UČEBNÍ TEXTY OSTRAVSKÉ UNIVERZITY

Přírodovědecká fakulta

ZÁKLADY PRAVDĚPODOBNOSTI A STATISTIKY

Josef Tvrdík

OSTRAVSKÁ UNIVERZITA 2010

OBSAH:

1 ÚVOD 4

1.1 CO JE STATISTIKA? 4
1.2 STATISTICKÁ DATA 5
1.3 MĚŘENÍ A TYPY ŠKÁL 7

2 POPISNÁ STATISTIKA 11

2.1 ČETNOST, ROZDĚLENÍ ČETNOSTI, GRAFICKÉ ZNÁZORNĚNÍ 11
2.2 CHARAKTERISTIKY POLOHY 20
2.3 CHARAKTERISTIKY VARIABILITY 27
2.4 DALŠÍ CHARAKTERISTIKY ROZDĚLENÍ POZOROVANÝCH HODNOT 31
2.5 NĚKTERÉ TECHNIKY POPISNÉ STATISTIKY 34
2.6 POPIS VZTAHU DVOU VELIČIN 38
2.7 PŘÍKLAD STATISTICKÉHO ZPRACOVÁNÍ DAT 43

3 ZÁKLADY POČTU PRAVDĚPODOBNOSTI 47

3.1 NÁHODNÝ POKUS, NÁHODNÝ JEV A PRAVDĚPODOBNOST 47
3.2 NÁHODNÁ VELIČINA A ROZDĚLENÍ PRAVDĚPODOBNOSTI 58
3.3 CHARAKTERISTIKY NÁHODNÝCH VELIČIN 62
3.4 PŘÍKLADY DISKRÉTNÍCH ROZDĚLENÍ 70
3.5 PŘÍKLADY SPOJITÝCH ROZDĚLENÍ 74
3.6 O CENTRÁLNÍ LIMITNÍ VĚTĚ 85

4 INDUKTIVNÍ STATISTIKA 89

4.1 ZÁKLADNÍ POJMY 89
4.2 STATISTICKÝ ODHAD 94
4.3 TESTOVÁNÍ HYPOTÉZ 101

5 LITERATURA - KOMENTOVANÝ SEZNAM 107

INTERAKTIVNÍ UČEBNICE PRO ZÁKLADNÍ KURS STATISTIKY: 109

6 STATISTICKÉ TABULKY 110

6.1 DISTRIBUČNÍ FUNKCE NORMOVANÉHO NORMÁLNÍHO ROZDĚLENÍ 110
6.2 VYBRANÉ KVANTILY ROZDĚLENÍ CHÍ-KVADRÁT 111
6.3 VYBRANÉ KVANTILY STUDENTOVA T-ROZDĚLENÍ 112
6.4 VYBRANÉ KVANTILY FISHEROVA-SNEDECOROVA F-ROZDĚLENÍ 113

Předmluva ke druhému vydání

Úpravy v tomto vydání vycházejí ze zkušeností v několikaletém užívání textu ve výuce. Byly vypuštěny nebo zjednodušeny některé úseky, které pro pochopení základních pojmů nebyly nezbytné. Na několika místech textu byly doplněny ilustrační příklady a obrázky. Část kapitoly o testování hypotéz o střední hodnotě (dvouvýběrové a párové *t*-testy) byla přesunuta do předmětu Analýza dat, který na předmět Základy matematické statistiky navazuje.

Byla přidána podkapitola 2.7 s příkladem využití jednoduchých metod popisné statistiky ve vyhodnocení dat o účinnosti čtyř stochastických algoritmů a doplněno vysvětlení a příklady hledání hodnot distribučních funkcí a kvantilů pomocí funkcí v Excelu. Kromě toho byly odstraněny některé drobné formální a typografické nedostatky a byl aktualizován seznam literatury, zejména o české knihy, učební texty a elektronické učebnice, které vyšly v posledních létech a jsou vhodné jako doplňující literatura.

Tak snad toto nové upravené vydání bude pro studenty příjemnější a srozumitelnější a bude dobrou pomůckou pro pochopení základních principů statistiky a jejich aplikaci v analýze dat.

1 Úvod

Po prostudování této kapitoly byste měli:

- vědět, čím se zabývá statistika a jaká data může zpracovávat,
- rozumět pojmům *objekt, veličina, datová matice, základní soubor, výběrový soubor*,
- chápat rozdíl mezi škálou nominální, ordinální, intervalovou a podílovou.

Čas potřebný k prostudování tohoto modulu je asi 2 hodiny.

1.1 Co je statistika?

Slovo ***statistika*** má původ v minulosti vzdálené několik století. Cítíme v něm latinský základ - *status*, tedy stav, a také *stát* - stav věcí veřejných. Nahlédneme-li do výkladového slovníku nebo do úvodních kapitol učebnic statistiky, dozvíme se, že "statistika se zabývá studiem zákonitostí hromadných jevů". Věta je to jistě pozoruhodná, ale nepřipravenému čtenáři mnoho nesděluje. Kromě toho se dočteme v učebnicích, že pod pojmem statistika je většinou míněna *matematická statistika*, což je obor matematiky, který se zabývá aplikacemi teorie pravděpodobnosti, (což je další obor matematiky), a že matematická statistika hledá správné metody usuzování z neúplných údajů, zatížených ještě navíc náhodným kolísáním. Vidíme, že je mnoho významů slova "statistika". Jedním z hlavních cílů tohoto předmětu a tohoto textu je vybudování základů pro správné pochopení významů slova "statistika" a pro využití některých statistických metod poznávání a chápání světa, který nás obklopuje, tedy pro *statistickou analýzu dat*.

Analýza je opakem syntézy**,** jak víme z křížovek. Také místo analýza můžeme užívat české slovo *rozbor*. Zde můžete tento pojem chápat jako postup rozdělení velkého celku na takové součásti, které nám ten nepřehledný celek pomáhají pochopit a porozumět mu.

Data jsou zobrazením jisté části reálného světa, často bývají vyjádřena číselně. Části světa můžeme zobrazovat různou formou - jako fotografii, mapu, kresbu - to všechno jsou data. V tomto textu však daty budeme rozumět především zobrazení do číselných hodnot.

Příklad 1-1:

Fotbalové mužstvo Baníku Ostrava jako jistý výsek z reálného světa může být zobrazeno třeba:

- skupinovou fotografií - tu jistě ocení běžný fanoušek, nebo snad ještě více mladá dáma hledající objekt hodný její pozornosti ,
- tabulkou, ve které bude u každého hráče zaznamenán věk, výška, váha, počet odehraných minut a vstřelených branek v této sezóně, datum ukončení smlouvy atd. Tato forma dat bude zřejmě užitečnější pro realizační tým zodpovědný za výkon mužstva. .

Ve statistické analýze rozumíme daty jen druhou možnost, tedy zobrazení ve formě tabulky.

1.2 Statistická data

Data jsou jistou formou zobrazení výseku z reálného světa, který nás obklopuje. Statistickými daty budeme rozumět číselné zobrazení takového výseku reálného světa, ve kterém se zobrazované objekty vyskytují hromadně, tzn. že různí jedinci (objekty) patřící *do stejné kategorie, kterou umíme jasně určit*, se objevují vícekrát.

Příklad 1-2:
Několik příkladů výseků z reálného světa s hromadným výskytem objektů:
a) ryby v přehradní nádrži Šance,
b) jabloně v ovocné zahradě pana Nováka,
c) občané České republiky k 1. lednu 2008.

Takové výseky z reálného světa, které zahrnují více objektů majících nějakou společnou vlastnost, a tedy patří do stejné kategorie, nazýváme *populace*. Výše uvedené příklady byly tedy příklady populací. Zobrazením buď *všech* nebo jen *některých* objektů populace vznikají *statistická data*. Ačkoliv u každého z uvedených příkladů nás budou zajímat zcela jiné vlastnosti sledovaných objektů, třeba v příkladu 1-2:
a) druh, délka, hmotnost, velikost šupin apod.,
b) stáří stromu a úroda v loňském roce vyjádřená v kilogramech.

Každé z těchto zobrazení bude mít stejnou strukturu, *strukturu tabulky*, ve které každý sloupec znamená jednu sledovanou vlastnost (veličinu) a každý řádek odpovídá jednomu objektu.

	veličina_1	*veličina_2*	...	*veličina_ j*	...	*veličina_p*
objekt_1	x_{11}	x_{12}	...	...	...	x_{1p}
objekt_2	x_{21}	x_{22}	...	...	...	x_{2p}
⋮	⋮	⋮	...	...	...	⋮
objekt_i	⋮	⋮	...	x_{ij}	...	⋮
⋮	⋮	⋮	...	...	...	⋮
objekt_n	x_{n1}	x_{n2}	...	...	...	x_{np}

Tabulka 1-1: Obvyklá struktura statistických dat

Uvnitř tabulky jsou číselné hodnoty veličin zjištěné na každém ze sledovaných objektů. Každý sloupec tabulky může být nadepsán jménem měřené veličiny, každý řádek lze označit tak, abychom jednoznačně poznali, kterému objektu je tento řádek přiřazen.

Příklad 1-3:
Data z příkladu 1-2 b) - jabloně pana Nováka, má-li všech svých 5 jabloní označeno čísly a jsou-li sledovány dvě veličiny, totiž stáří jabloně a množství sklizených jablek - mohou vypadat takto:

jabloň	*stáří (roky)*	*sklizeň (kg)*
1	20	65
2	21	80
3	15	25
4	10	12
5	24	55

Tabulka 1-2: Příklad statistických dat

Tabulka je základní a nejčastější strukturou statistických dat jako obrazu jisté části reálného světa. Její výhodou je to, že z ní snadno rozpoznáme, čeho je obrazem. Nevýhodou může být její velký rozsah, a tím i nepřehlednost, např. tabulka z příkladu 1-2 c) by měla více než deset miliónů řádků. Právě zpracování informací z takových rozsáhlých tabulek do přehlednější formy je jedním z úkolů statistické analýzy dat.

Číselné hodnoty uvnitř tabulky tvoří datovou strukturu o *n* řádcích a *p* sloupcích, která se v matematice nazývá matice. Proto se někdy o datech v tabulce hovoří jako o *datové matici*. Sloupce tabulky jsme dosud označovali jako *veličiny*. Někdy jsou však také označovány jako *znak*, *proměnná* (anglicky *variable*) a v některých vymezených souvislostech i celou řadou dalších názvů. Podobně i pro *objekty* existuje množství synonym: *jedinec*, (statistické) *individuum*, případ (anglicky *case*) atp. Protože však už rozumíme klíčovému konceptu, tj. statistickým datům ve struktuře tabulky, nemůže nás tato nadbytečná pestrost názvosloví nijak zmást.

Je však nutné rozlišovat jeden velice podstatný rozdíl mezi daty, která zobrazují *všechny* objekty z populace a daty, která zobrazují jenom *část objektů populace*. V případě, že data jsou obrazem celé populace, se tato data označují jako *základní soubor*. Analýzou základního souboru můžeme získat přehledněji a úsporněji uspořádaný popis dat, a tím i srozumitelnější popis sledovaného výseku reálného světa, číselné hodnoty *parametrů* populace. Takový postup označujeme jako ***popisnou (deskriptivní) statistiku***.

Základní soubor není vždy k dispozici. Třeba může být populace velice rozsáhlá a změřit všechny objekty je časově nebo finančně neúnosné nebo je dokonce takové měření nemožné. Např. měření je destruktivní, jako je třeba tlaková zkouška cihel a základní soubor můžeme získat jen tím, že v měřícím lisu rozdrtíme všechny vyrobené cihly. Tím bychom sice získali základní soubor, ale při tom bychom zničili tu část reálného světa, kterou má zobrazovat, a informace ze základního souboru už by přestaly být zajímavé.

Někdy data tedy zobrazují jen část objektů populace, avšak my bychom si rádi učinili obraz o celé populaci, o jejích parametrech. Je to podobná situace, jako když z několika útržků fotografie si chceme udělat obraz o krajině, která byla zachycena na celé fotografii. Je zřejmé, že naše úspěšnost v tomto úsilí bude záviset na tom, zda na útržcích budou přítomny všechny podstatné rysy krajiny, a také na tom, zda budeme správně usuzovat (odhadovat) z jednotlivostí na vlastnosti celku. Ve statistické analýze se taková část populace nazývá *výběr* a jeho zobrazení do dat *výběrový soubor*. Z výběrového souboru samozřejmě nemůžeme určit parametry populace, protože nemáme o populaci úplnou informaci, ale pouze *odhady parametrů populace*.

Metody správného usuzování z výběru na populaci, kdy z informací o části usuzujeme na celek a ze speciálního na obecné, nám poskytuje *matematická statistika*. Postup se označuje jako *statistická indukce* a aplikace takových metod se nazývají *induktivní statistika*.

Pojmy, s nimiž jste se seznámili v této kapitole, lze přehledně shrnout, jak je ukázáno v tabulce 1-3.

Tabulka 1-3: Přehled pojmů týkajících se statistických dat

	všechny objekty	jen část objektů
realita	*populace*	*výběr*
data	*základní soubor*	*výběrový soubor*
charakteristiky	*parametry*	*odhady (parametrů)*
metody	*deskriptivní statistika*	*induktivní statistika*

1.3 Měření a typy škál

K číselnému vyjadřování vlastností (a intenzity vlastností) jedinců, tedy ke kvantifikaci, slouží různé techniky měření. Měřením zjistíme pro jistý objekt číselnou hodnotu sledované veličiny, tím vlastně vytvoříme obraz objektu na číselné ose. Pokud chceme poznávat reálný svět z jeho obrazů (většinou se nám nic lepšího nenabízí), je jistě nutné, aby svět byl zobrazován nezkresleně. Měřící procedury musí mít řadu jasně definovaných vlastností, jako reprodukovatelnost, ověřitelnost atd.

Výsledky měření se vyjadřují číselnými hodnotami měřící stupnice, tzv. škály. Škálou jsou vymezeny všechny možné hodnoty, které měřená veličina může nabývat. Podle typu škály jsou definovány vztahy mezi hodnotami na škále. Rozeznáváme čtyři typy škál, a tedy i čtyři druhy měřených veličin (znaků). Uvedeme je v pořadí od nejhrubější, postihující nejméně detailů, po nejjemnější typ škály.

***Nominální** škála* klasifikuje objekty do určitých předem vymezených tříd či kategorií. Hodnoty v nominální škále se dají vyjádřit slovně a mezi různými hodnotami není definováno žádné uspořádání. Pokud jsou hodnoty nominální škály někdy označovány číselně, mějte na paměti, že toto číslo je pouze jakousi zkrat-

kou (kódem) slovní hodnoty*. O veličinách měřených v nominální škále hovoříme jako o *nominálních veličinách.*

Příklad 1-4:
V nominální škále se vyjadřují hodnoty veličin, jako jsou např.:
- pohlaví (s možnými hodnotami mužské, ženské),
- barva očí (modrá, hnědá, černá),
- výsledek léčby (uzdraven, zemřel),
- národnost (česká, slovenská, polská, německá, ...).

Ordinální (*pořadová*) *škála* umožňuje jedince podle sledované vlastnosti nejen rozlišovat, ale také uspořádat ve smyslu vztahů "je větší", "je menší" nebo "předchází", "následuje", aniž by však byla schopna vyjádřit číselně vzdálenost mezi větším a menším či mezi předcházejícím a následujícím. Veličiny měřené v ordinálních škálách se nazývají *ordinální veličiny*.

Nominální a ordinální veličiny jsou souhrnně označovány jako *kategoriální*.

Příklad 1-5:
V ordinální škále se měří znaky jako
- dosažené vzdělání (základní, střední, vysokoškolské),
- prospěch ve školním předmětu (výborně, velmi dobře, dobře, nevyhověl),
- důstojnická hodnost (podporučík, poručík, nadporučík, kapitán, ...),
- stav pacienta (vyléčen, remise, recidiva),
- hodnocení funkce technických zařízení (stupně závažnosti poruchy jaderné elektrárny),
- ohrožení povodní (stupně povodňové aktivity),
- hodnocení postojů v sociologických průzkumech (škála má hodnoty např. souhlasím, spíše souhlasím, spíše nesouhlasím, nesouhlasím),
- četnost výskytu (často, občas, zřídka, nikdy),
- chuť vína nebo jiné poživatiny podle degustátora atd.

Na ordinální škále se někdy měří i veličiny měřitelné kvantitativně jemnějšími škálami, pokud rozlišení ordinální škálou postačuje, např. postava člověka může být malá, střední nebo velká.

Intervalová *(rozdílová) škála* navíc umožňuje stanovit vzdálenost mezi hodnotami měřené veličiny. Je tedy oproti ordinální škále bohatší. Intervalová škála má definovánu jednotku měření, avšak nula byla definována s jistou libovůlí. Dovoluje proto počítat s rozdíly naměřených hodnot, nikoliv s jejich podíly.

Příklad 1-6:
Typickou veličinou měřenou v intervalové škále je teplota. Různé teplotní škály (Celsiova, Fahrenheitova) mají různě položené nuly (0 stupňů Celsia = 32 stupňů

*Současné programy pro statistickou analýzu dat většinou nevyžadují, aby data byla homogenní datová struktura, tedy matice s pouze číselnými hodnotami, a umí správně pracovat i s daty, kde hodnoty nominálních veličin jsou znakové řetězce.

Fahrenheita) a také rozdílné jednotky (jednotka Celsiovy stupnice = 1.8 jednotek Fahrenheitovy stupnice). Má-li těleso teplotu *C* stupňů Celsia, je zároveň teplé (32+1.8 *C*) stupňů Fahrenheita. Teploty dvou těles, lišících se o *d* stupňů Celsia, se zároveň liší o 1.8 *d* stupňů Fahrenheita, bez ohledu na to, v které části stupnice se tyto hodnoty nacházejí. Podíly teplot však tuto stálost nezachovávají. Např. dvojnásobnému zvýšení teploty z 10 na 20 stupňů Celsia odpovídá ve stupnici Fahrenheitově zvýšení 1.36 krát (z 50 na 68 stupňů), zatímco dvojnásobnému zvýšení teploty z 20 na 40 stupňů Celsia odpovídá ve stupnici Fahrenheitově zvýšení 1.53 krát (ze 68 na 104 stupně).

***Podílová** škála* zachovává nejen rozdíly (intervaly) mezi hodnotami, ale také podíly hodnot, neboť má nulu stanovenu absolutně a jednoznačně. Veličiny měřené v podílové škále mohou nabývat pouze kladných hodnot. Veličinám měřeným v podílové škále se říká také *kardinální veličiny*.

Příklad 1-7:
Podílovou škálou je např. Kelvinova teplotní stupnice, v níž všechny naměřené teploty jsou kladné, tzv. absolutní nula, tj. hodnota 0° K je fyzikálně nedosažitelná.

V podílových škálách se měří např.
- rozměry, objem a hmotnost těles,
- koncentrace, kapacity,
- fyzikální vlastnosti materiálu, doba trvání nějakého děje,
- počet mikroorganismů ve vzorku vody,
- počet elementů ve vzorku krve atd.

Veličiny měřené intervalovou nebo podílovou škálou se nazývají *metrické*. Při zpracování metrických dat většinou tyto veličiny považujeme za *spojité*, jako kdyby mohly nabývat kteroukoli hodnotu z číselného intervalu daného škálou, i když při praktickém měření tomu tak není, viz výše uvedené příklady, kdy hodnota se určuje načítáním, a tedy může být jen celočíselná. Dokonce i u veličin, které principiálně spojité jsou, jako délka nebo čas, musíme při praktickém měření volit konečnou jednotku rozlišení, takže i tyto veličiny se měří na diskrétní (nespojité) škále. Přesto však při statistickém zpracování většinou můžeme užívat pro metrické veličiny postupy matematicky odvozené pro veličiny spojité. Pro nominální a ordinální veličiny se naopak užívají techniky odvozené pro veličiny *diskrétní*, tj. veličiny nabývající jen určité od sebe vzdálené hodnoty. Obvykle takových možných hodnot nespojité veličiny bývá jen nevelký počet.

Shrnutí
- Data jsou zobrazením části reálného světa, většinou jsou vyjádřena číselně.
- Základní soubor jsou data zobrazující celou populaci. Jeho analýzou získáme přehledněji uspořádaný popis dat. Takový postup se označuje jako *popisná (deskriptivní) statistika*.
- Výběrový soubor jsou data zobrazující pouze část populace. Z výběrového souboru nemůžeme určit parametry populace, pouze jejich *odhady*.
- Metody správného usuzování z výběru na populaci, poskytuje *matematická statistika*.

- K číselnému vyjadřování vlastností jedinců (objektů) slouží měření. Měřením zjistíme pro jistý objekt číselnou hodnotu sledované veličiny, tím vytvoříme obraz objektu na číselné ose.
- Škálou jsou vymezeny všechny možné hodnoty, které měřená veličina může nabývat. Podle typu škály jsou definovány vztahy mezi hodnotami na škále.
- *Nominální škála* klasifikuje objekty do určitých předem vymezených kategorií. Mezi různými hodnotami není definováno žádné uspořádání. O veličinách měřených v nominální škále hovoříme jako o *nominálních veličinách.*
- *Ordinální* (*pořadová*) *škála* umožňuje jedince podle sledované vlastnosti nejen rozlišovat, ale také uspořádat ve smyslu vztahů "je větší", "je menší" nebo "předchází", "následuje", aniž by však byla schopna vyjádřit číselně vzdálenost mezi větším a menším či mezi předcházejícím a následujícím. Veličiny měřené v ordinálních škálách se nazývají *ordinální veličiny*. Nominální a ordinální veličiny jsou souhrnně označovány jako *kategoriální*.
- *Intervalová škála* umožňuje stanovit vzdálenost mezi hodnotami měřené veličiny. Má definovánu jednotku měření. Dovoluje počítat s rozdíly naměřených hodnot, nikoliv s jejich podíly.
- *Podílová škála* zachovává nejen rozdíly (intervaly) mezi hodnotami, ale také podíly hodnot, neboť má nulu stanovenu absolutně a všechny naměřené hodnoty jsou kladné.
- Veličiny měřené intervalovou nebo podílovou škálou se nazývají *metrické*. Při zpracování metrických dat většinou tyto veličiny považujeme za *spojité*. Pro nominální a ordinální veličiny se užívají techniky pro veličiny *diskrétní*.

Kontrolní otázky:

1. *Co je nejobvyklejší datová struktura v analýze dat?*
2. *Jaký význam mají v tabulce řádky a sloupce?*
3. *Charakterizujte pojmy základní soubor, výběrový soubor.*
4. *Vysvětlete rozdíl mezi škálou nominální, ordinální, intervalovou a podílovou.*

Pojmy k zapamatování:

- statistická data
- objekt, veličina
- škála
- základní soubor
- výběrový soubor
- deskriptivní statistika
- induktivní statistika

2 Popisná statistika

Tato kapitola je poměrně obsáhlá, proto se dělí do více částí. K prostudování celé této kapitoly budete potřebovat asi 10-12 hodin. Studium vám ulehčí četné ilustrativní příklady. K této kapitole se váže první korespondenční úkol.

2.1 Četnost, rozdělení četnosti, grafické znázornění

Cíl:
Po prostudování této části kapitoly byste měli umět:

- chápat rozdíly mezi absolutní a relativní četností,
- chápat, co je kumulativní četnost,
- graficky znázornit rozdělení četnosti.

Průvodce studiem:
Čas potřebný k prostudování základního učiva této části je asi 4 hodiny.

Nejprve se zabývejme diskrétními veličinami.

Příklad. 2-1: Pozorováním hnízd jistého druhu ptáků ve vymezené lokalitě byly zjištěny následující počty mláďat v jednotlivých hnízdech, tj. hodnoty x_j, $j = 1, 2, \ldots, n$:

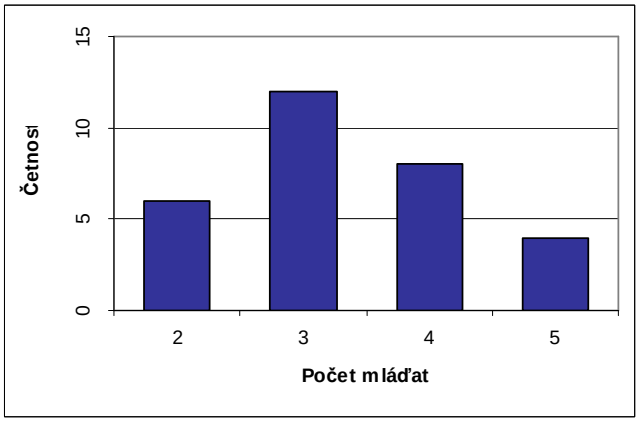
3, 4, 3, 5, 2, 3, 4, 2, 3, 5, 3, 4, 2, 5, 3, 3, 3, 4, 5, 2, 2, 2, 3, 3, 4, 4 ,3, 3, 4, 4

Uvedená řada 30 čísel obsahuje všechny pozorované informace, ale jejich vnímání je dosti obtížné. Porovnané údaje však můžeme snadno zpřehlednit. Uspořádejme data do následující tabulky, kde *i* je pořadové číslo, tj. index řádku tabulky, x_i^* je pozorovaná hodnota, n_i je počet hodnot x_i^*.

Tab. 2-1. Absolutní četnosti hodnot

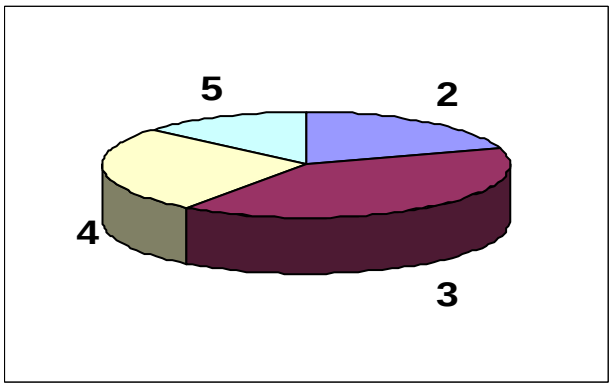
i	x_i^*	n_i
1	2	6
2	3	12
3	4	8
4	5	4
Celkem		$n = 30$

Tabulka obsahuje všechny informace jako řada čísel ve výše uvedeném příkladu (s výjimkou pořadí, ve kterém byly hodnoty zaznamenány), ale je pro vnímání podstatně snadnější. Navíc informace z tab. 2-1 můžeme snadno vyjádřit graficky, např. tak, že pro každou hodnotu x_i^* znázorníme hodnotu n_i výškou sloupečku (obr. 2-1).



Obr. 2-1: Sloupcový graf (bar plot)

Někdy se užívají pro grafické znázornění četnosti také výsečové grafy (pie plots), v nichž je četnost znázorněna plochou kruhové výseče (obr. 2-2). Tyto grafy mají v oblibě zejména novináři, v barevných variantách vypadají efektně. Jsou však méně informativní než sloupcové grafy, a proto se pokud možno jejich užívání v seriózních prezentacích vyhněte.



Obr. 2-2: Výsečový graf – četnosti počtu mláďat

Hodnoty n_i nazýváme *absolutními četnostmi*. Přívlastek „absolutní“ bývá často vynecháván, takže slyšíme-li četnost, chápeme to jako počet hodnot x_i^* zjištěný v datech, tedy absolutní četnost. Vidíme, že celkový počet všech pozorovaných údajů n je rozdělen (rozložen) mezi jednotlivé diskrétní pozorované hodnoty. Můžeme tedy hovořit o *rozdělení četnosti*. Platí triviální vztah

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

kde k je počet různých hodnot x_i zjištěných v datech. V uvedeném příkladu je $k = 4$. Tab. 2-1 můžeme nyní dále rozšířit - viz tab. 2-2.

Tab. 2-2: Relativní a kumulativní četnosti

i	x_i^*	n_i	f_i	N_i	F_i
1	2	6	6/30 = 0.20	6	0.20
2	3	12	12/30 = 0.40	18	0.60
3	4	8	8/30 = 0.27	26	0.87
4	5	4	4/30 = 0.13	30	1.00
Celkem		30	30/30 =1		

Tím jsme se dostali k dalším možnostem vyjadřování četností. Symbol f_i označuje *relativní četnost* definovanou jako

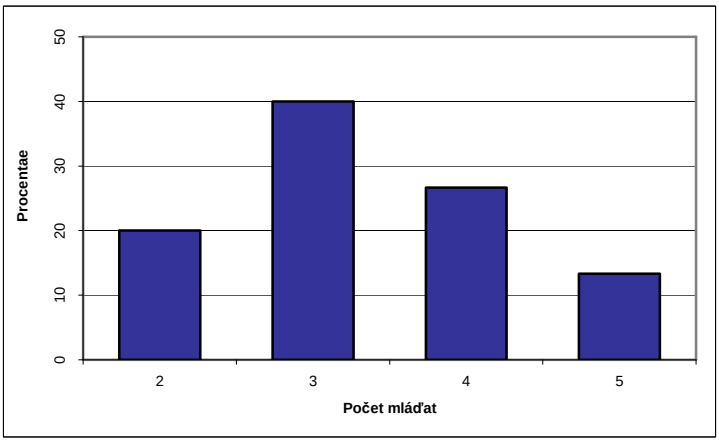
$$f_i = \frac{n_i}{n} ,$$

což představuje podíl počtu hodnot x_i^* v celkovém počtu všech pozorovaných hodnot. Ve sloupečku N_i jsou *kumulativní absolutní četnosti*, ve sloupečku F_i pak *kumulativní relativní četnosti*. Relativní kumulativní četnost F_i je definována jako podíl všech hodnot x_j, pro které platí $x_j \le x_i^*$. Spočítá se tak, že sečteme všechny relativní četnosti až do řádku *i*. Formálně to můžeme zapsat

$$F_i = \sum_{j=1}^{i} f_j .$$

Je zřejmé, že $f_i = F_i - F_{i-1}$. Analogické vztahy platí i pro absolutní kumulativní četnosti. Platí, že $F_i = \frac{N_i}{n}$.

Graf relativních četností je podobný grafu absolutních četností, jediná odlišnost je v měřítku svislé osy - viz obr. 2-3.



