na základě pravděpodobnostních pravidel.

Nepravděpodobnostní výběr je takový výběr, u kterého neznáme pravděpodobnost zařazení jednotek do výběru.

Kvalitu výběrových zjišťování ovlivňuje rozsah **výběrového souboru**, který se odvíjí od velikosti výběrové chyby; homogenity populace; typu proměnných a velikosti výchozích podmnožin; finančních, časových a organizačních faktorů.

Reprezentativnost závisí na dvou faktorech; validitě a reliabilitě.

Validita znamená platnost toho, že měříme to, co předpokládáme, že měříme.

Reliabilita znamená spolehlivost a přesnost, která je vyjádřena stupněm shody výsledků měření provedeného za stejných podmínek.

Literatura

GWET, Kilem L. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC, 2014. ISBN 978-0-9708062-8-4.

HENDL, Jan. *Přehled statistických metod zpracování dat: analýza a metaanalýza dat*. Vyd. 1. Praha: Portál, 2004. ISBN 8071788201.

HINDLS, Richard; Stanislava HRONOVÁ, Jan SEGER a Jakub FISCHER. *Statistika pro ekonomy*. 8. vyd. Praha: Professional Publishing, 2007. ISBN 978-80-86946-43-6.

KREIDL Martin. Zhodnocení vlivu práce výzkumných agentur na konstruktovou validitu škál. *Sociologický časopis/Czech Sociological.* Review, roč. 41 (2005), č.1, s. 103-124.

ŘEHÁK, Jan a Ondřej BROM. *SPSS – Praktická analýza dat*. 1. vydání. Brno: Computer Press, 2015. ISBN 978-80-251-4609-5.

VANACORE, Amalia a Maria PELLEGRINO. Benchmarking procedures for characterizing the extent of rater agreement: a comparative study. *Quality and Reliability Engineering International*, vol 38 (2021), no 3, s. 1404-1415.

7 Kontrolní otázky

1. Co je úkolem výběrového zjišťování a k čemu slouží?
2. Jaké jsou dva základní druhy výběrových zjišťování?
3. Co je to prostý pravděpodobnostní výběr a jaké jsou jeho výhody?
4. Jaký je rozdíl mezi úplným a neúplným zjišťováním?
5. Jaké typy pravděpodobnostních výběrů znáte a jaký je jejich hlavní princip?
6. Co je hlavním principem nepravděpodobnostního výběrového zjišťování?
7. Co je to reprezentativnost výběrového souboru?
8. Jaké faktory ovlivňují rozsah výběrového souboru?
9. Jak souvisí velikost výběrové chyby s rozsahem souboru?
10. Proč není vždy pravidlem, že větší výběrový soubor zaručuje lepší výsledky?
11. Jaké dva hlavní faktory ovlivňují, nakolik je soubor reprezentativní?
12. Co znamená termín validita?
13. Co znamená termín reliabilita?

8 TEORIE ODHADU

Teorie odhadu je jednou z úloh statistické indukce, kdy z informací o části populace chceme dospět k tvrzení, které se týká celé populace.

Pro výpočet výběrové charakteristiky (např. průměru, rozptylu) je typické, že skutečně naměřené hodnoty statistické proměnné jsou vždy získány na reálném a konečném výběru, a proto se její hodnota mění od jednoho náhodného výběru k druhému (příklad 8.1).

Příklad 8.1

Průměrný bodový zisk všech studentů ($\boldsymbol{N = 420}$) oboru ESM kurzu Statistika z testu byl 12,5 bodů. Tento údaj je **populační parametr** (konstanta) $\rightarrow \mu = 12{,}5$.

Průměrné bodové hodnocení jednotlivých studijních skupin bylo následující:

$n_1 = 30 \quad \bar{x}_1 = 9{,}5 \quad s_1 = 4{,}21$
$n_2 = 26 \quad \bar{x}_2 = 13 \quad s_2 = 1{,}75$
$n_3 = 50 \quad \bar{x}_3 = 11{,}5 \quad s_3 = 2{,}54$
$n_4 = 22 \quad \bar{x}_4 = 10{,}5 \quad s_4 = 3{,}48$

Výběrová charakteristika
Náhodná veličina s určitou variabilitou!

Teorie odhadu se zabývá metodami, jak na základě **výběrových charakteristik** neboli **statistik** získaných výpočtem z reprezentativních výběrů odhadnout co nejpřesnější a nejspolehlivější charakteristiky modelového rozložení dat (**populační parametry**).

Teorie odhadů, stejně jako testování statistických hypotéz o populačních parametrech, kterému se budeme věnovat v další kapitole, vycházejí z principů Centrální limitní věty, která nám říká, že provedeme-li mnoho náhodných výběrů a pro každý z nich vypočítáme určitou výběrovou charakteristiku (například průměr), pak se rozdělení těchto charakteristik bude s rostoucí velikostí výběrového vzorku blížit normálnímu rozdělení. To platí bez ohledu na to, jaké je původní rozdělení populace.

Odhady parametrů se provádí nejčastěji na základě:

- **momentové metody**, která je založena na výpočtu centrálních momentů (viz kapitola Číselné charakteristik NV). Tato metoda spočívá v porovnání teoretických a výběrových momentů.
- **metody maximální věrohodnosti,** která je univerzální metodou pro konstrukci odhadů parametrů různých typů rozdělení náhodných veličin. Její princip spočívá v nalezení takového odhadu parametru, pro který je pravděpodobnost, že pozorované hodnoty pocházejí z předpokládaného rozdělení, maximální.

Odhad neznámé populační charakteristiky může být buď **bodový**, nebo **intervalový**.

Bodový odhad spočívá v tom, že na základě zjištěných údajů z náhodného výběru dat odhadneme předem stanoveným způsobem jedno číslo, které považujeme za nestranný odhad parametru základního souboru. **Intervalový odhad** je náhodný interval, který s předem zvolenou pravděpodobností pokrývá skutečnou hodnotu populačního parametru. Takto sestrojený interval hodnot se nazývá **konfidenční interval (interval spolehlivosti).**

8.1 Bodové odhady parametrů rozdělení

Bodový odhad nebývá početně náročný a používáme ho obvykle v případě, je-li rozsah výběrového souboru vzhledem k rozsahu základního souboru dostatečně velký ($n \geq 30$).

Z náhodného výběru získáme konkrétní hodnoty x na jejichž základě definujeme statistiku $T(X)$. Jestliže cílová populace, ze které pochází náhodný výběr má rozdělení s hustotou pravděpodobností $f(x;\theta)$, pak se bodový odhad $\hat{\theta}$ neznámého parametru θ rovná:

$$\hat{\theta} = T(X = x) \tag{8.1}$$

Odhad musí, splňovat určité vlastnosti. Vzhledem k tomu, že při každé další realizaci náhodného výběru získáme jiné hodnoty veličin, bude i bodový odhad rozdílný. Takový odhad nám nemůže poskytnou přesnou hodnotu populačního parametru. Za předpokladu, že splňuje níže uvedené vlastnosti, ho však můžeme považovat za „věrohodného“ zástupce populačního parametru.

Mezi základní vlastnosti odhadů řadíme: nestrannost, konzistenci a vydatnost.

Nestrannost znamená, že střední hodnota odhadu $E(T)$ nevede k systematickému nadhodnocování či podhodnocování odhadované charakteristiky, tzn. kolísá systematicky kolem θ na obě strany. Střední hodnota kvadratické chyby odhadu od skutečné hodnoty parametru je nulová.

$$E(T) = \theta \leftrightarrow E[(T - \theta)^2] = 0 \tag{8.2}$$

Splňuje-li zvolená charakteristika tento požadavek, **nazýváme ji nestranným, nezkresleným či nevychýleným odhadem**. Odhady, které nesplňují požadavek (vztah 8.2), **nazýváme zkreslenými či vychýlenými odhady**, v takovém případě platí:

$$E(T) - \theta \neq 0 \leftrightarrow E(T) - \theta \rightarrow \textbf{vychýlenost odhadu} \tag{8.3}$$

Zkreslený odhad, pro který zkreslení mizí s rostoucím rozsahem výběru, se nazývá asymptotickým nezkresleným odhadem. Máme-li k dispozici výběr velkého rozsahu, lze pokládat asymptotickým nezkreslený odhad za rovnocenný s odhadem nezkresleným.

Konzistentním odhadem parametru θ se označuje takový odhad pro který platí, že s rostoucím počtem pozorování se zvyšuje pravděpodobnost (blíží se teoretické hodnotě 1), že odhad se bude co nejvíce blížit skutečné hodnotě odhadované charakteristiky základního souboru.

$$\operatorname{Lim}_{n\to\infty} P(|T-\theta|<\epsilon)=1 \leftrightarrow \lim_{n\to\infty} P(|T-\theta|>\epsilon)=0 \tag{8.4}$$

Odhad je konzistentní, jestliže je asymptoticky nestranný a jestliže **s rostoucím rozsahem se hodnota rozptylu blíží nule**.

Vydatnost odhadu znamená, že rozptyly odhadů okolo odhadovaného parametru při opakovaných výběrech jsou malé. Charakteristika T dává vydatný odhad charakteristiky θ, jestliže má ze všech nestranných odhadů **nejmenší variabilitu**.

Výše uvedené požadavky je **vhodné doplnit** o posouzení toho, zda je **odhad dostatečný a robustní.**

Dostatečný odhad (postačující) znamená, že charakteristika T shrnuje všechny informace o sledovaném parametru, které poskytují data z výběrového šetření. Neexistuje tedy žádná další charakteristika, která by měla o parametru θ nějakou další informaci.

Robustnost v tomto pojetí znamená, že odhadované robustní charakteristiky nejsou citlivé na odlehlé hodnoty, které nemohou být pro svoji důležitost ze souboru odstraněny, odhad není ovlivněn tím, zda data pocházejí či nepocházejí z normálního rozdělení a, v neposlední řadě, hodnoty vypočtené pomocí robustních (neparametrických) postupů, nejsou ovlivněny velkou variabilitou dat. Například robustním odhadem střední hodnoty rozdělení může být medián.

Při používání bodových odhadů je třeba si uvědomit, že i když za bodový odhad zvolíme charakteristiku, která má požadované vlastnosti, její hodnota vypočtená na základě údajů získaných náhodným výběrem se bude prakticky vždy lišit od odhadované charakteristiky základního souboru. Tato hodnota bude zatížena **výběrovou chybou,** která vzniká při odhadu na základě jednoho výběrového souboru.

Stanovení výběrové chyby, která vzniká při bodovém odhadu populačních parametrů posuzujeme nejčastěji pomocí **střední kvadratické chyby *MSE***[14].

$$MSE_T = \sigma_T^2 = D(T) = E(T-\theta)^2 \tag{8.5}$$

Kde: $E(T)$…. Střední hodnota odhadu,

$D(T)$…. Rozptyl odhadu.

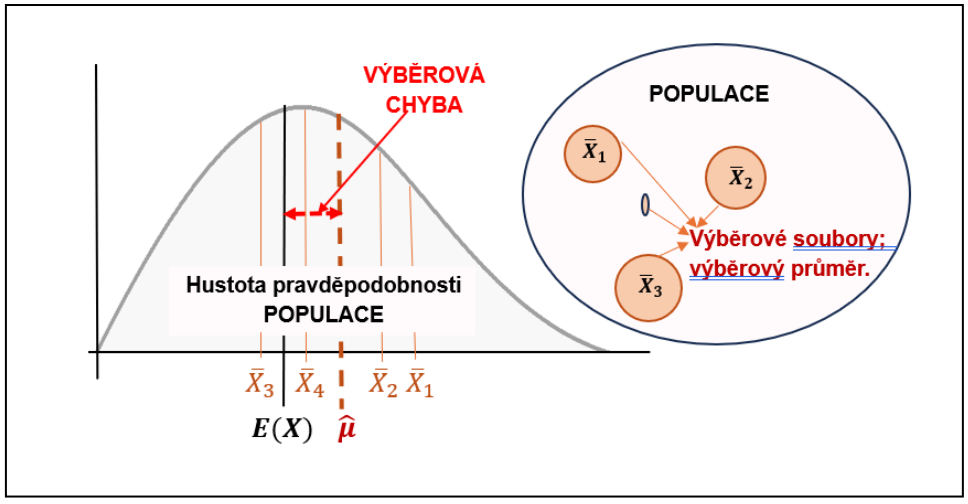
V případě, kdy posuzujeme přesnost bodového odhadu (odhad splňuje podmínku nestrannosti viz vztah 8.2) používáme k výpočtu **střední (směrodatnou) chybu odhadu *SE***[15]**,** kde je za měřítko přesnosti považována směrodatná odchylka σ_T.

$$SE_T = \sigma_T = \sqrt{D(T)} = \sqrt{E(T-\theta)^2} \tag{8.6}$$

Obecně střední chyba odhadu vyjadřuje míru nejistoty (variability) odhadovaného parametru způsobenou náhodným výběrem prvků z populace.

V případě průměrů nám střední chyba odhadu měří (obr. 8.1) rozptýlenost všech **výběrových průměrů**, které jsou vypočítané z různých výběrových souborů $X_I = (X_1, X_2, \dots, X_k)$ o rozsahu n prvků, které pocházejících z cílové populace (jednoho základního souboru). Střední chyba odhadu SE vyjadřuje kolísání výběrových průměrů kolem teoretické střední hodnoty $E(X)$ v celém základním souboru.

Obrázek 8.1: Podstata bodového odhadu teoretické střední hondoty

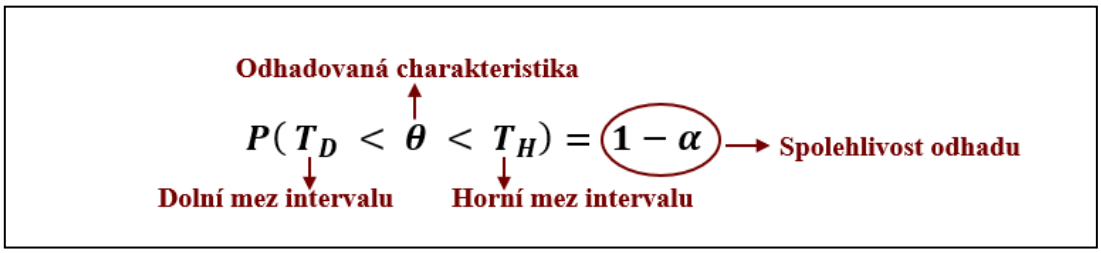


[14] MSE z anglického Mean squared error.
[15] SE z anglického Standardt error.

8.2 Intervalové odhady parametrů rozdělení

Intervalový odhad spočívá ve stanovení číselného intervalu, ve kterém leží neznámý populační parametr s předem zvolenou pravděpodobností (obr. 8.2). Tento interval bývá označován také jako konfidenční interval a slouží k zobecnění výsledků výběrového souboru na soubor základní.

Obrázek 8.2: Interval spolehlivosti



Šířka intervalového odhadu parametru θ je ovlivněna **koeficientem spolehlivosti** $\boldsymbol{P = 1 - \alpha}$.

Spolehlivost odhadu $\boldsymbol{1 - \alpha}$ udává míru pravděpodobnosti P, se kterou interval spolehlivosti pokryje parametr základního souboru při opakovaném provádění výběru – představuje **míru jistoty** odhadu. Nejčastěji používané hodnoty spolehlivosti odhadu jsou 90 %, 95 % nebo 99 %.

Provedeme-li 100 opakovaných výběrových šetření a následně na základě získaných údajů stanovíme 100 intervalových odhadů populačního parametru, pak při 95% spolehlivost odhadu to znamená, že ze 100 vypočtených intervalů spolehlivosti jich přibližně 95 pokryje hodnotu populačního parametru.

Příklad 8.3

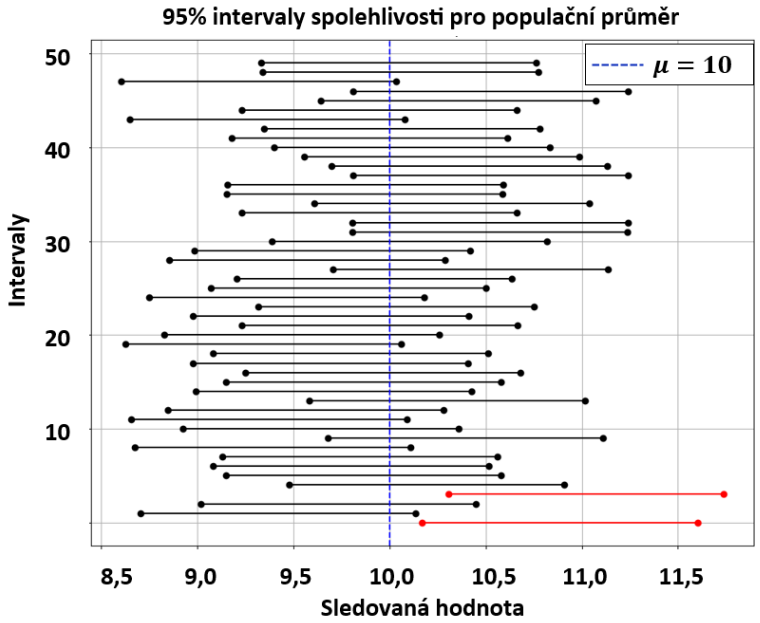
Představme si populaci, která má známý průměr μ = 10. Z této populace provedeme celkem 50 nezávislých náhodných výběrů, každý o rozsahu n = 30. Pro každý z těchto 50 výběrů vypočítáme průměr a následně sestavíme 95% interval spolehlivosti.

Řešení

Protože pracujeme s 95% spolehlivostí, teoreticky očekáváme, že 95 % intervalů (tj. přibližně 47) bude obsahovat skutečnou hodnotu populačního průměru (μ=10). Zároveň to znamená, že 5 % intervalů (tj. přibližně 2–3) populační průměr pokrýt nemusí.

Po provedení simulace všech 50 výběrů a výpočtu intervalů získáme výsledek, který je znázorněn v následujícím grafu.

Pokračování příkladu 8.3



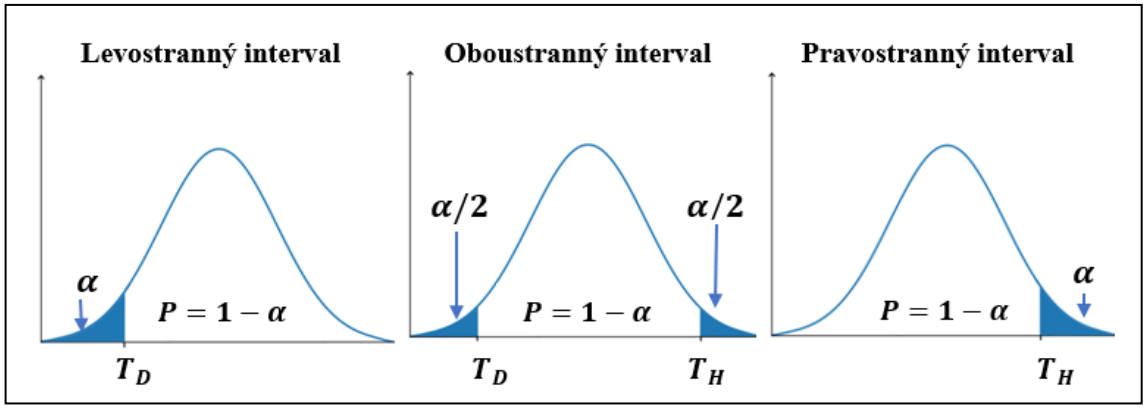
Modrá svislá čára představuje skutečnou hodnotu populačního průměru (μ=10). Vodorovné úsečky jsou jednotlivé 95% intervaly spolehlivosti vypočítané pro každý z 50 náhodných výběrů.

Jak můžete vidět z grafu, většina intervalů protíná modrou čáru, což znamená, že obsahují populační průměr. Některé z intervalů, které jsou zobrazeny červeně, však modrou čáru neprotínají. To názorně ilustruje, že i když je vysoká pravděpodobnost (95 %), že interval spolehlivosti bude obsahovat neznámý parametr, není to jistota.

Graf ilustruje, že **interval spolehlivosti je náhodná veličina** a že i při vysoké spolehlivosti (95 %) existuje šance, že některé intervaly skutečnou hodnotu nezachytí. Tento princip je klíčový pro statistické závěry. Ukazuje, že i při vysoké spolehlivosti musíme počítat s určitou mírou nejistoty.

Hladina významnosti α udává s jakou pravděpodobností nebude odhadovaná charakteristika základního souboru zahrnuta ve vypočteném intervalu spolehlivosti. Hladina významnosti vyjadřuje riziko toho, že náš odhad nebude správný – představuje **míru rizika** chyby odhadu (obr. 8.3). Její hodnota je dána dopočtem do celkové spolehlivosti, ve které se odhadovaná charakteristika může nacházet, nejčastěji nabývá hodnoty 0,05; 0,01.