Obr. 2-3: Sloupcový graf relativních četností v procentech

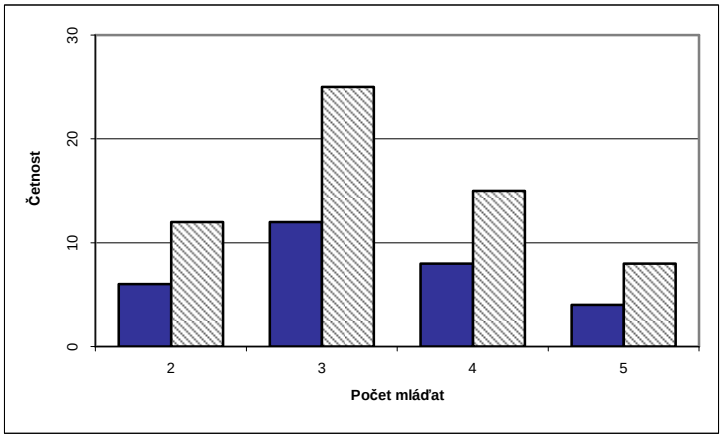
Opět vidíme, že relativní četnosti jsou rozděleny mezi jednotlivé pozorované hodnoty, ona jednička na řádku *Celkem* v tab. 2-2, která je součtem relativních četností, je rozložena podle podílu pozorovaných hodnot. Užitečnost relativních četností ukážeme dále, viz př. 2-2.

Příklad 2-2: V jiné lokalitě byly pozorovány tyto počty mláďat:

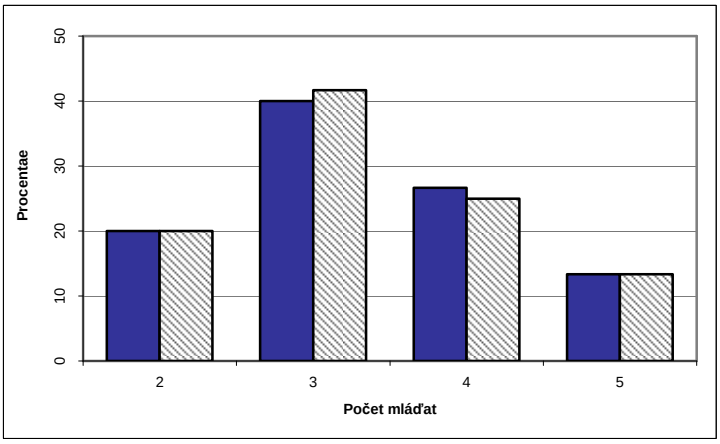
i	x_i^*	n_i
1	2	12
2	3	25
3	4	15
4	5	8
Celkem		$n = 60$

Porovnejte rozložení četností mláďat v obou lokalitách.

Pokud bychom zůstali u grafického znázornění absolutních četností, dostaneme graf na obr. 2-4. Četnosti se zřetelně liší, ale je tento závěr správný?



Obr. 2-4: Absolutní četnosti - srovnání dvou skupin



Obr. 2-5: Relativní četnosti v procentech - srovnání dvou skupin

Porovnáme-li relativní četnosti, dostaneme graf na obr. 2-5. Vidíme, že rozložení četností v obou lokalitách je velmi podobné. Prozatím se spokojíme s tímto subjektivním dojmem. Zda velmi podobné rozdělení četností znamená „prakticky stejné“ rozdělení četností, nemůžeme prostředky popisné statistiky objektivně

rozhodnout. K tomu potřebujeme znát jiné techniky, kterými se budeme zabývat v kapitole 4 a také v dalším semestru.

O trochu složitější je situace, kdy se zabýváme rozdělením četností v souvislosti se *spojitou veličinou* - viz př. 2-3.

Příklad 2-3: Hmotnost okurek (v gramech) posbíraných z pokusného záhonu byla následující:

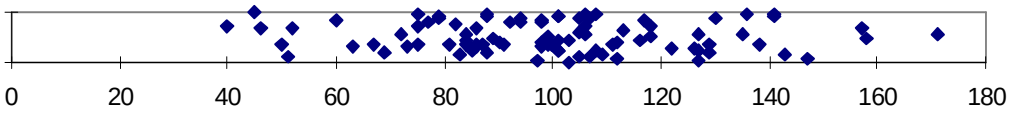
98	106	84	88	84	69	75	105	98	88	111	99
84	135	100	107	108	87	171	90	40	143	81	136
73	126	86	82	101	113	116	141	91	127	127	98
112	118	101	67	83	86	117	84	97	105	157	86
72	101	103	112	101	46	108	147	129	92	106	103
122	63	75	94	50	158	98	106	107	77	130	51
109	45	101	141	127	85	94	138	52	99	88	105
79	60	89	79	75	118	129					

Jaké je rozdělení četností?

Naměřené údaje můžeme graficky znázornit na číselné ose jako tzv. *diagram rozptýlení* - obr. 2-6:

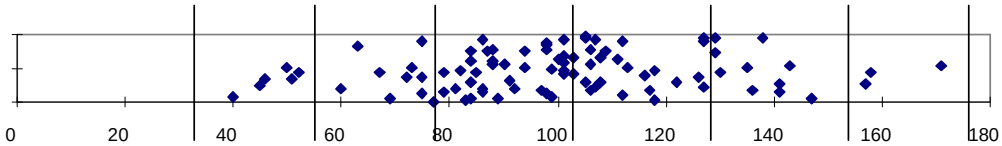
Obr. 2-6: Znázornění naměřených hodnot spojité veličiny - diagram rozptýlení

Vidíme, že v intervalu mezi nejmenší a největší porovnanou hodnotou jsou naměřené hodnoty různě *husté*, s největší hustotou v našem případě kolem prostředku intervalu, ale graf na obr. 2-6 příliš přehledný není, např. nemůžeme rozlišit, zda vyznačený bod na číselné ose znamená jednu či více naměřených hodnot. Mírného zlepšení dosáhneme tím, že naměřené body místo na číselnou osu znázorníme do obdélníku, ve kterém výšku zobrazovaného bodu volíme *náhodně*. Dostaneme tak *rozmítnutý diagram rozptýlení* (dot plot)- obr. 2-7:



Obr. 2-7: Znázornění naměřených hodnot spojité veličiny - rozmítnutý diagram rozptýlení

Ale zobrazené rozdělení četností stále není dost názorné. Nabízí se však další jednoduchý postup: Vyznačit na číselné ose hranice intervalů, viz obr. 2-8, a zjistit četnosti hodnot v každém intervalu.



Obr. 2-8: Znázornění naměřených hodnot spojité veličiny - intervaly

Dostaneme tak k intervalů (tříd), každý interval má šířku h_i, dolní hranici l_i, horní hranici u_i a svůj střed c_i. Z obr. 2-8 je zřejmé, že platí triviální vztahy

$$h_i = u_i - l_i, \qquad c_i = \frac{l_i + u_i}{2} = l_i + \frac{h_i}{2} = u_i - \frac{h_i}{2}, \qquad \text{pro} \quad i = 1, 2, ..., k$$

a $\quad l_i = u_{i-1} \quad$ pro $\; i = 2, 3, \ldots, k$

Prozatím jsme se nezabývali tím, jak volit počet a hranice intervalů a kam patří naměřená hodnota, která leží přesně na hranici dvou intervalů. U diskrétní veličiny jsme tyto problémy neměli, zde u spojité veličiny musíme tato svá subjektivní přání vyslovit, chceme-li naměřená data rozdělit do tříd podle příslušnosti k intervalům. Většinou se šířka všech intervalů volí shodná, tzn. $h_i = h$ pro $i = 1, 2, \ldots, k$. Pak hovoříme o *ekvidistantním* rozdělení tříd (intervalů). Počet intervalů by neměl být ani příliš malý (jeden interval nevypoví o rozdělení četnosti naměřených hodnot nic, dva intervaly málo), ani příliš velký (četnosti naměřených hodnot v intervalech by byly malé a tedy příliš silně ovlivněny náhodným kolísáním). Většinou je vhodné volit počet intervalů k někde mezi 5 a 20 s přihlédnutím k počtu naměřených hodnot n. V literatuře lze nalézt různé vztahy, které umožňují určit vhodný počet intervalů, např.

$$k = 1 + \log_2(n) \cong 1 + 3{,}3 \log_{10}(n),$$

kde $\log_2(n)$ znamená logaritmus n při základu *2*, $\log_{10}(n)$ je dekadický logaritmus.

Naměřená hodnota ležící na hranici intervalů by mohla být zařazena do kteréhokoli z obou sousedících intervalů. Většina programových prostředků, které nám pomáhají třídní uspořádání dat pohodlněji realizovat, zařazuje hraniční bod do levého intervalu, tedy do *i*-tého intervalu patří všechny naměřené hodnoty x_j, pro které platí $l_i < x_j \le u_i$. Z obr. 2-8 pak vidíme, že dolní hranice prvního intervalu l_1 musí být alespoň o trochu menší než nejmenší pozorovaná hodnota $x_{\min}$, tedy $l_1 = x_{\min} - \varepsilon_1$, $\varepsilon_1 > 0$. Podobně horní hranice posledního intervalu u_k může (ale nemusí) být větší než $x_{\max}$, $u_k = x_{\max} + \varepsilon_2$, $\varepsilon_2 \ge 0$. Pak šířku intervalu h určíme podle vztahu

$$h = \frac{u_k - l_1}{k} = \frac{x_{\max} + \varepsilon_2 - (x_{\min} - \varepsilon_1)}{k}$$

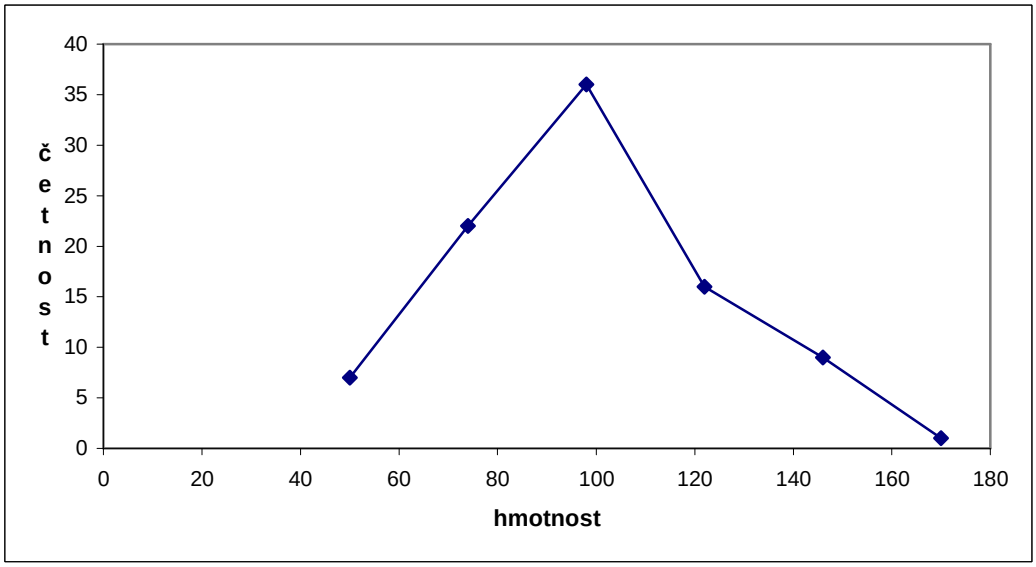
Hodnoty ε_1, ε_2 se většinou snažíme volit tak, aby hranice intervalu byly co nejzaokrouhlenější číselné hodnoty.
Předchozí poněkud zdlouhavé odstavce popisovaly jednoduchá přijatelná pravidla k řešení problémů spojených s rozdělením hodnot spojité veličiny do tříd.

Nyní se konečně můžeme vrátit k dořešení příkladu 2-3. Počet intervalů je $k = 1 + 3{,}3 \log_{10}(91) \cong 6$. Dostaneme tedy následující tabulku:

Tab. 2-3: Data z příkladu 2-3 uspořádaná do tříd

i	l_i	u_i	c_i	n_i	f_i
1	38	62	50	7	0.077
2	62	86	74	22	0.241
3	86	110	98	36	0.396
4	110	134	122	16	0.176
5	134	158	146	9	0.099
6	158	182	170	1	0.011
Celkem				91	1.000

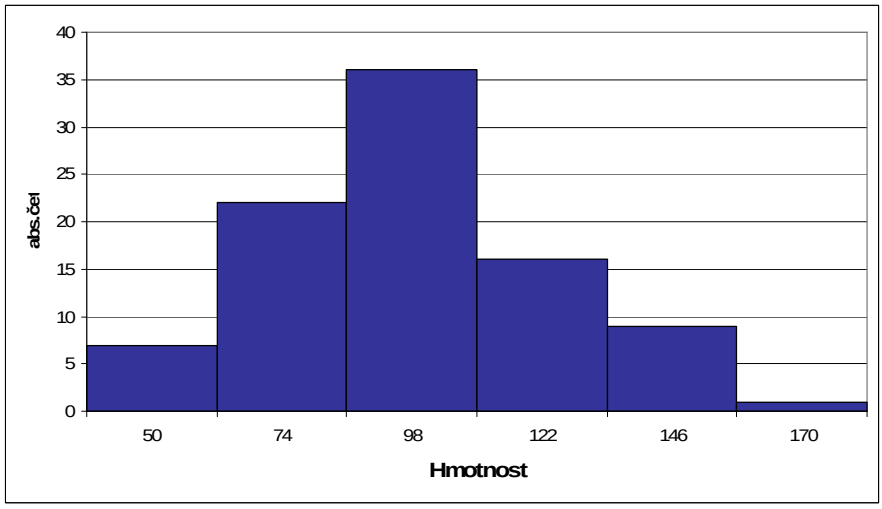
Informaci z tab. 2-3 můžeme přehledně zobrazit graficky. Pokud proti středům intervalu c_i vyneseme odpovídající četnosti n_i a body spojíme úsečkou, dostaneme *četnostní polygon* - obr. 2-9.



Obr. 2-9: Četnostní polygon

Zobrazíme-li četnosti v intervalech $\langle l_i, u_i \rangle$ vodorovnými úsečkami a vyznačíme sloupce pod těmito úsečkami, dostaneme *histogram*, viz obr. 2-10.

Pozor: Pokud ke kresbě histogramu užijeme Excel, položka *Histogram* v doplňku *Analýza dat*, dostaneme graf, ve kterém histogram není nakreslen bezvadně. Histogram zobrazuje rozdělení hodnot spojité veličiny, proto sloupce nemají být odděleny mezerami. proto před zařazením histogramu do prezentace výsledků je třeba obrázek patřičně upravit.

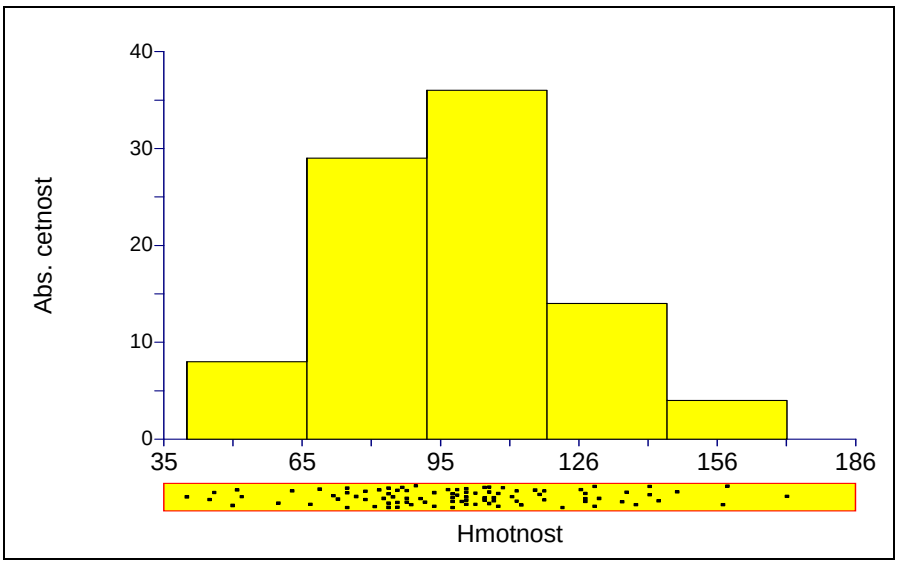


Obr. 2-10: Histogram

Všimněme si také, jak tvar histogramu je závislý na zvoleném počtu tříd (6 tříd na obr. 2-10, 5 tříd na obr. 2-11).

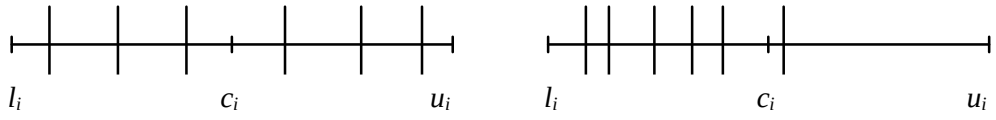
Histogram je nejčastěji používaný prostředek pro popis rozdělení četností hodnot spojité veličiny. V grafech na obr. 2-9 až 2-11 jsme místo absolutních četností n_i mohli užít relativní četnosti f_i. Tvar grafů by opticky samozřejmě zůstal stejný, jediná odlišnost by byla v měřítku svislé osy.

Znovu připomeňme souvislost tvaru histogramu s *hustotou* naměřených hodnot zobrazených na číselné ose. Čím vyšší počet bodů v intervalu (tedy čím je větší jejich hustota), tím je vyšší sloupeček histogramu - viz obr. 2-11, na kterém je kromě histogramu i rozmítnutý diagram rozptýlení:



Obr. 2-11: Histogram a diagram rozptýlení

Histogramy nám umožňují prezentovat rozdělení četností hodnot spojité veličiny přehlednou a snadno vnímatelnou formou - srovnej nepřehlednou řadu čísel v zadání př. 2-3 a histogram na obr. 2-10 nebo 2-11. Jak už to však v životě chodí, zpravidla tím, že něco získáme, většinou i něco ztrácíme. Zpracováním naměřených hodnot do tříd (tab. 2-3) ztrácíme informaci o tom, jak jsou data rozdělena uvnitř intervalů. Např. data v intervalech na obr. 2-12 a, b vedou ke stejné četnosti $n_i = 6$ a v obou případech je tato šestice naměřených bodů reprezentována středem intervalu c_i, což v případě b není nejpříhodnější reprezentant.



a) přibližně symetrické b) silně nesymetrické

Obr. 2-12: Rozdělení hodnot uvnitř intervalů

Naštěstí situace na obr. 2-12b představuje krajnost velmi nesymetrického rozdělení hodnot uvnitř intervalu, o které můžeme doufat, že se v empirických datech nevyskytuje příliš často.
Na závěr tohoto odstavce ještě potěšující poznámka: Popsané zpracování dat do intervalů a jejich grafické znázornění formou histogramů možná vyvolává představu nepřiměřené pracnosti a časové náročnosti. Máme však k dispozici celou řadu programových prostředků (tabulkové procesory, statistické programy), které tuto činnost velmi usnadňují a znalosti získané v tomto odstavci by měly usnadnit jejich ovládání a porozumění výsledkům.

Shrnutí:

- Pozorovaná data lze zpřehlednit uspořádáním do tabulky četností. Informace z tabulky můžeme vyjádřit graficky.
- *Absolutní četnost* n_i je počet hodnot x_i^*, zjištěný v datech.
- Počet všech pozorovaných údajů n je rozdělen (rozložen) mezi jednotlivé diskrétní pozorované hodnoty, hovoříme o *rozdělení četnosti*.
- *Relativní četnost* f_i je podíl počtu hodnot x_i^* z celkového počtu všech pozorovaných hodnot.
- Rozdělení spojité veličiny můžeme zobrazit *histogramem.*

Kontrolní otázky:

1. *Vysvětlete pojmy absolutní a relativní četnost.*
2. *Lze z výšky sloupců histogramu poznat, kde je hustota naměřených hodnot na číselné ose větší a kde je nízká?*

Pojmy k zapamatování:

- četnost absolutní a relativní
- kumulativní četnost
- sloupcový graf
- hustota naměřených hodnot
- histogram

2.2 Charakteristiky polohy

Cíl:
Po prostudování této části kapitoly byste měli vědět:
- co to je charakteristika polohy,
- základní vlastnosti aritmetického průměru,
- další charakteristiky polohy, jako medián, modus,
- co je to kvantil,
- co je uřezávaný průměr,
- co je geometrický průměr a kdy se používá.

Průvodce studiem:
Čas potřebný k prostudování základního učiva této části asi 3 hodiny

Charakteristikou polohy rozumíme takovou číselnou hodnotu, která vystihuje umístění pozorovaných hodnot na číselné ose. Z pohledu na obr. 2-6 je zřejmé, že to bude nějaké číslo z intervalu $\langle x_{\min}, x_{\max} \rangle$. Otázkou je, které číslo z tohoto intervalu *nejlépe* charakterizuje polohu pozorovaných hodnot na číselné ose a jakým postupem ho určit. Jedna z možností je polohu dat charakterizovat jejich těžištěm - viz obr. 2-13.

Obr. 2-13: Průměr je poloha „těžiště“ naměřených hodnot

Každou z naměřených hodnot si můžeme představit jako závaží jednotkové hmotnosti umístěné na dvojzvratné páce v místě, které odpovídá naměřené hodnotě, a hledáme polohu bodu, kolem kterého je tato páka v rovnováze. Takovou charakteristikou polohy je *průměr* (aritmetický průměr), $\bar{x}$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (2\text{-}1)$$

Průměr $\bar{x}$ je taková hodnota, která má tu vlastnost, že součet odchylek naměřených hodnot od průměru je roven nule (vyjádření rovnováhy na obr. 2-13 - součet momentů se rovná nule), $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$

Důkaz: $\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\,\bar{x} = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0.$ □

Další vlastnost průměru $\bar{x}$ je to, že suma čtverců (druhých mocnin) odchylek od průměru je minimální, tj. suma čtverců odchylek od jiné číselné hodnoty je větší.

Důkaz: Nechť $a \neq 0$. Pak $\bar{x} + a \neq \bar{x}$. Spočítejme tedy součet čtverců odchylek od čísla $\bar{x} + a$:

$$\sum_{i=1}^{n} [x_i - (\bar{x} + a)]^2 = \sum_{i=1}^{n} \left[(x_i - \bar{x}) - a \right]^2 = \sum_{i=1}^{n} \left[(x_i - \bar{x})^2 - 2a(x_i - \bar{x}) + a^2 \right] =$$

$$= \sum_{i-1}^{n} (x_i - \bar{x})^2 - 2a \sum_{i=1}^{n} (x_i - \bar{x}) + na^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 + na^2$$

Jelikož na^2 je vždycky kladné, je tedy součet čtverců odchylek od průměru minimální. □

Jsou-li data uspořádána v tabulce spolu s četnostmi (viz odst. 2.1), pak průměr můžeme snadno spočítat jako

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{k} n_i x_i^* = \sum_{i=1}^{k} f_i x_i^* , \qquad (2\text{-}2)$$

kde n je celkový počet naměřených hodnot $n = \sum_{i=1}^{k} n_i$, k je počet navzájem různých naměřených hodnot v případě diskrétní veličiny nebo počet intervalů v případě spojité veličiny (v obou případech je k počet řádků v tabulce četností) a n_i jsou absolutní, f_i relativní četnosti hodnot x_i^* v datech. O průměru počítaném podle (2-2) hovoříme jako o *váženém průměru*. Každá hodnota je vážena svou četností, tedy čím větší četnost, tím větší vliv na hodnotu průměru.
Pozorný čtenář si jistě povšimnul, že v případě, kdy tabulka četností vznikla uspořádáním hodnot spojité veličiny do k intervalů, se mohou hodnoty průměru spočítané podle vztahu (2-1) a (2-2) lišit. Do (2-2) za x_i^* dosazujeme hodnotu středu i-tého intervalu, tedy c_i, a jak víme, tato hodnota nemusí být vždy dobrým reprezentantem hodnot patřících do i-tého intervalu. Podmínkou k tomu, aby vážený průměr počítaný podle vztahu (2-2) byl roven průměru (2-1), tedy přesný, je, aby

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{k} n_i c_i$$

Naštěstí u většiny empirických dat je rozdělení hodnot uvnitř intervalu zhruba rovnoměrné, takže uvedený vztah bývá splněn s dostatečnou přesností a vážený průměr spočítaný podle (2-2) se od správné hodnoty průměru podle (2-1) liší nepodstatně. Průměr je vhodná charakteristika polohy tehdy, když je pro nás zajímavý i součet naměřených hodnot.

Příklad 2-4:
Je-li průměrná mzda 6 zaměstnanců firmy 10 000 Kč, pak celková měsíční vyplacená částka činí 6 x 10 000 = 60 000 Kč.
Průměr je však velice citlivý na odlehlé hodnoty (odlehlá hodnota je hodnota velmi vzdálená od průměru). Představte si, že v předchozím příkladu byly mzdy našich zaměstnanců 7 000, 8 000, 9 000, 11 000, 12 000, 13 000. Pak průměr opravdu je charakteristikou mzdy zaměstnanců, i když žádný z nich tuto průměr-

nou částku nedostává, avšak všichni mají mzdy poměrně blízké průměru, část jedinců o něco nižší, část o něco vyšší. Co se však stane, když váš nejlépe placený zaměstnanec bude mít místo 13 000 Kč mzdu ve výši 73 000 Kč? Pak $\sum x_i = 120000$ a průměr bude 20 000 Kč, tedy hodnota vzdálená jak od běžně placených pěti zaměstnanců s obvyklým příjmem, tak i od výjimečného platu experta.

Hodnota průměru je silně ovlivněna jednou odlehlou pozorovanou hodnotou. Průměr může dobře posloužit pro určení sumy měsíčně vyplácených peněz, ale o mzdě běžného zaměstnance nevypovídá téměř nic. Proto se užívají i jiné charakteristiky polohy, než je průměr.

Takovou jednoduchou charakteristikou je *modus*, $\hat{x}$. Užívá se především pro diskrétní veličiny a je definován jako hodnota, která je v datech nejčetnější. Tato definice nezaručuje, že modus je definován jednoznačně, v datech může být dvě nebo více hodnot, jejichž četnosti jsou shodné a současně žádná jiná hodnota není četnější. Pak říkáme, že data mají *bimodální* nebo *vícemodální* rozdělení. Modus je jediná charakteristika polohy vhodná pro nominální veličiny.

Další charakteristikou polohy je *medián*, $\tilde{x}$. Je to hodnota, která je uprostřed, uspořádáme-li naměřené hodnoty podle jejich velikosti. Počet hodnot menších než medián je stejný jako počet hodnot větších než medián.

Příklad 2-5:
Naměřené hodnoty jsou 15, 17, 20, 11, 14.
Uspořádáme je vzestupně: 11, 14, <u>15</u>, 17, 20. Medián je hodnota uprostřed, tedy $\tilde{x} = 15$.

Příklad 2-6:
Naměřené hodnoty jsou 15, 17, 21, 20, 11, 14.
Uspořádáme je vzestupně: 11, 14, <u>15</u>, <u>17</u>, 20, 21. Pokud je počet hodnot sudý, pak medián je průměr ze dvou prostředních hodnot, tedy $\tilde{x} = \frac{1}{2}(15+17) = 16$.

Oproti průměru má medián výhodu, že není citlivý na odlehlé hodnoty.

Příklad 2-7:
Pro data 11, 14, 15, 17, 20 je medián 15 a bude stejný i pro data 11, 14, 15, 17, 200, zatímco hodnota průměru se změní z 15,4 na 51,4.

Medián je vhodnou charakteristikou polohy pro ordinální veličiny, u nich by neměl být užíván průměr. Proto například běžně užívané studijní průměry v hodnocení žáků a studentů lze jen stěží brát vážně, neboť klasifikace má ordinální škálu, nemůžeme říci, že vzdálenost ve vědomostech mezi jedničkářem a dvojkařem je stejná jako mezi dvojkařem a trojkařem. Proto celkový prospěch by měl být hodnocen spíše mediánem než průměrem.

Analogicky můžeme zavést další charakteristiky založené na relativní četnosti hodnot v datech, které jsou menší nebo rovny této charakteristice. Označme tuto relativní četnost (podíl) *p*, $0 \le p \le 1$, a příslušnou charakteristiku *x(p)*. Pro medián bylo *p* rovno jedné polovině, tedy 0,5 a místo $\tilde{x}$ bychom mohli psát *x(0,5)*. Hodnotě *x(p)* se říká *p-kvantil* (nebo také *100p-percentil*). Některé často užívané kvantily mají zvláštní pojmenování:

$x(0{,}5)$	medián, $\tilde{x}$
$x(0{,}25)$, $x(0{,}75)$	dolní kvartil, horní kvartil
$x(0{,}1)$, $x(0{,}9)$	dolní decil, horní decil

Dolní kvartil určíme jako medián „dolní poloviny“ dat, horní kvartil jako medián „horní poloviny“ dat.

Příklad 2-8:
Pro data z předchozích příkladů 2-5 a 2-6 jsou dolní a horní kvartil podtržené hodnoty:

11, <u>14</u>, 15, <u>17</u>, 20 (Pokud počet hodnot je lichý, medián „patří“ jak do dolní, tak i do horní poloviny dat).

11, <u>14</u>, 15, 17, <u>20</u>, 21

K určení pořadí hodnoty, která je *p*-kvantilem můžeme užít jednoduchého vztahu
$z_p = np + 0.5$,
kde z_p je pořadí hodnoty v uspořádané posloupnosti $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$. Pokud z_p nevyjde celé, interpolujeme hledaný kvantil ze sousedních hodnot. Máme-li data uspořádaná do tříd, pak podle z_p a kumulativních četností určíme interval, ve kterém se hledaný kvantil nachází (tj. platí $N_{i-1} < z_p \le N_i$), a pak určíme *p*-kvantil lineární interpolací

$$x(p) = \frac{z_p - N_{i-1}}{n_i} h_i + l_i$$

Ve většině statistických programů se užívá poněkud pracnější, ale přesnější postup k určení p-kvantilu. Pozorované hodnoty se uspořádají do neklesající posloupnosti, $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$. Pak p-kvantil je určován podle vztahu

$$x(p) = (n+1)\left(p - \frac{i}{n+1}\right)\left(x_{(i+1)} - x_{(i)}\right) + x_{(i)}$$

kde hodnoty $x_{(i)}$ a $x_{(i+1)}$ se určují tak, aby platilo $\frac{i}{n+1} < p \le \frac{i+1}{n+1}$.