Obrázek 8.3: Podstata intervalového odhadu spolehlivosti



Příklad 8.2

V pracovní den jsme uskutečnili 15 měření ($n = 15$) času čekání na autobus městské linky č. 99.

S pravděpodobností 90 % byl stanoven interval, že průměrná doba čekání na městskou linku autobusu v pracovní den je v rozmezí od 6–9 minut.

Zvýšení spolehlivosti na 99 % při stejném rozsahu souboru, stanovilo interval v rozmezí 0–15 minut → je pro nás tato informace ještě relevantní?

Řešení

V případě, že požadujeme vysokou spolehlivost při odhadu doby čekání na autobus musíme zvýšit počet prováděných měření → rozsah souboru n.

Vysokou spolehlivost odhadu je možné zajistit zvýšením rozsahu výběrového souboru.

Intervaly pro odhad parametrů základního souboru

Jednostranný	**Levostranný**	$P(\theta \leq T_D) = \alpha$	
	$(\boldsymbol{T_D}, \infty)$	$P(\theta > T_D) = 1 - \alpha$	(8.7)

Kde: $1 - \alpha$ .. koeficient spolehlivosti,

α ..... riziko podhodnocení (hladina významnosti).

Jednostranný	**Pravostranný**	$P(\theta \geq T_H) = \alpha$	
	$(-\infty, \boldsymbol{T_H})$	$P(\theta < T_H) = 1 - \alpha$	(8.8)

Kde: α ..... riziko nadhodnocení (hladina významnosti).

Oboustranný

$(\boldsymbol{T_D}, \boldsymbol{T_H})$

$$P(\theta \leq T_D) = P(\theta \geq T_H) = \frac{\alpha}{2}$$
$$P(T_D \leq \theta < T_H) = 1 - \alpha \qquad (8.9)$$

Kde: α …… riziko odhadu (hladina významnosti).

Z hodnot oboustranného intervalu spolehlivosti můžeme vypočítat **maximální přípustná chybu odhadu $\boldsymbol{\Delta}$**, která představuje polovinu šířky tohoto intervalu. Vypočítáme ji jako rozdíl horní a dolní hranice intervalu, který vydělíme dvěma ($\frac{\boldsymbol{T_H - T_D}}{\boldsymbol{2}}$)**.**

Maximální přípustná chyba je hodnota, která nám říká, jakou největší chybu jsme ochotni akceptovat při odhadu populačního parametru na základě výběrového souboru.

Tato chyba je velmi důležitou charakteristikou, protože na jejím základě obvykle určujeme, jak velký vzorek potřebujeme, abychom dosáhli požadované přesnosti. Představuje cílovou hodnotu pro přesnost, kterou si stanovíme ještě před sběrem dat.

Interpretace intervalu spolehlivosti (konfidenčního intervalu) vychází z předpokladu, že náhodný je interval spolehlivosti, nikoliv parametr. Proto se výroky o pravděpodobnosti musí týkat intervalu, a ne parametru rozdělení, který je konstantní, neměnný, neznámý, a proto jej odhadujeme.

Nyní budeme využívat uvedené poznatky týkají se bodových a intervalových odhadů parametrů rozdělení na nejpoužívanější statistické charakteristiky za předpokladu, že sledovaná **náhodná veličina *X* byla vybrána z populace s normálním rozdělením.**

8.3 Bodový a intervalový odhad střední hodnoty

Nejlepším nestranným a konzistentním **bodovým odhadem střední hodnoty $\hat{\boldsymbol{\mu}}$** populace pocházející z normálního rozdělení je **výběrový průměr $\overline{\boldsymbol{X}}$.**

$$\hat{\mu} = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = E(X) \qquad (8.10)$$

8.3.1 Intervalový odhad střední hodnoty při známém populačním rozptylu

Sestrojení intervalu spolehlivosti pro různé výběrové charakteristiky se odvíjí od příslušné statistiky (výběrové charakteristiky), která vychází ze znalosti nebo neznalosti parametru a typu rozdělení náhodné veličiny.

Postup odvození intervalu spolehlivosti je vždy stejný. V této kapitole uvedeme jako vzor odvození intervalu spolehlivosti pro **střední hodnotu μ** za předpokladu, že **známe rozptyl základního souboru** (populační rozptyl).

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Výběrová statistika pro parametr populačního průměru je v tomto případě dána vztahem 6.1. (viz kapitola Náhodný výběr) a řídí se **kvantily normovaného normálního rozdělení**.

$$U = \frac{\bar{X} - \mu}{\sigma}\sqrt{n} \sim u_{\alpha}$$

Oboustranný interval spolehlivosti pro **parametr μ** vychází z předpokladu, že hustota pravděpodobnosti $f(x)$ normovaného normálního rozdělení výběrové náhodné veličiny je **symetrická funkce.**

$$P\left(u_{1-\frac{\alpha}{2}} < U < u_{1-\frac{\alpha}{2}}\right) = 1 - \alpha, \tag{8.11}$$

$$P\left(u_{1-\frac{\alpha}{2}} < \frac{\bar{X}-\mu}{\sigma}\sqrt{n} < u_{1-\frac{\alpha}{2}}\right) = 1 - \alpha, \tag{8.12}$$

$$P\left(\bar{X} - u_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + u_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha. \tag{8.13}$$

Levostranný interval spolehlivosti

$$P\left(\bar{X} - u_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}} < \mu\right) = 1 - \alpha. \tag{8.14}$$

Pravostranný interval spolehlivosti

$$P\left(\mu < \bar{X} + u_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha. \tag{8.15}$$

Jak již bylo uvedeno, postup tvorby intervalu spolehlivosti pro všechny základní číselné charakteristiky je obdobný. Proto nebudeme v následujícím textu opakovat celý postup, ale vždy si jen připomeneme, jakou výběrovou statistiku použít a z jakého typu rozdělení je odvozená.

8.3.2 Intervaly spolehlivosti pro μ při neznámém populačním rozptylu

V praxi se většinou setkáváme s tím, že při sestavování intervalu spolehlivosti pro střední hodnotu μ **neznáme populační rozptyl**. V takovém případě pracujeme s výběrovým rozptylem s^2 nebo jeho odmocninou směrodatnou odchylkou s.

Pokud máme dostatečný rozsah výběrového souboru $\boldsymbol{n > 30}$, je možné určit intervalový odhad průměru pomocí výběrové statistiky vztah 6.2 (viz kapitola Náhodný výběr), která se řídí **kvantily normovaného normálního rozdělení**.

Oboustranný interval $$P\left(\bar{X} - u_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{X} + u_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}\right) = 1 - \alpha. \quad (8.16)$$

Levostranný interval $$P\left(\bar{X} - u_{1-\alpha} \cdot \frac{s}{\sqrt{n}} < \mu\right) = 1 - \alpha. \quad (8.17)$$

Pravostranný interval $$P\left(\mu < \bar{X} + u_{1-\alpha} \cdot \frac{s}{\sqrt{n}}\right) = 1 - \alpha. \quad (8.18)$$

Jestliže nemáme dostatečný rozsah souboru $\boldsymbol{n < 30}$. Kvantily normovaného normálního rozdělení nahrazujeme, kvantily **Studentova t-rozdělení** a při výpočtu intervalu spolehlivosti využíváme výběrovou statistiku (vztah 6.3). Hustota pravděpodobnosti $f(x)$ Studentova t-rozdělení výběrové náhodné veličiny je **symetrická funkce.**

Oboustranný interval $$P\left(\bar{X} - t_{1-\frac{\alpha}{2}(n-1)} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{1-\frac{\alpha}{2}(n-1)} \cdot \frac{s}{\sqrt{n}}\right) = 1 - \alpha. \quad (8.19)$$

Levostranný interval $$P\left(\bar{X} - t_{1-\alpha(n-1)} \cdot \frac{s}{\sqrt{n}} < \mu\right) = 1 - \alpha. \quad (8.20)$$

Pravostranný interval $$P\left(\mu < \bar{X} + t_{1-\alpha(n-1)} \cdot \frac{s}{\sqrt{n}}\right) = 1 - \alpha. \quad (8.21)$$

Příklad 8.4

Na základě výpočtu popisných statistik (příklad 5.2) u proměnné cena notebooku, kterou jsem analyzovaly z datové matice *„Notebooky“* (datová matice je umístěna v kurzu v Moodle), jsme zjistili, že cena notebooků má normální rozdělení.

Z náhodného výběru o rozsahu 150 notebooků určete:

a) Se spolehlivostí 95 % intervalový odhad pro střední cenu notebooků.
b) Výběrovou chybu odhadu → SE_T.
c) Maximální přípustnou chybu odhadu → Δ.

Řešení

SPSS Analyze → Descriptive Statistics → Explore → proměnná *Cena* do okna Variable(s) → OK

Pokračování příkladu 8.4

			Statistic	Std. Error
Cena (Kč)	Mean		21492,44	728,775
	95% Confidence Interval for Mean	Lower Bound	20052,37	
		Upper Bound	22932,51	
	5% Trimmed Mean		21325,44	
	Median		22179,50	
	Variance		79666921,322	
	Std. Deviation		8925,633	

a) Z výpočtu, pomocí programu SPSS (kde je standardně nastavena spolehlivost 95 %) sestavíme interval spolehlivosti pro populační průměr

$$P\,(20052{,}37 < \mu < 22932{,}51) = 95\ \%.$$

Populační průměr ceny notebooku je s 95% spolehlivostí v rozmezí od 20 052 Kč do 22 932 Kč. To znamená, že s 95% jistotou se skutečná průměrná cena všech notebooků v populaci nachází v tomto rozmezí → kdybychom zjišťovaly ceny opakovaně, tak by 95 % nalezených intervalových odhadů obsahovalo skutečnou hodnotu populačního průměru ceny notebooku.

b) Stanovení výběrové chyby, která vzniká při bodovém odhadu populačních parametrů, provedeme pomocí střední chyby odhadu, jejíž hodnota je 728,78 Kč a kterou vyčteme přímo z výstupu SPSS.

c) Maximální přípustná chyba **Δ** je dána rozdílem horní a dolní meze intervalu spolehlivosti dělené dvěma ($\frac{T_H - T_D}{2}$). $\frac{22\,932{,}51 - 20\,052{,}37}{2} = 1\,440{,}07$

Průměrná cena notebooku, kterou jste vypočítal na základě hodnot z výběrového souboru (21 492 Kč), by se od skutečné průměrné ceny všech notebooků v populaci mohla lišit maximálně o 1 440 Kč oběma směry (nahoru i dolů) při dané spolehlivosti.

8.4 Bodový a intervalový odhad variability

Při odhadu **variability** se nejčastěji zaměřujeme na odhad dvou důležitých populačních parametrů: **rozptylu** a **směrodatné odchylky**. Tyto parametry nám poskytují důležité informace o tom, jak jsou hodnoty v populaci rozptýleny a jaká je jejich heterogenita.

8.4.1 Bodový a intervalový odhad populačního rozptylu

Nejlepším asymptoticky nestranným bodovým **odhadem populačního rozptylu $\hat{\sigma}^2$**, který pochází z normálního rozdělení, je **výběrový rozptyl s^2** .

$$\hat{\sigma}^2 = s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} = D(X) \qquad (8.22)$$

K modelování intervalového odhadu pro rozptyl za předpokladu, že populace má normální rozdělení, se používá výběrová statistika (vztah 6.4), která se řídí $chí$ −kvadrát rozdělení, jehož hustota pravděpodobnosti $f(x)$ výběrové náhodné veličiny je **asymetrická funkce.**

Oboustranný interval $$P\left(\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{\alpha/2}}\right) = 1-\alpha. \quad (8.23)$$

Levostranný interval $$P\left(\frac{(n-1)s^2}{\chi^2_{1-\alpha}} < \sigma^2\right) = 1-\alpha. \quad (8.24)$$

Pravostranný interval $$P\left(\sigma^2 < \frac{(n-1)s^2}{\chi^2_{\alpha}}\right) = 1-\alpha. \quad (8.25)$$

8.4.2 Bodový odhad směrodatné odchylky

Po odhadu populačního rozptylu se zaměříme na **směrodatnou odchylku**, která je neméně důležitou charakteristikou variability. Směrodatná odchylka je úzce spjata s rozptylem, ale má velkou výhodu v tom, že je vyjádřena ve **stejných jednotkách** jako původní data. Díky tomu je mnohem **snazší ji interpretovat** a porovnávat s ostatními statistikami, jako je například průměr.

Bodový odhad **směrodatné odchylky** základního souboru $\hat{\sigma}$ se určuje z odchylek jednotlivých pozorování od výběrového průměru pro n-1 stupňů volnosti[16].

$$\hat{\sigma} = s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} \quad (8.26)$$

Další úpravou tohoto vztahu a za předpokladu, že výběrová směrodatná odchylka s se vypočítá ze vztahu 5.27 můžeme odvodit bodový odhad pro **směrodatnou odchylku výběrových průměru $\hat{\sigma}_{\bar{X}}$.**

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \quad (8.27)$$

$$\hat{\sigma}_{\bar{X}} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{s}{\sqrt{n-1}} \quad (8.28)$$

POZOR! *Ačkoliv je princip **stupňů volnosti** a používání $n-1$ ve vzorcích pro správné a přesné odhady **zásadní z teoretického hlediska**, v našich praktických výpočtech ho nebudeme explicitně řešit. Důvodem je, že pracujeme s **velkými výběry** (většinou n>30), kde je rozdíl mezi dělením n a $n-1$ zanedbatelný. Navíc, statistické programy, jako je **SPSS**, provádějí tyto výpočty automaticky a s vysokou přesností, aniž bychom museli tuto úpravu zadávat ručně.*

[16] Stupně volnosti v teorii pravděpodobnosti udávají, kolik hodnot v souboru můžeme „volně měnit“, aniž by se porušila podmínka nestrannosti a nezkreslenosti odhadů.

8.5 Bodový a intervalový odhad relativní četnosti

V mnoha případech při řešení praktických úloh odhadujeme podíl jednotek s danou vlastností v celé populaci. Odhadujeme tedy parametr π alternativního rozdělení.

Při dostatečně velkém rozsahu výběru $n > \frac{9}{p(1-p)}$ je nejlepším nestranným, vydatným, konzistentním a dostatečným **bodovým odhadem relativní četnosti $\hat{\pi}$ výběrová relativní četnost p.**

Při konstrukci intervalu spolehlivosti relativní četnosti π postupujeme dvojím způsobem:

a) Při **malých rozsazích** vycházíme z toho, že při **výběrech s opakováním** se odhad relativní četnosti řídí **binomickým rozdělení,** při **výběrech bez opakování** pak **hypergeometrickým rozdělením.** V praxi obvykle krajní hodnoty intervalů v takovémto případě neurčujeme nebo v případě, že je to požadováno musíme využít speciální tabulky uvedených rozdělení anebo statistický software.

b) Vycházíme-li ze skutečnosti, že **rozsah výběrového souboru je dostatečně velký,** lze rozdělení výběrové relativní četnosti **aproximovat normálním rozdělením**. Výpočet intervalu spolehlivosti pro odhad relativní četnosti se pak provádí pomocí výběrové statistiky (vztah 6.5).

Takto vypočtený interval spolehlivosti je jen přibližný, a to z důvodu, že ve výběrové statistice není zohledněna tzv. oprava pro spojitost, kterou je třeba uvažovat tehdy, nahrazujeme-li nějaké diskrétní rozdělení rozdělením spojitým.

Oboustranný interval $$P\left(p - u_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{p(1-p)}}{\sqrt{n}} < \pi < p + u_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right) = 1 - \alpha. \quad (8.29)$$

Levostranný interval $$P\left(p - u_{1-\alpha} \cdot \frac{\sqrt{p(1-p)}}{\sqrt{n}} < \pi\right) = 1 - \alpha. \quad (8.30)$$

Pravostranný interval $$P\left(\pi < p + u_{1-\alpha} \cdot \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right) = 1 - \alpha. \quad (8.31)$$

Příklad 8.5

Zjišťujeme, jaký je podíl notebooků, které mají numerickou klávesnici. Data máme z matice *„Notebooky“*, která už je náhodným výběrem 150 kusů notebooků. Z průzkumové analýzy této kvalitativní proměnné jsme zjistili, že 68 notebooků má numerickou klávesnici.

a) Vypočítejte výběrový podíl (relativní četnost) notebooků s numerickou klávesnicí.

b) Proveďte 95% intervalový odhad podílu všech notebooků v populaci, které mají numerickou klávesnici.

c) Vypočítejte a interpretujte maximální přípustnou chybu odhadu.

Řešení

SPSS Analyze → Descriptive Statistics → Frequencies → proměnná *Klávesnice* do okna Variable(s) → OK

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	ano	68	45,3	45,3	45,3
	ne	82	54,7	54,7	100,0
	Total	150	100,0	100,0	

a) Výběrový podíl (neboli relativní četnost) se vypočítá jako podíl počtu notebooků s numerickou klávesnicí ku celkovému počtu notebooků ve výběru → $p_i = 45{,}3$.

SPSS Analyze → Compare Means and Proportions → One-Sample Proportions → do okna Test Variable(s) vložíme proměnnou *Klavesnice,* v okně Define Success vybereme Value(s) a napíšeme zde kód pro *ano* v našem případě hodnotu 1 → Confidence Intervals zvolíme Inetrval Type(s) Wald → Continue → OK

One-Sample Proportions Confidence Intervals

	Interval Type	Successes	Observed Trials	Proportion	Asymptotic Standard Error	95% Confidence Interval Lower	Upper
Numerická klávesnice = ano	Wald	68	150	,453	,041	,374	,533

b) Z výpočtu v programu SPSS jsme sestavili interval spolehlivosti pro odhad podílu všech notebooků v populaci, které mají numerickou klávesnici.

$$P\,(0{,}374 < \pi < 0{,}533) = 95\,\%$$

To znamená, že s 95% jistotou se skutečný podíl všech notebooků s numerickou klávesnicí v populaci nachází v rozmezí od 37,4 % do 53,3 %.

c) Maximální přípustná chyba odhadu (Δ) představuje polovinu šířky intervalu spolehlivosti $\Delta = \left(\frac{T_H - T_D}{2}\right) = \frac{0{,}534 - 0{,}375}{2} = 0{,}0795$.

To znamená, že náš odhad podílu notebooků s numerickou klávesnic se od skutečné hodnoty v populaci může lišit maximálně o 7,95 %.

Shrnutí kapitoly

Zobecnitelnost vyjadřuje, zda a s jakou jistotou platí výsledky získané z výběrového souboru na celou populaci.

Odhad neznámé populační charakteristiky může být buď **bodový**, nebo **intervalový**.

Bodový i **intervalový odhad** mají náhodně proměnlivý charakter. Výpočet odhadů pro každý náhodný výběr vede ke stanovení „trochu jiného“ bodového odhadu nebo intervalu spolehlivosti.

Bodový odhad spočívá v tom, že na základě zjištěných údajů z náhodného výběru odhadneme předem stanoveným způsobem jedno číslo, které považujeme za nestranný odhad parametru základního souboru. Vlastnosti bodového odhadu $\widehat{\theta}$ (nestrannost, konzistence a vydatnost) vypovídají o tom, že k odhadu populační charakteristiky θ byla použitá vhodná statistika T.

Intervalový odhad je číselný interval, ve kterém se s předem zvolenou pravděpodobností nachází odhadovaná hodnota populačního parametru. Takto sestrojený interval hodnot se nazývá **konfidenční interval** (neboli interval spolehlivosti).

Spolehlivost odhadu je dána pravděpodobností, s jakou se odhadovaná charakteristika základního souboru θ bude nacházet v intervalu vymezeném příslušnou statistikou $T(X)$.

Šířka intervalu spolehlivosti je závislá na rozsahu výběrového souboru. Čím větší rozsah výběrového souboru máme, tím přesnější je odhad populačního parametru.

Literatura

HASTIE, Trevor; Robert TIBSHIRANI a Jerome FRIEDMAN. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer, 2009. ISBN 978-0-387-84857-0.

FIELD, Andy. *Discovering Statistics Using IBM SPSS Statistics*. 5th ed. Los Angeles: SAGE Publications, 2018. ISBN 978-1526436566.

LITSCHMANNOVÁ Martina, *Úvod do statistiky.* Online. VŠB – Technická univerzita Ostrava, 2012. Dostupné z: https://mi21.vsb.cz/sites/mi21.vsb.cz/files/unit/interaktivni_uvod_do_statistiky.pdf. [citováno 2023-08-03]

ŘEHÁK, Jan a Petr SOUKUP. *Úvod do statistické analýzy dat v psychologii*. Praha: Portál, 2021. ISBN 978-80-262-1718-0.

8 Kontrolní otázky

1. Co je hlavním cílem teorie odhadu?
2. Jaký je rozdíl mezi populační charakteristikou (parametrem) a výběrovou charakteristikou (statistikou)?
3. Jaké jsou dva základní typy odhadů neznámé populační charakteristiky?
4. Definujte bodový odhad?
5. Co nám říká interval spolehlivosti?
6. Jaký je vztah mezi šířkou intervalu spolehlivosti a rozsahem výběrového souboru?
7. Jak ovlivní zvýšení spolehlivosti odhadu šířku intervalu, pokud se rozsah souboru nezmění?
8. Jaké základní vlastnosti by měl splňovat dobrý bodový odhad?
9. Stručně vysvětlete, co znamená nestrannost a konzistence.
10. Jak se vypočítá maximální přípustná chyba odhadu z intervalu spolehlivosti?
11. K čemu nám tato hodnota slouží?
12. Vysvětlete na příkladu 95% spolehlivosti, co znamená spolehlivost odhadu.
13. Jaký je rozdíl mezi predikčním intervalem a konfidenčním intervalem (intervalem spolehlivosti)?

8 Příklady k procvičení

Na základě datové matice *„Byty“*, kterou naleznete v kurzu v Moodle, odpovězte na následující otázky.

8.1 Bodový odhad a intervalový odhad populačního průměru

8.1.1 Vypočtěte a interpretujte bodový odhad pro průměrné měsíční nájemné za byt v populaci. Sestavte 95% interval spolehlivosti pro průměrné měsíční nájemné. Následně interpretujte význam tohoto intervalu a určete maximální přípustnou chybu odhadu.

$$[\hat{\mu} = 15\,531 \text{ Kč} ; P(14\,817 \text{ Kč} < \mu < 16\,245 \text{ Kč}) = 95\%;\ \Delta = 714 \text{ Kč}]$$

8.1.2 Z datové matice vytvořte podsoubor, který bude obsahovat pouze byty s dispozicí 3+1. Na základě tohoto podsouboru odpovězte na následující otázky. Vypočítejte a interpretujte bodový odhad průměrné plochy bytů 3+1 v populaci. Sestavte 95% interval spolehlivosti pro průměrnou plochu všech bytů 3+1 v populaci. Následně interpretujte význam tohoto intervalu a určete maximální přípustnou chybu odhadu.

$$[\hat{\mu} = 82{,}77\ m^2 ; P(80{,}13\ m^2 < \mu < 85{,}42\ m^2) = 95\%;\ \Delta = 2{,}645\ m^2]$$

8.2 Bodový odhad a intervalový odhad populačního podílu

8.2.1 Vypočtěte a interpretujte bodový odhad podílu bytů s výtahem v populaci. Sestavte 95% interval spolehlivosti pro podíl všech bytů v populaci, které mají výtah. Následně interpretujte význam tohoto intervalu a určete maximální přípustnou chybu odhadu.

$$[\hat{\pi} = 61{,}3\ \%; P(53{,}5\% < \pi < 69{,}1\ \%) = 95\%;\ \Delta = 7{,}79\ \%]$$

8.2.2 Z datové matice vytvořte podsoubor, který bude obsahovat pouze novostavby. Na základě tohoto podsouboru odpovězte na následující otázky. Vypočítejte a interpretujte bodový odhad podílu bytů s výtahem v novostavbách. Sestavte 95% interval spolehlivosti pro podíl všech bytů v novostavbách, které mají výtah. Následně interpretujte význam tohoto intervalu a určete maximální přípustnou chybu odhadu.

$$[\hat{\pi} = 81{,}8\ \%; P(72{,}5\ \% < \pi < 91{,}1\ \%) = 95\%;\ \Delta = 9{,}3\ \%]$$

9 TESTOVÁNÍ STATISTICKÝCH HYPOTÉZ

Úkolem testování statistických hypotéz je posoudit platnost či neplatnost správně[17] formulované hypotézy.

Testování statistických hypotéz je, spolu s odhady parametrů rozdělení základního souboru, jednou z hlavních úloh matematické statistiky. Testování nám umožňuje na základě informací o chování výběrové náhodné veličiny, usuzovat s předem stanovenou pravděpodobností na charakter celé populace. Stejně jako při odhadech spolehlivosti, i zde platí předpoklad, že výběrový soubor musí být reprezentativním zástupcem základního souboru, musí věrně reprezentovat známé parametry cílové populace – je tedy jejím odrazem.

Hypotéza znamená předpoklad, tvrzení či domněnku. V běžném využití je hypotéza neprokázané tvrzení, které by mělo být možné zkoumat, empiricky ověřovat. Odborná hypotéza je vědecky přijatelný předpoklad umožňující vědecké vysvětlení jevů. Na počátku vědeckého poznání stojí domněnka, kterou hypotéza rozpracovává a která musí být podložena celou řadou faktů vytyčujících nám další směr výzkumu. Hypotéza představuje vědecký předpoklad, který byl vyvozen z vědecké teorie.

Správně stanovené hypotézy mají klíčovou roli v rozhodování, jaké jevy mají být předmětem šetření, aby nedocházelo k získávání nadbytečných informací anebo naopak nebyly opomenuty informace důležité. Významnou funkcí hypotéz je propojení teoretické a empirické složky práce, tzv. operacionalizace hypotéz**,** tj. smysluplné a efektivní transformování obecných pojmů do podoby empiricky pozorovatelných znaků.

V našem textu se dále budeme podrobněji zabývat testování **statistických hypotéz**, které představují určitou podtřídu vědeckých hypotéz.

9.1 Statistická hypotéza

Statistickou hypotézou rozumíme jakýkoliv předpoklad či tvrzení, které se může týkat neznámých parametrů, tvaru rozdělení náhodné veličiny, a také dalších vlastností základního souboru – populace. Statistickou hypotézou může být například tvrzení, že dva náhodné výběry pocházejí ze stejného rozdělení, náhodný výběr pochází z normálního rozdělení, ověřování reprezentativnosti výběrového souboru vůči cílové populaci, zjišťování toho, zda dva výběrové soubory mají stejný populační parametr atd.

[17] Hypotéza představuje jednoznačné tvrzení, které se vyjadřuje oznamovací větou.

9.1.1 Parametrická hypotéza

Parametrické hypotézy se týkají výhradně hodnoty jednoho nebo několika parametrů rozdělení náhodné veličiny. K ověřování hypotéz o neznámých parametrech rozdělení pravděpodobnosti, kterým se uvažovaná náhodná veličina řídí, se používají **parametrické testy**. Tyto testy jsou založeny na určitých znalostech o charakteru pravděpodobnostního rozdělení studovaných náhodných veličin (např. t-test je založen na předpokladu, že základní soubory, z nichž byly provedeny náhodné výběry, mají normální rozdělení se stejnými rozptyly). V převážné většině jde o početně náročnější, ale silné testy.

POZOR! *Je třeba si uvědomit, že testování parametrických hypotéz v případě chybně určeného rozdělení pravděpodobnosti parametrické testové statistiky může vést k mylným závěrům!*

9.1.2 Neparametrická hypotéza

Neparametrické hypotézy se týkají tvrzení, které jsou zaměřena na obecné vlastnosti populace (např. o tvaru rozdělení základního souboru, o závislosti proměnných atd.) bez konkrétní znalosti parametrů rozdělení. K ověřování těchto tvrzení se používají **neparametrické testy.** Neparametrické testy nevyžadují znalost typu rozdělení např. normalitu rozdělení pravděpodobnosti, ale pouze jeho symetrii. Výpočet neparametrických testů ve velké většině případů nevychází ze skutečných hodnot náhodné veličiny, ale opírá se o pořadová čísla (ranky), která skutečné hodnoty nahrazují. Neparametrické testy mají menší sílu. Používají se především tehdy, jestliže není možné použit testy parametrických z důvodů nesplnění jejich předpokladů.

9.1.3 Nulová a alternativní hypotéza

Předpoklad, jehož platnost ověřujeme, se nazýváme **nulová hypotéza H_0**. Tato hypotéza představuje určitý rovnovážný stav a bývá vyjádřena rovnosti „= “. Je to tvrzení, které obvykle vyjadřuje „*nepřítomnost rozdílu*“ mezi testovanými náhodnými veličinami.

Nulovou hypotézou mohou být určitá tvrzení o parametrech rozdělení, nebo tvaru pravděpodobnostního rozdělení.

Jestliže populační parametr, odhadovaný na základě výběru označíme θ a hypotetickou hodnotu tohoto parametru θ_0, pak nulová hypotéza má tvar:

$$H_0: \theta = \theta_0 \tag{9.1}$$

Nulová hypotéza tvrdí, že odhadovaný populační parametr se statisticky významně **neliší** od hypotetické hodnoty.

V případě, že se nám nepodaří nulovou hypotézu vyvrátit, hovoříme o jejím **nezamítnutí**, a nikoliv o přijetí. Nezamítnutí nulové hypotézy neznamená absolutní potvrzení její platnosti. Pouze vyjadřuje skutečnost, že výsledek testu **neprokázal** dostatečně velkou **neshodu** mezi odhadovaným populačním parametrem a hypotetickou hodnotu tohoto parametru, která by byla důvodem k zamítnutí hypotézy.

Proti nulové hypotéze stavíme **alternativní hypotézu $\boldsymbol{H_1}$**. Ta představuje porušení rovnovážného stavu a je tvrzením o neznámých vlastnostech rozdělení pravděpodobnosti sledované náhodné veličiny, které popírá platnost nulové hypotézy. Alternativní hypotézu přijímáme v případě, že jsme nulovou hypotézu H_0 zamítli jako nesprávnou.

Tvar alternativní hypotézy záleží na typu nerovnosti mezi odhadovaným populačním parametrem θ a hypotetickou hodnotou tohoto parametru θ_0.

Oboustranná alternativa	$H_1: \theta \neq \theta_0$	(9.2)
Levostranná alternativa	$H_1: \theta < \theta_0$	(9.3)
Pravostranná alternativa	$H_1: \theta > \theta_0$	(9.4)

Slovní vyjádření alternativní hypotézy nám říká, že odhadovaný populační parametr se statisticky **významně liší** od hypotetické hodnoty.

9.2 Statistický test

Statistický test je rozhodovací postup, který nám na základě náhodného výběru umožní ověřit, zda hypotéza platí či nikoliv. Tento test reprezentuje takzvaná **výběrová statistika** (neboli **testová statistika**).

Testová statistika $T \rightarrow T(X = x)$ je funkcí náhodného výběru X, která nám charakterizuje stupeň nesouladu mezi předpokladem o datech (parametrech rozdělení, typu rozdělení náhodných veličin) a hodnotami získanými na základě výběrového šetření. Testová statistika je tedy také náhodnou veličinou a má určitý typ pravděpodobnostního rozdělení. Jinými slovy, testová statistika je transformace pozorovaných výběrových náhodných veličin, které pocházejí z určitého pravděpodobnostního rozdělení. Charakter rozdělení závisí na tom, jakou hypotézu testujeme. Je založena na odchylce výběrové charakteristiky od hypotetické hodnoty parametru.

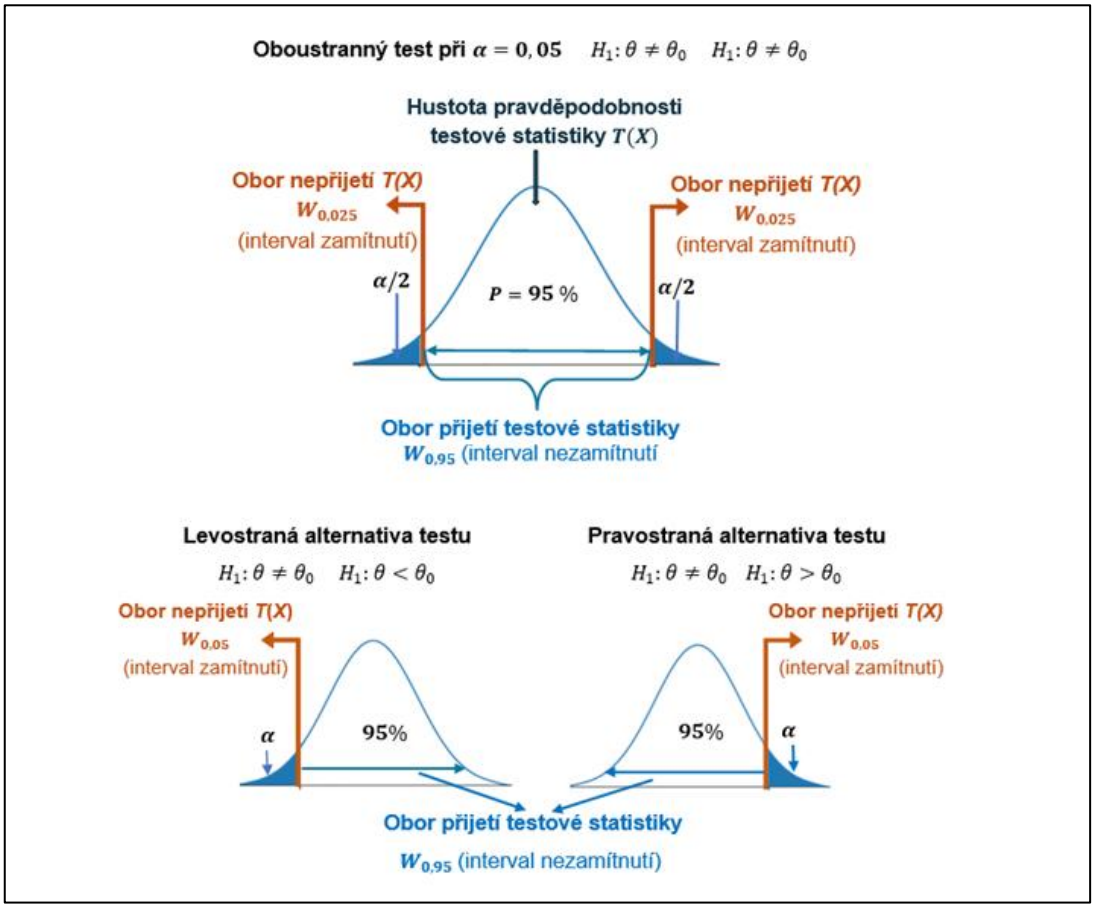
Interval (množinu) možných hodnot, testové statistiky *T(X)* rozdělujeme na dvě disjunktní podmnožiny (obr. 9.1).

Obor nepřijetí testové statistiky (kritický obor nebo interval zamítnutí), který značíme symbolem $\boldsymbol{W_\alpha}$, je ta část všech možných hodnot testové statistiky, kde je velmi nepravděpodobné, že by mohla ležet testová statistika za předpokladu, že platí nulová hypotéza.

Obor přijetí testové statistiky (interval nezamítnutí), který značíme symbolem $\boldsymbol{W_{1-\alpha}}$.

Body, jimiž jsou tyto podmnožiny odděleny, se nazývají **kritické hodnoty**. Konkrétní kritické hodnoty na zvolené **hladině významnosti** α se při praktickém provádění testů vyhledají ve statistických tabulkách nebo jsou určeny speciálními funkcemi ve statistických programech. Nejčastěji se jedná o **kvantil pravděpodobnostního rozdělení testové statistiky** $\boldsymbol{T(X)}$.

Obrázek 9.1: Vymezení oboru možných hodnot testové statistiky



Padne-li vypočtená hodnota testovacího kritéria $T(X)$ do kritického oboru W_α, zamítáme nulovou hypotézu H_0 a přijímáme alternativní hypotézu H_1.

Spadá-li vypočtená hodnota testovacího kritéria $T(X)$ do oboru přijetí $W_{1-\alpha}$, nezamítáme H_0. Nemůžeme však říct, že by to znamenalo její přijetí. Výsledek prováděného testu neurčuje, jak velká shoda je mezi skutečností a testovanou hypotézou.

Při testování statistických hypotéz provádíme rozhodnutí o přijetí nulové hypotézy (H_0) nebo alternativní hypotézy (H_1) na základě náhodného výběru. Je tedy zřejmé, že toto rozhodnutí nelze provést s absolutní jistotou a že vždy existuje riziko chybného rozhodnutí.

Jestliže je nulová hypotéza ve skutečnosti platná, a my ji přesto zamítneme, dopouštíme se **chyby 1. druhu (α)**. Riziko tohoto mylného rozhodnutí je velmi malé a jeho pravděpodobnost je rovna hladině významnosti α. Obvykle volíme 5% nebo 1% hladinu významnosti. Pokud nulová hypotéza platí a my ji nezamítneme, rozhodli jsme správně. Pravděpodobnost tohoto rozhodnutí se označuje jako $1-\alpha$ a nazývá se **spolehlivost**.

Jestliže nezamítneme nulovou hypotézu, která je ve skutečnosti nesprávná, dopouštíme se **chyby 2. druhu (β)**. Její doplněk do 1, tzn. $1-\beta$, vyjadřuje pravděpodobnost správného zamítnutí testované hypotézy (schopnost testu odhalit neplatnost nulové hypotézy) a nazývá se síla testu (obr. 9.2).

Platnost statistické hypotézy musí být vždy zkoumána na základě náhodného výběru z cílové populace!

Obrázek 9.2: Výsledky testu statistických hypotéz

		Výsledek testu	
		H_0 nezamítáme	**H_0 zamítáme**
Skutečnost	**H_0 je pravdivá**	Správné rozhodnutí Pravděpodobnost rozhodnutí: $1-\alpha \rightarrow$ **spolehlivost.**	**Chyba 1. druhu** Pravděpodobnost rozhodnutí: $\alpha \rightarrow$ hladina významnosti.
	H_0 není pravdivá	**Chyba 2. druhu** Pravděpodobnost rozhodnutí: β.	Správné rozhodnutí Pravděpodobnost rozhodnutí: $1-\beta \rightarrow$ **síla testu.**

Při testu hypotézy H_0 proti alternativě H_1 požadujeme, aby pravděpodobnost chyby 1. a 2. druhu byla co možná nejmenší.

Tato skutečnost je však v rozporu – pro daný rozsah výběru a pro dané testovací kritérium snížení hladiny významnosti α (tzn. snížení pravděpodobnosti chyby 1. druhu) má za následek zvýšení pravděpodobnosti chyby 2. druhu β a naopak. Způsob, jak současně snížit hodnotu pravděpodobnosti α a β spočívá ve zvýšení rozsahu náhodného výběru.

Doporučený obecný postup při testování statistických hypotéz

Klasický test

1. Formulace výzkumné otázky ve formě H_0 a H_1.
2. Volba hladiny významnosti $\alpha \rightarrow$ představuje pravděpodobnost rizika chybného zamítnutí nulové hypotézy.
3. Volba vhodného testu $\rightarrow$ volíme takovou testovou statistiku, u které známe rozdělení pravděpodobnosti při platnosti nulové hypotézy. Charakter rozdělení závisí na tom, jakou hypotézu testujeme.
4. Výpočet testové statistiky $T(X) \rightarrow$ která je funkcí náhodného výběru X a charakterizuje stupeň nesouladu mezi předpokladem o datech a hodnotách získaných na základě výběrových šetření.
5. Vymezení kritického oboru $\rightarrow$ na základě rozdělení pravděpodobnosti testové statistiky $T(X)$ a zvolené hladiny významnosti určíme kritickou hodnotu (konkrétní kvantil), který odděluje podmnožiny oboru přijetí a oboru zamítnutí nulové hypotézy.
6. Rozhodnutí o nulové hypotéze $\rightarrow$ posouzení toho, zdali vypočtená hodnota testové statistiky $T(X)$ spadá do oboru přijetí $W_{1-\alpha}$ nulové hypotézy či nikoli.
7. Interpretace výsledků provedeného testu $\rightarrow$ výsledek testu neurčuje, jak velká shoda je mezi skutečností a testovanou hypotézou.

Uvedený postup testování provádíme s určitou úpravou i tehdy, když data zpracováváme ve statistickém programu. V takovém případě není srovnávána testová statistika s kritickou hodnotou (klasický přístup), ale rozhodnutí o výsledku testu je spojeno s výpočtem *p*-hodnoty.

***P*-hodnota testu kvantifikuje pravděpodobnost výskytu hodnoty testovací statistiky za podmínky, že nulová hypotéza platí.** Udává nejnižší možnou hladinu významnosti α pro zamítnutí nulové hypotézy. Hodnota P-hodnota konstruovaná pro libovolný test se na rozdíl od hodnoty testového kritéria stává univerzálním prostředkem pro rozhodování o výsledku testování

Čistý test významnosti

1. Formulace výzkumné otázky ve formě H_0 a H_1.
2. Hladina významnosti $\rightarrow$ nepotřebujeme znát jako vstupní údaj. Výsledek testu umožňuje rozhodnout, na jaké hladině významnosti můžeme nulovou hypotézu zamítnout nebo nezamítnout.
3. Volba vhodného testu.

4. Výpočet testové statistiky.
5. Výpočet p-hodnoty $\rightarrow$ p-hodnota je nejnižší hladina významnosti, při které lze ještě zamítnout nulovou hypotézu.
6. Rozhodnutí o nulové hypotéze $\rightarrow$
   a) je-li p-hodnota $\leq \alpha$, pak H_0 zamítáme;
   b) je-li p-hodnota $> \alpha$, pak H_0 nezamítáme.
7. Interpretace výsledků provedeného testu.