Kompromisem mezi průměrem a mediánem jsou různé tzv. *robustní* charakteristiky polohy, které jsou nyní díky dostupnosti statistického programového vybavení stále častěji využívány. Povětšině jsou založeny na myšlence, že hodnoty vzdálené od mediánu mají mít ve výpočtu součtu pro průměr menší váhu. Zde uvedeme jen tzv. *α-uřezávaný průměr (*angl. *trimmed mean*). Vypočítává se tak, že se spočte průměr z *n (1-2α)* „vnitřních“ bodů, nejmenších $n\alpha$ hodnot a největších $n\alpha$ hodnot se prostě „uřízne“. Uříznuté body mají při výpočtu součtu hodnot váhu 0, všechny ostatní váhu 1. Medián je vlastně speciálním případem uříznutého průměru, kdy uřízneme všechny hodnoty až na jednu, je-li počet naměřených hodnot lichý, nebo až na dvě, je-li tento počet sudý.
Je zřejmé, že uříznutý průměr není citlivý na odlehlé hodnoty. Ze srovnání „obyčejného“ aritmetického průměru s uříznutým průměrem, případně s mediánem můžeme usuzovat o existenci či neexistenci odlehlých hodnot v datech.

V úvodu odst. 2.2 jsme říkali, že aritmetický průměr je vhodnou charakteristikou polohy v situaci, kdy je pro nás zajímavý i součet naměřených hodnot. V řadě úloh tato situace nenastává. Např. ekonomický vývoj bývá charakterizován tzv. tempy růstu. To znamená, že hodnota tohoto ukazatele v daném období se určuje poměrně ke stavu v období předchozím.

Příklad 2-9:
V období 1999 - 2005 bylo dosaženo objemu výroby Y_i a vypočtena tempa růstu x_i:

rok	i	Y_i	x_i
1999	0	1550	
2000	1	1535	0.99
2001	2	1228	0.80
2002	3	1105	0.90
2003	4	1215	1.10
2004	5	1361	1.12
2005	6	1525	1.12

Ptáme se, jaké bylo průměrné tempo růstu v tomto období. Je přirozené, že požadujeme, aby tato charakteristika měla tu vlastnost, že při každoročním průměrném růstu dosáhneme úrovně pozorované v roce 2005.

Tempa růstu jsou vypočtena jako $x_i = \frac{Y_i}{Y_{i-1}}$ pro $i = 1, 2, \ldots, n$.

V našem příkladu n = 6.

Platí tedy

$$Y_n = Y_{n-1}x_n = Y_0 x_1 x_2 \ldots x_n = Y_0 \prod_{i=1}^{n} x_i$$

Celkové tempo růstu za celé období je $\frac{Y_n}{Y_0} = \prod_{i=1}^{n} x_i$

Průměrné tempo růstu je pak taková hodnota $\bar{x}_G$, pro kterou platí

$(\bar{x}_G)^n = \prod_{i=1}^{n} x_i$ a tedy

$$\bar{x}_G = \sqrt[n]{\prod_{i=1}^{n} x_i} \qquad (2\text{-}3)$$

Charakteristice $\bar{x}_G$ říkáme *geometrický průměr*. Je vhodnou charakteristikou polohy tam, kde nás zajímá také součin pozorovaných hodnot (viz předchozí příklad) - hodnoty Y_n by bylo dosaženo také, kdyby každoroční tempo růstu bylo rovno $\bar{x}_G \doteq 0.997$. Aritmetický průměr tempa růstu v uvedeném příkladu je roven 1.01, přestože koncová hodnota Y_n je menší než počáteční hodnota Y_0. Ze vztahu (2-3) je zřejmé, že geometrický průměr můžeme užít pouze tehdy, když všechny pozorované hodnoty x_i jsou kladné ($x_i > 0$ pro $i = 1, 2, \ldots, n$).

Shrnutí:

- *Aritmetický průměr* je charakteristika polohy, jejíž hodnota má tu vlastnost, že součet odchylek naměřených hodnot od průměru je roven nule.
- Suma čtverců (druhých mocnin) odchylek od průměru je minimální.
- *Modus* je charakteristika polohy užívaná především pro diskrétní veličiny a je definován jako hodnota, která je v datech nejčetnější. Modus je jediná charakteristika polohy vhodná pro nominální veličiny.
- Další charakteristikou polohy je *medián* . Je to hodnota, která je uprostřed, uspořádáme-li naměřené hodnoty podle jejich velikosti. Počet hodnot menších než medián je stejný jako počet hodnot větších než medián.
- Hodnotě *x(p)* pod kterou leží *np* hodnot, se říká *p-kvantil* (nebo také *100p-percentil*).
- Některé často užívané kvantily mají zvláštní pojmenování: *medián, dolní kvartil, horní kvartil,* dolní decil, horní decil.
- Geometrický průměr je vhodnou charakteristikou tam, kde nás zajímá také součin pozorovaných hodnot.

Kontrolní otázky:

1. *Vysvětlete pojem charakteristika polohy.*
2. *Dokažte, že součet odchylek naměřených hodnot od průměru je roven nule.*
3. *Proč medián není citlivý na odlehlé hodnoty?*
4. *Co usoudíte o datech, pro která hodnota průměru je silně odlišná od mediánu a uříznutého průměru?*
5. *Co je dolní kvartil, co je horní kvartil?*

Pojmy k zapamatování:

- charakteristika polohy
- aritmetický průměr
- modus
- medián
- dolní kvartil, horní kvartil
- *p*-kvantil
- geometrický průměr

2.3 Charakteristiky variability

Cíl:
Po prostudování této kapitoly byste měli:

- chápat, co to je charakteristika variability a jak se liší od charakteristiky polohy,
- umět spočítat a interpretovat rozptyl a směrodatnou odchylku.

Průvodce studiem:
Prostudování této části kapitoly budete muset věnovat asi dvě hodiny.

Začneme příkladem.

Příklad 2-10:
Ve dvou studijních skupinách bylo dosaženo v testu těchto výsledků:

Skupina A:	10	12	15	18	20
Skupina B:	12	14	15	16	18

Průměr v obou skupinách je shodný $\bar{x}_A = \bar{x}_B = 15$, shodné jsou i mediány. Přesto na první pohled vidíme, že hodnoty zjištěné ve skupině A a B jsou odlišné. Abychom mohli tyto rozdíly jednoduše postihnout, potřebujeme ještě jiné charakteristiky než charakteristiky polohy. Vidíme, že odlišnost srovnávaných skupin je v tom, jak (do jaké míry) jsou na číselné ose rozházeny (rozptýleny) hodnoty okolo charakteristiky polohy. Právě tyto odlišnosti můžeme vyjádřit číselně pomocí charakteristik variability (rozptýlenosti, „rozházenosti“) naměřených hodnot.

Při letmém pohledu na data v příkladu nás asi napadne jedna z možných charakteristik variability, totiž rozdíl $x_{\max} - x_{\min}$, říkáme mu *rozpětí.* Tento rozdíl pro skupinu A činí 10, pro skupinu B jen 6, takže variabilita ve skupině A je zřetelně větší. Rozpětí má ovšem tu nevýhodu, že může být ovlivněno jednou extrémně odlišnou hodnotou. Pozorného čtenáře kap. 2.2 napadne další možná charakteristika rozptýlenosti, tzv. *mezikvartilové rozpětí,* $x(0,75) - x(0,25)$. Tato charakteristika variability je výrazně vhodnější, protože není ovlivněna jednou nebo několika málo extrémními hodnotami.

Naproti tomu variabilitu nemůžeme charakterizovat součtem odchylek od průměru, neboť je vždy rovna nule (viz odst. 2-2), takže variabilitu naměřených údajů nepostihuje.

Nejčastěji se užívají charakteristiky variability založené na součtu druhých mocnin (tzv. čtverců) odchylek od průměru. Charakteristika

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \qquad (2\text{-}4)$$

se nazývá *rozptyl*, anglicky *variance*. V některých souvislostech se můžete setkat s označením *výběrový rozptyl*, angl. *sample variance*. Vidíme, že s^2 je vždy větší nebo rovno nule. Nule je rovno jen v případě, kdy všechna x_i jsou konstantní, tedy $x_i = \bar{x}$ pro všechna $i = 1, 2, \ldots, n$. Platí, že čím více jsou data „rozházená“, tím je s^2 větší. To ilustrují hodnoty rozptylu pro výše uvedená data ($s_A^2 = 17$, $s_B^2 = 5$).

Nejužívanější charakteristika variability je odmocnina z rozptylu, tedy

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \qquad (2\text{-}5)$$

která má poněkud podivný název *směrodatná odchylka* (angl. *standard deviation*). Její výhodou oproti rozptylu je to, že má stejný rozměr (je ve stejných měrných jednotkách) jeho naměřené hodnoty x_i a jejich průměr $\bar{x}$.

V některých statistických příručkách a v dokumentaci statistických programových prostředků a kalkulaček se setkáváme ještě s jedním trochu odlišným vztahem pro výpočet rozptylu. Tento rozptyl bývá někdy označován jako *populační rozptyl* a je dán vztahem

$$M_2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \qquad (2\text{-}6)$$

Populační rozptyl M_2 je průměrný čtverec odchylky od průměru. Vidíme, že jediná odlišnost vztahu (2-6) od (2-4) je ve jmenovateli, zde je n místo $(n - 1)$. Platí tedy

$$s^2 = \frac{n}{n-1} M_2$$

a vždy $s^2 > M_2$, ale s rostoucím n se rozdíl $s^2 - M_2$ zmenšuje, takže pro větší hodnoty n je tento rozdíl nepodstatný.

Dvě různé, byť podobné, definice rozptylu občas působí nezkušeným uživatelům statistiky potíže. Obě charakteristiky s^2, M_2, se liší v některých statistických vlastnostech, jejichž vysvětlení přesahuje rámec této kapitoly. Prozatím přijmeme jednoduché praktické doporučení: Váháme-li, zda užít s^2 nebo M_2, tedy vzorec (2-4) nebo (2-6), je lepší užít s^2, tedy vztah (2-4) s $(n - 1)$ ve jmenovateli.

Pro pohodlnější výpočet můžeme tento vztah upravit

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right) =$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2\right] = \qquad (2\text{-}7)$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]$$

Nyní si na příkladu ukážeme, jak počítat rozptyl z dat uspořádaných do tabulky četností:

Příklad 2-11:

x_i	n_i	$n_i x_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$n_i(x_i - \bar{x})^2$
2	6	12	-4/3	16/9	96/9
3	12	36	-1/3	1/9	12/9
4	8	32	2/3	4/9	32/9
5	4	20	5/3	25/9	100/9
Součet	30	100			240/9

$$\bar{x} = \frac{100}{30} = \frac{10}{3} \cong 3{,}33$$

$$s^2 = \frac{1}{29}\cdot\frac{240}{9} = \frac{1}{29}\cdot\frac{80}{3} \cong 0{,}920$$

$$s = \sqrt{0{,}920} \cong 0{,}959$$

Postup můžeme vyjádřit vztahem

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i\left(x_i - \bar{x}\right)^2 \qquad (2\text{-}8)$$

kde k je počet řádků v tabulce četností. V našem příkladu bylo $k = 4$. V součtu čtverců jsou čtverce odchylek pozorovaných hodnot od průměru váženy četností pozorovaných hodnot n_i.

Výpočetní postup se zjednoduší, upravíme-li vzorec (2-8) do tvaru

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{k} n_i x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{k} n_i x_i\right)^2\right] \qquad (2\text{-}9)$$

Postup úprav vztahu (2-8) na (2-9) je podobný jako postup úprav vztahu (2-4) na (2-7).

Pak výpočet rozptylu bude vypadat takto:

x_i	n_i	$n_i x_i$	x_i^2	$n_i x_i^2$
2	6	12	4	24
3	12	36	9	108
4	8	32	16	128
5	4	20	25	100
	30	100		360

$$s^2 = \frac{1}{29}\left(360 - \frac{100\,.\,100}{30}\right) = \frac{1}{29}\,.\,\frac{80}{3} \cong 0{,}920$$

$$s = \sqrt{0{,}920} \cong 0{,}959$$

Vidíme, že ke stejnému výsledku jsme došli méně pracně, ale přece jen to vyžadovalo jakousi námahu. Potěšující je, že v současné době tuto výpočetní námahu většinou nemusíme vynakládat, neboť ji za nás vykonají různé programové prostředky (např. tabulkové procesory jako např. Excel, statistické programy) dostupné prakticky na každém počítači. Důležité však je, abychom si zapamatovali smysl a účel charakteristik variability.

Naše data z příkladu 2-11 můžeme ve zkratce popsat dvěma čísly:

- průměrem $\bar{x} = 10/3 \cong 3{,}33$, který charakterizuje polohu,
- směrodatnou odchylkou $s \cong 0{,}959$, která kvantifikuje variabilitu.

Shrnutí

- Kromě polohy je užitečné charakterizovat také variabilitu dat.
- Nejčastěji užívané charakteristiky variability se počítají ze součtu druhých mocnin odchylek pozorovaných hodnot od průměru.
- Charakteristika $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$ se nazývá *rozptyl*.
- *Směrodatná odchylka* je odmocnina z rozptylu.
- Pokud uvádíte ve výsledcích charakteristiku variability, uvažujte vždycky o tom, která z možných charakteristik je pro čtenáře užitečná. Většinou je nejvhodnější a dostatečnou charakteristikou směrodatná odchylka.

Kontrolní otázky:

1. *Mohou se při různých rozptylech shodovat charakteristiky polohy?*
2. *Proč jako charakteristiku variability nelze užít součet odchylek od průměru?*
3. *Vyjádřete slovně, co znamená populační rozptyl definovaný rovnicí (2-6).*

Pojmy k zapamatování:

- variabilita dat a její charakteristiky
- rozptyl, směrodatná odchylka

2.4 Další charakteristiky rozdělení pozorovaných hodnot

Cíl:
Po prostudování této kapitoly byste měli umět:

- co jsou empirické momenty,
- charakteristiky tvaru rozdělení (šikmost, špičatost),
- spočítat a interpretovat šikmost a špičatost rozdělení dat.

Průvodce studiem:
Prostudování této části kapitoly budete muset věnovat asi hodinu.

Kromě polohy a variability lze číselně vyjádřit i další charakteristiky postihující tvar rozdělení dat. Dříve než dvě takové charakteristiky uvedeme, seznámíme se s tzv. *empirickými momenty*, protože je pak při výpočtech charakteristik budeme potřebovat. Tzv. *k*-tý obecný moment M'_k je definován jako průměr *k*-tých mocnin

$$M'_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

První obecný moment je tedy $M'_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$, čili je to průměr $\bar{x}$, se kterým jsme se už setkali v kap. 2.2. Podobně existují i vyšší momenty, např. druhý moment, který je průměrnou hodnotou čtverců naměřených hodnot, tedy $M'_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$.

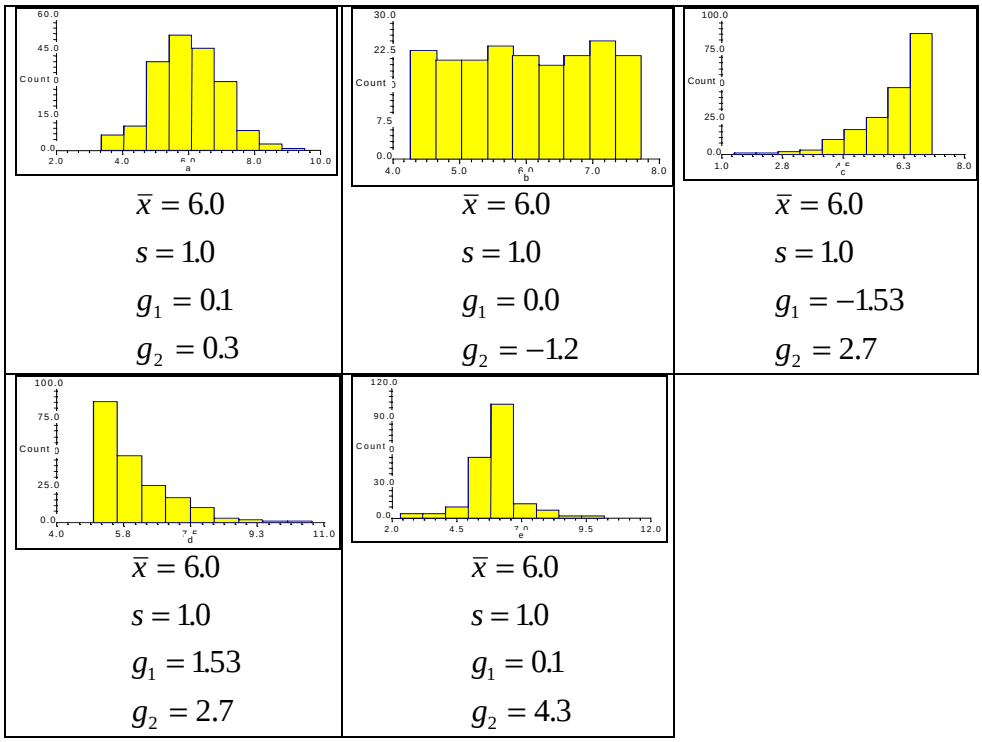
Dále se užívají také *centrální momenty* M_k, které vycházejí ze součtu mocnin odchylek od průměru.

$$M_k = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^k$$

- První centrální moment $M_1 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)$ není nijak užitečný, neboť je vždy $M_1 = 0$.
- Druhý centrální moment $M_2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$ je populační rozptyl, vždy platí $M_2 \geq 0$.
- Třetí centrální moment $M_3 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^3$. Vidíme, že M_3 může být i záporný.
- Čtvrtý centrální moment $M_4 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^4$, vždy platí $M_4 \geq 0$.

Příklad 2-12:

Než přejdeme k zavedení charakteristik tvaru rozdělení, podívejme se na následující histogramy.



Obr. 2-14: Různé tvary empirického rozdělení

Všech pět ukázek má stejné charakteristiky polohy i variability (průměry i směrodatné odchylky jsou shodné. Přesto na první pohled vidíme, že tvary rozdělení dat jsou různé. K číselnému vyjádření těchto rozdílů nám slouží další charakteristiky - šikmost (angl. skewness) a špičatost (angl. kurtosis).

Šikmost g_1 je definována jako

$$g_1 = \frac{M_3}{M_2\sqrt{M_2}} \tag{2-10}$$

Špičatost g_2 je definována vztahem

$$g_2 = \frac{M_4}{(M_2)^2} - 3 \tag{2-11}$$

Možná překvapuje, že v rov. (2-11) odečítáme na pravé straně trojku. Důvod je ten, že špičatost vztahujeme k nejčastěji vyskytujícímu se rozdělení, k tzv. normálnímu rozdělení (viz kap. 3), u kterého je poměr $M_4/(M_2)^2$ roven 3. Špičatost je tedy vztažena ke špičatosti normálního rozdělení, kladná špičatost zname-

ná špičatější rozdělení než normální, záporná špičatost znamená, že rozdělení pozorovaných hodnot je „placatější“ než normální.

Nulová šikmost znamená, že rozdělení dat je symetrické okolo průměru, kladná šikmost znamená, že rozdělení četností je zešikmeno vlevo (někdy říkáme, že rozdělení má těžší levý konec nebo levou stranu), záporná šikmost znamená zešikmení vpravo (těžší pravý konec). Čtveřice čísel ($\bar{x}$, s, g_1 , g_2) nám umožňuje udělat si představu o tvaru rozdělení dat a různá data porovnávat.

Shrnutí

- kromě polohy a variability lze tvar rozdělení dat popsat i dalšími charakteristikami tvaru rozdělení,
- tyto charakteristiky jsou založeny na třetím a čtvrtém centrálním empirickém momentu a nazývají šikmost a špičatost.

Kontrolní otázky:

1. *Co je znamená nulová šikmost?*
2. *Kdy je šikmost záporná?*
3. *Co je znamená nulová špičatost?*

Pojmy k zapamatování:

- empirické momenty
- centrální momenty
- šikmost, špičatost

2.5 Některé techniky popisné statistiky

Po prostudování této kapitoly byste měli umět:

- některé jednoduché techniky ručního zpracování méně rozsáhlých dat,
- zkonstruovat a především interpretovat krabicový graf (box plot).

Prostudování této části kapitoly budete muset věnovat asi hodinu.

2.5.1 Zaznamenávání četnosti

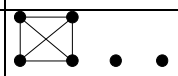
Při zjišťování a zpracování empirických dat jsme občas v situaci, kdy potřebujeme průběžně načítat četnosti výskytu pozorovaných hodnot (např. při vyhodnocování výsledků voleb, zaznamenávání počtu různých požadavků tazatelů v informační agentuře atd.). Vhodné techniky takového „ručního“ zjišťování četností si ukážeme na datech z př. 2-1:

Příklad 2-13:
Máme tyto pozorované hodnoty:
3, 4, 3, 5, 2, 3, 4, 2, 3, 5, 3, 4, 2, 5, 3, 3, 3, 4, 5, 2, 2, 2, 3, 3, 4, 4,3, 3, 4, 4

Na první pohled vidíme, že v datech se vyskytují hodnoty 2, 3, 4, 5. Každou hodnotu zapíšeme na samostatný řádek tabulky. Pak procházíme data a při každém výskytu hodnoty zapíšeme čárku na příslušný řádek tabulky. Abychom si usnadnili závěrečné sčítání čárek, tak vždy každou pátou zapíšeme vodorovně. Pokud se nám přihodí, že jsme při prvním pohledu výskytu některé hodnoty přehlédli, můžeme přidat do tabulky řádek s touto hodnotou.Výsledkem naší činnosti je následující tabulka, ze které lze určit velice snadno četnosti jednotlivých hodnot:

x_i		n_i
2	𝍸 /	6
3	𝍸 𝍸 //	12
4	𝍸 ///	8
5	////	4
		$n = 30$

V literatuře se můžeme setkat i s jinou technikou průběžného načítání četností. Místo čárek sdružených do pětic se postupně vyznačují jednotlivé části čtverců v pořadí vrcholy, hrany, úhlopříčky. Pak bychom dostali pro naše data tuto tabulku:

x_i		n_i
2		6
3		12
4		8
5		4

2.5.2 Zápis lodyha-list (Stem&Leaf)

Tento úsporný zápis získaných dat si ukážeme na příkladu.

Příklad 2-14:
Mějme tato pozorovaná data:

55 70 71 70 65 63 58 56 82 64 65 75 76 68 63 69 65 51

Tato data přepíšeme ve formě tabulky, v níž jeden sloupec (lodyha, angl. stem) budou tvořit číslice na desítkovém místě, ve druhém sloupci (list, angl. leaf), budou zapsány všechny číslice na místě jednotek (uspořádány vzestupně) z výskytů na příslušném řádku. Zápis stem&leaf tedy pro naše data vypadá takto:

lodyha (desítky)	listy (jednotky)
5	1568
6	33455589
7	00156
8	2

Tvar listu v tabulce charakterizuje rozdělení hodnot podobně jako histogram. Celá data jsou zaznamenávána ve vzestupném uspořádání, takže snadno můžeme určit medián, kvartily, případně i další kvantily:

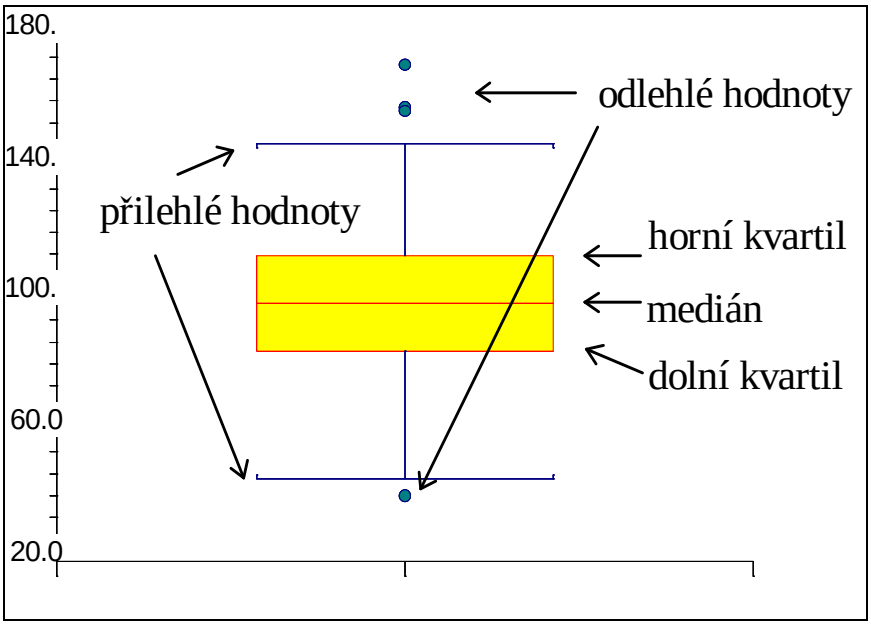
$$n=18$$
$$x(0,5)=65$$
$$x(0,25)=63$$
$$x(0,75)=70$$

Záznam dat formou STEM&LEAF je součástí volitelného výstupu některých statistických programů (NCSS ap.). Praktický význam má však zejména při ruční analýze dat malého rozsahu. Při ručním zpracování dat je vhodné vytvořit ten-

to zápis ve dvou krocích - nejprve zaznamenat číslice listu v pořadí, ve kterém se vyskytují v datech a ve druhém kroku teprve číslice na každém řádku setřídit.

2.5.3 Krabicový graf

Často užívanou grafickou formou prezentace rozdělení hodnot v datech je tzv. *krabicový (obdélníkový) graf*. Většinou se pro něj užívá původní anglický název *box plot*, někdy také *box & whiskers*, což bychom mohli přeložit jako krabička s vousy.
Krabicový graf vypadá takto:

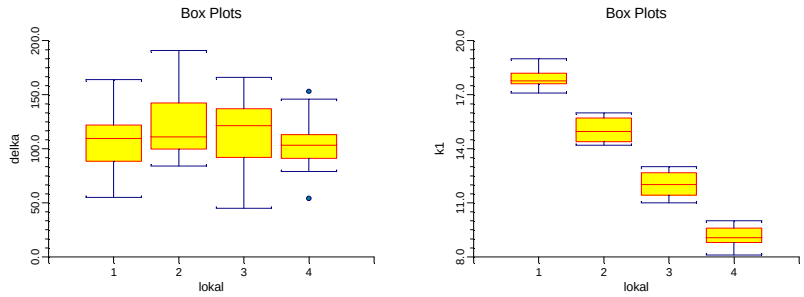


Vidíme, že mezikvartilové rozpětí je vyznačeno obdélníkem, uvnitř obdélníku je vyznačen medián. Úsečky („vousy“, angl. whiskers) končí v nejvzdálenější pozorované hodnotě ve vzdálenosti nejvýše 1,5 násobku mezikvartilového rozpětí od přilehlého kvartilu. Body vyznačené mimo vousy jsou hodnoty mimořádně vzdálené od mediánu, většinou je považujeme za odlehlé hodnoty. Z definice mezikvartilového rozpětí víme, že uvnitř krabičky leží 50 % pozorovaných hodnot. Z polohy mediánu uvnitř krabice okamžitě vidíme, zda těchto prostředních 50 % hodnot je rozděleno symetricky či sešikmeně.

Podobně na tvar rozdělení můžeme usuzovat z délky vousů. Pro většinu rozdělení by měla být naprostá většina pozorovaných hodnot uvnitř vousů. Např. u normálního rozdělení zde leží 99,3 % naměřených hodnot.

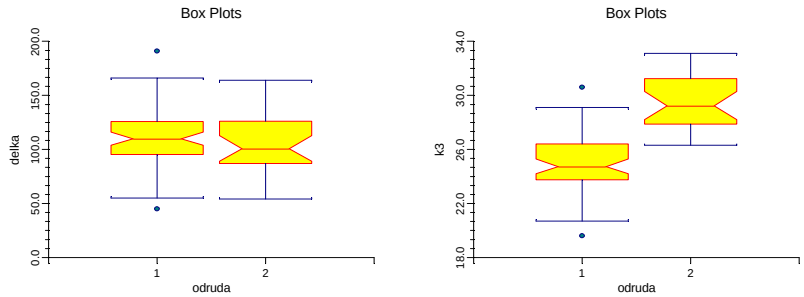
Krabicové grafy lze kreslit s pomocí běžného statistického software (např. NCSS, STATISTICA, SAS atd.). Tyto programy většinou dovolují alternativně zadat ještě další volbu tvaru krabicových grafů, tzv. notched box, tedy krabice s vrubem. Šířka (výška) vrubu vyznačuje interval spolehlivosti mediánu, takže porovnáním dvou krabic s vrubem můžeme rychle usuzovat, zda charakteristiky polohy skupin se liší významně (sešikmení se nepřekrývají) či jen nepodstatně (sešikmení mají společný úsek). Příklady takových krabicových grafů jsou na následujících obrázcích.

Příklad 2-15:



Krabicové grafy – porovnání čtyř skupin

Příklad 2-16:



Krabicové grafy s vruby – porovnání dvou skupin

Shrnutí:

- existují užitečné techniky umožňující rychlé nalezení tvaru rozdělení a základních charakteristik dat bez náročných výpočtů.
- pro méně rozsáhlá data zápis lodyha-list uchová úplnou informaci o datech, ukáže jejich rozdělení a usnadní výpočet mediánu a kvartilů.
- krabicové grafy na malé ploše poskytnou mnoho informací o rozdělení dat a charakteristikách polohy i variability.

Kontrolní otázky:

1. *Porovnejte krabicové grafy v příkladu 2-15. V čem se liší porovnávané skupiny ve veličině delka a v čem ve veličině k1?*
2. *Porovnejte krabicové grafy v příkladu 2-16. Liší se porovnávané skupiny ve veličině delka?*
3. *Porovnejte krabicové grafy v příkladu 2-16. Liší se porovnávané skupiny ve veličině k3?*

Pojmy k zapamatování:

- zaznamenávání četností
- zápis lodyha-list
- krabicový graf (box-plot)

2.6 Popis vztahu dvou veličin

Cíl:
Po prostudování této kapitoly byste měli:

- umět některé techniky popisné statistiky pro charakterizování vztahu dvou veličin,
- rozumět pojmu kovariance a korelace.

Průvodce studiem:
Prostudování této části kapitoly budete muset věnovat asi 2 hodiny.