Testy můžeme rozlišovat podle **počtu výběrových souborů** na testy:

1. **Jednovýběrové** $\rightarrow$ jeden náhodný výběr, který je pozorován za stejných podmínek.
2. **Dvouvýběrové** $\rightarrow$ dva náhodné výběry za dvojích různých podmínek;
   nezávislé (dva nezávislé náhodné výběry),
   závislé (jeden náhodný výběr z dvourozměrného rozdělení).
3. **Vícevýběrové** $\rightarrow$ více než dva náhodné výběry.

Dále testy dělíme podle znalosti **typu rozdělení a parametrů** daného rozdělení na:

1. **Parametrické** $\rightarrow$ vycházejí z předpokladu dostatečně velkého rozsahu výběru a znalosti jeho rozdělení.
2. **Neparametrické** $\rightarrow$ nezávisí na rozdělení (neznáme typ ani parametry rozdělení populace).

Shrnutí kapitoly

Testování statistických hypotéz, je jeden z důležitých úkolů matematické statistiky. Cílem testování je na základě informací z výběrového souboru usuzovat na charakter celé populace. Je důležité, aby výběrový soubor byl reprezentativním zástupcem základního souboru.

Statistická hypotéza je jakýkoliv předpoklad či tvrzení týkající se neznámých parametrů, tvaru rozdělení nebo dalších vlastností populace.

Nulová hypotéza (H_0) je předpoklad, jehož platnost ověřujeme. Obvykle vyjadřuje rovnovážný stav nebo nepřítomnost rozdílu.

Alternativní hypotéza (H_1) popírá platnost nulové hypotézy. Přijímáme ji v případě, že nulovou hypotézu zamítneme.

Statistický test je rozhodovací postup, který nám na základě náhodného výběru pomůže ověřit platnost hypotézy. Klíčovým prvkem je **testová statistika**, která kvantifikuje míru nesouladu mezi hypotézou a daty.

Při rozhodování existuje riziko chyb. **Chyba 1. druhu (α):** Zamítneme nulovou hypotézu, i když je pravdivá. **Chyba 2. druhu (β)** nastane tehdy, když nezamítneme nulovou hypotézu, i když je nepravdivá.

Síla testu $(1-\beta)$ je pravděpodobnost, že test správně odhalí neplatnost nulové hypotézy. Snížení chyby 1. druhu má za následek zvýšení chyby 2. druhu a naopak. Jediný způsob, jak snížit obě chyby současně, je zvětšit rozsah výběru.

Literatura

BUDÍKOVÁ, Marie; Maria KRÁLOVÁ a Bohumil MAROŠ. *Průvodce základními statistickými metodami*. vydání první. Praha: Grada Publishing, a.s., 2010. ISBN 978-80-247-3243-5.

CAMM, Jeffrey; James COCHRAN, David ANDERSON, Dennis SWEENEY, Thomas WILLIAMS et al. *Statistics for Business and Economics*. 15th Edition. Boston: Cengage Learning, 2023. ISBN 978-0357715857

HINDLS Richard; Markéta ARLTOVÁ, Stanislava HRONOVÁ, Ivana MALÁ, Luboš MAREK, et al. *Statistika v ekonomii*. První vydání. Professional Publishing, 2018. ISBN 978-80-88260-09-7.

HOŠKOVÁ, Pavla; Andrea JINDROVÁ, Marie PRÁŠILOVÁ a Rudolf ZEIPELT. *Statistika I*. Praha: PEF ČZU, 2013. ISBN: 978-80-213-2341-4.

JOHNSON, Richard A. a Dean W. WICHERN, *Applied multivariate statistical analysis. Sixth edition*. Harlow: Pearson, 2014. ISBN 978-1-292-02494-3.

9 Kontrolní otázky

1. Co je hlavním úkolem testování statistických hypotéz?
2. Jaký je rozdíl mezi parametrickou a neparametrickou hypotézou?
3. Vysvětlete rozdíl mezi nulovou (H_0) a alternativní (H_1) hypotézou.
4. Co vyjadřuje hladina významnosti?
5. Definujte chybu 2. druhu?
6. Jak spolu souvisí hladina významnosti α a pravděpodobnost chyby 2. druhu β?
7. Co je testová statistika $T(X)$?
8. Co se stane, pokud vypočtená hodnota testovacího kritéria $T(X)$ spadne do kritického oboru?
9. Jaký je doporučený obecný postup při testování statistických hypotéz?
10. Co je p-hodnota testu a jak se používá k rozhodnutí o nulové hypotéze?
11. Jaké typy alternativních hypotéz rozlišuje?

10 PARAMETRICKÉ TESTY

Parametrické testy jsou statistické metody, jež využíváme pro testování hypotéz o neznámých parametrech rozdělení náhodné veličiny. Klíčovým předpokladem pro jejich použití je znalost typu rozdělení sledované veličiny. U všech níže uvedených parametrických testů budeme předpokládat **normálně rozdělenou populaci**.

Ověřování tohoto předpokladu provádíme v rámci průzkumové analýzy dat pomocí grafických nástrojů (histogramy, Q-Q grafy, Box-ploty) a testů normality (Shapiro-Wilkův, Kolmogorův-Smirnovův test). Data, u kterých se prokáže porušení předpokladu, je možné normalizovat pomocí transformace. Data, která nelze transformovat, jsou asymetricky rozložená, nebo jsou diskrétní, analyzujeme pomocí neparametrický metod.

10.1 Jednovýběrové parametrické testy

Ve statistice často potřebujeme zjistit, zda se určitá hodnota (např. průměrná výška populace, příjem, nebo průměrný počet prodaných výrobků) skutečně liší od předpokládané či referenční hodnoty.

Jednovýběrové testy nám umožňují dělat závěry o charakteru chování celé populace na základě dat získaných z jednoho reprezentativního výběru. Na základě jednoho výběrového souboru rozhodujeme, zda neznámý populační parametr θ je nebo není roven určité předpokládané číselné hodnotě, či zda je (není) neznámý parametr větší (menší) než předpokládaná číselná hodnota θ_0 (předpoklad, norma).

10.1.1 Test o populačním rozptylu

Test o populačním rozptylu nám na základě dat z náhodného výběru umožňuje posoudit, zda se populační rozptyl σ^2, odhadnutý z výběrového rozptylu s^2, statisticky významně liší od předpokládané hodnoty σ_0^2. Jinými slovy, ověřujeme, jestli je zjištěný rozdíl způsobený náhodnými vlivy, nebo je statisticky významný (signifikantní).

Testová statistika se řídí ***chí*-kvadrát rozdělením** s $n-1$ stupni volnosti. Je to základní rozdělení pro testy o rozptylu.

H_0	H_1	p-hodnota	$T(X)$
$\sigma^2 = \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$p = P(T \leq x_K) = F_0(x_K)$	$\boldsymbol{K = \frac{n-1}{\sigma_0^2} s^2 \sim \chi^2_{\alpha(n-1)}}$ (10.1)
	$\sigma^2 > \sigma_0^2$	$p = P(T \geq x_K) = 1 - F_0(x_K)$	
	$\sigma^2 \neq \sigma_0^2$	$2 \cdot min\{F_0(x_K); 1 - F_0(x_K)\}$	

Kde: σ^2 … populační rozptyl,
σ_0^2 … předpokládána (hypotetická) hodnota rozptylu,
s^2 … výběrový rozptyl,
n … rozsah výběrového souboru,
χ^2_α … kritická hodnota *chí*-kvadrát rozdělení pro $(n-1)$ stupních volnosti.

10.1.2 Testy o populačním průměru

Máme-li normálně rozdělenou populaci s neznámou střední hodnotou μ s neznámým rozptylem σ^2, používáme k ověření předpokladu, že se populační průměr rovná určité hypotetické hodnotě jednovýběrový t-test.

Má-li populace normální rozdělení o známém rozptylu σ^2, používáme tzv. jednovýberový U-test. S takovou situací se v praxi většinou nesetkáme, test je zde uveden především pro úplnost.

H_0	H_1	p-hodnota	$T(X)$
$\mu = \mu_0$	$\mu < \mu_0$	$p = P(T \leq x_t) = F_0(x_t)$	$\boldsymbol{t = \frac{\bar{X} - \mu_0}{s}\sqrt{n} \sim t_{\alpha(n-1)}}$ (10.2)
	$\mu > \mu_0$	$p = P(T \geq x_t) = 1 - F_0(x_t)$	
	$\mu \neq \mu_0$	$2 \cdot min\{F_0(x_t); 1 - F_0(x_t)\}$	$\boldsymbol{U = \frac{\bar{X} - \mu_0}{\sigma}\sqrt{n} \sim u_\alpha}$ (10.3)

Kde: μ … populační průměr,
μ_0 … předpokládána (hypotetická) hodnota,
$\bar{x}$ … nestranný výběrový průměr,
s … výběrová směrodatná odchylka,
σ … populační směrodatná odchylka,
n … rozsah výběrového souboru,
t_α … kritická hodnota Studento t-rozdělení pro $(n-1)$ stupních volnosti,
u_α … kritická hodnota normovaného normálního rozdělení.

Příklad 10.1

Byl vysloven předpoklad, že průměrná cena notebooku je 22 000 Kč. Pomocí údajů z datového souboru *„Notebook"* ověřte na 5% hladině významnosti, zda se průměrná cena u sledovaných notebooků shoduje s vysloveným předpokladem.

Řešení

Proměnná cena notebooku se řídí normálním rozdělením. K testování nulové hypotézy H_0: $\mu = 22\,000$ oproti oboustranné alternativě H_1: $\mu \neq 22\,000$ použijeme **jednovýběrový *t*-test.**

SPSS Analyze → Compare Means and Proportions → One-Sample T Test → do okna Test Variable(s) vložíme proměnnou *Cena,* do okna Test Value napíšeme referenční hodnotu 22 000 → OK

Výstup z programu One-Sample Test je výsledek t-testu pro průměrnou cenu notebooků. Na tomto příkladu si podrobněji popíšeme, co jednotlivé hodnoty v tabulce znamenají a jak se interpretují. Obdobné výstupy budou prezentovat i výsledky dalších parametrických testů, ty už nebudeme popisovat takto detailně.

One-Sample Test

Test Value = 22000

	t	df	Significance One-Sided p	Significance Two-Sided p	Mean Difference	95% Confidence Interval of the Difference Lower	95% Confidence Interval of the Difference Upper
Cena notebooku (Kč) v prodejně	-,696	149	,244	**,487**	-507,560	-1947,63	932,51

Test Value = 22000 → hodnota, proti které testujeme nulovou hypotézu.

***t* = -0,696** → vypočtená **hodnota testové statistiky**.

df = 149 → počet **stupňů volnosti** pro tento test, který se vypočítá jako $n-1$, kde n je počet pozorování (v našem případě 150).

Significance One-Sided *p* = 0,244 → tato p-hodnota se používá pro **jednostranný test.**

Significance Two-Sided *p* = 0,487→ tato p-hodnota se používá pro **oboustranný test.** Protože je tato hodnota vyšší než 0,05 (naše hladina významnosti), **nezamítáme nulovou hypotézu**.

Mean Difference = -507,560 → průměrný rozdíl mezi naším výběrovým průměrem a testovanou hodnotou.

95% Confidence Interval of the Difference (Lower, Upper) → 95% interval spolehlivosti pro rozdíl mezi průměry. Pokud tento interval obsahuje nulu, nulovou hypotézu nezamítáme.

Pokračování příkladu 10.1

Na 5% hladině významnosti jsme **nezamítli nulovou hypotézu**, která tvrdila, že průměrná cena notebooku je 22 000 Kč. Znamená to, že ačkoli se průměrná cena notebooků v našem výběru (21 492,44 Kč) mírně liší od testované hodnoty, tento rozdíl **není statisticky významný**. Nelze vyloučit, že je způsoben pouze náhodou.

***Pozor! Zásada statistického testování**: Statistický test nedokazuje, že nulová hypotéza je pravdivá. Pouze vyhodnocuje, zda existuje dostatek důkazů pro její zamítnutí. To, že nemáme dostatek důkazů pro tvrzení, že se ceny liší, neznamená, že je dokázáno, že jsou stejné. Může se stát, že by větší nebo jiný vzorek dat tento rozdíl odhalil.*

10.1.3 Test o populačním podílu

Test o populačním podílů neboli parametru π, je zaměřený především na testování vlastností kategoriálních dat. Testujeme nulovou hypotézu, že pravděpodobnost nastoupení náhodného jevu v populaci (populační relativní četnost) je „rovna" předpokládané hypotetické pravděpodobnosti. Za předpokladu dostatečně velkého rozsah výběrového souboru $n > 5/\pi_0$ je možné testovou statistiku $T(X)$ aproximovat normovaným normálním rozdělením.

H_0	H_1	p-hodnota	$T(X)$
$\pi = \pi_0$	$\pi < \pi_0$	$p = P(T \leq x_u) = F_0(x_u)$	$U = \dfrac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)}}\sqrt{n} \sim u_\alpha$ (10.4)
	$\pi > \pi_0$	$p = P(T \geq x_u) = 1 - F_0(x_u)$	
	$\pi \neq \pi_0$	$2 \cdot min\{F_0(x_u); 1 - F_0(x_u)\}$	

Kde: π … populační podíl (relativní četnost),
π_0 … předpokládaný podíl (relativní četnost),
p … výběrový podíl (relativní četnost),
n … rozsah výběrového souboru,
u_α … kritická hodnota normovaného normálního rozdělení.

Pro menší rozsahy výběru, nebo v případě, že nejsou splněny podmínky pro aproximaci normálním rozdělením, se k **testování parametru** π používají testy založené na binomickém rozdělení. Tyto testy počítají p-hodnotu přímo z pravděpodobnostní funkce binomického rozdělení.

Příklad 10.2

Byl vysloven předpoklad, že podíl notebooků s herní grafickou kartou v celém souboru dat se rovná 15 %. S použitím údajů z výběrového datového souboru *„Notebook"* ověřte na 5% hladině významnosti, zda se podíl herních notebooků v základním souboru shoduje s vysloveným předpokladem.

K testování nulové hypotézy H_0: π = 15 % (podíl notebooku s herní grafikou v základním souboru je 15 %), oproti oboustranné alternativě H_1: $\pi \neq$ 15 % použijeme jednovýběrový test o populačního podílu.

Program SPSS nabízí k testování hypotézy o populačním podílu různé testy. Všechny mají stejné výchozí údaje, ale liší se v metodě výpočtu ***p***-hodnoty a testové statistiky. My budeme používat **test Score**, který poskytuje spolehlivé výsledky pro dostatečně velké rozsahy výběrového souboru (v našem případě ***n*** = 150) a vychází z normálního rozdělení, které aproximuje binomické rozdělení**.**

SPSS

Analyze → Compare Means and Proportions → One-Sample Proportions → do okna Test Variable(s) vložíme proměnnou *Grafika,* v okně Define Success vybereme Value(s) a napíšeme zde kód herní grafické karty v našem případě hodnotu 3 → Tests zvolíme test Score a do okna Test Value napíšeme referenční hodnotu 0,15 → Continue → OK

One-Sample Proportions Tests

		Observed			Observed - Test Value[a]	Asymptotic Standard Error		Significance	
	Test Type	Successes	Trials	Proportion			Z	One-Sided p	Two-Sided p
Grafická karta = Herní	Score	19	150	,127	-,023	,027	-,800	,212	**,424**

a. Test Value = ,15

Test Value = 0,15 → předpokládaná hodnota podílu.

Successes = 19 → počet jednotek, které mají sledovanou vlastnost (19 notebooků má herní grafickou kartou).

Trials = 150→ celkový počet pozorování ve vzorku (n = 150)

Observed Proportion = 0,127 → podíl NB s herní grafickou kartou na celkovém rozsahu.

Observed-Test Value = -0,023 → rozdíl mezi pozorovaným podílem a předpokladem.

Asymptotic Standard Error = -0,027 → směrodatná chyba odhadu podílu.

Z = -0,800 → vypočtená hodnota testové statistiky.

Significance One-Sided p = 0,212 → p-hodnota pro jednostranný test.

Significance Two-Sided p = 0,424 → p-hodnota pro oboustranný test. Protože je tato hodnota vyšší než 0,05 (naše hladina významnosti), **nezamítáme nulovou hypotézu**.

Na základě provedeného testů se nepodařilo zamítnout nulovou hypotézu, že podíl notebooků s herní grafikou v populaci činí 15 %.

10.2 Dvouvýběrové parametrické testy

V mnoha praktických situacích nás zajímá nejen to, zda se číselné charakteristiky jedné populace liší od určité pevné hodnoty, ale především to, zda se **dvě různé skupiny** liší navzájem ve vztahu k nějaké kvantitativní charakteristice, jako je například průměr, nebo rozptyl. A právě pro zodpovězení takovýchto otázek jsou určeny **dvouvýběrové parametrické testy**.

10.2.1 Testy shody dvou populačních rozptylu

Testy o rovnosti dvou rozptylů slouží k ověření toho, zda dva výběrové soubory pocházejí z rozdělení se stejným rozptylem. V podstatě ověřují, zda oba soubory vykazují přibližně stejný rozptyl sledované náhodné veličiny.

Při testování shody dvou rozptylů, které jsou odhadnuty z nezávislých náhodných výběrů, se používá **Fischerovo-Snedecorovo rozdělení**. Toto rozdělení je **asymetrické**, což znamená, že p-hodnota se nepočítá symetricky na obě strany od středu, ale jako součet ploch, které jsou stejně extrémní jako pozorovaná hodnota.

Nejpoužívanějším testem určeným pro testování shody rozptylů je **Fisherův F-test**. Test je založený na porovnání většího a menšího z obou odhadů rozptylů. Testuje se, zda je statisticky významný rozdíl v rozptylech, nikoli zda je rozptyl první skupiny větší než druhé skupiny.

H_0	H_1	p-hodnota	$T(X)$
$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 < \sigma_2^2$	$p = P(T \leq x_F) = F_0(x_F)$	$F = \frac{s_1^2}{s_2^2} \sim F_{(n_1-1;n_2-1)}$ (10.5)
	$\sigma_1^2 > \sigma_2^2$	$p = P(T \geq x_F) = 1 - F_0(x_F)$	

Kde: $\sigma_1^2,\ \sigma_2^2$ … populační rozptyly,
s_1^2, s_2^2 … výběrové rozptyly,
n_1, n_2 … rozsahy náhodných výběrových souboru,
F_α … Fischer-Snedecorovo rozdělení pro $(n_1 - 1; n_2 - 1)$ stupňů volnosti.

Kromě Fisherova F-testu se v praxi často používá **Leveneův test**, který je na rozdíl od F-testu méně citlivý na porušení předpokladu normality dat. Z tohoto důvodu je Leveneův test považován za robustnější a je často standardním výstupem v mnoha statistických programech. Zásadní výhodou Leveneova testu oproti F-testu je jeho univerzálnost: zatímco F-test je omezen pouze na porovnání dvou výběrů, Leveneův test umožňuje testovat shodu rozptylů nejen u dvou, ale i u více nezávislých výběrových souborů. Pro shodu rozptylu používáme termín **homoskedasticita**, rozdílnost rozptylu označujeme jako **heteroskedasticitu**.

Testy hypotéz o shodě dvou rozptylů se také nazývají **testy homogenity** a využívají se především jako první krok před testy shody dvou nezávislých průměrů.

10.2.2 Testy shody dvou nezávislých populačních průměru

Testy o shodě dvou nezávislých populačních průměru nám umožňují na základě nezávislých náhodných výběru porovnat střední hodnoty u dvou základních souborů (populací). Formulace nulové hypotézy vychází z rovnosti mezi průměry jednotlivých nezávislých výběrů.

Nezávislé výběry jsou takové výběry, kde se zjištěné údaje netýkají stejných prvků (spotřeba potravin u domácnosti v různých okresech, výše příjmů mužů a žen atd.).

H_0	H_1	p-hodnota
$\mu_1 = \mu_2$	$\mu_1 < \mu_2$	$p = P(\mathrm{T} \leq x_t) = F_0(x_t)$
	$\mu_1 > \mu_2$	$p = P(\mathrm{T} \geq x_t) = 1 - F_0(x_t)$
	$\mu_1 \neq \mu_2$	$2 \cdot min\{F_0(x_t); 1 - F_0(x_t)\}$

Kde: μ_1, μ_2 … populační průměry.

Při srovnávání průměrů dvou nezávislých skupin se nejčastěji používá **dvouvýběrový *t*-test**. Ten má však jeden klíčový předpoklad **shodu rozptylů** v obou populacích.

V případě, nesplnění předpokladu shody rozptylů (ověření pomocí výše uvedeného F-testu nebo Levenova testu) používáme pro porovnání středních hodnot dvou normálních populací **Aspin-Welchův test**, který je robustní alternativa ke Studentovu dvouvýběrovému t-testu v případě **různých rozptylů** (heterogenity rozptylů).

	Test	$T(X)$
Shoda rozptylu	**Dvouvýběrový *t*-test**	$t = \dfrac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\dfrac{s_1^2(n_1-1) + s_2^2(n_2-1)}{n_1+n_2-2}}} \sqrt{\dfrac{n_1 n_2}{n_1+n_2}} \sim t_{\alpha(n_1+n_2-2)}$ (10.6)
Různé rozptyly	**Aspin-Welchův test**	$t = \dfrac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_{\alpha(\nu)}$, kde $\nu \cong \dfrac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2 \frac{1}{n_1-1} + \left(\frac{s_2^2}{n_2}\right)^2 \frac{1}{n_2-1}}$ (10.7)

Kde: $\bar{x}_1, \bar{x}_2$ … nestranné výběrové průměry,

s_1^2, s_2^2 … výběrové rozptyly,

n_1, n_2 … rozsahy náhodných výběrových souboru,

t_α … Studento *t*-rozdělení pro $(n_{1+}n_2 - 2)$ stupních volnosti,

$t_{\alpha(\nu)}$ … Studento *t*-rozdělení pro (ν) stupňů volnosti.

Příklad 10.3

Byl vysloven předpoklad, že průměrná cena notebooků s hliníkovým rámem je vyšší než průměrná cena notebooků s plastovým rámem. Pomocí údajů z datového souboru *„Notebook"* ověřte na 5% hladině významnosti, zda existuje statisticky významný rozdíl mezi průměrnými cenami obou skupin ve prospěch hliníkového rámu.

Řešení

Postup testování hypotézy o shodě dvou průměrů si rozdělíme do dvou kroků.

V prvním kroku budeme testovat nulovou hypotézu o shodě rozptylů $H_0 : \sigma_1^2 = \sigma_2^2$. Na základě výsledků se rozhodneme, jaký test použít pro testování dvou populačních průměrů.

V druhém kroku budeme testovat nulovou hypotézu o shodě dvou populačních průměrů $H_0: \mu_1 = \mu_2$ oproti jednostranné alternativě $H_1: \mu_1 < \mu_2$.

SPSS

Analyze → Compare Means and Proportions → Independent-Samples T Test → do okna Test Variable(s) vložíme proměnnou *Cena,* do okna Grouping Variable vložíme proměnnou *Materiál* → Define Groups zde zkontrolujeme kódy proměnných → Continue → OK

Group Statistics

	Materiál z jakého je vyrobený rám NB	N	Mean	Std. Deviation	Std. Error Mean
Cena notebooku (Kč)	plast	96	20645,65	10649,196	1086,879
	hliník	54	22997,85	4133,719	562,528

Deskriptivní statistiky (Group Statistics) → z tohoto výstupu vidíme, že průměrná cena notebooků s **hliníkovým rámem** (22 997,85 Kč) je vyšší než průměrná cena notebooků s **plastovým rámem** (20 645,65 Kč). Zde ale nevíme, jestli je tento rozdíl statisticky významný, nebo je jen náhodný.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
						Significance				95% Confidence Interval of the Difference
		F	Sig.	t	df	One-Sided p	Two-Sided p	Mean Difference	Std. Error Difference	Lower	Upper
Cena notebooku (Kč)	Equal variances assumed	**89,680**	**<,001**	-1,557	148	,061	,122	-2352,206	1511,082	-5338,289	633,877
	Equal variances not assumed			**-1,922**	135,309	**,028**	,057	-2352,206	1223,823	-4772,502	68,090

Testování hypotéz (Independent Samples Test)

Test shody rozptylů **(Leveneův test)** potvrdil, že **rozptyly cen** u obou typů materiálů **nejsou shodné** ($p < 0{,}001$).

Pokračování příkladu 10.3

Z testu shody průměrů (**Aspin-Welchův test**) vyplývá, že $p = 0{,}028$. Protože je tato hodnota menší než 0,05 (naše hladina významnosti), **zamítáme nulovou hypotézu**.

Na 5% hladině významnosti bylo prokázáno, že průměrná cena notebooků s hliníkovým rámem je statisticky významně vyšší než průměrná cena notebooků s plastovým rámem. Zjištěný rozdíl v ceně tedy není náhodný.

POZOR! *V případě, že by byla prokázána* ***shoda rozptylů*** *(homogenita) je nutné provádět testování shody průměrů pomocí* ***dvouvýběrového t-testu.***

10.2.3 Test o shodě dvou závislých populačních průměru

V předchozích případech jsme se zabývali porovnáváním dvou nezávislých souborů. Nyní se zaměříme na situace, kdy chceme testovat střední hodnoty **dvou závislých výběrů**. To znamená, že srovnáváme vlivy dvou různých faktorů na jednom měřeném objektu. Vycházíme přitom z předpokladu, že jeden náhodný výběr obsahuje znaky opakovaně měřené na stejných statistických jednotkách (tzv. párová měření), a proto se data řídí dvourozměrným rozdělením.

Při hodnocení rozdílů mezi porovnávanými výběry používáme **párový *t*-test,** který je ekvivalentní jednovýběrovému t-testu pro výběrový soubor diferencí.

Na základě párových výběrů získáváme dvojice hodnot x_i ($i = 1, 2,.. , n$), y_i ($i = 1, 2,.. , n$). Pro každou dvojici pozorování (x_i, y_i vypočteme diferenci $d_i = x_i - y_i$ ($i = 1,.., n$). Získaný soubor diferencí d_1, d_1, …, d_n považujeme za náhodný výběr o rozsahu n z normálně rozděleného základního souboru.

H_0	H_1	p-hodnota	$T(X)$
$\mu_d = 0$	$\mu_d < 0$	$p = P(T \leq x_t) = F_0(x_U)$	$t = \frac{\bar{d}}{s_d}\sqrt{n} \sim t_{\alpha(n-1)}$ (10.8)
	$\mu_d > 0$	$p = P(T \geq x_t) = 1 - F_0(x_t)$	
	$\mu_d \neq 0$	$2 \cdot min\{F_0(x_t); 1 - F_0(x_t)\}$	

Kde: μ_d … populační průměr diferencí,

$\bar{d}$ … nestranný průměr diferencí,

s_d … výběrová směrodatná odchylka diferencí,

n … rozsah náhodného výběrového souboru,

t_α … Studento t-rozdělení pro $n - 1$ stupňů volnosti.

Příklad 10.4

Bylo vysloveno tvrzení, že průměrná prodejní cena notebooků se liší od doporučené ceny výrobcem. Pomocí dat z výběrového souboru „*Notebook"*, který obsahuje dvojice cen pro stejné modely, ověřte na 5% hladině významnosti, zda existuje statisticky významný rozdíl v cenách.

Řešení

K testování nulové hypotézy H_0: $\mu_d = 0$ oproti oboustranné alternativě H_1: $\mu_d \neq 0$ použijeme **párový *t*-test.**

SPSS

Analyze → Compare Means and Proportions → Paired-Samples T Test → do okna Paired Variable(s) pod Variable 1 vložíme proměnnou *Cena,* pod Variable 2 vložíme proměnnou *Doporučená cena* → OK

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Cena notebooku (Kč) v prodejně	21492,44	150	8925,633	728,775
	Doporučená cena výrobcem	21572,67	150	8911,962	727,659

Z prvního výstupu vidíme, že průměrná cena notebooku v prodejně (21 492,44 Kč) je mírně nižší než doporučená cena výrobcem (21 572,67 Kč). Tento rozdíl činí přibližně 80 Kč. Stejně jako u jiných deskriptivních statistik, ani zde nevíme, jestli je tento rozdíl statisticky významný.

Paired Samples Test

		Paired Differences							Significance	
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper	t	df	One-Sided p	Two-Sided p
Pair 1	Cena notebooku (Kč) v prodejně - Doporučená cena výrobcem	-80,227	143,232	11,695	-103,336	-57,117	**-6,860**	149	<,001	**<,001**

Z výsledků párového t-testu vyplývá, že testová statistika má hodnotu -6,860 a p-hodnota (pro oboustranný test) je < 0,001.

Protože je p-hodnota menší než standardní hladina významnosti 0,05, zamítáme nulovou hypotézu. Na 5% hladině významnosti existují dostatečné statistické důkazy, které potvrzují, že průměrná prodejní cena notebooků se statisticky významně liší od doporučené ceny výrobcem.

10.2.4 Test o populačních podílech

Test o populačních podílech, je zaměřený na testování rozdílů mezi dvěma populacemi s alternativním rozdělením s parametry π_1 a π_2. Za předpokladu dostatečně velkých rozsahů výběrových souboru n_1 a n_2 ($n_1 > 100$, $n_2 > 100$) je možné testovou statistiku $T(X)$ aproximovat normovaným normálním rozdělením.

H_0	H_1	p-hodnota
$\pi_1 = \pi_2$	$\pi_1 < \pi_2$	$p = P(T \leq x_u) = F_0(x_u)$
	$\pi_1 > \pi_2$	$p = P(T \geq x_u) = 1 - F_0(x_u)$
	$\pi_1 \neq \pi_2$	$2 \cdot min\{F_0(x_u); 1 - F_0(x_u)\}$
$T(X)$		
$U = \frac{p_1 - p_2}{\sqrt{\frac{m_1 + m_2}{n_1 + n_2}\left(1 - \frac{m_1 + m_2}{n_1 + n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim u_\alpha$ (10.9)		

Kde: π_1, π_2 … populační podíly (relativní četnosti),

p_1, p_2 … výběrové podíly (relativní četnosti),

n_1, n_2 … rozsahy výběrových souborů,

m_1, m_2 … rozsahy souborů, které jsou nositelem sledovaného znaku,

u_α … kritická hodnota normovaného normálního rozdělení.

Pro menší rozsahy výběru, nebo v případě, že nejsou splněny podmínky pro aproximaci normálním rozdělením, se stejně jako u jednovýběrových testů o parametru π používají testy založené na binomickém rozdělení.

Příklad 10.5

Výzkum trhu s notebooky si klade za cíl ověřit zajímavý předpoklad. Zajímá nás, zda podíl notebooků, které nemají numerickou klávesnici, je odlišný v závislosti na materiálu, ze kterého jsou vyrobeny. S použitím dat z náhodného výběrového souboru *„Notebook"* ověřte na 5% hladině významnosti, zda je tento předpoklad správný.

Řešení

Testujeme nulovou hypotézu o shodě dvou populačních podílů $H_0: \pi_1 = \pi_2$ oproti oboustranné alternativě $H_1: \pi_1 \neq \pi_2$.

Program SPSS nabízí k testování hypotézy o populačním podílu různé testy. Všechny mají stejné výchozí údaje, ale liší se v metodě výpočtu p-hodnoty a testové statistiky. My budeme

Pokračování příkladu 10.5

používat **Waldův test**, který je při velkých výběrech založen na asymptotickém rozdělení testové statistiky, které se blíží normálnímu rozdělení.

SPSS

Analyze → Compare Means and Proportions → Independent-Samples Proportions → do okna Test Variable(s) vložíme proměnnou *Klavesnice,* v okně Define Success vybereme Value(s) a napíšeme zde kód pro hodnotu *ne* v našem případě hodnotu 2 → do okna Grouping Variable vložíme proměnnou *Materiál,* v okně Define Success vybereme Value(s) a napíšeme zde kódy pro Group 1: 1 pro Group 2: 2; → Tests zvolíme test Wald → Continue → OK

Independent-Samples Proportions Group Statistics

	Materiál z jakého je vyrobený rám NB	Successes	Trials	Proportion	Asymptotic Standard Error
Numerická klávesnice = ne	= plast	50	96	,521	,051
	= hliník	32	54	,593	,067

Z prvního výstupu vidíme, že z celkového počtu 150 notebooků je 96 plastových a 54 hliníkových. Z 96 plastových notebooků bylo zjištěno, že jich 50 (52,1 %) nemá numerickou klávesnici. U hliníkových modelů byl podíl notebooků bez numerické klávesnice vyšší a dosáhl hodnoty 59,3 % (32 z 54).

Independent-Samples Proportions Tests

	Test Type	Difference in Proportions	Asymptotic Standard Error	Z	Significance One-Sided p	Significance Two-Sided p
Numerická klávesnice = ne	Wald	-,072	,084	**-,853**	,197	**,393**

Druhý výstup znázorňuje výsledky dvouvýběrového testu. Hodnota testové statistiky je -0,853 a p-hodnota pro oboustrannou alternativu, kterou porovnáváme s hladinou významnosti alfa, je 0,393.

Z těchto výsledků plyne, že na 5% hladině významnosti nelze zamítnout nulovou hypotézu. To znamená, že neexistuje statisticky významný rozdíl v podílu notebooků bez numerické klávesnice mezi plastovými a hliníkovými modely.

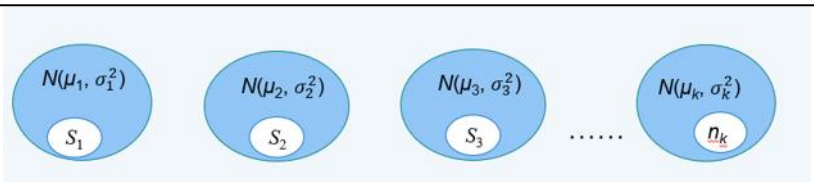
10.3 Vícevýběrové parametrické testy

Dosud jsme se věnovali porovnávání jednoho nebo dvou náhodných výběrů. V praxi se však často setkáváme se situacemi, kde je potřeba porovnávat tři nebo více výběrů najednou. Právě k tomu slouží **vícevýběrové parametrické testy.**

Jejich hlavním cílem je zjistit, zda se parametry mezi více než dvěma náhodnými výběry statisticky významně liší, nebo zda pocházejí ze stejné populace. Tyto testy jsou důležité,

protože nám umožňují komplexně analyzovat vliv jedné kategoriální proměnné s více než dvěma úrovněmi na kvantitativní proměnnou (např. vliv velikosti úhlopříčky na cenu notebooků). To vše bez nutnosti provádět sérii opakovaných testů, což by zvyšovalo riziko chybného závěru a také by došlo k zvýšení chyby 1. druhu nad stanovenou mez (obr. 10.3).

Obrázek: 10.1 Proč nelze používat opakované dvouvýběrové *t* - testy?

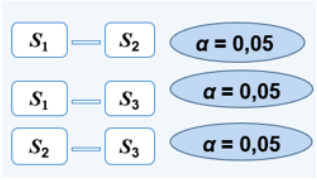


Náhodný výběr 1 (S_1), náhodný výběr 2 (S_2), náhodný výběr 3 (S_3), …….

Porovnáním dvojici souborů pomocí *t*-testu.

Pravděpodobnost nesprávného zamítnutí H_0 je pro každý test 5 %.

Pravděpodobnost, že **nedojde k chybě 1. druhu** je pak 0,95 · 0,95 · 0,95 = **0,857**



Pravděpodobnost, že dojde alespoň k chybě 1. druhu je 1 – 0,857 = 0,143!

10.3.1 Test o shodě více jak dvou populačních rozptylů

Testy k ověřování hypotéz o rovnosti více než dvou rozptylů (homoskedasticity) jsou zobecněním dvouvýběrového testu o shodě rozptylů. Tyto testy se zaměřují na analýzu variability dat uvnitř jednotlivých skupin a tvoří nedílnou součást předběžné analýzy dat před prováděním složitějších parametrických testů, jako je například analýza rozptylu (ANOVA).

Nejčastěji používaným testem pro ověření homogenity rozptylů je Leveneův test. Jak již bylo uvedeno, jeho velkou výhodou je robustnost vůči odchylkám od normality dat, což umožňuje jeho použití i při mírném porušení normality. Testová statistika Leveneova testu je založena na analýze rozptylu absolutních hodnot odchylek od skupinového mediánu.

H_0	H_1	*p*-hodnota	$T(X)$
$\sigma_1^2 = \sigma_2^2 = \ ... = \sigma_k^2$	neplatí H_0	$p = P(T \leq x_F) = F_0(x_F)$ $p = P(T \geq x_F) = 1 - F_0(x_F)$	$F_{Levene} \sim F_{(k-1;n-k)}$

Pokud výsledek testu prokáže, že rozptyly nejsou shodné (tedy dojde k heteroskedasticitě), je při dalším testování (například průměrů) nutné zvolit alternativní postup. Můžeme použít

buď testy neparametrické, nebo provést úpravu dat například pomocí jejich transformace. Další možností je použít parametrické testy s korekcí pro nerovné rozptyly (např. Welchův F-test pro ANOVA).

10.3.2 Test o shodě více jak dvou populačních průměrů

Analýza rozptylu (ANOVA) slouží k porovnání **průměrů tří a více nezávislých náhodných výběrů.** Cílem je zjistit, zda existuje statisticky významný rozdíl mezi průměry těchto skupin.

ANOVA test zkoumá, zda je variabilita mezi výběry (tj. rozdíly mezi průměry výběrů) větší než variabilita uvnitř jednotlivých výběrů (rozdíly mezi daty v rámci jednoho výběru). Pokud je rozdíl mezi výběry velký, je pravděpodobné, že průměry nejsou stejné.

Analýza rozptylu je statistická metoda, které nám umožňují provádět vícenásobné porovnávání středních hodnot. ANOVA byla původně vypracována pro vyhodnocení výsledků polních pokusů v zemědělství. V současnosti představuje velmi obecný statistický postup, jehož se využívá v přírodních i společenských vědách, a také při zpracování kvantitativních výsledků experimentů.

Správné použití analýzy rozptylu je vázáno na splnění následujících předpokladů:

- Normalita rozdělení náhodných veličin – obecný předpoklad všech parametrických testů. Ověřování normality se provádí pomocí testů normality a grafického znázornění dat.
- Statistická nezávislost náhodných chyb – obecný předpoklad všech nezávislých parametrických testů. Hodnoty jednoho výběru nejsou ovlivňovány hodnotami jiného výběrů (např. výška studentů na jedné škole nemá žádný vliv na výšku studentů na druhé škole – výběry jsou nezávislé, protože jsou prováděny v oddělených populacích). Ověřování normality je možné provádět například pomocí testů rezidui.
- Homogenita rozptylů neboli homoskedasticita – variabilita dat je konzistentní. Rozptyly náhodných výběrů jsou přibližně stejné. Tento předpoklad se ověřuje testy homogenity.

ANOVA je poměrně robustní statistická metody vhledem k částečnému porušeni předpokladů normality či homoskedasticity, zejména v případě většího rozsahu sledovaných dat. Normalitu potažmo narušenou homoskedasticitu, lze řešit například pomocí logaritmické transformace dat. Zvýšenou pozornost je však třeba věnovat nesplnění předpokladu statistické nezávislosti náhodných chyb, které může vést často k chybným závěrům. Z uvedeného vyplývá, že pokud dojde k „mírnému“ porušení předpokladu je možné ANOVU použít. Obecně platí, že čím je rozsah výběru větší, tím je možné očekávat vyšší robustnost vůči nesplnění podmínek.

Podle počtu sledovaných faktorů rozdělujeme analýzu rozptylu na:

Analýzu rozptylu jednoduchého třídění – posuzujeme vliv působení jednoho faktoru (nejčastěji používaná).

Analýzu rozptylu dvojného třídění – posuzujeme vliv působení dvou faktorů na sledovanou charakteristiku (parametr) výběrového souboru.

Vícefaktorové modely analýzy rozptylu – posuzuje působení více než dvou faktorů (obtížné sestavení vhodného modelu a interpretace výsledků).

Podle počtu pozorování na:

Nevyvážený model – neortogonální model (počty pozorování v každém z náhodných výběrů jsou různé),

Vyvážený model – ortogonální model (počty pozorování v každém z náhodných výběrů jsou shodné).

POZOR! *Analýza rozptylu se nezabývá přímo porovnáváním rozptylů, ale pomocí rozkladu celkové variability dat usuzuje o rozdílech mezi průměry.*

JEDNOFAKTOROVÁ ANALÝZA ROZPTYLU

Analýza rozptylu při jednoduchém třídění předpokládá, že na k-náhodných výběrů, které jsou na sobě nezávislé působí **jeden faktor**, který sledujeme na několika jeho úrovních (skupinách). Úrovně faktoru:

- určité množství kvantitativního faktoru,
- určitá variantu kvalitativního faktoru.

Obrázek 10.2: Schéma jednofaktorové analýzy rozptylu

		TŘÍDY = pozorované hodnoty						
		1	2	…	j	…	rozsah	průměr
FAKTOR	1	x_{11}	x_{12}		x_{1j}	x_{1n}	n_1	$\overline{x_{1.}}$
	2	x_{21}					n_2	$\overline{x_{2.}}$
	.							
	i	x_{i1}			x_{ij}		n_i	$\overline{x_{i.}}$
	.							
	k	x_{k1}	x_{k2}		x_{kj}	x_{kn}	n_k	$\overline{x_{k.}}$

i-tá úroveň sledovaného faktoru
j-té opakování měření proměnné x

počet pozorování v i-tém náhodném výběru $\sum_{i=1}^{k} n_i = n$

Naměřené hodnoty uspořádáme podle jednoho třídícího kritéria (obr. 10.2), tzn. podle úrovní sledovaného faktoru (k), do tolika tříd (n_i), na kolika úrovních (skupinách) tento faktor sledujeme.