Dosud jsme se zabývali charakteristikami a rozdělením četnosti hodnot jen jedné veličiny. Většinou však na každém objektu měříme více veličin a zajímá nás nejen každá veličina zvlášť, ale také vzájemné vztahy veličin. Hledáme odpovědi na otázky, zda hodnoty jedné veličiny souvisí s hodnotami veličiny jiné, či zda jsou hodnoty veličin na sobě nezávislé. V následujících odstavcích si ukážeme některé jednoduché techniky popisné statistiky, které nám umožňují vztahy dvou veličin postihnout. Uvidíme, že možnosti popisu vztahu dvou veličin jsou závislé na tom, zda sledované veličiny jsou spojité (či aspoň jako na spojité na ně můžeme pohlížet přesto, že jejich škála spojitá není, jako u veličin, jejichž hodnoty jsou vyjadřovány jen celými čísly, ale interval jejich hodnot je dosti široký, např. od 0 do 100), nebo zda obě či jedna ze sledovaných veličin je nominální či ordinální, ale s úzkou škálou.

2.6.1 Kontingenční tabulka

Kontingenční tabulka, tj. dvourozměrná tabulka rozdělení četnosti je základní možnost, jak zachytit vztah dvou nominálních veličin. Máme-li dvě nominální veličiny X, Y, kde X může nabývat hodnot $x_1, x_2, \dots, x_C$ a veličina Y může nabývat hodnot $y_1, y_2, \dots, y_R$, pak rozdělení četnosti pozorovaných hodnot můžeme vyjádřit následující tabulkou:

		X						
		x_1	x_2	...	x_j	...	x_C	$n_{i.}$
	y_1	n_{11}	n_{12}		n_{1j}		n_{1C}	$n_{1.}$
	y_2	n_{21}	n_{22}				n_{2C}	$n_{2.}$
Y	:	:	:		:			:
	y_i	n_{i1}			n_{ij}		n_{iC}	$n_{i.}$
	:	:	:		:			:
	y_R	n_{R1}	n_{R2}		n_{Rj}		n_{RC}	$n_{R.}$
		$n_{.1}$	$n_{.2}$		$n_{.j}$		$n_{.C}$	$n_{..} = n$

Hodnoty n_{ij} jsou absolutní četnosti, tzn. počty sledovaných objektů, kdy veličina Y má hodnotu y_i a současně veličina X má hodnotu x_i. Kromě toho do kontingenční tabulky můžeme zaznamenat tzv. *marginální četnosti* $n_{i.}$ a $n_{.j}$. Jsou definovány jako řádkové, resp. sloupcové součty.

$$n_{i\bullet} = \sum_{j=1}^{C} n_{ij}\,, \qquad n_{\bullet j} = \sum_{i=1}^{R} n_{ij}$$

Celkový počet objektů n je samozřejmě součet přes všechna políčka tabulky:

$$n = \sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij} = \sum_{i=1}^{R} n_{i\bullet} = \sum_{j=1}^{C} n_{\bullet j}$$

Podobnou tabulku můžeme vytvořit i z relativních četností. Obvykle se relativní četnosti vyjadřují v procentech. Vidíme, že jsou tři možnosti, jak počítat relativní četnosti:

- tzv. celková (tabulková) procenta: $T_{ij} = \frac{n_{ij}}{n} 100$
- řádková procenta $R_{ij} = \frac{n_{ij}}{n_{i\bullet}} 100$, u kterých řádkový součet je 100%.
- sloupcová procenta $C_{ij} = \frac{n_{ij}}{n_{\bullet j}} 100$, u kterých sloupcový součet je 100%

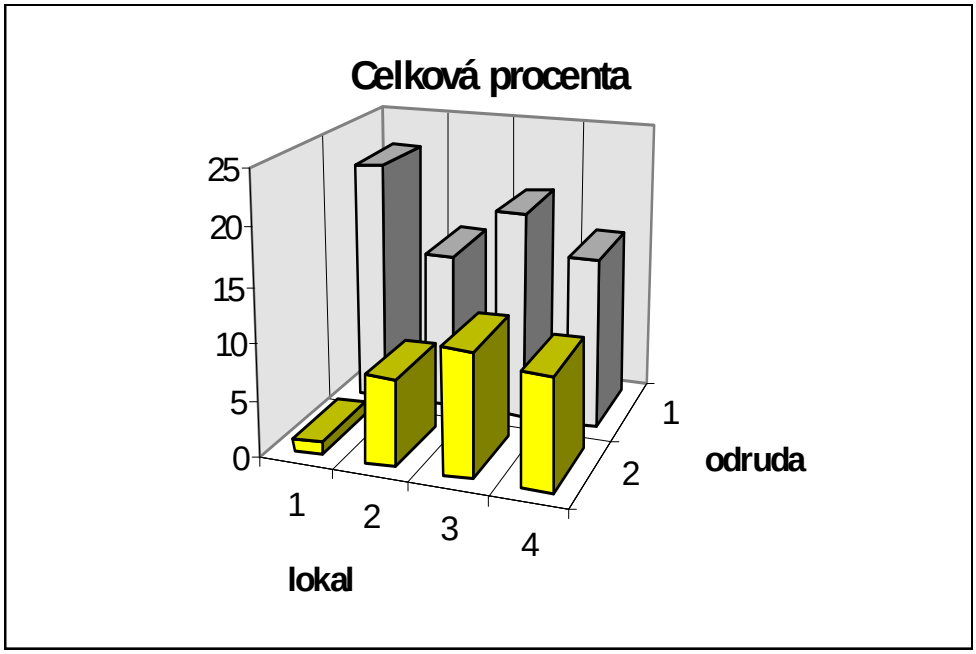
Četnosti z kontingenční tabulky můžeme znázornit trojrozměrným grafem. Z grafu pak můžeme poměrně snadno usuzovat na souvislost či nezávislost veličin.

Příklad 2-17:

Pro dvě nominální veličiny *lokal* a *odrůda* byla z dat spočtena tato tabulka četností.

	lokal				
odruda	1	2	3	4	Total
1	20	13	17	14	64
2	1	7	10	9	27
Total	21	20	27	23	91

Tyto četnosti jsou znázorněny v následujícím grafu.



Graf závislosti dvou nominálních veličin (četnosti z kontingenční tabulky)

Je zřejmé, že stejným způsobem jako vztah dvou nominálních veličin lze i popsat vztah dvou ordinálních veličin. Dokonce i u spojitých veličin můžeme použít dvojrozměrnou tabulku četností, pokud spojité veličiny předtím uspořádáme do tříd. Takovou tabulku pak nazýváme *korelační tabulkou*.

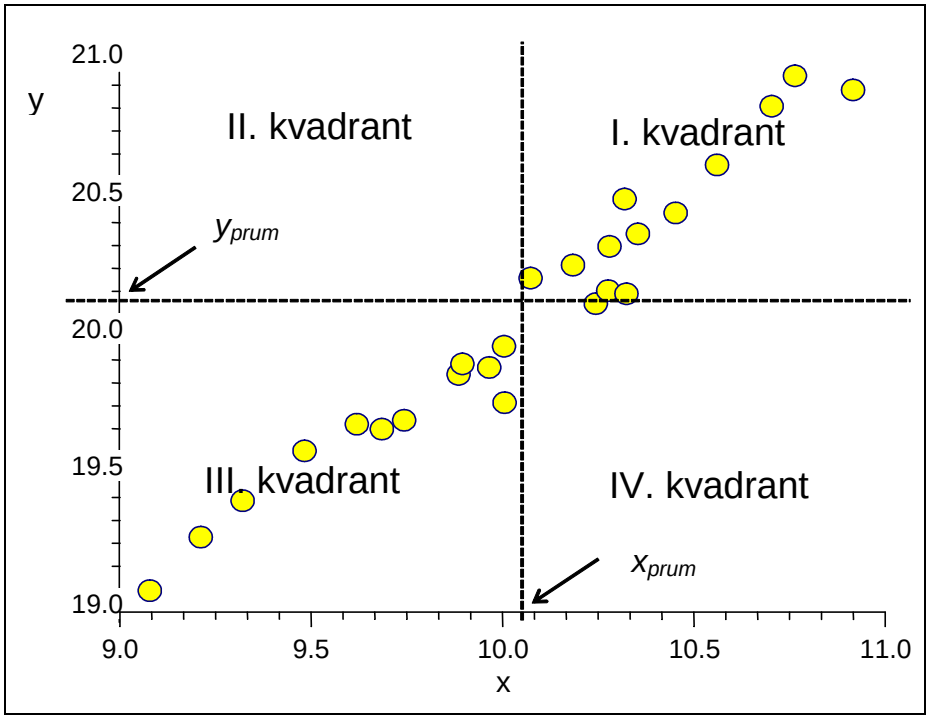
V kap. 4 si ukážeme některé další možnosti posouzení vztahu dvou nominálních veličin, které však vyžadují porozumění pojmům přesahujícím popisnou statistiku.

2.6.2 Nominální a spojitá veličina

Pro takovou dvojici je vhodné charakterizovat polohu, variabilitu, příp. rozdělení četností hodnot spojité veličiny pro každou z pozorovaných hodnot nominální veličiny. Další velmi názornou možností je zobrazení závislosti spojité veličiny na veličině nominální je krabicový graf pro jednotlivé kategorie nominální veličiny, jak bylo ukázáno v odst. 2.5.3.

2.6.3 Dvě spojité veličiny

Nejjednodušší způsob, jak znázornit vztah dvou veličin, je nakreslit jejich *bodový graf* (angl. *xy-plot* nebo *scatter plot*). Z grafu většinou okamžitě vidíme, zda hodnoty jedné veličiny mají tendenci růst s hodnotami druhé veličiny nebo klesat či spolu nesouvisí. Na obr. 2-14 máme graficky znázorněny naměřené hodnoty dvou veličin a také vyznačeny dvě přímky odpovídající průměrům každé veličiny a kvadranty, na které tyto přímky dělí rovinu, ve které jsou zobrazeny naměřené body.



Obr. 2-14: Graf pozorované závislosti dvou spojitých veličin

Závislost těchto dvou veličin můžeme charakterizovat číselně pomocí odchylek od průměrů:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\cdot\left(y_i - \overline{y}\right) \qquad (2\text{-}12)$$

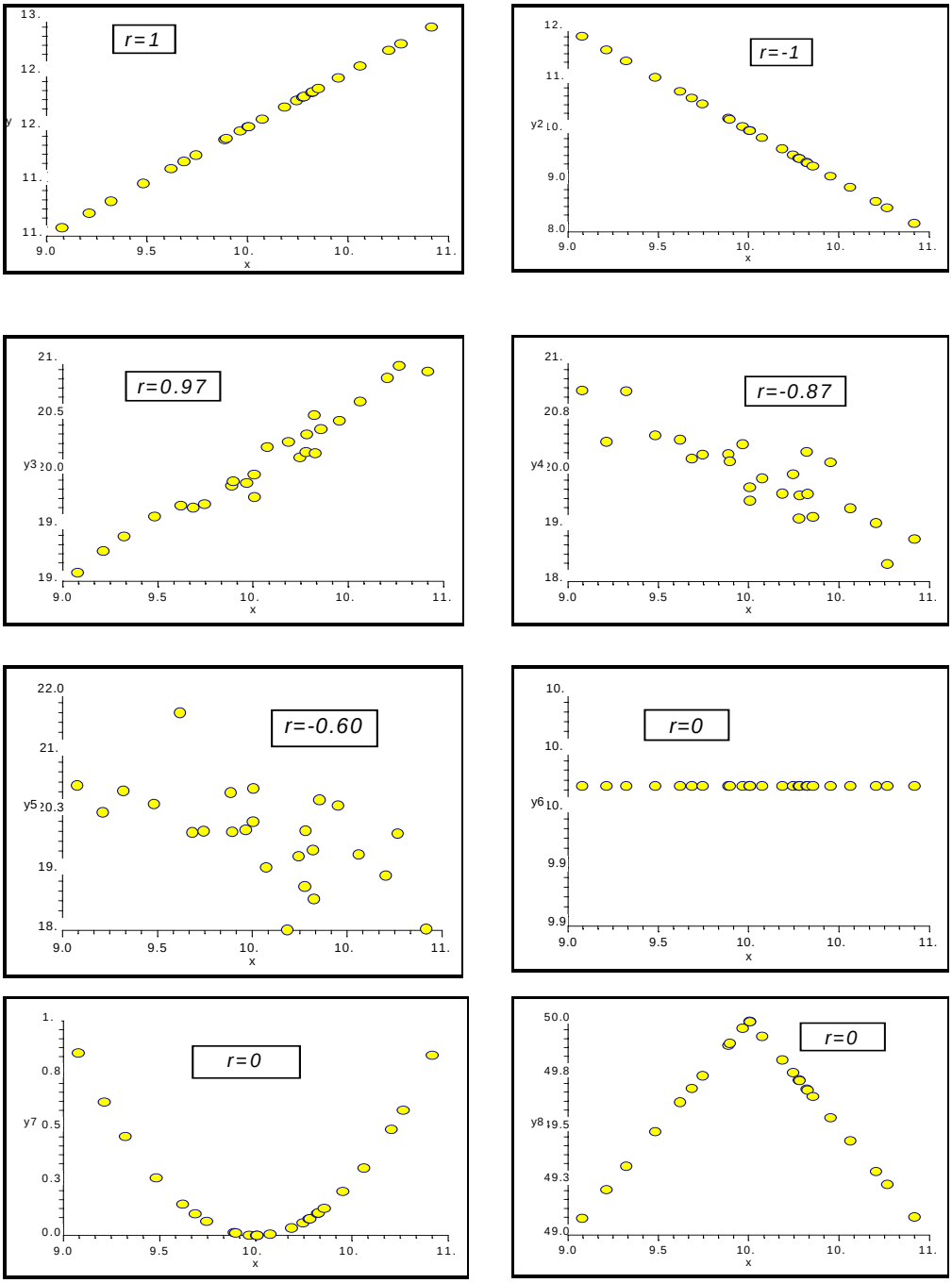
Charakteristice s_{xy} se říká *kovariance*. Vidíme, že body (x_i , y_i) z I. a III. kvadrantu zvětšují hodnotu součtu ve výrazu na pravé straně rovnice (2-12), zatímco body z II. a IV. kvadrantu hodnotu součtu zmenšují, neboť pro ně součin $\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)$ je záporný. Můžeme tedy usoudit, že pokud kovariance je kladná, je mezi veličinami kladná souvislost (s rostoucím *x* roste *y*), pokud kovariance je záporná, je vztah opačný. Pokud je kovariance blízká nule, není mezi veličinami lineární závislost. Trochu obtíže však způsobuje to, že lze těžko posoudit, co znamená, že kovariance je blízká nule. Kovariance není omezena zdola ani shora, a proto těžko posoudit, jaké hodnoty jsou dostatečně blízké nule. Problém lze částečně vyřešit zavedením jiné charakteristiky, *korelačního koeficientu*

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \qquad (2\text{-}13)$$

kde s_x a s_y jsou směrodatné odchylky veličin *x* a *y*.

Pro korelační koeficient platí $-1 \le r_{xy} \le 1$, přičemž hodnoty $\left|r_{xy}\right| = 1$ znamenají přesnou lineární závislost (body v grafu leží v přímce)- viz obr. 2-15.

Příklad 2-18: Grafy závislosti dvou veličin a hodnoty korelačního koeficientu



Obr. 2-15: Různé tvary závislosti a hodnoty korelačního koeficientu

Vidíme, že korelační koeficient je charakteristikou těsnosti *lineárního* vztahu dvou veličin. Hodnoty r_{xy} blízké nule nemusí nutně znamenat nezávislost veličin, znamenají pouze to, že mezi veličinami není lineární závislost.

2.7 Příklad statistického zpracování dat

Zadání:
Čtyři stochastické algoritmy pro globální optimalizaci byly ověřovány na 6 testovacích funkcích. Vzhledem ke stochastické povaze těchto algoritmů je nutno testy provádět opakovaně, proto u každé úlohy bylo provedeno 100 opakování. Časová náročnost hledání je vyjádřena počtem vyhodnocení funkce (veličina *ne*), přesnost přiblížení správnému řešení je vyjádřena počtem platných číslic nalezeného řešení shodných s řešením správným (veličina *lambda*). Výsledky numerických testů algoritmů jsou v souboru ALG07_d10.xls, který je dostupný na http://albert.osu.cz/tvrdik/down/vyuka.html.
Za přijatelné přiblížení správnému řešení je považováno takové přiblížení, kdy *lambda* > 4. Zpracujte přehlednou tabulku základních charakteristik algoritmů a úloh (průměrná časová náročnost, spolehlivost hledání globálního minima vyjádřená jako počet opakovaní splňujících podmínku *lambda* > 4. Pomocí krabicových grafů porovnejte časovou náročnost algoritmů.

Řešení:
Vždycky je dobré na začátku nahlédnout do dat a zjistit obory jejich hodnot.

V úloze jsou dvě nominální veličiny (*algoritmus*, *funkce*), jejich hodnoty jsou znakové řetězce. Následující tabulka nám poskytne informaci o celkovém počtu pozorovaných hodnot a o rozdělení četností.

Counts Section

	algoritmus				
funkce	**8hc1**	**BREST**	**DER**	**debr18**	**Total**
ackley	100	100	100	100	400
dejong1	100	100	100	100	400
griewangk	100	100	100	100	400
rastrig	100	100	100	100	400
rosen	100	100	100	100	400
schwefel	100	100	100	100	400
Total	600	600	600	600	2400

Vidíme, že četnost výskytu jednotlivých hodnot odpovídá zadání, tj. 100 opakování algoritmu na každé úloze.

Základní charakteristiky dvou číselných veličin (*ne*, *lambda*) jsou v následující tabulce.

Variable Summary Section

Variables	**Count**	**Mean**	**Standard Deviation**	**Minimum**	**Maximum**
lambda	2400	6.69	1.04	0	8.5
ne	2400	30357.39	36564.99	6220	185900

Z počtu pozorovaných hodnot (2400) vidíme, že v datech není žádná chybějící hodnota, na všech řádcích datové tabulky jsou hodnoty jak *lambda*, tak *ne*. Vidíme, že minimum veličiny *lambda* je rovno 0, tedy nejméně jeden běh algoritmu se nepřiblížil dostatečně ke správnému řešení.

Tyto tabulky jsou pouze pracovní, slouží nám jen ke kontrole dat a získání základního přehledu o jejich obsahu. Nejsou součástí prezentace výsledků statistické analýzy.

Výsledky:
Spolehlivost algoritmů je uvedena v tabulce 1, číselné hodnoty jsou počty běhů, v nichž se hledání dostatečně přiblížilo správnému řešení. Vzhledem k tomu, že pro každou úlohu bylo provedeno 100 opakování, je to současně i spolehlivost v procentech.

Tabulka 1: Spolehlivost v procentech

	Algoritmus			
funkce	8hc1	BREST	DER	debr18
ackley	100	100	99	100
dejong1	100	100	100	100
griewangk	94	100	78	99
rastrig	99	100	82	100
rosen	100	100	100	100
schwefel	100	100	96	99

Z tabulky 1 je zřejmé, že jedině algoritmus BREST dosáhl 100% spolehlivosti na všech šesti testovacích funkcích, algoritmus DER byl naopak výrazně nejméně spolehlivý.

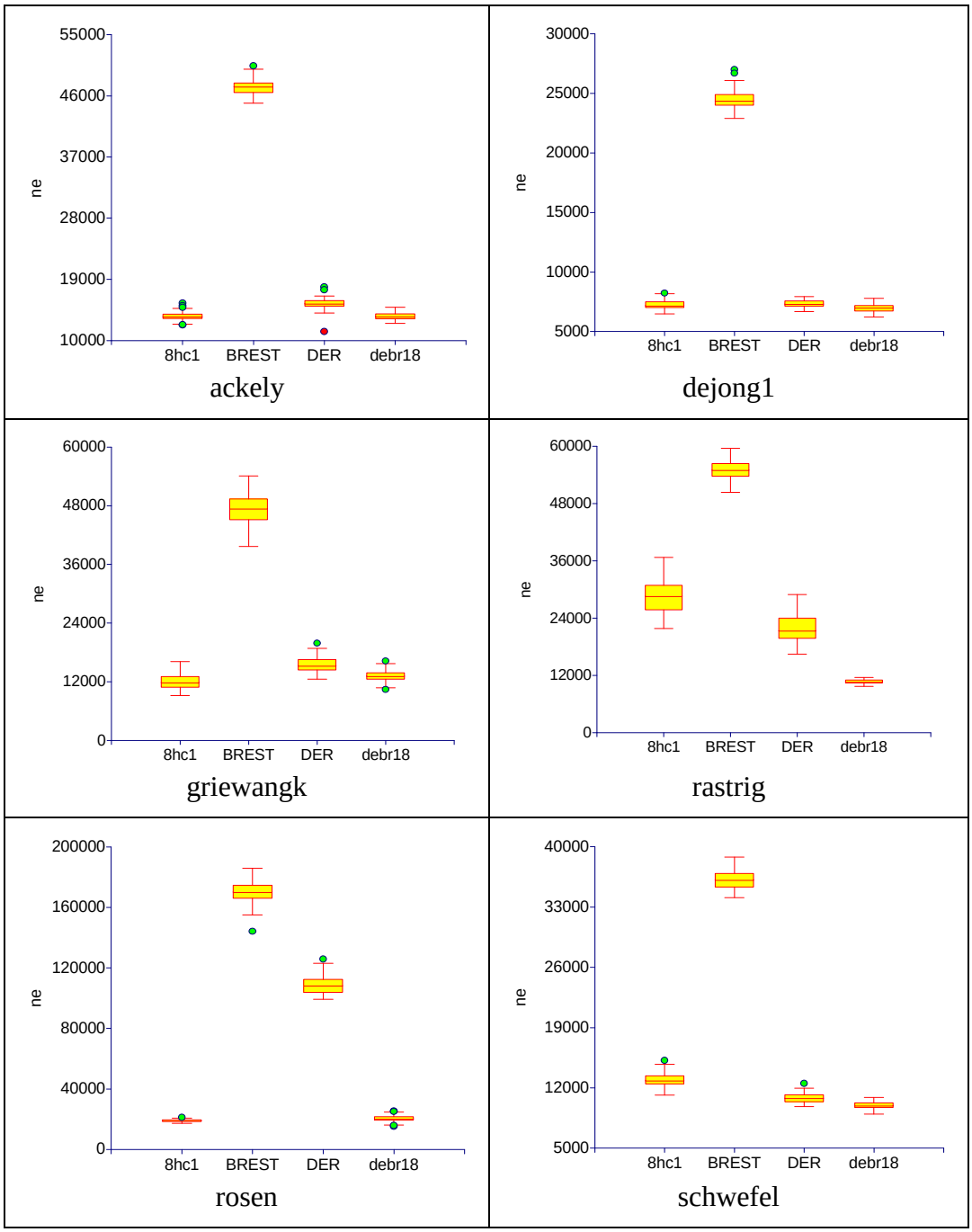
Časové nároky vyjádřené jako průměrný počet vyhodnocení účelové funkce potřebný k dosažení podmínky ukončení hledání jsou uvedeny v tabulce 2.

Tabulka 2: Časová náročnost (průměr veličiny *ne*)

	Algoritmus			
funkce	8hc1	BREST	DER	debr18
ackley	13554	47265	15431	13569
dejong1	7257	24511	7357	6973
griewangk	12145	47333	15503	13153
rastrig	28529	55082	21813	10711
rosen	19132	170179	108593	20524
schwefel	12936	36223	10838	9964

Porovnání časové náročnosti algoritmů na každé z testovaných funkcí je na obrázku 1 na další straně. Z obrázku vidíme, algoritmus BREST byl na všech úlohách výrazně nejpomalejší s časovou náročností několikrát vyšší než ostatní algoritmy. Nejméně spolehlivý algoritmus DER nebyl na žádné z úloh nejrychlejší.

Závěr z naší analýzy tedy je, že spolehlivý algoritmus BREST je příliš časově náročný. Algoritmy *8hc1* a *debr18* jsou rychlejší a spolehlivější než algoritmus DER. Proto považujeme algoritmy *8hc1* a *debr18* za nejúspěšnější v tomto testu a je možno je doporučit pro další zkoumání a využití v řešení problémů hledání globálního minima.



Obrázek 1: Porovnání časové náročnosti algoritmů na jednotlivých funkcích

Shrnutí

- Vztah dvou veličin lze přehledně zobrazit prostředky popisné statistiky.
- Důležité je si uvědomit, v jakých škálách byly sledované veličiny měřeny a podle toho volit vhodný způsob vyjádření jejich vztahu.
- Graf je názornější než číselné charakteristiky.
- Pro korelační koeficient platí $-1 \leq r_{xy} \leq 1$.
- Korelační koeficient je charakteristikou těsnosti lineárního vztahu dvou veličin.
- Hodnoty korelačního koeficientu blízké nule nemusí nutně znamenat nezávislost veličin, znamenají pouze, že mezi veličinami není lineární závislost.

Kontrolní otázky:

1. *Porovnejte krabicové grafy v příkladu 2-16. Liší se mediány veličiny k3 v porovnávaných skupinách?*
2. *Proč tři poslední závislosti na obr. 2-15 mají všechny hodnotu korelačního koeficientu nulovou, ač tvar závislosti je odlišný?*

Pojmy k zapamatování:

- kontingenční tabulka
- kovariance
- korelační koeficient

Korespondenční úlohy budou zadávány vždy na začátku semestru.

Z popisné statistiky budou tři příklady.

3 Základy počtu pravděpodobnosti

Prozatím jsme se ve výkladu analýzy dat a deskriptivní statistiky obešli bez znalosti jakýchkoli pojmů z teorie (počtu) pravděpodobnosti. K porozumění základům induktivní statistiky v kap. 4 však takové znalosti budou nezbytné. Takže nám nezbývá než se pokusit nutné elementy této matematické, tedy formální a abstraktní disciplíny zvládnout. Povzbuzením nám může být, že mnoho impulsů k zavedení základních pojmů v počtu pravděpodobnosti vychází z každodenního života. Jeden z prvních podnětů ke vzniku počtu pravděpodobnosti vyšel v 17. století z hazardních her. Navíc slova pravděpodobnost užívá kdekdo, a většinou správně, v hodnocení každodenních jevů, aniž by znal formální definici tohoto pojmu, vystačí s jeho intuitivním pochopením. Bohužel s intuicí nevystačíme, chceme-li užívat metody induktivní statistiky. A bez těchto metod se neobejdeme v žádném vědním či technickém oboru zkoumajícím svět, ve kterém žijeme, ale ani v mnoha praktických činnostech, které zdánlivě nemají s vědou nic společného.

Průvodce studiem: Kapitola o základech počtu pravděpodobnosti je vzhledem ke své obsáhlosti a náročnosti rozdělena do čtyř částí. Celkově její studium zabere 15-25 hodin.

První část kapitoly vám zabere asi čtyři až pět hodin. Pochopení učiva vám usnadní četné ilustrační příklady. Vztahy, které se často užívají pro výpočet pravděpodobnosti, jsou uvedeny v rámečcích. Nezapomeňte, že každý z těchto vztahů platí jen za určitých podmínek, které musí splňovat jevy, jichž se vztah týká.

3.1 Náhodný pokus, náhodný jev a pravděpodobnost

Každodenně se setkáváme s ději, u kterých nevíme s jistotou, jakým výsledkem skončí. Příkladem je třeba

- zkouška k získání řidičského průkazu (projdeme nebo neprojdeme?),
- zkoumání vzorku říční vody (kolik v něm nalezneme mikroorganismů?),
- těhotenství (narodí se kluk nebo holka nebo dokonce více dětí ?).

Obecně se takový děj s nejistým výsledkem nazývá *náhodný pokus*. Společným rysem náhodných pokusů je, že

- výsledkem musí být právě jeden z množiny alespoň dvou možných výsledků,
- uvažovaný pokus je možno nezávisle a za stejných podmínek opakovat.

Druhou vlastnost výše uvedené příklady beze zbytku nesplňují, ale v této chvíli se tím nebudeme trápit. Příkladem takového snadno představitelného náhodného pokusu je hod hrací kostkou, který se právě pro tuto jednoduchost tradičně užívá k výkladu základů počtu pravděpodobnosti.

Výsledkem náhodného pokusu je *náhodný jev*. U hodu kostkou je to na př. „padla jednička“ nebo „padla sudá“ nebo „padlo více než 4“ atd. Náhodné jevy označujeme velkými písmeny ze začátku abecedy, případně velkými písmeny s indexem. Označme tedy možné výsledky hodu kostkou takto:

E_1	padla jednička
E_2	padla dvojka
E_3	padla trojka
E_4	padla čtyřka
E_5	padla pětka
E_6	padla šestka

Jiný výsledek nastat nemůže, kostka spadnout musí. Žádný z jevů E_i, $i = 1, 2,...,6$, není složen z jiných jevů, nelze jej dále rozložit, ani nemohou nastat žádné dva takové jevy současně. Říkáme, že jevy E_i jsou *elementární jevy*.

Ale jev *B*, „padne sudá“ je složen z jevů E_2, E_4 a E_6, říkáme, že je *sjednocením* těchto jevů, což zapisujeme $B = E_2 \cup E_4 \cup E_6$.

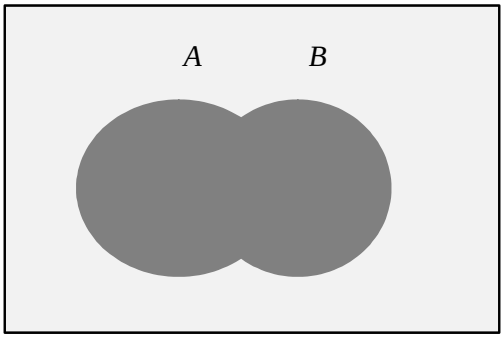
Sjednocením všech elementárních jevů dostaneme *jev jistý* - označíme jej symbolem *U*, tedy v našem příkladu $U = E_1 \cup E_2 \cup \ldots \cup E_6$.

Jev, který nastat nemůže (např. na kostce nemůže padnout sedmička), nazýváme *jevem nemožným* a značíme jej $\emptyset$.

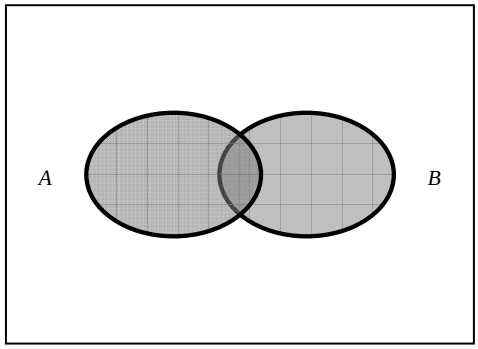
Uvažujme jev *B* - „padne sudá“. O jevu *A* - „nepadne sudá“ říkáme, že je *opačným (komplementárním) jevem* k jevu B, označujeme jej $\overline{B}$ (non *B*), takže můžeme psát $A = \overline{B}$. Je zřejmé, že sjednocením jevu B a jevu opačného, tj. $\overline{B}$, je jev jistý, $B \cup \overline{B} = U$.

Jev *B*, „padne sudá“ a jev *C*, „padne lichá“, nemají žádný společný jev. Říkáme, že jevy *B* a *C* jsou *neslučitelné* (*disjunktní*). Naopak, jev *B* a jev *D*, „padne více než 4“ neslučitelné nejsou, protože mají společný jev E_6. *Průnik* neslučitelných jevů je jev nemožný, což zapíšeme $B \cap C = \emptyset$, zatímco průnikem jevů B a D je jev E_6, $B \cap D = E_6$, a jevy *B*, *D* tedy opravdu disjunktní nejsou.

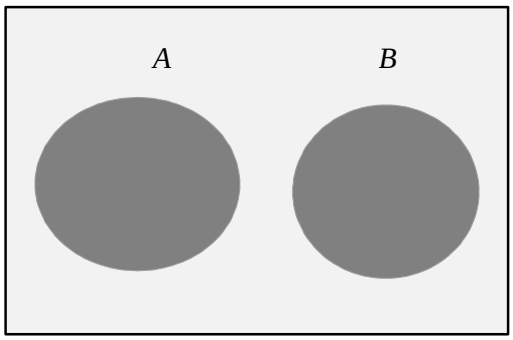
Vztahy jevů podobně jako vztahy množin můžeme vyjádřit názorně pomocí Vennových diagramů:



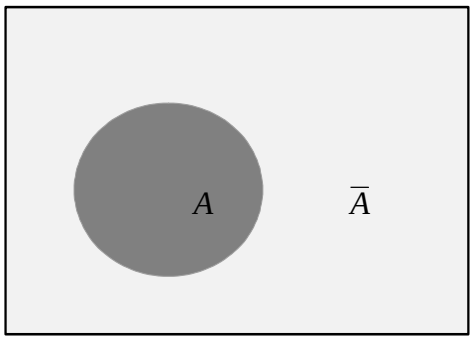
Sjednocení jevů, $A \cup B$



Průnik jevů, $A \cap B$



Neslučitelné jevy *A*, *B*



Jev *A* a jev opačný

Uvažujme nyní elementární náhodné jevy $E_1, E_2, \ldots, E_k$, pro které platí:

a) $E_i \cap E_j = \emptyset$ pro $i \neq j,\ i, j = 1, 2, \ldots, k$ (každá dvojice různých elementárních jevů jsou jevy neslučitelné)

b) $E_1 \cup E_2 \cup \ldots \cup E_k = U$ (jeden z těchto elementárních jevů musí nastat)

Množinu $\Omega = \{E_1, E_2, \ldots, E_k\}$ pak nazýváme *systémem elementárních jevů*.

Náhodným jevem pak je libovolná podmnožina množiny Ω. Lze vytvořit 2^k různých podmnožin (včetně prázdné množiny a celé množiny Ω). Prázdná množina odpovídá jevu nemožnému, celá množina Ω pak jevu jistému. Podle toho, jak jemně (podrobně) zvolíme systém elementárních jevů, tak podrobně dokážeme tímto matematickým modelem náhodného jevu popsat reálný pokus. Zde jsme uvedli systém konečného počtu *k* elementárních náhodných jevů. Je však možné modelovat náhodné pokusy pomocí systému nekonečného počtu náhodných jevů, ale pro výklad základů pravděpodobnosti vystačíme s konečným počtem elementárních jevů.

Jelikož výsledky náhodného pokusu (tj. náhodné jevy) modelujeme jako systém podmnožin, můžeme zavést některé číselné funkce náhodných jevů a matematicky odvodit (dokázat) pravidla, jak s těmito funkcemi počítat. Jednou z takových

funkcí je *pravděpodobnost*. Pro každý náhodný jev A je pravděpodobnost $P(A)$ funkce (míra) jevu s těmito vlastnostmi:

(a) $0 \le P(A) \le 1$ (pravděpodobnost je nezáporná a normovaná funkce)

(b) $P(U) = 1$ (pravděpodobnost jevu jistého je rovna jedné)

(c) Je-li $A \cap B = \emptyset$, pak $P(A \cup B) = P(A) + P(B)$
(pravděpodobnost sjednocení disjunktních jevů je rovna součtu pravděpodobností jevů)

Tvrzení (a), (b), (c) označujeme jako axiomy teorie pravděpodobnosti. Pravděpodobnost $P(A)$ měří (ohodnocuje) možnost výskytu jevu A v náhodném pokusu.

Je však otázkou, jak určit číselnou hodnotu $P(A)$. Existují dvě vcelku jednoduché možnosti. První způsob je omezen na tzv. jednoduché náhodné pokusy, kdy všechny elementární jevy jsou stejně pravděpodobné. Pak se tak zvaná *klasická pravděpodobnost* počítá jako podíl počtu výsledků příznivých n_A (ve kterých nastane jev A) ku počtu všech možných výsledků n, tj. $P(A) = \frac{n_A}{n}$.

Příklad 3-1: Uvažujme náhodný pokus hod hrací kostkou. Je zřejmé, pokud má kostka těžiště uprostřed v průsečíku tělesových úhlopříček (přesněji řečeno, je homogenní a isotropní), že $P(E_1) = P(E_2) = \ldots = P(E_6)$. Nechť jev A je „padne sudá“. Pak $P(A) = \frac{3}{6} = 0{,}5$, neboť je šest možných elementárních jevů $E_1, E_2, \ldots, E_6$, ale jen tři (E_2, E_4, E_6) jsou příznivé, kdy nastane jev A.

U jiných než klasických náhodných pokusů se musí pravděpodobnosti odhadovat z pozorování relativní četnosti výskytu jevu A v n nezávislých opakováních náhodného pokusu, $f_A = \frac{n_A}{n}$, n_A je počet pokusů, kdy nastal jev A. Pravděpodobnost $P(A)$ je dána vztahem

$$\boxed{P(A) = \lim_{n \to \infty} \left(\frac{n_A}{n} \right)} \qquad (3\text{-}1)$$

Tomuto vztahu se říká *statistická definice pravděpodobnosti*.

Pokud bychom pravděpodobnost jevu A v předchozím příkladu nebyli schopni určit uvedeným klasickým postupem, nezbývalo by nic jiného než n-krát hodit kostkou (n je pokud možno velké) a zaznamenat počet výsledků n_A, kdy padla sudá. Výsledkem by pak při $n = 600$ mohlo být třeba $n_A = 303$. Pak bychom $P(A)$ mohli odhadnout jako $\frac{n_A}{n} = \frac{303}{600} = 0{,}505$.

Podobně chceme-li zjistit, jaká je pravděpodobnost jevu, že náhodně vybraný muž z dospělé populace měří alespoň 2 metry, nezbývá než vybrat náhodně *n* dospělých mužů a zjistit, jaká je relativní četnost dvoumetrových dlouhánů.
Z axiomů definice pravděpodobnosti (a), (b), (c) bezprostředně vyplývají další vztahy pro počítání pravděpodobnosti:

$P(U) = P(A \cup \overline{A}) = P(A) + P(\overline{A}) = 1$, a tedy

$$\boxed{P(A) = 1 - P(\overline{A})}. \qquad (3\text{-}2)$$

Pro libovolné dva jevy A, B platí (viz Vennův diagram sjednocení dvou jevů)

$$A \cup B = (A \cap \overline{B}) \cup (\overline{A} \cap B) \cup (A \cap B) \qquad (3\text{-}3)$$

Na pravé straně rovnice (3) je sjednocení tří disjunktních jevů, takže podle axiomu (c) dostaneme

$$P(A \cup B) = P(A \cap \overline{B}) + P(\overline{A} \cap B) + P(A \cap B) \qquad (3\text{-}4)$$

Zároveň $A = (A \cap B) \cup (A \cap \overline{B})$ a $B = (A \cap B) \cup (\overline{A} \cap B)$. Na pravých stranách jsou opět sjednocení disjunktních jevů a tedy podle axiomu (c) platí

$$P(A) = P(A \cap \overline{B}) + P(A \cap B) \quad \text{a} \quad P(B) = P(\overline{A} \cap B) + P(A \cap B)$$

a po dosazení do rovnice (4) dostaneme

$$\boxed{P(A \cup B) = P(A) + P(B) - P(A \cap B)} \qquad (3\text{-}5)$$

Často nás zajímá pravděpodobnost jevu *A* za podmínky, že nastal jiný jev, jev *B*. Např. pravděpodobnost, že padne šestka za podmínky, že padla sudá nebo praktičtější příklad, jaká je pravděpodobnost onemocnění za podmínky, že pacient je očkován. Zkusme se na tuto situaci podívat nejdříve přes relativní četnosti.

V *n* pokusech nastal jev B n_B-krát. Současně s jevem B nastal jev A $n_{A \cap B}$-krát. Relativní četnost jevu A za podmínky, že nastal jev B je

$$f_{A|B} = \frac{n_{A \cap B}}{n_B}, \qquad (3\text{-}6)$$

tedy také $f_{A|B} = \dfrac{n_{A \cap B} / n}{n_B / n} = \dfrac{f_{A \cap B}}{f_B}$. (3-7)

Víme už, že pravděpodobnost je vlastně jakýmsi abstraktnějším pohledem na relativní četnost, takže *podmíněná pravděpodobnost* $P(A|B)$ jevu *A* za podmínky *B* je definována podobně

$$\boxed{P(A \mid B) = \frac{P(A \cap B)}{P(B)}} \tag{3-8}$$

S pomocí podmíněné pravděpodobnosti můžeme zavést pojem *nezávislosti jevů*. Jev *A* je nezávislý na jevu *B* a naopak, jev *B* je nezávislý na jevu *A*, tedy jevy *A*, *B* jsou nezávislé, když podmíněná pravděpodobnost $P(A|B)$ na jevu *B* nezávisí, tedy $P(A \mid B) = P(A)$, podobně i $P(B \mid A) = P(B)$. Pak z definice podmíněné pravděpodobnosti pro *nezávislé jevy* platí

$$\boxed{P(A \cap B) = P(A) \cdot P(B)} \tag{3-9}$$

Vztah (3-9) je návodem, jak počítat pravděpodobnosti průniku *nezávislých* jevů.

Příklad 3-2: Jaká je pravděpodobnost, že ve dvou hodech kostkou padne dvakrát šestka?

Nechť *A* je jev, že šestka padne v prvním hodu, *B* je jev, že šestka padne v druhém hodu. Jelikož jde zřejmě o nezávislé jevy (kostka nemá paměť, takže druhý hod není ovlivňován výsledkem prvního hodu),

$$P(A \cap B) = P(A) \cdot P(B) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}.$$

Dalším užitečným vztahem je *věta o úplné pravděpodobnosti*. Máme-li jevy $A_1, A_2, \ldots, A_k$ (nemusí být elementární), pro které platí:

(a) $A_i \cap A_j = \emptyset$ pro $i \neq j, \quad i, j = 1,2, \ldots, k$ (jevy v každé dvojici různých jevů jsou jevy neslučitelné)

(b) $A_1 \cup A_2 \cup \ldots \cup A_k = U$
(Pokud jevy $A_1, A_2, \ldots, A_k$ splňují podmínky (a), (b), říkáme, že tyto jevy tvoří *rozklad* jevu jistého nebo že tvoří *systém jevů*)

(c) $P(A_i) > 0$ pro všechna $i = 1, 2, \ldots, k$,

pak pro libovolný jev *C* platí

$$\boxed{P(C) = \sum_{i=1}^{k} P(C \mid A_i) \cdot P(A_i)} \tag{3-10}$$

Tuto větu o úplné pravděpodobnosti můžeme snadno dokázat:

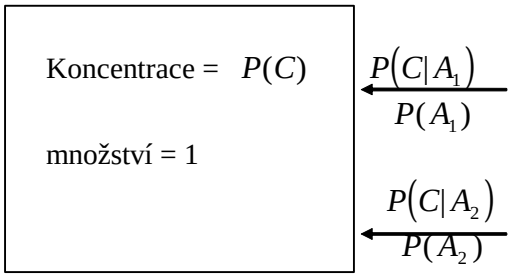
$$C = C \cap U = C \cap (A_1 \cup A_2 \cup \ldots \cup A_k) = (C \cap A_1) \cup (C \cap A_2) \cup \ldots \cup (C \cap A_k)$$

Podle pravidla o sčítání pravděpodobností neslučitelných jevů je

$$P(C) = \sum_{i=1}^{k} P(C \cap A_i)$$

a po dosazení z definice podmíněné pravděpodobnosti (3-8) dostaneme vztah (3-10).

Ke stejnému vztahu (10) dojdeme i zcela odlišnou úvahou. Uvažujme míchání k vstupujících množství obsahující látku C a nechť relativní množství i-tého vstupu je $P(A_i)$, $\sum_{i=1}^{k} P(A_i) = 1$, koncentrace látky C v i-tém vstupu nechť je $P(C|A_i)$, $P(C)$ je pak koncentrace látky C ve výsledné směsi - viz obrázek pro k = 2. Vyjádříme-li koncentraci $P(C)$ z látkové bilance (aplikujeme zákon zachování hmoty), dostaneme vztah (3-10).



Bilance složky C je vyjádřena rovnicí
$P(C) \cdot 1 = P(C \mid A_1) \cdot P(A_1) + P(C \mid A_2) \cdot P(A_2)$, což odpovídá rovnici (10).

Příklad 3-3: Smícháme 3 litry 50% slivovice s 2 litry 60% slivovice. Jaká bude výsledná koncentrace etanolu ve výsledné směsi (za předpokladu, že při míchání nedochází ke změně objemu)?

$$P(C) = P(C \mid A_1) \cdot P(A_1) + P(C \mid A_2) \cdot P(A_2) =$$
$$= \frac{50}{100} \cdot \frac{3}{5} + \frac{60}{100} \cdot \frac{2}{5} = \frac{27}{50} = 0,54$$

Takže výsledkem je 54% slivovice.

Z podmíněné pravděpodobnosti dojdeme i k dalšímu často užívaného vztahu, Bayesovu vzorci (*Bayesově větě*). Pokud jevy $A_1, A_2, \ldots, A_k$ jsou rozkladem jevu jistého, $P(A_j) > 0$ a $P(C) > 0$, pak pro libovolné $j = 1, 2, \ldots, k$ platí

$$\boxed{P(A_j \mid C) = \frac{P(C \mid A_j) \cdot P(A_j)}{\sum_{i=1}^{k} P(C \mid A_i) \cdot P(A_i)}}. \qquad (3\text{-}11)$$

Bayesův vzorec můžeme snadno dokázat, Jelikož jak $P(A_j) > 0$, tak i $P(C) > 0$, z definice podmíněné pravděpodobnosti dostaneme

$$P(A_j \cap C) = P\left(A_j \mid C\right) \cdot P(C) = P\left(C \mid A_j\right) \cdot P(A_j),$$

tedy $$P\left(A_j \mid C\right) = \frac{P\left(C \mid A_j\right) \cdot P(A_j)}{P(C)}.$$

Když za $P(C)$ dosadíme z věty o úplné pravděpodobnosti (3-10), dostaneme Bayesův vzorec (3-11).

Pokusme se trochu vysvětlit, k čemu se Bayesův vzorec používá. Někdy se říká, že s jeho pomocí počítáme pravděpodobnost příčin. Vraťme se k našemu příkladu o míchání slivovice. Jevem C je „náhodně vybraná molekula z výsledné směsi je molekula etanolu“, pak $P\left(A_j|C\right)$ je pravděpodobnost, že tato molekula pochází z nádoby j.

Uvažujme analogický následující příklad (poznámka pro biology - příklad je smyšlený, takže údaje z něho neodkazujte jako pozorovaná fakta).

Příklad 3-4: Čápi k nám přilétají třemi cestami, přes Bospor (přiletí tak 20% všech čápů, z toho je 3% černých), přes Sicílii (přiletí tak 30% všech čápů, z toho je 4% černých) a přes Gibraltar (přiletí tak 50% všech čápů, z toho je 5% černých). Relativní četnost černých čápů u nás je vlastně odhad pravděpodobnosti jevu C „náhodně vybraný čáp na území naší republiky je černý“. Zpozorujeme-li u nás černého čápa (nastal jev C), samozřejmě nemůžeme s jistotou určit, kterou ze tří cest přiletěl (pokud to není jeden z několika málo čápů, kteří jsou vybaveny vysílačkou a jsou sledováni v rámci projektu Africká Odysea), ale dosazením do Bayesova vzorce můžeme spočítat podmíněné pravděpodobnosti pro každou z těchto tří cest - $P\left(A_1|C\right)$, $P\left(A_2|C\right)$, $P\left(A_3|C\right)$.

$$P(A_1 \mid C) = \frac{P(C \mid A_1) \cdot P(A_1)}{\sum_{i=1}^{3} P(C \mid A_i) \cdot P(A_i)} =$$

$$= \frac{0,03 \cdot 0,20}{0,03 \cdot 0,20 + 0,04 \cdot 0,30 + 0,05 \cdot 0,50} = \frac{0,006}{0,043} \cong 0,14$$

$$P(A_2 \mid C) = \frac{P(C \mid A_2) \cdot P(A_2)}{\sum_{i=1}^{3} P(C \mid A_i) \cdot P(A_i)} =$$

$$= \frac{0,04 \cdot 0,30}{0,03 \cdot 0,20 + 0,04 \cdot 0,30 + 0,05 \cdot 0,50} = \frac{0,012}{0,043} \cong 0,28$$

$$P(A_3 \mid C) = \frac{P(C \mid A_3) \cdot P(A_3)}{\sum_{i=1}^{3} P(C \mid A_i) \cdot P(A_i)} =$$

$$= \frac{0,05 \cdot 0,50}{0,03 \cdot 0,20 + 0,04 \cdot 0,30 + 0,05 \cdot 0,50} = \frac{0,025}{0,043} \cong 0,58$$

Vidíme, že pravděpodobnost toho, že černý čáp přilétl přes Gibraltar je zhruba čtyřikrát větší než pravděpodobnost, že přiletěl přes Bospor a zhruba dvakrát větší než pravděpodobnost, že přiletěl přes Sicílii.

Příklad 3-5: (Komenda, Biometrie, str. 30-31): Jedno promile populace trpí určitou chorobou, kterou je možné prokázat bakteriologicky. Je žádoucí rozpoznat nositele, aby se zabránilo infekčnímu šíření a přehradily mechanismy přenosu. Bakteriologický test dává pozitivní výsledek u skutečně nakažených s pravděpodobností 0,98 (tzv. senzitivita testu), negativní výsledek u zdravých jedinců s pravděpodobností 0,99 (tzv. specificita testu). Spolehlivost a účinnost testu se hodnotí podle podílu nositelů infekce zjištěných mezi jedinci s pozitivním testem a podle podílu zdravých mezi jedinci, u nichž je výsledek testu negativní.

Náhodné jevy označíme následujícím způsobem:

C	jedinec je infikován (nositel nákazy, nemocný)
$\overline{C}$	jedinec je zdráv (komplementární jev k jevu C)
$+$	jedinec reagoval v testu pozitivně
$-$	jedinec reagoval v testu negativně (komplementární jev k jevu +)

Ze zadání úlohy platí

$$P(C) = 0,001 \ , \qquad P(+ \mid C) = 0,98 \ , \qquad P(- \mid \overline{C}) = 0,99$$

Pravděpodobnosti $P(C \mid +)$ a $P(\overline{C} \mid -)$ se pak spočítají podle Bayesova vzorce

$$P(C \mid +) = \frac{P(+ \mid C) \cdot P(C)}{P(+ \mid C) \cdot P(C) + P(+ \mid \bar{C}) \cdot P(\bar{C})} =$$

$$= \frac{0,98 \cdot 0,001}{0,98 \cdot 0,001 + 0,01 \cdot 0.999} = \frac{0,00098}{0,01097} = 0,0893$$

$$P(\bar{C} \mid -) = \frac{P(- \mid \bar{C}) \cdot P(\bar{C})}{P(- \mid \bar{C}) \cdot P(\bar{C}) + P(- \mid C) \cdot P(C)} =$$

$$= \frac{0,99 \cdot 0,999}{0,99 \cdot 0,999 + 0,02 \cdot 0.001} = \frac{0,98901}{0,98903} = 0,9998$$

Zatímco jedinec náhodně vybraný z populace je nosičem choroby s pravděpodobností 0,001, bude subjekt s pozitivním nálezem nosičem choroby s pravděpodobností 0,089, tedy s možností téměř devadesátkrát vyšší. Test funguje jako metoda „zhušťování podezřelých.

Shrneme-li naše dosavadní poznatky, vidíme, že Bayesovu větu můžeme užít pro zpřesnění *apriorních pravděpodobností* $P(A_1)$, $P(A_2)$, ... , $P(A_k)$, známe-li podmíněné pravděpodobnosti $P(C|A_1)$, $P(C|A_2)$, ... , $P(C|A_k)$. Zpřesnění je založeno na tom, že už víme, že nastal jev C a počítáme podmíněné pravděpodobnosti $P(A_1|C)$, $P(A_2|C)$, ... , $P(A_k|C)$. Těmto pravděpodobnostem se říká *aposteriorní pravděpodobnosti*.

Obecně můžeme říci, že použití Bayesova vzorce je jeden z postupů, jak řešit diagnostickou úlohu, totiž určit pravděpodobnou příčinu pozorovaného jevu *C*. Podle Bayesova vzorce můžeme spočítat pravděpodobnost všech možných příčin pozorovaného jevu *C* a příčinu nejpravděpodobnější pak považovat za příčinu skutečnou.

Bayesovské metody se v současné době stále často užívanými postupy v různých demografických, epidemiologických a environmentálních grafických informačních systémech, zejména k časovému a prostorovému vyhlazování empirických četností.

Souhrn:

- Výsledkem náhodného pokusu musí být právě jeden z množiny alespoň dvou možných výsledků.
- Uvažovaný náhodný pokus je možno nezávisle a za stejných podmínek opakovat.
- Výsledkem náhodného pokusu je náhodný jev.
- Elementární jev není složen z jiných jevů (nelze jej napsat jako sjednocení dvou elementárních jevů).
- Vztahy mezi jevy lze znázornit Vennovými diagramy jako vztahy množin
- Náhodnému jevu přiřazujeme pravděpodobnost.
- Pravděpodobnost musí splňovat vlastnosti dané axiomy teorie pravděpodobnosti.

- Pravděpodobnost sjednocení neslučitelných jevů se spočítá jako součet pravděpodobností jednotlivých jevů.
- $P(A) = 1 - P(\overline{A})$
- Pravděpodobnost jevu *A* se počítá jako podíl počtu výsledků příznivých (ve kterých nastane jev *A*) ku počtu všech možných výsledků.
- Podmíněná pravděpodobnost je definována jako $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$.
- Pravděpodobnost průniku nezávislých jevů se spočítá jako součin pravděpodobností jednotlivých jevů.

Kontrolní otázky:

1. *Co je jev jistý? Jaká je jeho pravděpodobnost?*
2. *Co je jev opačný? Co vznikne sjednocením jevu s jevem jemu opačným?*
3. *Kdy se pravděpodobnost sjednocení jevů spočítá jako součet pravděpodobností jednotlivých jevů?*
4. *Jakou podmínku musí splňovat jevy, abychom pravděpodobnost jejich průniku mohli spočítat jako součin pravděpodobností jednotlivých jevů?*
5. *Co musí platit o jevech, abychom mohli užít vztah pro úplnou pravděpodobnost a Bayesův vzorec?*
6. *Co jsou nezávislé jevy? Uveďte příklady nezávislých jevů.*

Pojmy k zapamatování:

- náhodný pokus, náhodný jev
- elementární jev
- jev opačný, neslučitelné jevy
- pravděpodobnost a její vlastnosti
- podmíněná pravděpodobnost
- nezávislé jevy
- počítání pravděpodobností jevů, klasická pravděpodobnost
- statistická definice pravděpodobnosti
- úplná pravděpodobnost, Bayesův vzorec a jeho užití

Korespondenční úlohy budou zadávány vždy na začátku semestru.

Z klasické pravděpodobnosti budou čtyři příklady.

Druhá část kapitoly vám zabere také asi čtyři až pět hodin. Obsahuje řadu klíčových pojmů, důležitých pro správné pochopení základů teorie pravděpodobnosti a jejich pozdější aplikaci ve statistice. Počítejte s tím, že k této kapitole se budete vracet a některé věci pochopíte důkladněji až při opakovaném studiu, zejména když bude motivováno pochopením aplikace těchto poznatků, se kterým se setkáte v dalších částích této kapitoly a v kapitolách o induktivní statistice.

3.2 Náhodná veličina a rozdělení pravděpodobnosti

Náhodná veličina je kromě pravděpodobnosti další abstraktní představou, která dovoluje náhodnému jevu (tentokrát *jen elementárnímu*) přiřadit číselnou hodnotu. Formálně *náhodná veličina* je funkce (zobrazení) X v systému elementárních jevů Ω, která každému elementárnímu jevu $E \in \Omega$ přiřadí právě jediné reálné číslo.

Náhodné veličiny většinou označujeme velkými písmeny z konce abecedy - X, Y, Z, W apod., zatímco hodnoty, kterých náhodné veličiny nabývají, se označují odpovídajícími malými písmeny - x, y, z, w ap. Zápis $X = x$ pak čteme náhodná veličina X má hodnotu x, podobně $Y < y$ čteme hodnota náhodné veličiny Y je menší než y atd.

Pro pochopení pojmu náhodná veličina považujme náhodnou veličinu za jakýsi abstraktní pohled na měření. Měření totiž splňuje představu náhodného pokusu - v okamžiku, kdy vstupujeme na váhu, nevíme přesně, jaká bude naše hmotnost (78 kg, 79 kg či jiná?); nevíme, jaká bude koncentrace oxidu siřičitého ve vzorku ovzduší; jaký bude počet druhů ptáků odchycených ke kroužkování atp. Pozorovaná hodnota není deterministická, je ovlivňována shodou náhod, některé hodnoty jsou pravděpodobnější, jiné méně pravděpodobné.

Tím, že náhodnou veličinou umíme zobrazit výsledky náhodného pokusu na číselnou osu, umíme elementární jevy uspořádat. Jelikož jevu je přiřazena pravděpodobnost, umíme pak definovat i *rozdělení pravděpodobnosti*. Volně můžeme říci, že pravděpodobnost jevu jistého, tedy 1, je rozdělena (rozložena) nad body nebo intervaly číselné osy. Toto rozdělení pravděpodobnosti lze jednoznačně popsat *distribuční funkcí*

$$F(x) = P(X < x) \tag{1}$$

Distribuční funkce je definována pro všechny body číselné osy, tedy pro $x \in (-\infty, +\infty)$. Jelikož distribuční funkce je pravděpodobnost, je jasné, že pro její hodnoty musí platit $0 \le F(x) \le 1$.

Distribuční funkce je *neklesající*, tj. pro $x_1 < x_2$ platí

$$F(x_1) \le F(x_2)\,. \tag{2}$$

Toto tvrzení snadno dokážeme. Jev $X < x_2$ je sjednocením disjunktních jevů $X < x_1$ a $x_1 \le X < x_2$, takže

$$F(x_2) = P(X < x_2) = \\ = P(X < x_1) + P(x_1 \le X < x_2) = F(x_1) + P(x_1 \le X < x_2) \qquad (3)$$

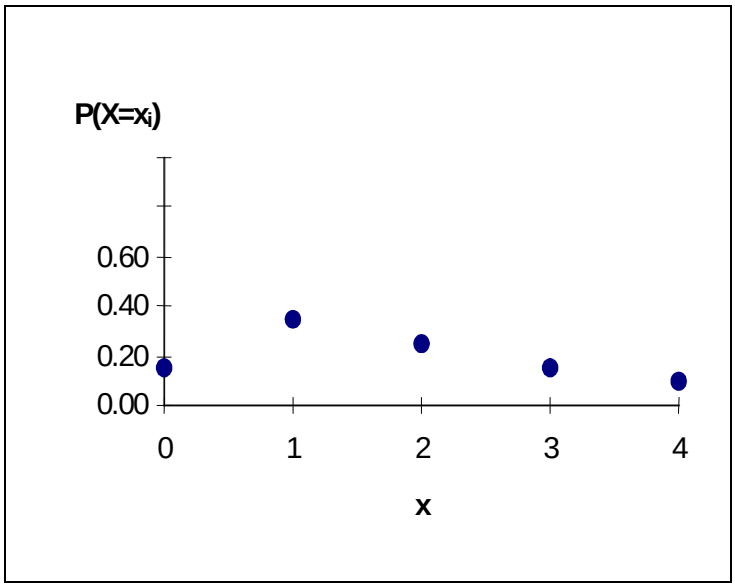
Jelikož $P(x_1 \le X < x_2) \ge 0$ (je to pravděpodobnost), platí tudíž $F(x_1) \le F(x_2)$, tzn. že distribuční funkce je neklesající.

Podobně jako v odstavci 1.3 jsme rozlišovali spojité a diskrétní škály (a měřené veličiny), je účelné podobně rozlišovat i náhodné veličiny na spojité a diskrétní.

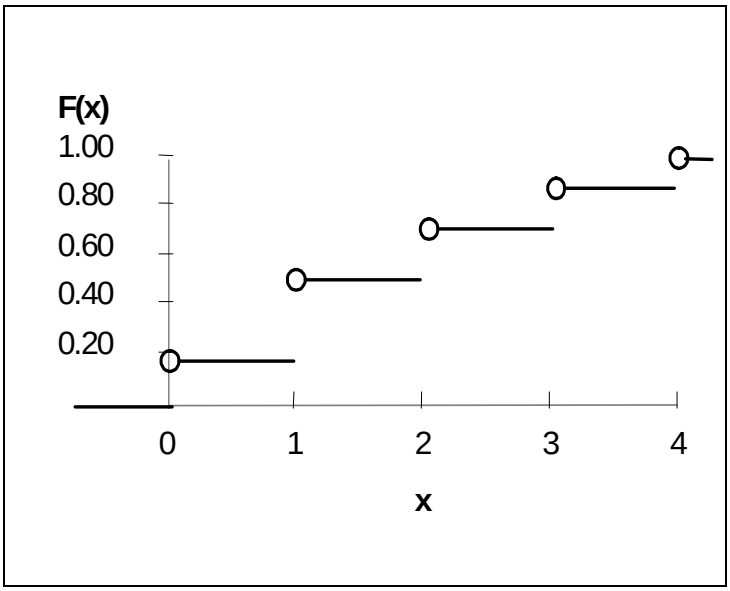
Diskrétní (nespojitá) náhodná veličina může nabývat pouze diskrétních (tj. od sebe oddělených) hodnot $x_1, x_2, \dots, x_k$. Pravděpodobnostní rozdělení (a tím i distribuční funkce) je jednoznačně určena dvojicemi hodnot $\left[x_i, P(X = x_i)\right]$, $i = 1, 2, \dots, k$, tj. tabulkou o dvou sloupcích a k řádcích. Této funkci $P(X = x_i)$ definované pro všechny hodnoty $x_1, x_2, \dots, x_k$, se říká *pravděpodobnostní funkce*.