Podstatou ANOVY je, rozklad celkové variability do dvou částí, na variabilitu mezi skupinami a variabilitu uvnitř skupin.

$$\boldsymbol{S = S_1 + S_r}$$

S . . **celková variabilita** (celkový součet čtverců) variability celého pokusu.

S_1 . . **variabilita mezi skupinami** část variability, charakterizuje **vliv faktoru.**

S_r . . **variabilita uvnitř skupin** část variability, charakterizuje **působení náhodných vlivů, případně jiných neuvažovaných faktorů.**

Zdroj variability	**Součet čtverců odchylek**	**Stupně volnosti**	**Výběrový rozptyl**
Mezi skupinami	$\boldsymbol{S_1 = \sum_{i=1}^{k} n_i (x_i - \bar{x})^2}$	$\boldsymbol{k - 1}$	$\boldsymbol{s_1^2 = \frac{S_1}{k-1}}$
Uvnitř skupin	$\boldsymbol{S_r = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}$	$\boldsymbol{\sum_{i=1}^{k} n_i - k}$	$\boldsymbol{s_r^2 = \frac{S_r}{\sum_{i=1}^{k} n_i - k}}$
Celková	$\boldsymbol{S = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2}$	$\boldsymbol{\sum_{i=1}^{k} n_i - 1}$	

Kde: $\mu_1, \mu_2, .., \mu_k$ … neznámé populační průměry,

$\bar{x}_1, \bar{x}_2, .., \bar{x}_k$ … výběrových průměrů,

s_1^2 … výběrový rozptyl mezi skupinami,

s_r^2 … výběrový rozptyl uvnitř tříd (reziduální rozptyl),

n … celkový rozsah náhodných výběrů,

k … počet náhodných výběrů,

F_α … Fischer-Snedecorovo rozdělení pro $(k-1; n-k)$ stupňů volnosti.

Na základě k-nezávislých náhodných výběrů testujeme nulovou hypotézu, která uvádí, že mezi skupinovými průměry nejsou žádné rozdíly. Testování nulové hypotézy je založeno na zjišťování rozdílů průměru mezi více nezávislými výběry pomocí **testovacího kritéria *F*.**

$\boldsymbol{H_0}$	$\boldsymbol{H_1}$	***p*-hodnota**	$\boldsymbol{T(X)}$
$\boldsymbol{\mu_1 = \mu_2 = .. = \mu_k}$	**neplatí** $\boldsymbol{H_0}$	$p = P(T \geq x_F) = 1 - F_0(x_F)$	$\boldsymbol{F = \frac{s_1^2}{s_r^2} \sim F_{(k-1;n-k)}}$ (9.10)

V případě, že na zvolené hladině významnosti α **zamítneme nulovou hypotézu o shodě průměrů, činíme závěr, že alespoň jeden průměr se významně odlišuje od ostatních**. To znamená, že alespoň dva výběrové soubory nepocházejí ze stejného základního souboru (populace). V takovém případě je potřeba provést podrobnější vyhodnocení výsledků a zjistit, mezi kterými výběry je statisticky významný rozdíl. K podrobnějšímu vyhodnocení výsledků slouží metody mnohonásobného porovnávaní.

POZOR! *Název metody „Analýza rozptylu“ může být zavádějící, neboť by mohl naznačovat, že se zaměřuje na srovnávání rozptylů mezi jednotlivými výběry. Hlavním cílem této metody je však srovnání středních hodnot (průměrů). Název vychází ze skutečnosti, že testová statistika vychází z porovnání rozptylu mezi skupinami a rozptylu uvnitř skupin.*

METODY MNOHONÁSOBNÝCH POROVNÁNÍ (POST HOC ANALÝZA)

Tyto metody jsou založeny na testování vzájemných rozdílů mezi skupinovými průměry. Metody mnohonásobného porovnávání jsou statistické testy, kterými porovnáváme vzájemné rozdíly mezi skupinovými středními hodnotami a posuzujeme jejich statistickou významnost těchto rozdílů. Cílem metod je zjistit, mezi kterými konkrétními dvojicemi výběrů se průměry statisticky významně liší.

Jejich princip vychází z testů shody dvou středních hodnot.

$$\boldsymbol{H_0}: \boldsymbol{\mu_{1\cdot} = \mu_{2\cdot}}$$

proti oboustranné alternativě

$$\boldsymbol{H_1}: \boldsymbol{\mu_{1\cdot} \neq \mu_{2\cdot}}$$

Kde: $\mu_{1\cdot}, \mu_{2\cdot}, .., \mu_{k\cdot}$ … neznámé řádkové populační průměry.

Mezi nejčastěji používané metody patří:

Tukeyho metoda (*T*-metoda)

T-metoda se používá v případě, kdy pracujeme s vyváženým modelem, to znamená, že máme stejný počet pozorování v každém z výběrů. Je určena speciálně pro **párová srovnávání** všech kombinací průměrů.

Tato metoda je *konzervativní*, tzn. že vyžaduje větší rozdíl mezi průměry, aby byl označen za statisticky významný. Má **nižší pravděpodobnost chyby 1. druhu** (tzn. zamítnutí nulovou hypotézu, ačkoli je platná). Tukeyho metody se snaží **minimalizovat riziko falešně**

pozitivního výsledku, což ji činí spolehlivou, ale na druhou stranu může přehlédnout menší, ale reálné rozdíly mezi průměry.

Scheffeho metoda (*S*-metoda)

Scheffého metoda mnohonásobného porovnávání se také nazývá testem násobných kontrastů. Je nejvšestrannější a umožňuje testovat nejen párová srovnání, ale i složitější kombinace průměrů.

S-metoda je **velmi přísná** a vyžaduje **velký rozdíl mezi průměry**, aby ho označila za statisticky významný. Díky této přísnosti je **účinná v ochraně proti chybě 1. druhu** (falešně pozitivní závěr).

Tukeyho metoda je spolehlivější pro párová srovnání, kdežto Scheffého metoda je bezpečnější a všestrannější pro komplexnější srovnání. V praxi se Tukeyho metoda používá častěji, protože většina analýz se zaměřuje právě na párové srovnání.

Příklad 10.6

Na trhu s výpočetní technikou se často spekuluje o vlivu jednotlivých komponent na celkovou cenu zařízení. Jako manažer produktu chcete na základě dat z 150 náhodně vybraných notebooků (datová matice „*Notebook*“) ověřit (na 5% hladině významnosti), zda existuje statisticky významný rozdíl v **průměrné ceně notebooku** v závislosti na **typu displeje**. *U **parametrických testů**, a to včetně jednofaktorové analýzy rozptylu (ANOVA), nesmíme nikdy zapomenout na ověření předpokladů. Zejména je klíčové **ověření normality dat** v každé testované skupině. Bez splnění tohoto předpokladu by mohly být výsledky testu zkreslené a neplatné.*

Řešení

Cílem je ověřit, zda existuje statisticky významný rozdíl v průměrné hodnotě kvantitativní proměnné (*Cena*) v závislosti na kategoriální proměnné (faktoru – *Displeje*) s několika úrovněmi (Matný, Lesklý, Antireflexní).

K testování nulové hypotézy o populačních průměrech H_0: $\mu_1 = \mu_2 = \mu_3$ oproti oboustranné alternativě H_1: neplatí H_0, použijeme vícevýběrový test o hodnotě průměru – **Analýzu rozptylu (ANOVA)**.

Nejdříve se však musíme zaměřit na ověření předpokladů této metody, a to především na posouzení normality a následně na ověření shody rozptylů.

Pokračování příkladu 10.6

Normalitu budeme testovat pomocí Shapiro-Wilkova testu.

SPSS Analyze → Descriptive Statistics → Explore → do okna Dependent List: proměnná *Cena;* do okna Factor List: proměnná *Displej* → Plots → Normality plots with tests → Continue → OK

Tests of Normality

		Kolmogorov-Smirnov			Shapiro-Wilk		
	Typ displeje	Statistic	df	Sig.	Statistic	df	Sig.
Cena notebooku (Kč) v prodejně	Matný	,073	55	,200	,963	55	**,087**
	Lesklý	,084	76	,200	,973	76	**,107**
	Antireflexní	,131	19	,200	,954	19	**,464**

Protože p-hodnoty (signifikance) Shapiro-Wilkova testu jsou ve všech třech skupinách (Matný, Lesklý, Antireflexní) větší než 0,05, nulovou hypotézu nemůžeme zamítnout, tzn. že data v každé z těchto skupin pocházejí z normálního rozložení.

Shodu rozptylů $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$ budeme testovat na základě Levenova testu homogenity.

SPSS Analyze → Compare Means and Proportions → One-Way ANOVA → do okna Dependent List vložíme proměnnou *Cena,* do okna Factor vložíme proměnnou *Displej* → Options označíme Homogeneity of variance test → Continue → OK

Tests of Homogeneity of Variances

		Levene Statistic	df1	df2	Sig.
Cena notebooku (Kč) v prodejně	Based on Mean	**1,427**	2	147	**,243**
	Based on Median	1,491	2	147	,229

Z Levenova test homogenity vyplynulo, že neexistuje statisticky významný rozdíl v rozptylech – **rozptyly jsou homogenní**. *P*-hodnota (p = 0,243) je vyšší než stanovená hladinu významnosti 0,05, a proto nezamítáme nulovou hypotézu.

Dále pokračujeme interpretací výstupu analýzy rozptylů.

ANOVA

Cena notebooku (Kč) v prodejně

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	316536746,07	2	158268373,03	**2,014**	**,137**
Within Groups	11553834531	147	78597513,816		
Total	11870371277	149			

Závěry z analýzy rozptylu (ANOVA) ukazují, že hodnota F-statistiky je 2,014 a p-hodnota (p = 0,137) je větší než zvolená hladina významnosti α = 0,05. To znamená, že neexistují statisticky významné rozdíly v průměrných cenách notebooků v závislosti na typu displeje. Zjištěné rozdíly v průměrných cenách jsou s největší pravděpodobností způsobeny náhodou.

10.4 Vztah mezi testováním hypotéz a intervalovými odhady parametrů

Spolehlivost testu $(1-\alpha)$, tedy pravděpodobnost, že nezamítneme nulovou hypotézu, pokud je skutečně platná, označuje zároveň pravděpodobnost, že parametr populace leží v příslušném intervalu spolehlivosti. Z toho plyne, že pokud testovaná hodnota parametru spadá do $(1-\alpha)$ intervalu spolehlivosti, nemůžeme příslušnou nulovou hypotézu zamítnout na hladině významnosti α. Interval spolehlivosti tak lze chápat jako množinu všech možných hodnot testovaných parametrů pro, které na dané hladině významnosti nemůžeme nulovou hypotézu zamítnout.

Příklad 10.7

Představte si, že firma vyrábí balíčky cukrovinek a tvrdí, že průměrná hmotnost balíčku je 100 gramů. Na 5 % hladině významnosti ověřujeme, zda toto tvrzení platí. K dispozici máme náhodný vzorek 30 balíčků. Průměrná hmotnost balíčku je 102 gramů.

Testujeme hypotézy

H_0: $\mu = \mu_0$ → průměrná hmotnost balíčku je 100 gramů.
H_1: $\mu \neq \mu_0$ → průměrná hmotnost balíčku se liší od 100 gramů.

Místo přímého testování hypotézy můžete vypočítat **95% interval spolehlivosti** pro průměrnou hmotnost. Předpokládejme, že výpočet intervalu spolehlivosti z náhodného výběru (s výběrovým průměrem $\bar{x}$ = 102 gramů) je v intervalu:

$$P(98{,}5\ \text{g} < \mu < 105{,}5\ \text{g}) = 0{,}95.$$

Naše testovaná hodnota z nulové hypotézy (100 gramů) spadá do vypočteného intervalu spolehlivosti (98,5 g až 105,5 g). Hodnota **100 gramů se nachází uvnitř** tohoto intervalu. Z tohoto důvodu **nulovou hypotézu nezamítnete**.
Tento výsledek znamená, že na 5% hladině významnosti neexistuje dostatek důkazů pro tvrzení, že by se průměrná hmotnost balíčků lišila od 100 gramů. Jinými slovy, rozdíl mezi 100 g (hypotéza) a 102 g (náš vzorek) není statisticky významný. Pokud by se interval spolehlivosti pohyboval například od 101 g do 106 g, hodnota 100 g by do něj nespadala a nulovou hypotézu bychom zamítli.

Shrnutí kapitoly

Parametrické testy slouží k ověřování hypotéz o populačních parametrech, jako jsou průměry, rozptyly a podíly. Jedná se silné testy.

Použití těchto testů je možné jen po **důsledném ověření jejich předpokladů** ještě před samotným testováním, jelikož porušení předpokladů může vést ke zkresleným a neplatným závěrům.

Jednovýběrové testy slouží k porovnání jednoho výběru s hypotetickou hodnotou populace. Na základě jednoho výběrového souboru rozhodujeme, zda neznámý populační parametr odpovídá určité předpokládané číselné hodnotě (předpokladu, normě).

Dvouvýběrové testy se používají k porovnání populačních parametrů dvou nezávislých nebo závislých výběrů.

Nezávislé výběry – každý výběr obsahuje znaky měřené na jiných statistických jednotkách.

Závislé výběry – jeden náhodný výběr obsahuje znaky opakovaně měřené na stejných statistických jednotkách → dvourozměrné rozdělení.

Vícevýběrové testy slouží k porovnávání více než dvou výběrů najednou, což je klíčové pro zamezení zvyšování chyby 1. druhu, ke které by docházelo při opakovaných párových srovnáních.

Literatura

CAMM, Jeffrey; James COCHRAN, David ANDERSON, Dennis SWEENEY, Thomas WILLIAMS et al. *Statistics for Business and Economics*. 15th Edition. Boston: Cengage Learning, 2023. ISBN 978-0357715857

HENDL, Jan. *Přehled statistických metod zpracování dat: analýza a metaanalýza dat*. Vyd. 1. Praha: Portál, 2004. ISBN 8071788201.

HINDLS Richard; Markéta ARLTOVÁ, Stanislava HRONOVÁ, Ivana MALÁ, Luboš MAREK, et al. Statistika v ekonomii. První vydání. Professional Publishing, 2018. ISBN 978-80-88260-09-7.

HOŠKOVÁ, Pavla; Andrea JINDROVÁ, Marie PRÁŠILOVÁ a Rudolf ZEIPELT. *Statistika I*. Praha: PEF ČZU, 2013. ISBN: 978-80-213-2341-4.

JOHNSON, Richard A. a Dean W. WICHERN, Applied multivariate statistical analysis. Sixth edition. Harlow: Pearson, 2014. ISBN 978-1-292-02494-3.

10 Kontrolní otázky

1. Co je hlavním předpokladem pro použití parametrických testů?
2. Jaké kroky je třeba provést pro ověření předpokladu normality dat?
3. Kdy se místo parametrických testů používají metody neparametrické?
4. Jaký je účel jednovýběrových parametrických testů?
5. Jaký test použijete pro ověření hypotézy o populačním rozptylu a jaké rozdělení se řídí jeho testová statistika?
6. Kdy se používá jednovýběrový t-test a jak se liší od jednovýběrového U-testu?
7. Co znamená p-hodnota a jakou roli hraje při rozhodování o zamítnutí nulové hypotézy?
8. Jaký je rozdíl mezi testem o populačním podílu pro malý a velký rozsah výběrového souboru?
9. Co je hlavním cílem dvouvýběrových parametrických testů?
10. Kdy se používá Fisherův F-test a v čem se liší od Leveneova testu?
11. Jaký je rozdíl mezi homoskedasticitou a heteroskedasticitou?
12. Kdy se pro srovnání dvou nezávislých průměrů použije Aspin-Welchův test a proč?
13. Vysvětlete rozdíl mezi nezávislými a závislými výběry a uveďte příklad testu pro závislé výběry.
14. V jakém případě se pro testování hypotézy o shodě dvou nezávislých průměrů použije dvouvýběrový t-test?
15. Jaký je hlavní rozdíl mezi vícevýběrovými a jednovýběrovými testy?
16. K čemu slouží metody mnohonásobného porovnávání?

10 Příklady k procvičení

Na základě datové matice *„Byty“*, kterou naleznete v kurzu v Moodle, odpovězte na následující otázky.

10.1 Jednovýběrové testy

10.1.1 Na základě náhodného výběru z datové matice *„Byty“* ověřte, zda se průměrná cena za metr čtvereční statisticky významně liší od referenční hodnoty 15 000 Kč. Použijte hladinu významnosti α = 0,05. [t = 1,471; p = 0,144]

10.1.2 Předpokládáte, že 65 % bytů v populaci má výtah. Ověřte na 5% hladině významnosti, zda podíl bytů s výtahem ve vašem výběrovém souboru statisticky významně odpovídá tomuto předpokladu. [Z = -0,942; p = 0,346]

10.2 Dvouvýběrové testy

10.2.1 Testujte na hladině významnosti α=0,05, zda existuje statisticky významný rozdíl v průměrné ceně za metr čtvereční mezi byty, které jsou po rekonstrukci, a novostavbou.

$$[F = 8{,}365\,; p = 0{,}005; t = -16{,}947; p < 0{,}001]$$

10.2.2 Ověřte, zda se podíl bytů s výtahem statisticky významně liší mezi novostavbami a byty po rekonstrukci. Pro testování použijte hladinu významnosti α=0,05.

$$[Z = 0{,}108\,; p = 0{,}914]$$

10.3. Vícevýběrové testy

10.3.1 Na 5 % hladině významnosti ověřte, zda existuje statisticky významný rozdíl v průměrné ceně bytu v závislosti na jeho stavu (novostavba, po rekonstrukci, původní).

$$[SW_R = 0{,}925\,; p = 0{,}005; SW_N = 0{,}972\,; p = 0{,}142; SW_P = 0{,}897\,; p = 0{,}002]$$

10.3.2 Zjistěte, zda má město, ve kterém se byt nachází, statisticky významný vliv na průměrnou cenu za metr čtvereční (α=0,05). V případě statisticky významného rozdílu proveďte post-hoc analýzu.

$$[SW_P = 0{,}942\,; p = 0{,}050; SW_B = 0{,}952\,; p = 0{,}024; SW_O = 0{,}981\,; p = 0{,}541]$$

$$[F_{Levene} = 2{,}938\,; p = 0{,}056; F = 6{,}521\,; p = 0{,}002;\ \mu_P \neq \mu_B]$$

11 NEPARAMETRICKÉ TESTY

Neparametrické testy jsou statistické metody, které se používají v případě, že nejsou splněny předpoklady pro použití parametrických testů. Tyto testy vycházejí z velmi obecných předpokladů a nezávisí na konkrétním tvaru rozdělení základního souboru.

Výhodou těchto testů je jejich použitelnost bez ohledu na typ rozdělení a také skutečnost, že jsou méně citlivé na extrémní hodnoty.

Nevýhodou neparametrických testů je jejich menší síla (nižší schopnost rozpoznat neplatnou nulovou hypotézu). Tento nedostatek je však kompenzován širšími možnostmi použití neparametrických testů a vhodnou volbou testu jej lze téměř eliminovat.

Neparametrické testy se používají především v těchto případech:

- Data nemají normální rozdělení.
- U malých výběrových souborů.
- Při analýze dat s extrémními hodnotami.

Neparametrické testy se týkají **neparametrických hypotéz**, které jsou zaměřeny na obecné vlastnosti populace (např. tvar rozdělení nebo závislost proměnných) bez nutnosti znát konkrétní parametry rozdělení. Často požadují pouze to, aby rozdělení náhodné veličiny bylo spojité.

Neparametrické testy vycházejí z různých principů výpočtů

Testy shody – testová statistika měří velikost rozdílu mezi pozorovanými a očekávanými četnostmi. Tyto testy jsou založeny na porovnávání pozorovaných četností s očekávanými četnostmi (četnosti, které by nastaly, kdyby platila např. nulová hypotéza o nezávislosti nebo rovnoměrném rozdělení). Patří sem ***chí*-kvadrát test dobré shody**, který se používá k ověření hypotézy, zda pozorované rozdělení četností odpovídá určitému teoretickému rozdělení. Dále pak ***chí*-kvadrát test nezávislosti**, který se používá k ověření toho, zda existuje závislost mezi dvěma kategorickými proměnnými.

Testy odchylek – testová statistika je založena na odchylkách hodnot od určité předem dané hodnoty. Patří zde například **znaménkový test,** který se používá pro párová data. Pro každé párové pozorování se zaznamená, zda je rozdíl kladný, záporný, nebo nulový. Testová statistika je založená na sledování a posouzení četnosti (počtu) kladných a záporných odchylek (znamének) a nezohledňuje velikost rozdílu, pouze jejich směr. Dále pak **Dixonův test extrémních odchylek**, který je určený k identifikaci odlehlých hodnot v malých souborech dat.