Příklad pravděpodobnostní funkce je uveden v tabulce a její grafické znázornění vidíme na následujícím obrázku.

x_i	$P(X=x_i)$	$F(x)$
0	0.15	0.00
1	0.35	0.15
2	0.25	0.50
3	0.15	0.75
4	0.10	0.90
>4		1.00



Pravděpodobnostní funkce



Distribuční funkce

Hodnoty distribuční funkce diskrétní náhodné veličiny pak jsou určeny vztahem

$$F(x) = \sum_{x_i < x} P(X = x_i)\,, \qquad (4)$$

čili distribuční funkce je schodovitá funkce s výškou „schodu“ rovnou hodnotě $P(X = x_i)$ v bodě x_i.

Spojitá náhodná veličina může nabývat všech reálných hodnot nebo alespoň všech hodnot z nějakého konečného intervalu. Hodnoty náhodné veličiny pokrývají interval hustě, tedy je jich nespočetně mnoho.

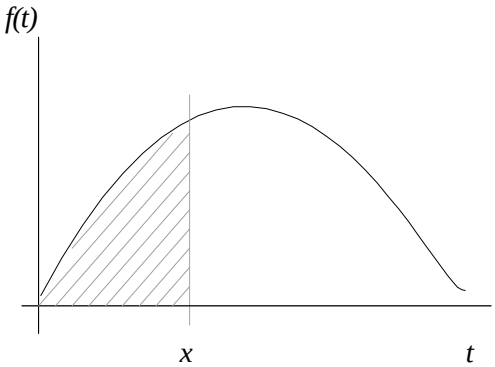
Distribuční funkce spojité náhodné veličiny (také říkáme distribuční funkce spojitého rozdělení) se vyjádří ve tvaru

$$F(x) = \int_{-\infty}^{x} f(t)dt\ . \qquad (5)$$

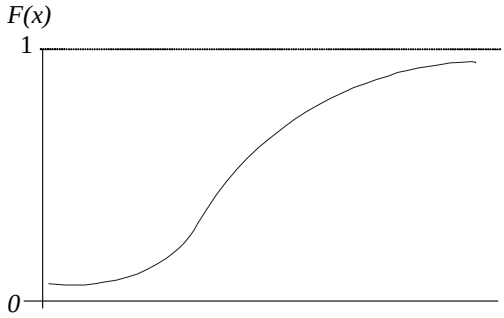
kde $f(t)$ je nezáporná funkce zvaná *hustota* (nebo hustota pravděpodobnosti). Ze vztahu (5) můžeme odvodit i další vlastnosti hustoty

$$\int_{-\infty}^{+\infty} f(x)dx = 1 \qquad \text{a} \qquad f(x) = \frac{dF(x)}{dx}\text{, pokud derivace existuje.} \qquad (6)$$

Význam vztahu (5) lze ilustrovat následujícím obrázkem, hodnota distribuční funkce v bodě *x* je rovna obsahu vyšrafované plochy vlevo od svislé přímky *t* = *x*.



Hustota spojité náhodné veličiny

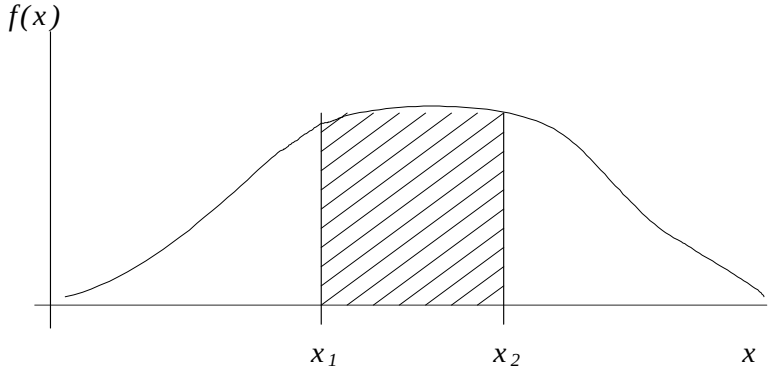


Distribuční funkce spojité náhodné veličiny

Jak ukazuje vztah (3), pravděpodobnost, že hodnota náhodné veličiny je v intervalu $\langle x_1, x_2)$, $x_1 < x_2$, lze určit jako rozdíl hodnot distribuční funkce

$$P(x_1 \le X < x_2) = F(x_2) - F(x_1) = \int_{-\infty}^{x_2} f(x)dx - \int_{-\infty}^{x_1} f(x)dx = \int_{x_1}^{x_2} f(x)dx \qquad (7)$$

Tuto pravděpodobnost můžeme znázornit jako velikost vyšrafované plochy na následujícím obrázku.



Povšimněme si, že bude-li se zmenšovat rozdíl $(x_2 - x_1)$, bude se zmenšovat i pravděpodobnost $P(x_1 \le X < x_2)$, až pro $x_2 = x_1$ bude $P(X = x_1) = 0$, takže platí $P(x_1 \le X < x_2) = P(x_1 < X < x_2)$.

Porovnejme graf pravděpodobnostní funkce se sloupcovým grafem relativních četností v kap. 2, resp. graf hustoty pravděpodobnosti s histogramem relativních četností. Vidíme, že obě dvojice grafů popisují téměř totéž - rozdělení četnosti hodnot, rozdíl je jen v tom, že grafy v kap. 2 popisují rozdělení pozorovaných hodnot (tzv. empirické rozdělení), zatímco grafy v této kapitole popisují teoretické (modelové) rozdělení pravděpodobnosti spojené s abstraktní představou náhodné veličiny. Také vidíme, že distribuční funkce je obdobou kumulativní relativní četnosti.

3.3 Charakteristiky náhodných veličin

V kapitole 2 jsme zavedli pro popis pozorovaných dat charakteristiky polohy, variability atd. Analogické charakteristiky existují i pro náhodné veličiny. Analogií průměru je *střední hodnota* náhodné veličiny, *E(X).* Pokud je význam jasný, závorky můžeme vynechat a psát *EX*. Pro diskrétní náhodnou veličinu *X* je střední hodnota definována jako

$$E(X) = \sum_i x_i P(X = x_i) \ . \qquad (1)$$

Vidíme, že vztah (1) je přesnou obdobou vztahu pro výpočet váženého průměru, kdy využíváme relativních četností hodnot x_i.

Pro spojitou veličinu s hustotou *f(x)* je střední hodnota definována vztahem

$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx \ . \qquad (2)$$

Jestliže máme nějakou reálnou funkci *g(x)* - např. logaritmus, druhá mocnina ap. - pak tato funkce náhodné veličiny *X* je opět náhodná veličina, *Y = g(X)* a její střední hodnota je

$$E(Y) = E\left[g(X)\right] = \sum_i g(x_i)\ P(X = x_i) \qquad (3)$$

pro diskrétní veličinu *X* (pochopitelně i veličina *Y* je diskrétní).

Pro spojitou náhodnou veličinu Y = g(X) je pak střední hodnota dána vztahem

$$E(Y) = E\left[g(X)\right] = \int_{-\infty}^{\infty} g(x)\ f(x)\ dx\,, \qquad (4)$$

f(x) je hustota pravděpodobnosti náhodné veličiny *X*.

Charakteristikou variability je *rozptyl*, *var(X)*, definovaný jako střední hodnota druhé mocniny (někdy říkáme čtverce) odchylky od střední hodnoty *E(X)*, tedy

$$\mathrm{var}(X) = E\left[X - E(X)\right]^2 \qquad (5)$$

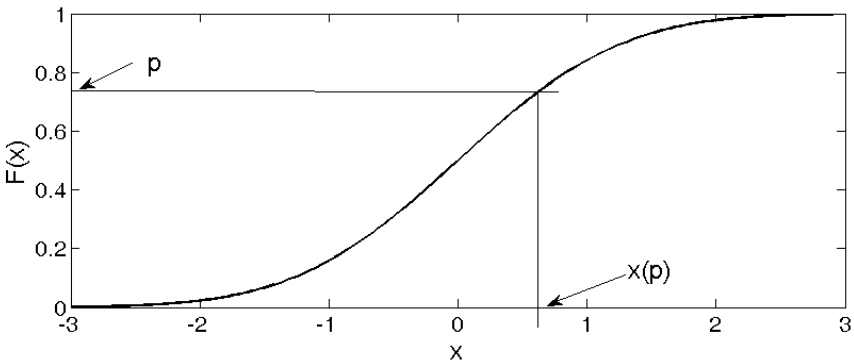
Odmocnina z rozptylu, $\sqrt{\mathrm{var}(X)}$, se nazývá *směrodatná odchylka*.

Podobně jako jsme v kapitole 2 zavedli empirické kvantily, jsou definovány i kvantily pro náhodnou veličinu. Kvantil (říkáme *p*-kvantil) je taková hodnota *x(p)*, pro kterou platí

$$P\left[X \le x(p)\right] \ge p \quad \text{a současně} \quad P\left[X \ge x(p)\right] \ge 1 - p \qquad (6)$$

Kvantil *x(0,5)* se nazývá (teoretický) *medián*, kvantily *x(0,25)* a *x(0,75)* jsou dolní a horní *kvartil*. Kvantily, kdy *p = 0,1; 0,2; ... ; 0,9* jsou *decily* atd.

Kvantil spojitého rozdělení s rostoucí distribuční funkcí je inverzní funkce k funkci distribuční, což ukazuje následující obrázek. Pro zvolenou hodnotu *p* nalezneme na vodorovné ose hodnotu kvantilu *x(p)*.



Obrázek – kvantil jako inverzní funkce k distribuční funkci

Další charakteristikou polohy podobně jako u empirického rozdělení je *modus*, což je hodnota, ve které má pravděpodobnostní funkce, resp. hustota maximum.

Dále se k charakterizování rozdělení náhodné veličiny užívají momenty. *Obecný k*-tý *moment* je definován jako

$$\mu'_k = E(X^k), \qquad k = 1, 2, \dots, \tag{7}$$

k-tý *centrální moment* je

$$\mu_k = E\left[(X - EX)^k\right]. \tag{8}$$

Šikmost rozdělení náhodné veličiny se charakterizuje hodnotou

$$\gamma_1 = \frac{\mu_3}{\mu_2\sqrt{\mu_2}} = \frac{E\left[(X - EX)^3\right]}{\mathrm{var}(X)\sqrt{\mathrm{var}(X)}} \tag{9}$$

a *špičatost* rozdělení je charakterizována jako

$$\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3 = \frac{E\left[(X - EX)^4\right]}{\left[\mathrm{var}(X)\right]^2} - 3 \tag{10}$$

Pro charakteristiky náhodných veličin můžeme odvodit celou řadu užitečných vztahů. Některé z těchto vztahů zde uvedeme bez důkazu, který je ponechán pro samostatná cvičení.

$$E(X + Y) = E(X) + E(Y). \tag{11}$$

Střední hodnota součtu náhodných veličin je rovna součtu středních hodnot. Je zřejmé, že podobný vztah platí i pro více než dva sčítance. Pro diskrétní veličiny *X,Y* můžeme vztah (11) snadno dokázat:

$$E(X+Y)=\sum_i\sum_j (x_i+y_j)\ P\left[(X=x_i)\cap(Y=y_j)\right]=$$
$$=\sum_i x_i\sum_j P\left[(X=x_i)\cap(Y=y_j)\right]+\sum_j y_j\sum_i P\left[(X=x_i)\cap(Y=y_j)\right]=$$
$$=\sum_i x_i P(X=x_i)+\sum_j y_j P(Y=y_j)=E(X)+E(Y)$$

Jsou-li a, b konstanty, pak

$$E(a+bX)=a+bE(X) \tag{12}$$

neboť

$$E(a+bX)=\sum_i (a+bx_i)P(X=x_i)=$$
$$=a\sum_i P(X=x_i)+b\sum_i x_i P(X=x_i)=a+bE(X)$$

a pro rozptyl platí

$$\mathrm{var}(a+bX)=b^2\ \mathrm{var}(X) \tag{13}$$

neboť

$$\mathrm{var}(a+bX)=E\left[(a+bX)-E(a+bX)\right]^2=$$
$$=E\left[a+bX-a-bE(X)\right]^2=E\left[b\ (X-E(X)\right]^2=$$
$$=b^2E\left[X-E(X)\right]^2=b^2\mathrm{var}(X)$$

Je-li $b=0$, pak dostaneme $E(a)=a$ a $\mathrm{var}(a)=0$.

Pro *normovanou* náhodnou veličinu $U=\dfrac{X-E(X)}{\sqrt{\mathrm{var}(X)}}$ platí

$$E(U)=0 \quad \text{a} \quad \mathrm{var}(U)=1\ , \tag{14}$$

neboť

$$E(U)=E\left[\frac{X-E(X)}{\sqrt{\mathrm{var}(X)}}\right]=\frac{1}{\sqrt{\mathrm{var}(X)}}E\left[X-E(X)\right]=\frac{1}{\sqrt{\mathrm{var}(X)}}\left[E(X)-E(X)\right]=0$$

a pro rozptyl *normované* náhodné veličiny platí

$$\mathrm{var}(U)=\mathrm{var}\left[\frac{X-E(X)}{\sqrt{\mathrm{var}(X)}}\right]=\frac{\mathrm{var}\left[X-E(X)\right]}{\mathrm{var}(X)}=\frac{\mathrm{var}(X)}{\mathrm{var}(X)}=1.$$

Dále platí

$$\mathrm{var}(X)=E(X^2)-\left[E(X)\right]^2, \tag{15}$$

neboť

$$\mathrm{var}(X)=E\left[X-E(X)\right]^2=E\left[X^2-2\ X\ E(X)+(E(X))^2\right]=$$
$$=E(X^2)-2\ E(X)\ E(X)+(E(X))^2=E(X^2)-\left[E(X)\right]^2$$

Rozptyl náhodné veličiny můžeme tedy vyjádřit jako rozdíl střední hodnoty jejího čtverce a čtverce střední hodnoty.

V odstavci 2.6 jsme se zabývali vztahem dvou veličin (dvou sloupců datové matice). Podobně můžeme popsat vztah dvou náhodných veličin *(X,Y)*. Této dvojici se říká dvourozměrný *náhodný vektor*. Rozdělení náhodného vektoru je popsáno *sdruženou distribuční funkcí*

$$F_{XY}(x,y) = P(X < x, Y < y) \tag{17}$$

Označení $(X < x, Y < y)$ znamená náhodný jev, že náhodná veličina X nabývá hodnot menších než x a současně Y nabývá hodnot menších než y.

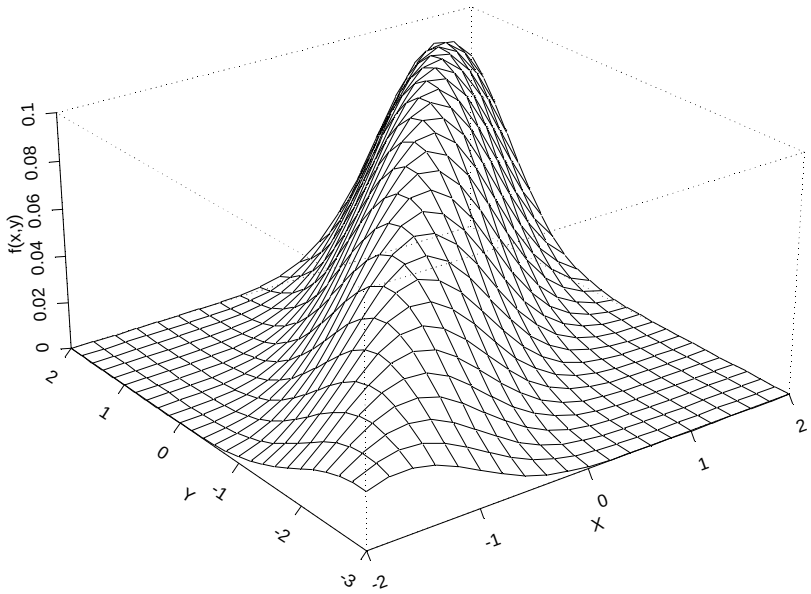
Rozdělení diskrétních náhodných vektorů lze popsat i *sdruženou pravděpodobnostní funkcí* $P(X = x_i, Y = y_j)$ definovanou pro všechny možné dvojice hodnot (x_i, y_j), jichž náhodný vektor může nabývat. Sdružená pravděpodobnostní funkce je tedy dvourozměrná tabulka obsahující hodnoty pravděpodobnosti.

		X				
		x_1	x_2	…	x_C	marginální
	y_1	$P(X=x_1,Y=y_1)$	$P(X=x_2,Y=y_1)$	…	$P(X=x_C,Y=y_1)$	$P(Y=y_1)$.
	y_2	$P(X=x_1,Y=y_2)$	$P(X=x_2,Y=y_2)$	…	$P(X=x_C,Y=y_2)$	$P(Y=y_2)$.
Y	:	:	:	:	:	:
	y_R	$P(X=x_1,Y=y_R)$	$P(X=x_2,Y=y_R)$	…	$P(X=x_C,Y=y_R)$	$P(Y=y_R)$.
	marginální	$P(X=x_1)$	$P(X=x_2)$	…	$P(X=x_C)$	1

Rozdělení spojitého vektoru popisuje *sdružená hustota* $f_{XY}(x,y)$, pro kterou platí

$$F_{XY}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{XY}(u,v)\,du\,dv\,. \tag{18}$$

Příklad grafického znázornění sdružené hustoty dvourozměrného náhodného vektoru je na následujícím obrázku.



Hodnota distribuční funkce v bodě (x, y) v souladu s (18) je rovna objemu tělesa, které vznikne pravoúhlým vykrojením právě v tomto bodu (x, y). Celkový objem tělesa pod plochou sdružené hustoty je samozřejmě roven jedné.

Podobně jako jsme v kapitole 2 zavedli marginální četnosti, i zde dojdeme k *marginálnímu rozdělení*.

Marginální distribuční funkce jsou

$$F(x) = \lim_{y\to\infty}\left[F_{XY}(x,y)\right] \qquad \text{a} \qquad F(y) = \lim_{x\to\infty}\left[F_{XY}(x,y)\right]. \tag{19}$$

Marginální pravděpodobnostní funkce pro diskrétní náhodný vektor dostaneme sčítáním pravděpodobností přes celý sloupec, resp. řádek, tedy

$$P(X = x_i) = \sum_j P(X = x_i, Y = y_j) \text{ a } P(X = y_j) = \sum_i P(Y = x_i, Y = y_j) \tag{20}$$

a podobně pro spojitý vektor jsou *marginální hustoty*

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x,y)dy \qquad \text{a} \qquad f_Y(y) = \int_{-\infty}^{+\infty} f_{XY}(x,y)dx\,. \tag{21}$$

Nyní můžeme zavést důležitý pojem - *nezávislost dvou náhodných veličin*. Náhodné veličiny *X, Y* jsou *nezávislé*, když jsou nezávislé dva náhodné jevy, jev $X < x$ a $Y < y$. Jak víme z odst. 3.1, pravděpodobnost současného nastání dvou nezávislých jevů se vypočítá jako součin pravděpodobností každého z těchto jevů. Tedy

$$P\left[(X < x) \cap (Y < y)\right] = P(X < x) \cdot P(Y < y)\,. \tag{22}$$

Pravděpodobnost na levé straně rovnice (22) definuje sdruženou distribuční funkci, pravděpodobnosti na pravé straně pak marginální distribuční funkce, takže rovnici můžeme přepsat ve tvaru (23)

$$F_{XY}(x,y) = F_X(x) \cdot F_Y(y). \tag{23}$$

Pro nezávislé náhodné veličiny platí, že sdružená distribuční funkce je rovna součinu marginálních distribučních funkcí.

Také sdružená pravděpodobnostní funkce dvou *nezávislých* veličin je rovna součinu marginálních pravděpodobnostních funkcí:

$$P(X = x_i, Y = y_j) = P(X = x_i) \cdot P(Y = y_j) \tag{24}$$

a podobný vztah platí v případě spojitých nezávislých veličin i pro sdruženou hustotu

$$f_{XY}(x,y) = f_X(x) \cdot f_Y(y). \tag{25}$$

Nejsou-li dvě náhodné veličiny nezávislé, existuje mezi nimi nějaká závislost. Tato závislost není deterministická, jde o náhodné veličiny. Říkáme, že závislost je *stochastická*. Vztah dvou náhodných veličin lze číselně charakterizovat *kovariancí* (teoretickou kovariancí, srovnej s odstavcem 2.6), která je definována

$$\mathrm{cov}(X,Y) = E\left[(X - EX)\,(Y - EY)\right] \tag{26}$$

Pro diskrétní náhodné veličiny se kovariance spočítá jako

$$\mathrm{cov}(X,Y) = \sum_i \sum_j (x_i - EX)\,(y_j - EY)\; P(X = x_i, Y = y_j) \tag{27}$$

a pro spojité veličiny je kovariance

$$\mathrm{cov}(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - EX)\,(y - EY)\; f_{XY}(x,y)\; dx\,dy \tag{28}$$

Z definice kovariance (26) vidíme, že

$\mathrm{cov}(X,Y) = \mathrm{cov}(Y,X)$ .a $\mathrm{cov}(X,X) = \mathrm{var}(X)$.

Kovariance může nabývat libovolných reálných hodnot, nezáporných i záporných. Pomocí kovariance však můžeme definovat jinou charakteristiku závislosti dvou náhodných veličin (jejichž rozptyly jsou kladné), *korelační koeficient*:

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)}\,\sqrt{\mathrm{var}(Y)}}, \tag{29}$$

pro který platí $|\rho_{X,Y}| \le 1$, čili korelační koeficient může nabývat hodnot jen z intervalu $\left[-1, 1\right]$. Náhodné veličiny X, Y se nazývají *nekorelované*, jestliže $\mathrm{cov}(X,Y) = 0$, a tedy i korelační koeficient je nulový. Jsou-li veličiny nezávislé, jsou i nekorelované. Opačně to neplatí, nulový korelační koeficient nemusí nutně znamenat nezávislost veličin.

Pro *nezávislé* náhodné veličiny X ,Y platí, že

$$E(XY) = E(X)\,E(Y). \tag{30}$$

Rozptyl součtu dvou náhodných veličin je

$$\operatorname{var}(X+Y) = E\left[X+Y-E(X+Y)\right]^2 = E\left[(X-EX)+(Y-EY)\right]^2 =$$
$$= E\left[(X-EX)^2 + 2(X-EX)(Y-EY) + (Y-EY)^2\right] =$$
$$= \operatorname{var}(X) + 2\operatorname{cov}(X,Y) + \operatorname{var}(Y) =$$
$$= \operatorname{var}(X) + 2\rho_{X,Y}\sqrt{\operatorname{var}(X)}\sqrt{\operatorname{var}(Y)} + \operatorname{var}(Y)$$

Podobně pro rozdíl náhodných veličin dostaneme

$$\operatorname{var}(X-Y) = E\left[X-Y-E(X-Y)\right]^2 = E\left[(X-EX)-(Y-EY)\right]^2 =$$
$$= E\left[(X-EX)^2 - 2(X-EX)(Y-EY) + (Y-EY)^2\right] =$$
$$= \operatorname{var}(X) - 2\operatorname{cov}(X,Y) + \operatorname{var}(Y) =$$
$$= \operatorname{var}(X) - 2\rho_{X,Y}\sqrt{\operatorname{var}(X)}\sqrt{\operatorname{var}(Y)} + \operatorname{var}(Y)$$

Vidíme, že rozptyl součtu, resp. rozdílu dvou náhodných veličin závisí i na tom, zda jsou veličiny korelovány. Specielně pro *nekorelované* veličiny vidíme, že platí

$$\operatorname{var}(X+Y) = \operatorname{var}(X-Y) = \operatorname{var}(X) + \operatorname{var}(Y) \tag{31}$$

Dále, pokud náhodná veličina Y je lineární funkcí náhodné veličiny X, tzn. $Y = aX + b$, $a \neq 0$, pak platí

$$\rho_{X,Y} = \rho_{X,aX+b} = \begin{cases} 1 & je-li \quad a > 0 \\ -1 & je-li \quad a < 0 \end{cases} \tag{32}$$

Platnost tohoto vztahu snadno dokážeme. Z definice kovariance (26) dostaneme

$$\operatorname{cov}(X, aX+b) = E\left\{(X-EX)\left[(aX+b)-E(aX+b)\right]\right\} =$$
$$= E\left[(X-EX)(aX+b-aEX-b)\right] = a\,E\left[(X-EX)(X-EX)\right] = a\operatorname{var}(X)$$

a po dosazení do definice korelačního koeficientu vidíme ($\operatorname{var}(X) > 0$), že

$$\rho_{X,Y} = \rho_{X,aX+b} = \frac{a\operatorname{var}(X)}{\sqrt{\operatorname{var}(X)\,a^2\operatorname{var}(X)}} = \frac{a}{\sqrt{a^2}} = \begin{cases} 1 & je-li \quad a > 0 \\ -1 & je-li \quad a < 0 \end{cases}$$

Vyjádříme-li tento výsledek slovně, znamená to, že pro přesnou deterministickou lineární závislost dvou veličin je jejich koeficient korelace v absolutní hodnotě roven jedné.

Souhrn:

- Náhodná veličina je zobrazením elementárních jevů na číselnou osu.
- Rozdělení pravděpodobnosti náhodné veličiny je jednoznačně definováno distribuční funkcí.
- Hodnota distribuční funkce v bodu x je pravděpodobnost jevu, že náhodná veličina je menší než x, $F(x) = P(X < x)$.
- Distribuční funkce diskrétní náhodné veličiny je definována jako $F(x) = \sum_{x_i < x} P(X = x_i)$, kde $P(X = x_i)$ je pravděpodobnostní funkce.

- Distribuční funkce spojité náhodné veličiny je definována jako $F(x) = \int_{-\infty}^{x} f(t)dt$, kde $f(t)$ je hustota.
- Pravděpodobnost, že hodnota náhodné veličiny je v intervalu $\langle x_1, x_2)$, $x_1 < x_2$, lze určit jako rozdíl hodnot distribuční funkce $P(x_1 \le X < x_2) = F(x_2) - F(x_1)$.
- Střední hodnota diskrétní náhodné veličiny je $E(X) = \sum_i x_i P(X = x_i)$.
- Střední hodnota spojité náhodné veličiny je $E(X) = \int_{-\infty}^{+\infty} x f(x)dx$.
- Rozptyl je definován jako $\mathrm{var}(X) = E\left[X - E(X)\right]^2$.
- *p*-kvantil je taková hodnota *x(p)*, pro kterou platí $P\left[X \le x(p)\right] \ge p$ a současně $P\left[X \ge x(p)\right] \ge 1 - p$.
- Pro spojitou veličinu je kvantil inverzní funkce k distribuční funkci.
- Střední hodnota součtu náhodných veličin je rovna součtu středních hodnot, $E(X+Y) = E(X) + E(Y)$.
- $E(a+bX) = a + bE(X)$, $\mathrm{var}(a+bX) = b^2\ \mathrm{var}(X)$.
- $\mathrm{var}(X) = E(X^2) - \left[E(X)\right]^2$.
- Pro *nezávislé* náhodné veličiny platí, že sdružená distribuční funkce je rovna součinu marginálních distribučních funkcí..
- *Stochastickou* závislost dvou náhodných veličin lze číselně charakterizovat *kovariancí*, nebo *korelačním koeficientem*, $\rho_{X,Y} = \dfrac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)}}$, $|\rho_{X,Y}| \le 1$.
- Náhodné veličiny *X*, *Y* se nazývají *nekorelované*, jestliže $\mathrm{cov}(X,Y) = 0$, tedy i korelační koeficient je nulový.
- *Nezávislé* veličiny jsou *nekorelované*, naopak to nemusí platit.

Kontrolní otázky:

1. *Co je to náhodná veličina a jak je definováno její pravděpodobnostní rozdělení?*
2. *Jaký je vztah distribuční a pravděpodobnostní funkce, resp. distribuční funkce a hustoty?*
3. *Jak se spočítá pravděpodobnost, že hodnota náhodné veličiny je v intervalu* $\langle x_1, x_2)$, $x_1 < x_2$ *?*
4. *Může být hodnota distribuční funkce záporná?*
5. *Může být hodnota distribuční funkce větší než jedna?*
6. *Co je to střední hodnota náhodné veličiny?*
7. *Co je to rozptyl náhodné veličiny?*
8. *Na grafu distribuční funkce spojité náhodné veličiny si ujasněte, jak se určí p- kvantil.*

9. *Jsou-li dvě náhodné veličiny nezávislé, co platí pro sdruženou distribuční funkci, pro sdruženou pravděpodobnostní funkci, resp.pro sdruženou hustotu?*
10. *Kdy o dvou veličinách říkáme, že jsou nekorelované?*

Pojmy k zapamatování:

- náhodná veličina
- diskrétní náhodná veličina, spojitá náhodná veličina
- rozdělení pravděpodobnosti
- distribuční funkce, pravděpodobnostní funkce, hustota
- střední hodnota
- rozptyl
- kvantil
- šikmost a špičatost rozdělení
- náhodný vektor, pravděpodobnostní rozdělení náhodného vektoru
- sdružená distribuční funkce, marginální distribuční funkce
- nezávislé veličiny
- stochastická závislost
- kovariance, korelační koeficient

Následující část kapitoly o pravděpodobnosti vám zabere asi tři až čtyři hodiny. Můžete na tuto část kapitoly nahlížet jako na aplikaci poznatků z části předcházející. Opět počítejte s tím, že k této kapitole se budete vracet, neboť její důkladné pochopení je potřebné při aplikacích induktivní statistiky.

3.4 Příklady diskrétních rozdělení

3.4.1 Alternativní rozdělení

Toto rozdělení má náhodná veličina, která nabývá pouze hodnot 0 a 1 s pravděpodobnostmi $P(X=1)=p$ a $P(X=0)=1-p$. Hodnota p, $0<p<1$, se nazývá *parametr* rozdělení.

Příkladem takové náhodně veličiny je počet lvů při hodu jednou mincí, kdy buď padne jeden lev nebo žádný.

Střední hodnotu alternativní náhodné veličiny snadno určíme podle definice
$E(X)=\sum_i x_i\ P(X=x_i)=0\,.\,(1-p)+1\,.\,p=p$.

Podobně rozptyl
$\mathrm{var}(X)=E(X^2)-\left[E(X)\right]^2=0^2\,.\,(1-p)+1^2\,.\,p-p^2=p-p^2=p\,(1-p)$.

3.4.2 Binomické rozdělení

Toto rozdělení má náhodná veličina Y, která vznikne jako součet n nezávislých alternativně rozdělených náhodných veličin se stejným parametrem p, tedy

$$Y = X_1 + X_2 + \ldots + X_n$$

Příkladem takové náhodné veličiny je počet lvů při hodu n mincemi, při čemž pro každou minci je pravděpodobnost, že padne lev, rovna p.

Střední hodnota binomicky rozdělené náhodné veličiny je součtem středních hodnot jednotlivých sčítanců

$$E(Y) = E(X_1 + X_2 + \ldots + X_n) = \sum_{i=1}^{n} E(X_i) = np \qquad (1)$$

a rozptyl je opět součet rozptylů jednotlivých sčítanců (veličiny jsou nezávislé)

$$\mathrm{var}(Y) = \mathrm{var}(X_1 + X_2 + \ldots + X_n) = \sum_{i=1}^{n} \mathrm{var}(X_i) = np(1-p) \qquad (2)$$

Hodnoty n a p jsou parametry binomického rozdělení. Skutečnost, že náhodná veličina Y má binomické rozdělení, budeme vyjadřovat zkratkou $Y \sim Bi(n, p)$.

K pravděpodobnostní funkci binomicky rozdělené veličiny dospějeme následující úvahou. Náhodná veličina Y může nabývat hodnoty 0, 1, 2, ... , n. Představme si, že k lvů padne tak, že na prvních k mincích bude lev, na zbývajících $(n - k)$ bude rub. Při tomto výsledku náhodného pokusu bude $Y = k$, pravděpodobnost tohoto jevu můžeme spočítat jako $p^k(1-p)^{n-k}$ - jde o nezávislé jevy, tedy násobíme pravděpodobnosti. Stejnou hodnotu náhodné veličiny, však můžeme dostat i tak, že lev padne na jiných k mincích než právě na k prvních. Těchto k mincí, na kterých musí být lev, aby $Y = k$, můžeme vybrat $\binom{n}{k}$ způsoby, a tak pravděpodobnostní funkci náhodné veličiny $Y \sim Bi(n, p)$ lze vyjádřit jako

$$P(Y = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k=0, 1, 2, \ldots, n$$

Poznámka: Zápis $\binom{n}{k}$ čteme „n nad k“. Platí, že

$$\binom{n}{k} = \frac{n!}{k!(n-k)!},$$

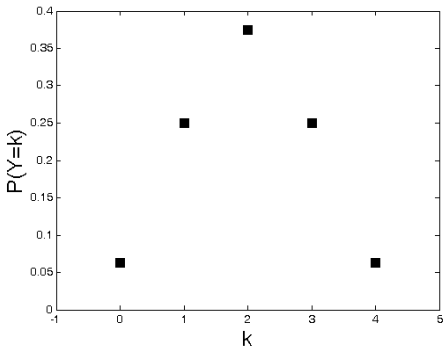
kde $n! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot (n-1) \cdot n$ (čti „n-faktoriál“).

Pro $k=0$ je definováno $\binom{n}{0} = 1$.

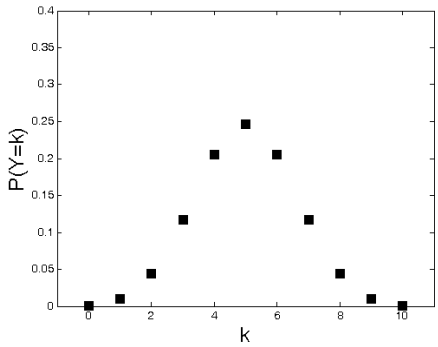
Pak zjevně platí $\binom{n}{k} = \binom{n}{n-k}$.

Výraz $\binom{n}{k}$ udává počet možností výběru k prvků z n různých prvků, $0 \le k \le n$, počet kombinací bez opakování. *Konec poznámky.*

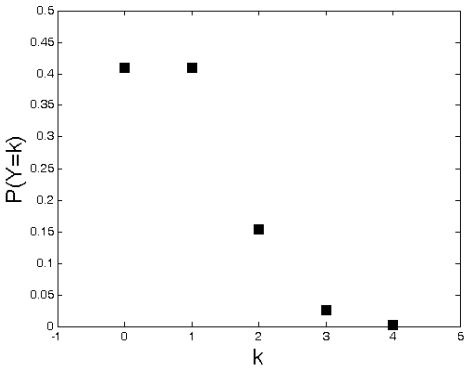
Grafické znázornění pravděpodobnostních funkcí binomického rozdělení pro různé hodnoty parametrů je na následujícím obrázku.



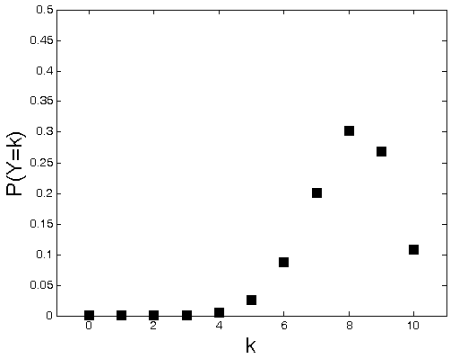
$Y \sim Bi(4, 0.5)$



$Y \sim Bi(10, 0.5)$



$Y \sim Bi(4, 0.2)$



$Y \sim Bi(10, 0.8)$

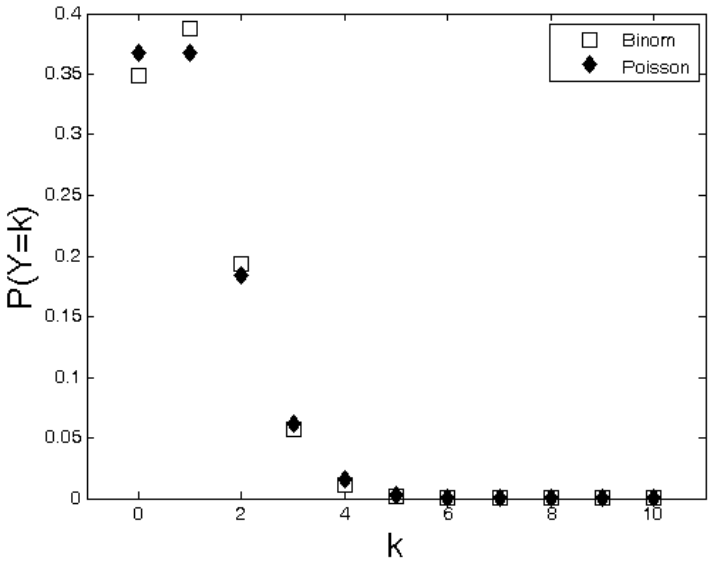
3.4.3 Poissonovo rozdělení

Toto rozdělení má náhodná veličina Y, která může nabývat hodnoty k=0, 1, 2, ... s pravděpodobností

$$P(Y = k) = e^{-\lambda} \frac{\lambda^k}{k!}, \qquad (1)$$

λ je jediný parametr tohoto rozdělení. Střední hodnota je $E(Y) = \lambda$, rozptyl je $\mathrm{var}(Y) = \lambda$. Poissonovo rozdělení s parametrem $\lambda = n\,p$ se často užívá k aproximaci binomického rozdělení $Y \sim Bi(n, p)$, když n je velké a p je malé. Doporučuje se, aby bylo $n > 30$ a $p < 0.1$. Smysl této aproximace je zejména v usnadnění výpočtu pravděpodobnostní funkce v aplikacích, neboť Poissonova rozdělení se užívá k modelování požadavků hromadné obsluhy, počtu poruch technického zařízení atd.

Následující obrázek ukazuje, že i pro menší hodnoty n je shoda obou rozdělení docela těsná.



Pravděpodobnostní funkce binomického a Poissonova rozdělení

$X \sim Bi(10, 0.1)$, parametr Poissonova rozdělení $\lambda = 10 \cdot 0{,}1 = 1$

3.4.4 Rovnoměrné diskrétní rozdělení

Toto rozdělení má náhodná veličina *X*, která může nabývat k různých hodnot $x_1, x_2, \ldots, x_k$, přičemž každá hodnota je stejně pravděpodobná, tj. pravděpodobnost jevu jistého je rozdělena rovnoměrně mezi všechny elementární jevy. Pravděpodobnostní funkce má tedy tvar

$$P(X = x_i) = \frac{1}{k}, \qquad i=1, 2, \ldots, k \qquad (1)$$

Toto rozdělení je modelem pokusů házení mincí *(k=2)* nebo házení hrací kostkou *(k = 6)*.

Střední hodnota rovnoměrně rozdělené diskrétní náhodné veličiny je pak

$$E(X) = \sum_{i=1}^{k} x_i \; P(X = x_i) = \frac{1}{k}\sum_{i=1}^{k} x_i \qquad (2)$$

a rozptyl je

$$\operatorname{var}(X) = E[X - E(X)]^2 = E(X^2) - [E(X)]^2 =$$

$$= \frac{1}{k}\sum_{i=1}^{k} x_i^{\,2} + \frac{1}{k^2}\left(\sum_{i=1}^{k} x_i\right)^2 = \frac{1}{k}\left[\sum_{i=1}^{k} x_i^{\,2} - \frac{1}{k}\left(\sum_{i=1}^{k} x_i\right)^2\right] \qquad (3)$$

Specielně pro hod kostkou střední hodnota (2) je

$$E(X) = \frac{1}{6}\sum_{i=1}^{6} x_i = \frac{21}{6} = 3{,}5$$

a rozptyl (3) je

$$\operatorname{var}(X) = \frac{1}{6}\left[\sum_{i=1}^{6} x_i^{\,2} - \frac{1}{6}\left(\sum_{i=1}^{6} x_i\right)^2\right] = \frac{1}{6}\left[91 - \frac{1}{6}(21)^2\right] = \frac{35}{12}.$$

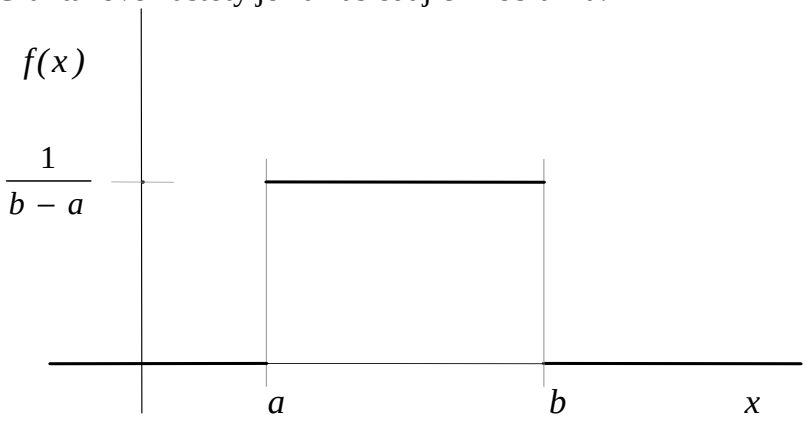
3.5 Příklady spojitých rozdělení

3.5.1 Rovnoměrné spojité rozdělení

Spojitá náhodná veličina X má rovnoměrné rozdělení, jestliže hustota pravděpodobnosti je na intervalu hodnot (a, b) konstantní a mimo tento interval nulová, tj.

$$f(x) = \begin{cases} \dfrac{1}{b-a} & pro \quad a < x < b \\ 0 & jinak \end{cases} \tag{1}$$

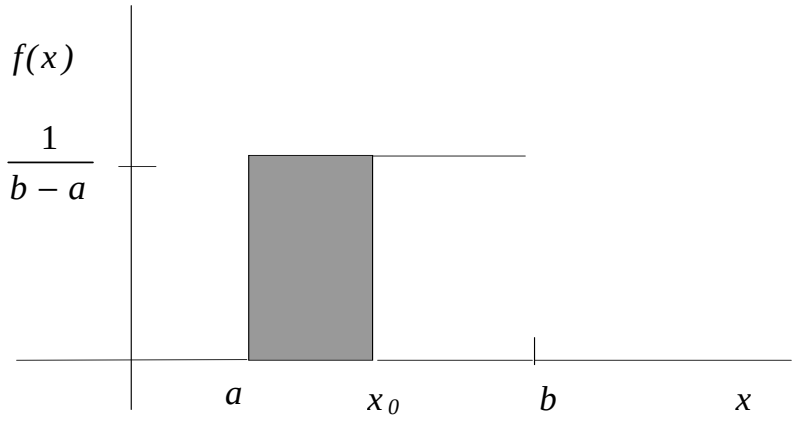
Graf takové hustoty je na následujícím obrázku.



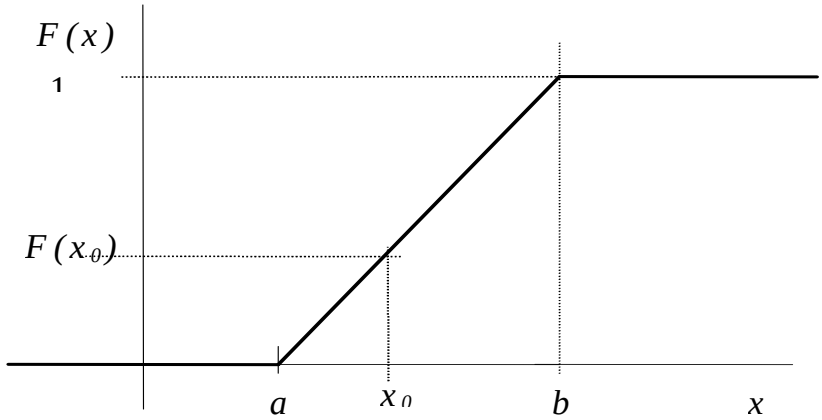
Distribuční funkce rovnoměrně rozdělené náhodné veličiny X je

$$\begin{aligned} &F(x) = 0 \quad pro \quad x \le a \\ &F(x) = P(X < x) = \int_a^x f(t)\, dt = \frac{1}{b-a}(x-a) \qquad pro \quad a < x < b \\ &F(x) = 1 \quad pro \quad x \ge b \end{aligned} \tag{2}$$

tedy pro zvolenou hodnotu $x_0 \in (a, b)$ je to plocha obdélníku pod grafem funkce hustoty vlevo od hodnoty x_0 - viz následující obrázek.



Hodnota distribuční funkce v bodu x_0 je obsah šedé plochy pod hustotou



Graf distribuční funkce rovnoměrného rozděleni

Základní charakteristiky rovnoměrně rozdělené náhodné veličiny jsou

$$E(X) = \frac{1}{b-a}\int_a^b x\,dx = \frac{a+b}{2} \tag{3}$$

$$\mathrm{var}(X) = EX^2 - (EX)^2 = \frac{1}{b-a}\int_a^b x^2\,dx - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12} \tag{4}$$

Je zřejmé, že rovnoměrné rozdělení je symetrické vzhledem ke střední hodnotě, a tedy medián je roven střední hodnotě. Modus není definován. Jelikož distribuční funkce na intervalu $[a,b]$ roste lineárně, jsou i mezi po sobě následujícími percentily stejné vzdálenosti.

U různých programových produktů (tabulkové procesory, programovací jazyky, statistické a simulační programy) je dostupný tzv. *generátor náhodných čísel*. Je to funkce, jejímž voláním lze získat hodnoty náhodné veličiny s rovnoměrným rozdělením pravděpodobnosti. Běžně se setkáváme s tím, že tato funkce generuje hodnoty veličiny U z intervalu $[0,1)$, pokud potřebujeme hodnoty z jiného intervalu $[a,b)$, $a < b$, snadno je získáme lineární transformací $X = a + (b-a)\,U$. Některé programové produkty dovolují i generování hodnot diskrétní náhodné veličiny s rovnoměrným rozdělením, jinak tyto hodnoty můžeme získat vhodnou transformací (zaokrouhlením) spojité veličiny *X*. Je nutno mít na paměti, že tzv. generátory náhodných čísel jsou deterministické algoritmy, tzn., že jednou vygenerovanou řadu hodnot jsme schopni při stejném počátečním zadání přesně zopakovat. Vygenerované hodnoty tedy nejsou, přísně vzato, náhodné. Proto se někdy takto vygenerovaným hodnotám říká *pseudonáhodná* čísla. Při použití těchto generátorů je proto namístě jistá opatrnost a ověření toho, zda rozdělení pseudonáhodných hodnot lze opravdu považovat za rovnoměrné.

3.5.2 Normální rozdělení

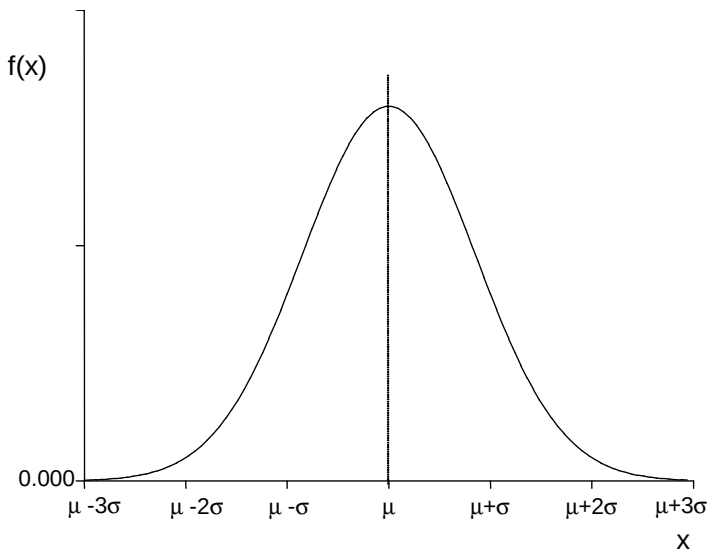
Spojitá náhodná veličina má *normální* (Gaussovo) *rozdělení*, jestliže její hustota má tvar

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad (1)$$

kde $-\infty < x < +\infty$, μ, σ jsou reálná čísla, $\sigma > 0$. Říkáme, že náhodná veličina *X* má normální rozdělení s parametry μ a σ^2, což ve zkratce zapisujeme

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

Graf hustoty normálního rozdělení je na následujícím obrázku.



Vidíme, že hustota normálního rozdělení je symetrická kolem přímky x = μ, takže platí $f(\mu - y) = f(\mu + y)$.

a medián $x(0{,}5) = \mu$. Hustota je největší v bodě μ (modus je roven μ) a od tohoto bodu na obě strany hustota rychle klesá. Tvar hustoty ukazuje, že hodnoty blízké μ jsou velmi pravděpodobné, zatímco hodnoty od μ vzdálené jsou málo pravděpodobné. Tuto funkci užil před dvěma staletími Gauss k popisu rozdělení chyb astronomických měření. V průběhu let se toto rozdělení ukázalo být vhodným popisem i v mnoha dalších situacích a získalo zásadní pozici v aplikacích statistiky. Pro toto rozdělení se začalo užívat označení *normální rozdělení* (někdy také *Gaussovo).* Lze ukázat, že pro střední hodnotu a rozptyl platí $EX = \mu$ a $\mathrm{var}(X) = \sigma^2$, tedy parametry tohoto rozdělení znamenají střední hodnotu a rozptyl.

Má-li náhodná veličina X normální rozdělení, $X \sim N(\mu, \sigma^2)$, potom náhodná veličina $Y = aX + b, \quad a \neq 0$, (říkáme, že veličina Y vznikne lineární transformací veličiny X) má opět normální rozdělení, avšak hodnoty parametrů jsou v důsledku lineární transformace odlišné, totiž $Y \sim N(a\mu + b, a^2\sigma^2)$.

Zvolíme-li specielně $a = \dfrac{1}{\sigma}$, $b = -\dfrac{\mu}{\sigma}$, pak náhodná veličina

$U = \dfrac{X - \mu}{\sigma}$ má rozdělení $U \sim N(0,\ 1)$.

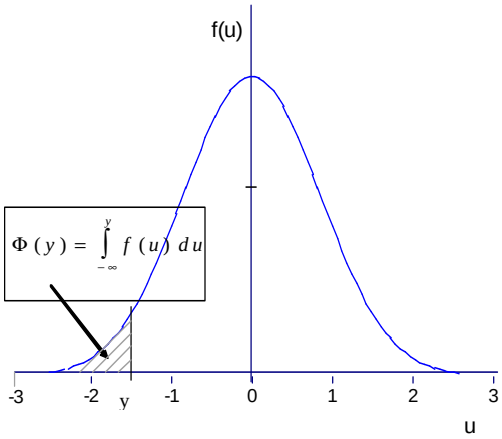
Tomuto rozdělení říkáme *normované normální rozdělení*, náhodná veličina U vznikla normováním veličiny X, tj. takovou lineární transformací, aby $EU = 0$ a $\mathrm{var}(U) = 1$. *Hustotu normovaného normálního rozdělení* můžeme vyjádřit po dosazení do (1) jako

$$f(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{u^2}{2}), \qquad -\infty < u < +\infty \tag{3}$$

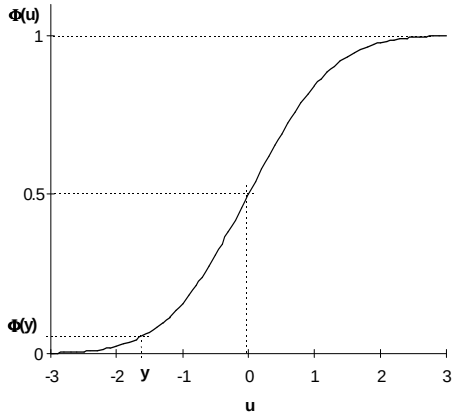
a *distribuční funkce normovaného normálního rozdělení*, pro kterou se vzhledem k jejímu stěžejnímu postavení ve statistice užívá zvláštní symbol Φ, je pak

$$\Phi(u) = P(U < u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} \exp(-\frac{t^2}{2})\, dt\ , \qquad -\infty < u < +\infty\ . \tag{4}$$

Graf hustoty a distribuční funkce normovaného normálního rozdělení vidíme na následujících obrázcích.



Hustota normovaného normálního rozdělení



Distribuční funkce normovaného normálního rozdělení

Distribuční funkci normovaného normálního rozdělení nelze vyjádřit aritmetickým výrazem, který by umožňoval jednoduché vyhodnocení funkce $\Phi(u)$ v bodě u a naopak z hodnoty funkce $\Phi(u)$ zjistit hodnotu argumentu u, jako to bylo možné u distribuční funkce rovnoměrného rozdělení. U normovaného normálního rozdělení je nutno tyto výpočty provádět docela komplikovanými numerickými metodami. Pro ušetření práce je však funkce $\Phi(u)$ tabelována a tyto tabulky jsou součástí většiny statistických učebnic včetně těchto skript. Kromě toho numerické postupy k vyhodnocení distribuční funkce normovaného normálního rozdělení a také mnoha dalších rozdělení jsou součástí běžných programových prostředků pro statistiku (Excel, NCSS atd.), a tím je jejich využívání usnadněno. Statistické tabulky pak nejsou potřeba.

Jak vidíme z obrázků, hustota normovaného normálního rozdělení je symetrická vzhledem k ose $u = 0$, takže platí také

$$\Phi(-u) = 1 - \Phi(u) \tag{5}$$

Pomocí distribuční funkce normovaného normálního rozdělení $\Phi(u)$ můžeme vyjádřit hodnoty distribuční funkce normálního rozdělení pro libovolné dovolené hodnoty parametrů. Když $X \sim N(\mu, \sigma^2)$, pak pro distribuční funkci náhodné veličiny X platí

$$F(x) = P(X < x) = P\left(\frac{X-\mu}{\sigma} < \frac{x-\mu}{\sigma}\right) = P\left(U < \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right) \tag{6}$$

Tedy známe-li hodnoty parametrů μ a σ^2, pak pro známou hodnotu x umíme určit hodnotu distribuční funkce v bodě x.

Příklad: Z dlouholetých antropometrických výzkumů je známo, že tělesná výška dospělých mužů i žen má normální rozdělení. Naším úkolem je zjistit, jaká je v dospělé mužské populaci relativní četnost mužů menších než 170 cm, jestliže

známe parametry této populace $\mu = 175$ cm a $\sigma^2 = 49$ cm^2. Podobné úlohy jsou velmi užitečné např. pro řízení výroby konfekce, navrhování nábytku atd.

Řešením naší úlohy je vlastně zjistit hodnotu distribuční funkce rozdělení $X \sim N(175, 49)$ v bodě 170.

$$F(170) = \Phi\left(\frac{170-\mu}{\sigma}\right) = \Phi\left(-\frac{5}{7}\right) = 1 - \Phi\left(\frac{5}{7}\right)$$

V tabulkách nalezneme $\Phi(5/7) = 0,762$ a tedy $F(170) = 0{,}238$. V populaci je zhruba 24% mužů menších než 170 cm.

Pokud bychom využili funkci NORMSDIST v Excelu, která vrací hodnotu distribuční funkce normovaného normálního rozdělení, tak zadáním NORMSDIST(5/7) dostaneme hodnotu 0,762475. Dokonce můžeme užít funkci NORMDIST, která má čtyři parametry. První je hodnota argumentu, pro který chceme určit hodnotu distribuční funkce, další dva parametry jsou střední hodnota a směrodatná odchylka (!!!) normálního rozdělení. Poslední parametr je logická hodnota, pokud chceme získat hodnotu distribuční funkce, je potřeba zadat hodnotu tohoto parametru PRAVDA nebo nenulové číslo, jinak bychom dostali hustotu. Zadáním NORMDIST(170;175;7;PRAVDA) dostaneme hodnotu distribuční funkce v bodě 170, $F(170) = 0.237525$.

Časté jsou úlohy, kdy známe hodnotu distribuční funkce $F(x)$ normálního rozdělení $N(\mu, \sigma^2)$ a hledáme hodnotu argumentu *x*. Z odst. 3.3 víme, že hodnotě *x(p)*, pro kterou platí $F(x(p)) = p$, se říká *p*-kvantil.

Z definice je zřejmé, že platí $F(x(p)) = p$ a také $\Phi(u(p)) = p$, kde $u(p) = \dfrac{x(p)-\mu}{\sigma}$. Odtud pak $x(p) = \sigma\, u(p) + \mu$, což je návod, jak určit *p*-kvantil náhodné veličiny $X \sim N(\mu, \sigma^2)$, známe-li hodnoty parametrů.

Pokud bychom chtěli nalézt *p*-kvantil rozdělení z předchozího příkladu pro *p* = 0.238, pak v tabulce 6.1 nalezneme $u(0,762) \doteq 0.72$, ze symetrie rozdělení je $u(0.238) \doteq -0.72$ a po dosazení do $x(p) = \sigma\, u(p) + \mu$ dostaneme $x(0,238) = 7 \cdot (-0,72) + 175 = 169,96 \doteq 170$. V Excelu funkce NORMINV s parametry *p*, μ, σ vrátí hodnotu příslušného kvantilu, tedy NORMINV(0,238; 175; 7) vrátí hodnotu 170,01.

3.5.3 Rozdělení Chí-kvadrát

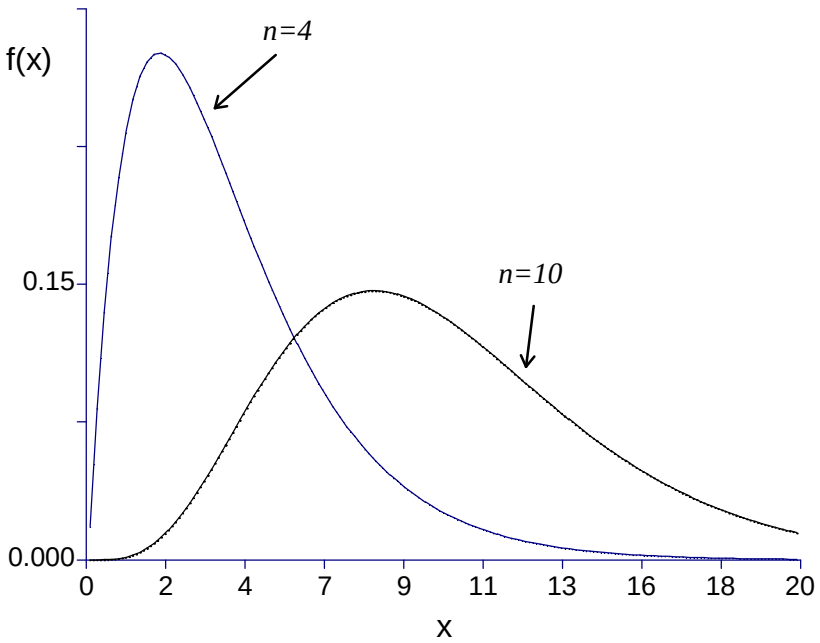
Toto rozdělení patří mezi rozdělení odvozená od normálně rozdělených náhodných veličin. Taková rozdělení se velmi často užívají v úlohách induktivní statistiky. Rozdělení χ^2 (čteme chí-kvadrát) má náhodná veličina, která vznikne součtem druhých mocnin nezávislých náhodných veličin normálně rozdělených.

Přesněji, nechť $U_1, U_2, \ldots, U_n$ jsou *nezávislé* náhodné veličiny a *každá* má rozdělení $N(0,\ 1)$. Potom náhodná veličina $X = \sum_{i=1}^{n} U_i^{\,2}$ má rozdělení χ^2 s n stupni volnosti, což zkráceně zapisujeme $X \sim \chi_n^2$. Hodnota n je jediný parametr tohoto rozdělení.

Střední hodnota je $EX = n$, rozptyl je $\mathrm{var}(X) = 2n$.

Hustota je graficky znázorněna na následujícím obrázku. Je zřejmé, že hustota rozdělení χ^2 pro hodnoty $x \le 0$ je nulová. Distribuční funkci podobně jako u normálního rozdělení nelze vyjádřit jednoduchým výrazem (ostatně i hustota je komplikovaný výraz), proto je tabelována, podobně i kvantily rozdělení χ^2, viz tab. 6.2. V Excelu pro určení kvantilů rozdělení χ^2 můžeme užít funkci CHIINV, její parametry jsou *1-p* a počet stupňů volnosti, takže např. zadáním CHIINV(0,05; 1) dostaneme hodnotu 0,95-kvantilu rozdělení χ^2 = 3,84145.