Posuzuje, zda je extrémní hodnota v datovém souboru pouze náhodná nebo, zda je zatížena hrubou chybou. Používá se tam, kde je potřeba rychle posoudit validitu několika málo opakovaných měření.

Permutačními testy – tyto metody jsou modernější a početně náročnější, ale nabízejí velkou flexibilitu a nevyžadují žádné předpoklady o rozdělení dat. Jsou založeny na opakovaném vzorkování neboli náhodném přeskupování naměřeného souboru dat. Cílem je posoudit variabilitu možných výsledků při opakovaném přeskupování (permutaci) hodnot. Principem permutačního testování je srovnání pozorované testové statistiky s testovými statistikami, které by bylo možno teoreticky získat ze stejného datového souboru, kdyby přiřazení jednotlivých pozorovaných hodnot do sledovaných skupin bylo náhodné. Z takto vytvořeného empirického rozdělení testové statistiky se odhaduje p-hodnota. Velmi robustní testy, nevyžadují náhodný výběr. Nemusíme znát žádné předpoklady o rozdělení v populaci. Testy jsou založeny na randomizaci získaných hodnot patří zde například **Fisherův exaktní test**.

Pořadové testy – jedná se o testy u kterých výpočet nevychází ze skutečných hodnot náhodné veličiny, ale s pořadových čísel (ordinální škály), které skutečné hodnoty nahrazují (nejmenší hodnota má pořadové číslo 1, největší pak n). Princip těchto testů tedy spočívá v nahrazení stávajících hodnot náhodné veličiny pořadovými čísly. Výhodou je, že převedeme-li hodnoty na pořadová čísla, odstraníme tím vliv odlehlých hodnot.

V dalším textu se budeme věnovat přehledu nejčastěji používaných neparametrických testů. Vzhledem k odlišným principům jejich výpočtu se zaměříme pouze na ty nejdůležitější a nejběžnější metody, aniž bychom uváděli kompletní výčet všech dostupných testů.

11.1 Testy shody rozdělení

Testy shody umožňují srovnání empirického (výběrového) rozdělení s jistým rozdělením teoretickým. Jejich význam spočívá zejména v tom, že umožňují potvrdit domněnku o rozdělení pravděpodobností náhodné veličiny, a tedy použít statistické metody, které jsou tímto rozdělením podmíněny.

11.1.1 *Chí*-kvadrát test dobré shody

***Chí*-kvadrát test dobré shody** se používá k ověření hypotézy, že náhodný výběr pochází z rozdělení určitého typu, jehož parametry jsou dány (např. normální rozdělení). Častěji se však

používá k testování, zda výběr pochází z populace, kde jsou četnosti jednotlivých variant rozděleny podle kvantitativního či kvalitativního znaku do k skupin (tříd), a podíly těchto variant v populaci jsou rovny hodnotám libovolného pravděpodobnostního rozdělení. Pro jednotlivé skupiny vypočteme teoretické (očekávané) četnosti n_{oj} odvozené za předpokladu platnosti nulové hypotézy. Na základě distribuční funkce normálního rozdělení a parametrů daného rozdělení stanovíme pravděpodobnost p_j, že hodnota náhodné veličiny X padne do j-tého intervalu.

$\boldsymbol{H_0}$	$\boldsymbol{H_1}$	$\boldsymbol{T(X)}$
$F(x) = F_0(x)$	$F(x) \neq F_0(x)$	$\boldsymbol{\chi^2 = \sum_{j=1}^{k} \frac{(n_j - n_{oj})^2}{n_{oj}} \sim \chi^2_{\alpha(k-c-1)}}$ (11.1) $\boldsymbol{n_{oj} = n \cdot p_j}$ (11.2)

Kde: $F(x)$ … distribuční funkce,

$F_0(x)$ … hypotetická distribuční funkce,

n … celkový rozsah souboru,

n_j … empirické (skutečné, pozorované) četnosti,

n_{oj} … teoretické (očekávané) četnosti,

p_j … pravděpodobnost, že hodnota veličiny X bude v j-tém intervalu,

χ^2_α … kritická hodnota *chí*-kvadrát rozdělení pro $(k - c - 1)$ stupnů volnosti, kde k je počet tříd výběrového souboru a c je počet parametrů, které neznáme, a které odhadujeme z výběrového souboru.

Kritický obor je vymezen nerovností $\chi^2 > \chi^2_{\alpha(k-c-1)}$. Z tohoto vztahu vyplývá zamítnutí nulové hypotézy H_0 (náhodný výběr je z populace s daným rozdělením pravděpodobností) ve prospěch alternativní hypotézy H_1, která říká, že náhodný výběr není z populace s daným rozdělením pravděpodobností.

Chí-kvadrát test dobré shody je vhodný pro diskrétní i spojitá rozdělení, ovšem za předpokladu, že jsou splněny určité podmínky. Pro jeho použití musí být celkový rozsah výběrového souboru **větší než 50 jednotek** a všechny teoretické četnosti musí být **větší než 5**. Pokud teoretické četnosti této podmínce nevyhovují, je možné dosáhnout jejího splnění sloučením několika sousedních tříd. Tímto krokem ovšem dojde také ke snížení počtu stupňů volnosti, protože počet tříd po sloučení je nižší.

Příklad 11.1

Velkoobchodní distributor, který nakupuje notebooky, si vytvořil předpokládaný poměr pro nákup na základě trendů na trhu: **50 % černé, 30 % stříbrné, 10 % zlaté a 10 % bílé**. Pro otestování, zda jeho nákupní strategie odpovídá skutečné poptávce, má k dispozici data ze vzorku 150 notebooků (datová matice *„Notebook")*, kde je zastoupení barev následující: 82 černých, 49 stříbrných, 8 zlatých a 11 bílých. Naším úkolem je ověřit, zda je rozdíl mezi skutečným a předpokládaným rozdělením statisticky významný ($\alpha = 0{,}05$).

Řešení

K testování hypotézy použijeme ***chí*-kvadrát test dobré shody**. Než k němu přistoupíme, musíme nejprve ověřit jeho předpoklady. Velikost souboru n =150 předpoklad splňuje, ale pro náš test je důležité, aby v každé kategorii (barvě) byla teoretická četnost větší než 5.

Výpočet teoretických četností vychází ze vztahu 11.2.

Barva černá (50 %) $\rightarrow 150 \cdot 0{,}5 = \mathbf{75} \;\; \rightarrow n \cdot p_j = \; n_{oj}$

Barva stříbrná (30 %) $\rightarrow 150 \cdot 0{,}3 = \mathbf{45}$

Barva zlatá (10 %) $\rightarrow 150 \cdot 0{,}1 = \mathbf{15}$

Barva bílá (10 %) $\rightarrow 150 \cdot 0{,}1 = \mathbf{15}$

Všechny vypočítané teoretické četnosti jsou **větší než 5**. Předpoklad je tedy splněn a můžeme pokračovat v provedení *chí*-kvadrát testu.

Testujeme nulovou hypotézu $H_0: F(x) = F_0(x)$, že skutečné rozložení barev odpovídá předpokládanému poměru, proti alternativní hypotéze $H_1: F(x) \neq F_0(x)$, že skutečné rozložení barev se od předpokládaného poměru statisticky významně liší.

SPSS

Analyze → Nonparametric Tests → Legacy Dialogs → Chi-square → do okna Test Variable List vložíme proměnnou *Barva,* v okně Expected Values označíme Values a vložíme zde očekávané pravděpodobnosti pro jednotlivé barvy (pozor ve stejném pořadí, jak jsou barvy zakódované) → OK

Barva notebooku

	Observed N	Expected N	Residual
Černý	82	75,0	7,0
Stříbrný	49	45,0	4,0
Bílý	11	15,0	-4,0
Zlatý	8	15,0	-7,0
Total	150		

V žádném z políček tabulky není teoretická četnost není menší než 5 nejmenší je 15.

Test Statistics

	Barva notebooku
Chi-Square	5,342[a]
df	3
Asymp. Sig.	,148

a. 0 cells (0,0%) have expected frequencies less than 5. The minimum expected cell frequency is 15,0.

Pokračování příkladu 11.1

Na základě výstupu *chí*-kvadrát testu (χ^2= 5,342) a vypočtené p-hodnoty (p = 0,148), která je vyšší než stanovená hladina významnosti 0,05, nezamítáme tedy nulovou hypotézu. To znamená, že neexistuje statisticky významný rozdíl mezi pozorovaným rozložením barev notebooků ve vzorku a ideálním poměrem, který si stanovil distributor.

11.1.2 Kolmogorovův-Smirnovův test

Kolmogorov-Smirnovův test shody se používá k ověření hypotézy, že náhodný výběr pochází z rozdělení se spojitou distribuční funkcí $F(x)$, která je plně specifikována, tzn. že známe její typ i příslušné parametry. Tento test používá k testování hypotézy o tvaru rozdělení zkoumané náhodné veličiny X přímo jednotlivé naměřené hodnoty $x_1, x_2, \ldots, x_n$ a tím nedochází ke ztrátě informace, obsažené ve výběru. V případech výběru malého rozsahu je síla Kolmogorov-Smirnovova testu větší než síla χ^2 testu dobré shody.

$\boldsymbol{H_0}$	$\boldsymbol{H_1}$	$\boldsymbol{T(X)}$
$F(x) = F_0(x)$	$F(x) \neq F_0(x)$	$D_n = \max\left\{\left\lvert F(x_i) - \frac{i-1}{n}\right\rvert, \left\lvert F(x_i) - \frac{i}{n}\right\rvert\right\} \sim D_{\alpha(n)}$ (11.3)

Kde: $F(x)$ … distribuční funkce,
$F_0(x)$ … hypotetická distribuční funkce,
n … celkový rozsah souboru,
x_i … pozorované hodnoty pro $i = 1, 2, \ldots, n$,
$D_{n(\alpha)}$ … tabelovaná kritická hodnota pro Kolmogorov-Smirnovův test.

Kritický obor je vymezen nerovností $D_n > D_{n(\alpha)}$. Jestliže hodnota testovacího kritéria D_n překročí tabelovanou kritickou hodnotu $D_{n(\alpha)}$, zamítáme nulovou hypotézu o shodě mezi empirickým a teoretickým rozdělením na hladině významnosti α.

Speciálně k ověření normality dat se dále používá Shapirův-Wilkův test, o kterém jsme již hovořili v rámci průzkumové analýzy dat.

11.1.3 Shapiro-Wilkův test

Shapirův-Wilkův test se používá pro testování nulové hypotézy, která tvrdí, že náhodný výběr $X = x_1, x_2, \ldots, x_n$ pochází z normálního rozdělení s $N(\mu,\sigma^2)$. Tento test je založen na zjištění,

zda se body sestrojeného kvantil-kvantilového grafu (Q-Q plotu) významně liší od regresní přímky proložené těmito body.

Shapirův-Wilkův test byl původně navržen pro malé soubory dat ($n < 50$), protože jeho výpočetní náročnost byla v době jeho vzniku (1965) limitujícím faktorem. Nicméně s rozvojem výpočetní techniky a statistického softwaru se jeho použití výrazně rozšířilo. V dnešní době je Shapirův-Wilkův test běžně používán pro soubory až do cca 2000 pozorování. Test je citlivější na odchylky od normality než Kolmogorovův-Smirnovův test.

Z důvodu komplexnosti výpočtu testové statistiky je matematické pozadí testu poměrně složité, a proto se Shapirův-Wilkův test běžně nepočítá ručně, ale využívá se jeho implementace v softwarových nástrojích.

Příklad 11.2

V rámci studie trhu s notebooky je naším cílem zjistit, zda hmotnost notebooků pochází z normálního rozdělení. Tato informace je klíčová pro další analýzu dat, protože normalita je jedním z hlavních předpokladů pro použití mnoha statistických metod.

Řešení

K testování hypotézy použijeme **Shapiro-Wilkův test**, protože je pro soubory, jako je ten náš (s 150 jednotkami), považován za jeden z nejsilnějších a nejvhodnějších testů normality. Ve výstupu z programu bude také uveden výsledek Kolmogorova-Smirnovova testu, postup výpočtu je identický.

Testujeme nulovou hypotézu $H_0: F(x) = N(\mu, \sigma^2)$, že hmotnost notebooků pochází z normálního rozdělení, proti alternativní hypotéze $H_1: F(x) \neq \text{s } N(\mu, \sigma^2)$., že hmotnost notebooků nepochází z normálního rozdělení.

SPSS Analyze → Descriptive Statistics → Explore → proměnná *Hmotnost* do okna Variable(s) → Plots označíme Normality plots with tests → Continue → OK

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Hmotnost (kg)	,079	150	,022	,978	150	,017

a. Lilliefors Significance Correction

Z výstupu vyplývá, že p-hodnota ($p = 0{,}017$) Shapiro-Wilkova ($S\text{-}W = 0{,}978$) testu je menší než zvolená hladina významnosti (0,05). Proto zamítáme nulovou hypotézu. To znamená, že data o hmotnosti notebooků nepocházejí z normálního rozdělení a pro další analýzu je nutné použít neparametrické testy.

11.2 Pořadové testy

Pořadové testy se používají k ověření hypotéz o shodě mediánů v populaci, nebo přesněji, že data z obou výběrů pocházejí ze stejné populace. Využívají se především v případě, kdy data nesplňují předpoklady parametrických testů. Tyto testy pracují s pořadím hodnot neboli s ordinálními škálami. Ordinální data mohou být přímo výstupem měření, nebo mohou být na pořadí převedena kvantitativní, spojitá data.

11.2.1 Mannův-Whitneyho test

Mann-Whitney *U* test pro dva nezávislé výběry je neparametrickou obdobou t-testu pro dva nezávislé výběry.

Předpokládejme, že dva základní soubory mají spojité rozdělení s distribučními funkcemi $F(x)$ a $F(y)$. Z těchto souborů byly pořízeny dva nezávislé náhodné výběry o rozsazích m: x_1, x_2, …, x_m a n: y_1, y_2, …, y_n , kde ($m \leq n$). Je třeba ověřit hypotézu, že tyto nezávislé výběry mají stejnou polohu rozložení základního souboru proti alternativní hypotéze, že se oba svojí polohou významně liší.

Oba výběry (X, Y) spojíme do jednoho souboru (sdružený výběr) a hodnoty uspořádáme podle ve vzestupném pořadí podle velikosti. Jednotlivým hodnotám tohoto souboru přiřadíme pořadová čísla R_{x_m}, R_{y_n} (tzn. hodnoty očíslujeme od nejmenší k největší přirozenými čísly, přičemž stejně velkým hodnotám přiřadíme průměrné pořadí). Následně sečteme pořadová čísla prvního výběru $R_{x_1} + R_{x_2} + \ldots + R_{x_m} = T_x$ a druhého výběru $R_{y_1} + R_{y_2} + \ldots + R_{y_n} = T_y$.

H_0	H_1	$T(X)$	
$F(x) = F(y)$	$F(x) \neq F(y)$	$U_x = mn + \frac{m(m+1)}{2} - T_x$	(11.4)
$\tilde{\mu}_x = \tilde{\mu}_y$	$\tilde{\mu}_x \neq \tilde{\mu}_y$	$U_y = mn + \frac{n(n+1)}{2} - T_y$	(11.5)
		$U = \min(U_x, U_y) \sim U_{\alpha(m,n)}$	(11.6)

Kde: $F(x), F(y)$ … distribuční funkce náhodných výběrů X a Y,

$\tilde{\mu}_x$, $\tilde{\mu}_y$ … populační mediány náhodných výběrů X a Y,

$U_{\alpha(m,n)}$ … kritická hodnota rozdělení Mannův-Whitneyho testu.

Pokud $U < U_{\alpha(m,n)} \rightarrow$ nulovou hypotézu na hladině významnosti α zamítáme.

11.2.2 Dvouvýběrový Wilcoxonův test

Dvouvýběrový Wilcoxonův test je také neparametrickou obdobou t-testu pro nezávislé soubory. Postup pro stanovení testovacího kritéria je obdobný jako u Mann-Whitneyho testu, ale je jednodušší, protože se jako testová statistika používá menší ze součtů pořadí.

H_0	H_1	$T(X)$
$\tilde{\mu}_x = \tilde{\mu}_y$	$\tilde{\mu}_x \neq \tilde{\mu}_y$	$t = \min(T_x, T_y) \sim T_{\alpha(m,n)}$ (11.7)

Kde: $\tilde{\mu}_x,\ \tilde{\mu}_y$ … populační mediány náhodných výběrů X a Y,

T_x, T_y … součty pořadových čísel pro náhodné výběry X a Y,

$T_{\alpha(m,n)}$ … kritická hodnota rozdělení pro dvouvýběrový Wilcoxonův test.

Pokud $T < T_{\alpha(m,n)} \rightarrow$ nulovou hypotézu o shodě populačních mediánů na hladině významnosti α zamítáme.

Příklad 11.3

Student si chce koupit nový notebook. Pro každodenní nošení do školy je pro něj klíčová nízká hmotnost, ale zároveň preferuje prémiový materiál, jako je hliník, namísto běžného plastu. Student si není jistý, zda notebooky s hliníkovým rámem nejsou významně těžší než notebooky s plastovým rámem. Na hladině významnosti 5 % ověřte, zda existuje statisticky významný rozdíl v hmotnosti mezi notebooky, které mají hliníkové rámy, a těmi, které mají plastové rámy.

Řešení

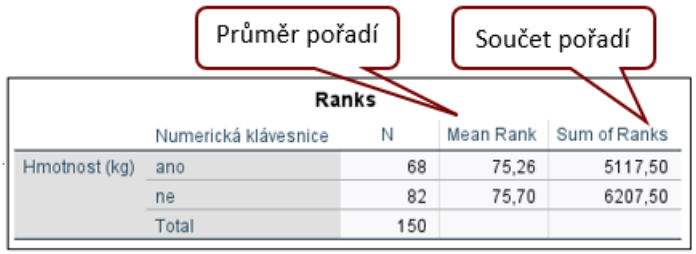
Na základě Shapiro-Wilkova testu bylo zjištěno, že proměnná hmotnost notebooků nemá normální rozdělení, a to jak pro notebooky s hliníkovým, tak i s plastovým rámem. Z tohoto důvodu nelze použít parametrické testy, které vyžadují normální rozdělení dat. Pro porovnání průměrné hmotnosti dvou nezávislých skupin (notebooky s hliníkovým a plastovým rámem) proto použijeme Mann-Whitneyho U test.

Testujeme nulovou hypotézu $H_0: \tilde{\mu}_x = \tilde{\mu}_y$, že mediány hmotnosti notebooků jsou v obou skupinách (hliníkový a plastový rám) stejné, proti alternativní hypotéze $H_1: \tilde{\mu}_x \neq \tilde{\mu}_y$, že mediány hmotnosti notebooků se v obou skupinách statisticky významně liší.

Pokračování příkladu 11.3

SPSS

Analyze → Nonparametric Tests → Legacy Dialogs → Two-Independent-Samples → do okna Test Variable List vložíme proměnnou *Hmotnost,* do okna Grouping Variable vložíme proměnnou *Materiál* → Define Groups → do okna Group 1 napíšeme 1 kód pro hodnotu *plast* do okna Group 2 napíšeme 2 kód pro hodnotu *hliník* → Continue → OK



Ranks

	Numerická klávesnice	N	Mean Rank	Sum of Ranks
Hmotnost (kg)	ano	68	75,26	5117,50
	ne	82	75,70	6207,50
	Total	150		

Test Statistics[a]

	Hmotnost (kg)
Mann-Whitney U	2226,000
Wilcoxon W	3711,000
Z	-1,438
Asymp. Sig. (2-tailed)	,150

a. Grouping Variable: Materiál z jakého je vyrobený rám NB

Z výstupu analýzy je patrné, že hodnota U statistiky je 2226 a p-hodnota = 0,150. Protože je p-hodnota větší než zvolená hladina významnosti (0,05), nezamítáme nulovou hypotézu. To znamená, že neexistuje statisticky významný rozdíl v mediánech hmotnosti mezi notebooky, které mají hliníkový rám, a těmi, které mají rám plastový.

11.2.3 Wilcoxonův test

Wilcoxonův test, neboli Wilcoxonův test pro párová data, je neparametrický pořadový test, který se používá k porovnání mediánů dvou závislých (párových) souborů dat. Je to neparametrická alternativa k párovému t-testu a je vhodný v případech, kdy data nesplňují předpoklad normality nebo se jedná o ordinální data.

Tento test ověřuje nulovou hypotézu H_0, která předpokládá, že populační medián diferencí d_i je roven nule (medián rozdílů je nulový), proti alternativní hypotéze H_1, která říká, že medián rozdílů je různý od nuly.

Postup testování spočívá v tom, že pro každou dvojici n závislých pozorování x_i, y_i, vypočteme diferenci d_i ($d_i = x_i - y_i$) pro $i = 1, 2, \dots n$. Nenulovým diferencím v absolutní hodnotě přiřadíme vzestupně pořadová čísla, které následně rozdělíme do dvou skupin podle znaménka diferencí d_i (shodným diferencím dáváme průměrné pořadí). Následně sečteme pořadová čísla pro skupinu záporných diferencí W^-a pro skupinu kladných diferencí W^+.

H_0	H_1	$T(X)$
$\tilde{\mu}_d = 0$	$\tilde{\mu}_d \neq 0$	$W = \min(W^-, W^+) \sim W_{\alpha(n)}$ (11.8)

Kde: $\tilde{\mu}_d$ .. populační medián diferencí,

$W_{\alpha(n)}$ .. kritická hodnota rozdělení Wilcoxonův párový test pro n ... počet nenulových diferencí.

Jestliže W < $W_{\alpha,(n)}$) → zamítáme nulovou hypotézu na hladině významností α.

Příklad 11.4

Majitel obchodu s elektronikou chce zjistit, jak se liší jeho prodejní ceny hliníkových notebooků od doporučených cen výrobcem. Na hladině významnosti 5 % (α = 0,05) zjistěte, zda existuje statisticky významný rozdíl mezi prodejní cenou a doporučenou cenou u notebooků s hliníkovým rámem.

Řešení

Na základě Shapiro-Wilkova testu bylo zjištěno, že proměnné cena a doporučená cena u notebooků s hliníkovým rámem nemají normální rozdělení. Pro testování tedy použijeme neparametrický test, který je vhodný pro párová data, která nesplňují předpoklad normality. Tímto testem je Wilcoxonův test pro párová data.

Testujeme nulovou hypotézu $\tilde{\mu}_d = 0$, zda je medián rozdílů nulový. Neexistuje statisticky významný rozdíl mezi prodejní a doporučenou cenou u notebooků s hliníkovým rámem.

SPSS Analyze → Nonparametric Tests → Legacy Dialogs → Two-Related-Samples → do okna Test Pairs: vložíme do Variable1 proměnnou *Cena,* do okna Variable2 proměnnou *Doporučená_cena* → OK

Ranks

		N	Mean Rank	Sum of Ranks
Doporučená cena výrobcem - Cena notebooku (Kč) v prodejně	Negative Ranks	16[a]	47,97	767,50
	Positive Ranks	134[b]	78,79	10557,50
	Ties	0[c]		
	Total	150		

a. Doporučená cena výrobcem < Cena notebooku (Kč) v prodejně
b. Doporučená cena výrobcem > Cena notebooku (Kč) v prodejně
c. Doporučená cena výrobcem = Cena notebooku (Kč) v prodejně

Test Statistics[a]

	Doporučená cena výrobcem - Cena notebooku (Kč) v prodejně
Z	-9,184[b]
Asymp. Sig. (2-tailed)	<,001

a. Wilcoxon Signed Ranks Test
b. Based on negative ranks.

Protože je p-hodnota ($p < 0{,}001$) menší než zvolená hladina významnosti 0,05, zamítáme nulovou hypotézu. Existuje statisticky významný rozdíl v mediánech cen mezi prodejní cenou a doporučenou cenou u sledovaných hliníkových notebooků. Ceny, za které se hliníkové notebooky prodávají, se významně liší od cen doporučených výrobcem.

11.2.4 Kruskal-Wallisův test

Kruskal-Wallisův test je neparametrická obdoba jednofaktorové analýzy rozptylu (ANOVY). Používá se především tehdy, jsou-li výrazně narušeny základní předpoklady pro provedení ANOVY (normalita dat a homogenita rozptylu) a také tehdy, mají-li jednotlivé výběry velmi málo pozorování (nebo jsou jejich počty silně nevyvážené) a normalitu není možné spolehlivě stanovit.

Kruskal-Wallisův test slouží k ověření nulové hypotézy, že $k > 2$ nezávislých náhodných výběrů o rozsazích $n = n_1, n_2, .., n_k$ pochází z jednoho základního souboru (ze stejného rozdělení). Předpokládáme, že tyto náhodné výběry byly pořízeny ze základních souborů se spojitými distribučními funkcemi.

Pro testování Kruskal-Wallisova testu se všechny výběrové soubory nejprve sloučí do jednoho souboru. Následně se každé hodnotě přiřadí vzestupně pořadové číslo. Pokud se vyskytnou stejné hodnoty, přiřadí se jim průměrné pořadí. Poté se součty pořadových čísel sečtou zvlášť pro každý výběrový soubor. Tyto součty se označují jako T_1,T_2,…,T_k, kde T_i je součet pořadových hodnot pro i-tý výběr.

$\boldsymbol{H_0}$	$\boldsymbol{H_1}$	$\boldsymbol{T(X)}$
$F_1(x) = F_2(x) = .. = F_k(x)$ $\tilde{\mu}_1 = \tilde{\mu}_2 = \cdots = \tilde{\mu}_k$	Neplatí H_0	$KW = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(n+1) \sim \chi^2_{\alpha,k-1}$ (11.9)

Kde: $F_1(x),\ F_2(x),..,\ F_k(x)$ … distribuční funkce náhodných výběru $i = 1, 2, .., k$,

$\tilde{\mu}_1, \tilde{\mu}_2, \ldots, \tilde{\mu}_k$ … populační mediány náhodných výběrů $i = 1, 2, .., k$,

n … celkový rozsah souboru = $\sum_{i=1}^{k} n_i$

$\chi^2_{\alpha,k-1}$ … kritická hodnota *chí*-kvadrát rozdělení pro k-1 nezávislých výběrů.

Jestliže $KW \geq \chi^2_{\alpha,k-1} \rightarrow$ zamítáme nulovou hypotézu, hodnoty nejméně dvou pozorovaných výběrových souborů se od sebe výrazně liší (nepocházejí ze stejného rozdělení).

Pro zjištění podrobnějších výsledků neboli pro mnohonásobné srovnání (tzv. **post-hoc testy**) jednotlivých dvojic skupin, můžeme v kombinaci s Kruskal-Wallisovým testem použít **Neményiho metodu** a **Dunnovu metodu.**

Neményiho metoda srovnávání nezávislých výběrů, je konstruována pro vyvážené modely (platí $n_1 = n_2 = \cdots = n_k = N$). Postup při této metodě spočívá v porovnávání všech

párových diferencí $|T_i - T_j|$s kritickou hodnotou $N_{\alpha(k,N)}$. Jestliže $|T_i - T_j| \geq N_{\alpha(k,N)}$, zamítá se hypotéza, že i-tý a j-tý výběr pocházejí z téhož rozdělení. Kritické hodnoty jsou tabelovány pro $N \leq 25$ a $k \leq 10$, kde k je počet porovnávaných tříd, N celkový počet opakování v každé třídě.

Dunnova metoda se používá v případě, že rozsahy jednotlivých výběrových souborů nejsou stejné (**nevyvážený model**). Jestliže $|T_i - T_j| > u_{\frac{2\alpha}{k(k-1)}} \sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$, kde: $u_{\frac{2\alpha}{k(k-1)}}$ je tabelovaná kritická hodnota rozdělení $N(0,1)$, zamítáme na hladině významnosti α hypotézu, že i-tý a j-tý výběr pocházejí z téhož rozdělení.

Příklad 11.5

Představte si, že v rámci seminární práce provádíme analýzu dat z trhu s notebooky. Máte za úkol zjistit, zda existuje statisticky významný rozdíl v průměrné výdrži baterie u notebooků, které jsou vybaveny různým typem grafické karty.

Řešení

Proměnná baterie nemá ani pro jedu kategorii grafické karty normální rozdělení, musíte pro použít **Kruskal-Wallisův test**, který je vhodný pro porovnání tří a více nezávislých skupin.

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Grafická karta	Statistic	df	Sig.	Statistic	df	Sig.
Výdrž baterie (hod)	Integrovan	,200	55	<,001	,883	55	<,001
	Dedikovná	,158	76	<,001	,912	76	<,001
	Herní	,313	19	<,001	,783	19	<,001

a. Lilliefors Significance Correction

Testujeme nulovou hypotézu $H_0: F_1(x) = F_2(x) = F_3(x)$, že všechny nezávislé výběry pocházejí z populace ze stejným rozdělení, proti alternativní hypotéze $H_1: \overline{H}_0$, že se alespoň jeden výběr v rozdělení mediánu statisticky významně liší od ostatních výběrů.

SPSS

Analyze → Nonparametric Tests → Independent-Samples → záložka Fieds do okna Test Fields vložíme proměnnou *Baterie,* do okna Groups vložíme proměnnou *Grafika* → záložka Settings označíme Customize tests a zaškrtneme Kruskal-Wallis → Run

Independent-Samples Kruskal-Wallis Test Summary

Total N	150
Test Statistic	26,203[a]
Degree Of Freedom	2
Asymptotic Sig.(2-sided test)	<,001

a. The test statistic is adjusted for ties.

Pairwise Comparisons of Grafická karta

Sample 1-Sample 2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.[a]
Dedikovná-Integrovan	12,559	7,687	1,634	,102	,307
Dedikovná-Herní	-56,980	11,137	-5,116	<,001	,000
Integrovan-Herní	-44,422	11,554	-3,845	<,001	,000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is ,050.
a. Significance values have been adjusted by the Bonferroni correction for multiple tests.

Pokračování příkladu 11.5

Z výstupu Kruskl-Wallisova testu vyplývá, že p-hodnota ($p < 0{,}001$) je menší než hladina významnosti ($\alpha = 0{,}05$), proto **zamítáme nulovou hypotézu**. To znamená, že existuje statisticky významný rozdíl v mediánech výdrže baterie mezi alespoň dvěma skupinami typů grafických karet. Pro zjištění, mezi kterými konkrétními skupinami rozdíl existuje, je nutné provést podrobnější hodnocení výsledků pomocí metod párového srovnávání.
V případě, že je počet pozorování v každé skupině jiný, hovoříme o nevyváženém modelu. Pro správné porovnání v takovýchto situacích je nutné použít korigovanou p-hodnotu (Adj. Sig.). Tato korekce je nezbytná proto, že při vícenásobném srovnávání se zvyšuje pravděpodobnost, že náhodně získáme statisticky významný výsledek.

Dedikovaná – Integrovaná grafika: $\boldsymbol{p = 0{,}307}$ je **větší** než hladina významnosti 0,05. **Nezamítáme nulovou hypotézu**. To znamená, že mezi notebooky s dedikovanou a integrovanou grafikou neexistuje statisticky významný rozdíl v mediánu výdrže baterie.

Dedikovaná – Herní grafika: $\boldsymbol{p < 0{,}0001}$. Tato hodnota je menší než hladina významnosti 0,05. **Zamítáme nulovou hypotézu**. Existuje statisticky významný rozdíl v mediánu výdrže baterie.

Integrovaná – Herní grafika: $\boldsymbol{< 0{,}0001}$. Tato hodnota je také menší než 0,05. **Zamítáme nulovou hypotézu.** Existuje statisticky významný rozdíl v mediánu výdrže baterie.

Na základě Kruskal-Wallisova testu a následných post-hoc testů jsme zjistili, že medián výdrže baterie se statisticky významně liší mezi herní grafickou kartou a ostatními dvěma typy (dedikovanou a integrovanou).

Shrnutí kapitoly

Neparametrické testy jsou statistické metody, které se používají v situacích, kdy nejsou splněny předpoklady pro použití parametrických testů. Mezi jejich hlavní výhody patří, že jsou **použitelné bez ohledu na typ rozdělení** a jsou méně citlivé na extrémní hodnoty. Nevýhodou je jejich menší síla, tedy nižší schopnost rozpoznat neplatnou nulovou hypotézu.

Testy shody rozdělení srovnávají empirické rozdělení s určitým teoretickým rozdělením. Vycházejí z porovnávání **pozorovaných četností** s **očekávanými četnostmi**.

- ***Chí*-kvadrát test dobré shody**
- **Kolmogorovův-Smirnovův test**
- **Shapirův-Wilkův test**

Testy odchylek slouží k posouzení, zda se naměřené hodnoty významně liší od určité předem dané hodnoty.

Pořadové testy vycházejí z toho, že původní hodnoty dat jsou nahrazeny pořadovými čísly. Hlavní výhodou tohoto postupu je, že se odstraní vliv odlehlých hodnot.

- **Mannův-Whitneyho *U* test**
- **Wilcoxonův test pro párová data**
- **Kruskal-Wallisův test**

Permutační testy se používají k posouzení variability možných výsledků náhodným přeskupováním (permutací) naměřených dat.

LITERATURA

CONOVER, William Jay. *Practical nonparametric statistics.* John wiley & sons, 1999. ISBN: 978-0-471-16068-7

HOŠKOVÁ, Pavla; Andrea JINDROVÁ, Marie PRÁŠILOVÁ a Rudolf ZEIPELT. *Statistika I*. Praha: PEF ČZU, 2013. ISBN: 978-80-213-2341-4.

JAMES, Gareth; Daniela WITTEN, Trevor HASTIE a Robert TIBSHIRANI. *An Introduction to Statistical Learning: with Applications in R.* 2. vydání. Springer, 2021. ISBN 978-10-71618-19-6.

NEUBAUER, Jiří; Marek SEDLAČÍK a Oldřich KŘÍŽ. *Základy statistiky: Aplikace v technických a ekonomických oborech.* 3. vydání. Praha: Grada Publishing, 2021. ISBN 978-80-271-3421-2.

11 Kontrolní otázky

1. Vysvětlete rozdíl mezi parametrickými a neparametrickými testy.
2. V jakých situacích je vhodné použít neparametrické testy?
3. Jaké jsou hlavní výhody neparametrických testů a jaké jsou jejich nevýhody?
4. K čemu slouží testy shody rozdělení?
5. Co jsou to očekávané četnosti?
6. Jaké jsou hlavní předpoklady $chí$-kvadrát testu dobré shody rozdělení?
7. Jaký neparametrický test slouží k ověření, zda dva nezávislé výběry pocházejí ze stejné populace?
8. Kdy se používá Kruskal-Wallisův test?
9. V jakých případech je vhodné použít Wilcoxonův test pro párová data a jaká je jeho parametrická obdoba?
10. Jaký je princip výpočtu pořadových testů? Co se stane s původními hodnotami dat?
11. Co je p-hodnota (signifikantní hodnota) a jaký je její vztah k hladině významnosti?

11 Příklady k procvičení

Na základě datové matice *„Byty“*, kterou naleznete v kurzu v Moodle, odpovězte na následující otázky.

11.1.1 Realitní makléř chcete zjistit, zda je rozložení bytů podle dispozice na trhu rovnoměrné. Na základě informací z výběrového souboru ověřte, zda rozložení bytů podle jejich dispozice odpovídá tomuto předpokladu. Pro testování hypotéz použijte hladinu spolehlivosti 95 %.

$[\chi^2 = 8{,}320\,; p = 0{,}016]$

11.1.2 Pomocí vhodného testu ověřte, zda proměnná plocha bytu pochází z normálního rozdělení.

$[S\text{-}W = 0{,}898\,; p < 0{,}001]$

11.1.3 Na základě informací z datového souboru o bytech a ověřte, zda existuje statisticky významný rozdíl v ploše bytu mezi domy, které jsou vybaveny výtahem, a těmi, které ho nemají. Pro testování hypotéz použijte hladinu významnosti α=0,05.

$[M\text{-}W = 821{,}00; p < 0{,}001]$

11.1.4 Na hladině významnosti 5 % ověřte, zda existuje statisticky významný rozdíl v ploše bytu v závislosti na jeho stavu (např. novostavba, po rekonstrukci, původní stav)

$[K\text{-}W = 129{,}180; p < 0{,}001]$

SEZNAM LITERATURY

ANDĚL, Jiří. *Základy matematické statistiky*. Vyd. 3. Praha: Matfyzpress, 2011. ISBN 978-80-7378-162-0.

BUDÍKOVÁ, Marie; Maria KRÁLOVÁ a Bohumil MAROŠ. *Průvodce základními statistickými metodami*. vydání první. Praha: Grada Publishing, a.s., 2010. ISBN 978-80-247-3243-5.

CAMM, Jeffrey; James COCHRAN, David ANDERSON, Dennis SWEENEY, Thomas WILLIAMS et al. *Statistics for Business and Economics*. 15th Edition. Boston: Cengage Learning, 2023. ISBN 978-0357715857.

CONOVER, William Jay. *Practical nonparametric statistics.* John wiley & sons, 1999. ISBN: 978-0-471-16068-7

FIELD, Andy. *Discovering Statistics Using IBM SPSS Statistics*. 5th ed. Los Angeles: SAGE Publications, 2018. ISBN 978-1526436566.

GWET, Kilem L. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC, 2014. ISBN 978-0-9708062-8-4.

HASTIE, Trevor; Robert TIBSHIRANI a Jerome FRIEDMAN. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer, 2009. ISBN 978-0-387-84857-0.

HEBÁK, Petr. *Statistické myšlení a nástroje analýzy dat*. 2. vydání. Praha: Informatorium, 2015. ISBN 978-80-7333-118-4.

HEBÁK, Petr a Jana KAHOUNOVÁ. 2014. *Počet pravděpodobnosti v příkladech*. Praha: Informatorium. ISBN 978-80-7333-109-2.

HENDL, Jan. *Přehled statistických metod zpracování dat: analýza a metaanalýza dat*. Vyd. 1. Praha: Portál, 2004. ISBN 8071788201.

HENDL, Jan. *Kvalitativní výzkum*. 5. přeprac. vyd. Praha: Portál. 2023. 978-80-262-1968-2.

HINDLS, Richard; Stanislava HRONOVÁ, Jan SEGER a Jakub FISCHER. *Statistika pro ekonomy*. 8. vyd. Praha: Professional Publishing, 2007. ISBN 978-80-86946-43-6.

HINDLS Richard; Markéta ARLTOVÁ, Stanislava HRONOVÁ, Ivana MALÁ, Luboš MAREK, et al. *Statistika v ekonomii*. První vydání. Professional Publishing, 2018. ISBN 978-80-88260-09-7.

HOŠKOVÁ, Pavla; Andrea JINDROVÁ, Marie PRÁŠILOVÁ a Rudolf ZEIPELT. *Statistika I.* Praha: PEF ČZU, 2013. ISBN: 978-80-213-2341-4.

JAMES, Gareth; Daniela WITTEN, Trevor HASTIE a Robert TIBSHIRANI. *An Introduction to Statistical Learning: with Applications in R.* 2. vydání. Springer, 2021. ISBN 978-10-71618-19-6.

JOHNSON, Richard A. a Dean W. WICHERN. *Applied multivariate statistical analysis*. Sixth edition. Harlow: Pearson, 2014. ISBN 978-1-292-02494-3.

KREIDL Martin. Zhodnocení vlivu práce výzkumných agentur na konstruktovou validitu škál. *Sociologický časopis/Czech Sociological.* Review, roč. 41 (2005), č.1, s. 103-124.

LITSCHMANNOVÁ Martina, *Úvod do statistiky.* Online. VŠB – Technická univerzita Ostrava, 2012. Dostupné z: https://mi21.vsb.cz/sites/mi21.vsb.cz/files/unit/interaktivni_uvod_do_statistiky.pdf. [citováno 2023-08-03]

MELOUN, Milan, Jiří MILITKÝ a Martin HIL. *Statistická analýza vícerozměrných dat v příkladech*. Vyd. 2. Praha: Academia, 2012. ISBN 978-80-2002-071-0.

NEUBAUER, Jiří; Marek SEDLAČÍK a Oldřich KŘÍŽ. *Základy statistiky: aplikace v technických a ekonomických oborech*. 2., rozšířené vydání. Praha: Grada, 2016. ISBN 978-80-247-5786-5.

ŘEHÁK, Jan a Petr SOUKUP. *Úvod do statistické analýzy dat v psychologii*. Praha: Portál, 2021. ISBN 978-80-262-1718-0.

ŘEZANKOVÁ, Hana a Tomáš LÖSTER. *Základy statistiky*. Vyd. 1. Praha: Oeconomica, 2013. ISBN 978-80-245-1957-9.

VANACORE, Amalia a Maria PELLEGRINO. Benchmarking procedures for characterizing the extent of rater agreement: a comparative study. *Quality and Reliability Engineering International*, vol 38 (2021), no 3, s. 1404-1415.

VOJTÍŠEK, Petr. *Výzkumné metody: Metody a techniky výzkumu a jejich aplikace v absolventských pracích vyšších odborných škol.* Praha: Vyšší odborná škola sociálně právní, 2012. ISBN 978-80-905109–3-7.

Název: ZÁKLADY DATOVÉ ANALÝZY
Autor: Ing. Andrea Jindrová, Ph.D.
Vydavatel: Česká zemědělská univerzita v Praze
Adresa: Česká zemědělská univerzita v Praze, Kamýcká 129, Praha –Suchdol, 165 00
Pořadí vydání: 1.
Rok vydání: 2026, online
ISBN 978-80-213-3530-1