S rostoucím n se rozdělení χ_n^2 blíží normálnímu rozdělení s parametry $\mu = n$ a $\sigma^2 = 2n$, $\chi_n^2 \to N(n, 2n)$.



Hustota rozdělení χ_n^2 pro n = 4 a n = 10

5.5.4 Studentovo t-rozdělení

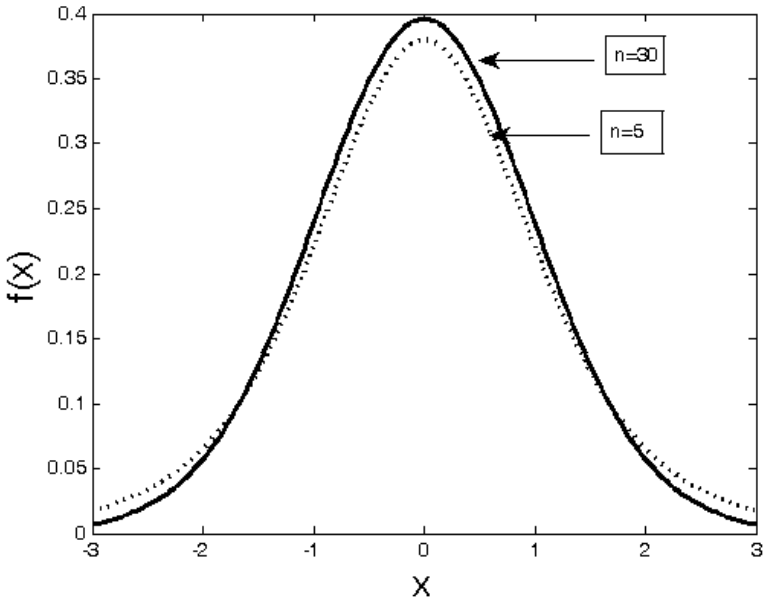
I toto rozdělení patří mezi rozdělení odvozená od normálního rozdělení. Když náhodná veličina U má normované normální rozdělení, $U \sim N(0,\ 1)$, náhodná veličina X má rozdělení χ^2 s n stupni volnosti, $X \sim \chi_n^2$ a U a X jsou nezávislé náhodné veličiny, potom náhodná veličina

$$T = \frac{U}{\sqrt{X / n}}$$

má t-rozdělení s n stupni volnosti, což ve zkratce zapisujeme $T \sim t_n$. Hodnota n je jediný parametr tohoto rozdělení. Toto rozdělení se také někdy nazývá Studentovo rozdělení podle pseudonymu Student, kterým na začátku 20. století podpisoval své statistické práce chemik pivovaru Guiness v Dublinu William Sealy. Gosset, jeden ze zakladatelů aplikací induktivní statistiky, a to v oblasti nesporně významné – v zabezpečení kvality piva.

Pro $n > 2$ platí, že střední hodnota je $ET = 0$, rozptyl je $\mathrm{var}(T) = \dfrac{n}{n-2}$.

S rostoucím n se t-rozdělení blíží normovanému normálnímu rozdělení, $t_n \to N(0,1)$, pro $n > 30$ je tvar obou rozdělení prakticky shodný. Tvar grafu hustoty t rozdělení pro různé počty stupňů volnosti vidíme na následujícím obrázku.



Hustota rozdělení t_n pro $n = 5$ a $n = 30$

Kvantily t–rozdělení jsou tabelovány nebo je můžeme určit s pomocí software.
V Excelu funkce TINV s parametry 1-2p a počtem stupňů volnosti vrací hodnotu p-kvantilu, např. TINV(0,05; 25) vrátí hodnotu 2,0595.

3.5.5 Fisherovo-Snedecorovo F-rozdělení

Nechť X_m a X_n jsou nezávislé náhodné veličiny, které mají rozdělení $X_m \sim \chi^2_m$ a $X_n \sim \chi^2_n$. Potom náhodná veličina

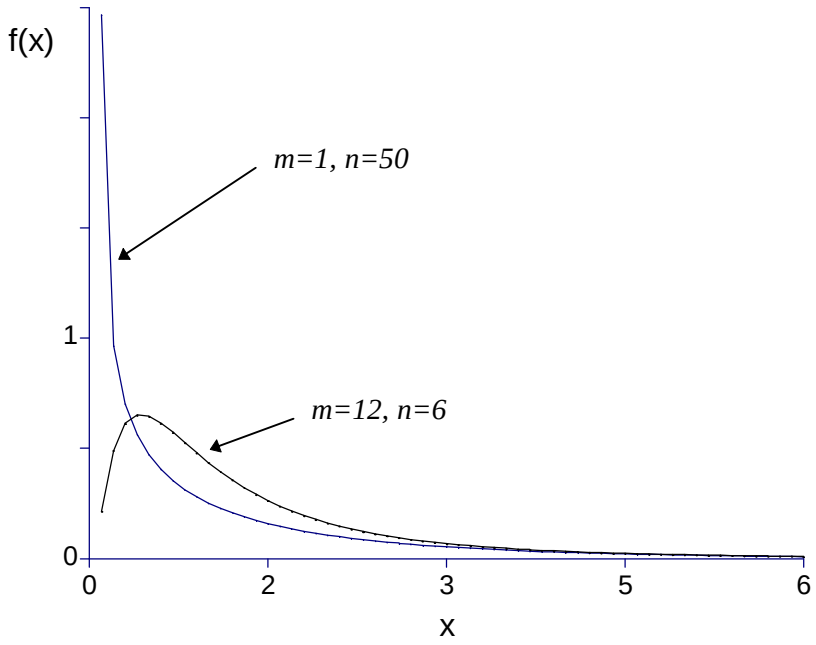
$$Y = \frac{X_m / m}{X_n / n}$$

má F-rozdělení s m a n stupni volnosti, ve zkratce to zapisujeme $Y \sim F_{m,n}$. Hodnoty m a n jsou parametry rozdělení, m je počet stupňů volnosti pro čitate-

le, n je počet stupňů volnosti pro jmenovatele, na pořadí parametrů tvar rozdělení pochopitelně závisí.

Pro $n > 4$ platí $EY = \dfrac{n}{n-2}$ a $\operatorname{var}(Y) = \dfrac{2\,n^2(m+n-2)}{m\,(n-2)^2\,(n-4)}$

Hustota F-rozdělení je graficky zobrazena na následujícím obrázku.



Hustoty F-rozdělení pro různé hodnoty parametrů

Vzhledem k tomu, že náhodná veličina Y je podílem veličin X_m / m a X_n / n, pro kvantily F - rozdělení platí

$$F_{m,n}(p) = \frac{1}{F_{n,m}(1-p)}$$

a dále také platí $F_{1,n}(p) = \left[t_n(1-p/2)\right]^2 = \left[t_n(p/2)\right]^2$.

Vybrané kvantily F-rozdělení jsou v tabulce 6.4. V Excelu je počítá funkce FINV, p-kvantil dostaneme při zadání parametrů 1-p, m, n, např. FINV(0,05; 10; 20) vrátí hodnotu 2,347875, což je 0,95-kvantil.

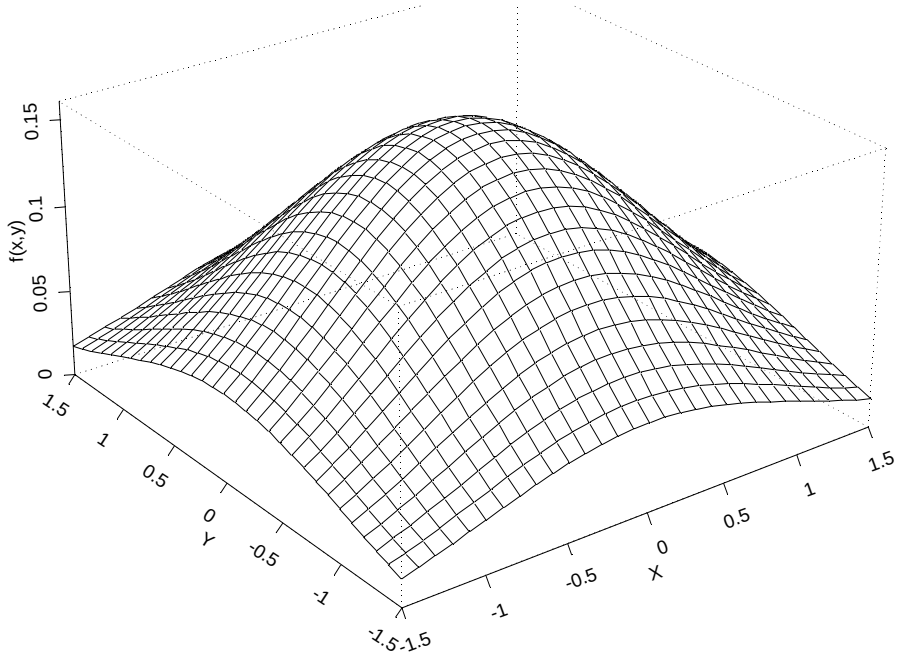
3.5.6 Dvourozměrné normální rozdělení

Náhodný vektor (X,Y) má dvourozměrné normální rozdělení s parametry $\mu, \nu, \sigma^2, \tau^2, \rho$, $\sigma^2 > 0, \tau^2 > 0$, $|\rho| < 1$, jestliže má sdruženou hustotu

$$f_{XY}(x,y) =$$

$$= \frac{1}{2\pi\sigma\tau\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu}{\sigma}\right)^2 - \frac{2\rho(x-\mu)(y-\nu)}{\sigma\tau} + \left(\frac{y-\nu}{\tau}\right)^2\right]\right\}$$

pro všechna reálná x a y. Příklad takové sdružené hustoty pro parametry $\mu = 0, \nu = 0, \ \sigma^2 = 1, \tau^2 = 1, \ \rho = 0$ je na následujícím obrázku.



Graf sdružené hustoty dvourozměrného normálního rozdělení
$\mu = 0, \nu = 0, \ \sigma^2 = 1, \tau^2 = 1, \ \rho = 0$

Potom i marginální rozdělení jsou normální, $X \sim N(\mu, \sigma^2)$, $Y \sim N(\nu, \tau^2)$,
$EX = \mu$, $EY = \nu$,
$\mathrm{var}(X) = \sigma^2$, $\mathrm{var}(Y) = \tau^2$.

Hodnota parametru ρ je rovna hodnotě korelačního koeficientu ρ_{XY}. Pro dvourozměrné *normální* rozdělení platí, že je-li $\rho = 0$ (tedy náhodné veličiny X a Y jsou *nekorelované*), pak jsou i *nezávislé*.

Souhrn:

- alternativní rozdělení, výpočet střední hodnoty a rozptylu
- binomické rozdělení, jeho pravděpodobnostní funkce, střední hodnota, rozptyl
- Poissonovo rozdělení
- rovnoměrné diskrétní rozdělení, jeho pravděpodobnostní funkce, střední hodnota, rozptyl
- parametry rozdělení
- rovnoměrné spojité rozdělení, vztah mezi hustotou a distribuční funkcí
- normální rozdělení, parametry, hustota, kvantily
- normované normální rozdělení
- dvourozměrné normální rozdělení, jeho parametry

Kontrolní otázky:

1. *Bylo potřeba pro určení střední hodnoty a rozptylu binomicky rozdělené náhodné veličiny užít její pravděpodobnostní funkci?*
2. *Jaký je vztah distribuční funkce a hustoty rovnoměrného spojitého rozdělení?*
3. *Jaká je pravděpodobnost, že hodnota náhodné veličiny, která má normované normální rozdělení, je v intervalu* $\langle -1,0)$ *?*
4. *Určete 0,975 kvantil normovaného normálního rozdělení a stejný kvantil t-rozdělení s 5 a pak i 100 stupni volnosti. Porovnejte tyto hodnoty a zdůvodněte jejich rozdíly.*

Pojmy k zapamatování:

- binomické rozdělení
- diskrétní rovnoměrné rozdělení
- spojité rovnoměrné rozdělení
- parametry rozdělení
- normální rozdělení, normované normální rozdělení
- rozdělení χ^2, *t*-rozdělení, *F*-rozdělení
- stupeň volnosti

Tato část kapitoly o pravděpodobnosti vám zabere asi dvě až tři hodiny. Seznámíte se v ní s centrální limitní větou a pomůže vám pochopit, proč je normální rozdělení často ve statistice využíváno jako vhodný model sledované reality.

3.6 O centrální limitní větě

Normální rozdělení má pro své vlastnosti klíčový význam v mnoha aplikacích statistiky. Jak jsme již uvedli, má-li náhodná veličina $X \sim N(\mu, \sigma^2)$, potom náhodná veličina $Y = aX + b,\ a \neq 0$, má opět normální rozdělení, $Y \sim N(a\mu + b, a^2\sigma^2)$. V předchozích odstavcích jsme viděli, že k normálnímu rozdělení se pro velká n blíží rozdělení χ^2_n a t-rozdělení. Další důležitou vlastností normálního rozdělení je to, že součet konečného počtu nezávislých normálně rozdělených náhodných veličin má opět normální rozdělení.

Specielně pro $X_1, X_2, \ldots, X_n$ *nezávislých* náhodných veličin se stejným rozdělením $N(\mu,\ \sigma^2)$ platí

a) $E\left[\frac{1}{n}(X_1 + X_2 + \ldots + X_n)\right] = \frac{n\mu}{n} = \mu$,

b) $\mathrm{var}\left[\frac{1}{n}(X_1 + X_2 + \ldots + X_n)\right] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$,

c) je-li $Y = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$, pak $Y \sim N(\mu, \frac{\sigma^2}{n})$.

K normálnímu rozdělení se však přibližuje i součet nezávislých náhodných veličin z jakéhokoliv rozdělení. Je to důsledek tzv. *centrální limitní věty*. Jsou-li $X_1, X_2, \ldots, X_n$ vzájemně nezávislé náhodné veličiny téhož (ale jinak libovolného rozdělení) se střední hodnotou μ a rozptylem σ^2, pak pro každé reálné x platí

$$\lim_{n\to\infty} P\left[\frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}(X_i - \mu) < x\right] = \Phi(x).$$

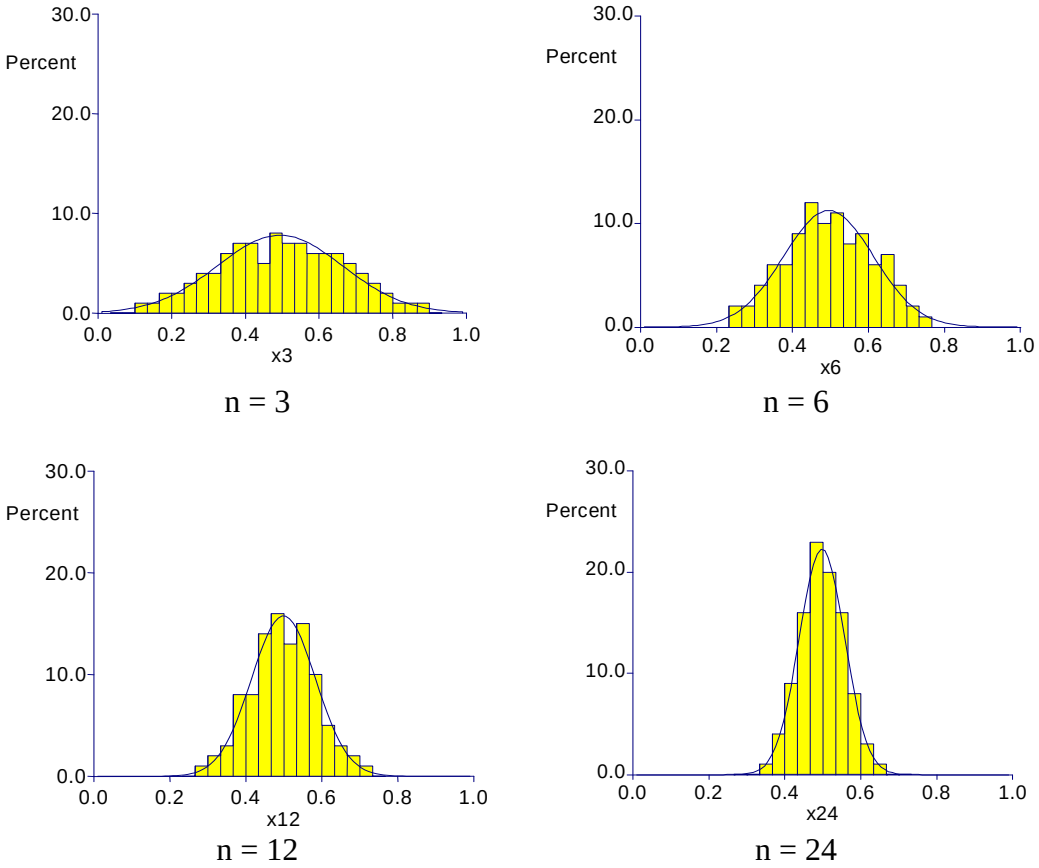
Tzn., že pro dostatečně velké n se distribuční funkce náhodné veličiny

$$Z_n = \frac{\sum_{i=1}^{n} X_i - E\left(\sum_{i=1}^{n} X_i\right)}{\sqrt{\mathrm{var}\left(\sum_{i=1}^{n} X_i\right)}} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}} = \frac{\sum_{i=1}^{n}(X_i - \mu)}{\sigma\sqrt{n}}$$

jen nepatrně liší od distribuční funkce normovaného normálního rozdělení. Volně řečeno, součet (a tedy i průměr) většího počtu nezávislých stejně rozdělených náhodných veličin má přibližně normální rozdělení.

Tuto skutečnost ilustruje následující příklad na obr. 3.6-1, ve kterém jsou znázorněna empirická rozdělení hodnot získaných z 1000 nezávislých realizací ná-

hodné veličiny $Y = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$, kdy náhodné veličiny X_i, $i = 1, 2, \ldots, n$ měly rovnoměrné spojité rozdělení na intervalu $< 0, 1)$, a n bylo postupně rovno 3, 6, 12 a 24. Z histogramů na obrázku vidíme, že s rostoucím n se empirické rozdělení stále těsněji blíží k normálnímu rozdělení a také se zmenšuje rozptyl.



Obr. 3.6-1: Rozdělení výběrových průměrů z rovnoměrného rozdělení pro různé rozsahy výběru

Centrální limitní věta má také další často prakticky využívanou formulaci - *aproximaci binomického rozdělení normálním rozdělením*: Když náhodná veličina $Y_n \sim Bi(n,p)$, pak pro všechna reálná x

$$\lim_{n\to\infty} P\left(\frac{Y_n - np}{\sqrt{np(1-p)}} < x\right) = \Phi(x).$$

Označíme-li relativní četnost úspěchu $f_n = \frac{Y_n}{n}$, pak po krácení zlomku v závorce číslem n dostaneme

$$\lim_{n\to\infty} P\left(\frac{f_n - p}{\sqrt{p(1-p)/n}} < x\right) = \Phi(x)\,,$$

takže pro dostatečně velké n distribuční funkce náhodné veličiny

$$\frac{f_n - p}{\sqrt{p(1-p)/n}}$$

se jen nepatrně liší od distribuční funkce normovaného normálního rozdělení.

Příklad: V průzkumu volebních preferencí dotazem na 900 náhodně vybraných potenciální voličů bylo zjištěno, že politickou stranu ABC by volilo 25 % dotazovaných voličů. Jaká je pravděpodobnost, že stranu ABC v celé populaci preferuje alespoň 27 % voličů?

Jde tedy o to, jaká je pravděpodobnost, že náhodná veličina $f_n \geq 0{,}27$ za předpokladu, že $p = 0{,}25$.

Tuto pravděpodobnost lze zapsat jako

$$P\left(U \geq \frac{f_n - p}{\sqrt{p(1-p)/n}}\right) = 1 - \Phi\left(\frac{f_n - p}{\sqrt{p(1-p)/n}}\right), \quad \text{kde } U \sim N(0,1)\,.$$

Spočítáme hodnotu argumentu distribuční funkce

$$\frac{f_n - p}{\sqrt{p(1-p)/n}} = \frac{0,27 - 0,25}{\sqrt{0,25 \cdot (1-0,25)/900}} = \frac{0,02}{\sqrt{0,25 \cdot (1-0,25)/900}} = 1,39$$

a v tabulkách nalezneme hodnotu distribuční funkce normovaného normálního rozdělení v tomto bodě - $\Phi(1{,}39) \cong 0{,}92$. Hledaná pravděpodobnost, že strana ABC získá ve volbách alespoň 27% hlasů, je 0,08.

Souhrn:

- Jsou-li $X_1, X_2, \ldots, X_n$ *nezávislé* náhodné veličiny se stejným rozdělením, střední hodnotou μ a rozptylem σ^2, pak platí

  $$E\left[\frac{1}{n}(X_1 + X_2 + \ldots + X_n)\right] = \frac{n\mu}{n} = \mu \quad \text{a}$$

  $$\operatorname{var}\left[\frac{1}{n}(X_1 + X_2 + \ldots + X_n)\right] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

- Jsou-li $X_1, X_2, \ldots, X_n$ *nezávislé* náhodné veličiny se stejným *normálním* rozdělením se střední hodnotou μ a rozptylem σ^2, pak i jejich průměr $Y = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$, má normální rozdělení $Y \sim N(\mu, \frac{\sigma^2}{n})$.
- Rozdělení součtu a průměru $X_1, X_2, \ldots, X_n$ *nezávislých* náhodných veličin se stejným rozdělením se blíží normálnímu rozdělení.
- Pro velké hodnoty parametru n lze binomické rozdělení aproximovat normálním rozdělením.

Kontrolní otázky:

1. *Když z populace opakovaně vybereme n objektů, jaké bude rozdělení průměrů těchto výběrů?*
2. *Jaké budou parametry normálního rozdělení, který aproximujeme binomické rozdělení s parametry n a p?*

Pojmy k zapamatování:

- rozdělení součtu a průměru nezávislých veličin stejného rozdělení
- centrální limitní věta
- aproximace binomického rozdělení normální rozdělením

Korespondenční úlohy budou zadávány vždy na začátku semestru.

Náhodných veličin se budou týkat čtyři příklady.

4 Induktivní statistika

První část této kapitoly je věnována základním pojmům induktivní statistiky, především náhodnému výběru a dalším souvisejícím pojmům a vysvětlíme si, co je to statistický odhad. Studium této části vám zabere asi tři hodiny.

4.1 Základní pojmy

Metody induktivní statistiky (matematická statistika, statistická indukce) se užívají tam, kde chceme dojít k nějakým tvrzením o *populaci* (vyslovit nějakou „obecnou pravdu“), ale k dispozici máme data jen o části jedinců této populace, tzv. *výběr*. Intuitivně je zřejmé (a matematicky dokazatelné, viz Havránek, 1993), že máme-li data pouze o části populace, je vyjádření „obecné pravdy“ o celé populaci zatíženo rizikem nesprávného úsudku. Kdykoliv induktivním uvažováním zobecňujeme (generalizujeme) zjištění z dílčího pozorování na tvrzení o celku, vždy je toto tvrzení zatíženo nejistotou, že může být nepravdivé. Ale na druhou stranu induktivní uvažování je nepostradatelným postupem v poznávání světa, ve kterém žijeme.

Můžeme uvést bezpočet příkladů takových nesprávných nebo přinejmenším zpochybnitelných závěrů získaných induktivními úsudky:

- Po zkušenostech z několika kontaktů s německými turisty uzavřeme: „(Všichni) Němci jsou hluční a přehnaně sebevědomí“.
- Z letmého porovnání několika českých a moravských vesnic, kterými projíždíme na dovolené, usoudíme: „Moravané jsou pracovitější než Češi“.

Podobně povrchním induktivním úsudkem můžeme dojít k závěrům typu:

- „Slováci se rádi perou“,
- „Absolventi Ostravské university jsou horší než absolventi University Karlovy“,
- „V Maďarsku nejsou blondýnky“,
- „Češi jsou rasisté“,
- „Poláci umějí jen kšeftovat“,
- „Rakušané jsou příliš opatrní“,
- „Dánové jsou opilci“,
- „Ženy jsou slabší než muži“,
- „Chlapi nic nevydrží“.

Raději skončíme s ukázkami povrchního a problematického induktivního usuzování, možná několika uvedenými příklady jsme se nepříjemně dotkli kdekoho v širokém okolí. Snad však tyto příklady dostatečně zřetelně ukazují, že s povrchností v induktivním uvažování je nutno tvrdě bojovat a hledat takové postupy, které riziko nesprávného úsudku minimalizují nebo alespoň snižují.

Jednou z takových cest snížení rizika chybného závěru induktivního úsudku jsou metody induktivní statistiky. Tyto metody se opírají o výsledky teorie pravděpodobnosti. V mnoha situacích vědeckého zkoumání, řešení technických a ekono-

mických problémů či v mnoha dalších úlohách jsou tyto metody standardními postupy, neboť právě ony minimalizují pravděpodobnost nesprávného úsudku.

K tomu, abychom metody induktivní statistiky mohli použít a dojít tak k co nejspolehlivějším „obecným pravdám“ o populaci, je nutné, abychom měli k dispozici pozorování o n jedincích (objektech) z této populace, při čemž těchto n jedinců musí být z populace vybráno *náhodně*. Pak říkáme, že naše pozorování jsou realizací *náhodného výběru* (angl. *random sample*). Jelikož v matematické statistice se většinou o jiném než náhodném výběru neuvažuje, často se užívá jen *výběr* (angl. *sample*).

Realizace náhodného výběru vznikne tak, že

- o zařazení jedince do výběru rozhoduje *náhoda* (nikoliv naše či cizí vůle, rozmar nebo záměr),
- každý jedinec z populace má *stejnou pravděpodobnost* zařazení do výběru.

Máme-li data, která jsou realizací takového náhodného výběru, pak můžeme v induktivním uvažování využít výsledky teorie pravděpodobnosti, tzn. kvantifikovat riziko omylu, případně vybrat metodu, která riziko omylu minimalizuje.

Náhodný výběr jedinců z populace lze pořídit postupem, který známe například z losování Sportky. Do osudí vložíme reprezentanta každého jedince z populace (v případě Sportky je to 49 stejných míčků označených čísly 1 až 49, ty tvoří tzv. *oporu výběru*), zamícháme a vybereme n jedinců (v případě Sportky 6 + 1 míčků), a ty tvoří náhodný výběr. Mechanismus výběru nemusí být realizován fyzickým zařízením, které můžeme několikrát měsíčně vidět na televizní obrazovce, ale může být simulován počítačem nebo vytvořen myšlenkově - pořídíme seznam všech jedinců populace (*opora výběru*) a jedince do náhodného výběru zařazujeme pomocí tabulky náhodných čísel - viz např. Laga, Likeš, 1978.

Uvedený způsob konstrukce náhodného výběru je možný jen u konečné populace, kdy oporu výběru jsme schopni vytvořit. To není možné vždy, např. nejsme s to utvořit oporu výběru populace mravenců v České republice ani molekul v ovzduší Ostravského regionu. Ale i při výběrech z takových populací je nutné respektovat uvedené požadavky, tj. o zařazení jedince do výběru musí rozhodovat *náhoda* a každý jedinec z populace musí mít *stejnou pravděpodobnost* zařazení do výběru.

V mnoha výzkumech bývají tyto podmínky opomíjeny a tím jsou pak znehodnoceny výsledky statistické analýzy. Tak např. pacienti jednoho zdravotního zařízení nejsou náhodným výběrem z populace v dané lokalitě, neboť o zařazení do výběru nerozhoduje náhoda, ale pacientova volba lékaře a další nenáhodné vlivy, navíc dlouhodobě zdraví lidé vůbec nemají šanci se do výběru dostat. Podobně „náhodně odchycení“ lidé na ulici stěží splňují podmínky náhodného výběru z městské populace. Do takového výběru se nemohou dostat lidé, kteří nevycházejí ven a volba místa a doby „odchytu“ ovlivňuje složení výběru, neboť zastoupení lidí v ulicích je časově a místně závislé. Např. z „náhodného“ výběru pořízeného u vchodu do menzy v době oběda bychom došli k závěru, že téměř všich-

ni obyvatelé Ostravy mají maturitu a že naprostá většina jsou studenti. Rovněž stěží lze považovat za náhodný výběr z populace určitého druhu brouků ty, které se podařilo chytit - možná ty zdatnější se chytit nepodařilo. Vzorek z vagónu uhlí odebraný z povrchu není náhodný výběr, neboť uvnitř může být kamení a tomu jsme nedali šanci k zařazení do výběru.

Postup výběru analogický losování Sportky je vhodný pro situace, kdy počet jedinců v populaci je značně větší než počet jedinců ve výběru (tzv. *rozsah výběru*). Pokud počet jedinců v populaci není výrazně větší než rozsah výběru, měl by být vylosovaný jedinec vždy vrácen do osudí. O tomto postupu konstrukce náhodného výběru říkáme, že je to *výběr s vracením*.

Existuje ještě jeden způsob výběru jedinců z populace, kterým lze získat náhodný výběr. Je to tzv. *stratifikovaný výběr*. Ten můžeme použít tehdy, kdy známe relativní četnosti jednotlivých vrstev v populaci, např. četnosti věkových kategorií, sociálních úrovní apod. Pak můžeme pořídit tolik opor výběru, kolik je vrstev v populaci a z každé vrstvy pořídíme náhodný výběr takového rozsahu, aby relativní četnost jedinců z každé vrstvy stratifikovaného výběru odpovídala relativní četnosti vrstvy v populaci. Stratifikovaný výběr je tedy sjednocením náhodných výběrů ze všech vrstev populace a rozsahy těchto výběrů jsou určeny relativní četností jednotlivých vrstev v populaci.

Nechť tedy máme náhodně vybráno n jedinců z populace a na každém jedinci zjišťujeme hodnotu jedné veličiny (znaku). Naměřené hodnoty tohoto znaku (odpovídají jednomu sloupci v datové matici, viz kap. 1) jsou *realizací* náhodného výběru.

Z pohledu matematické statistiky je náhodný výběr abstraktní pojem, který dovoluje zobecnit tvrzení o všech možných jeho realizacích. *Náhodný výběr* je vektor o n složkách $(X_1, X_2, \ldots, X_n)$, kde složky tohoto vektoru jsou *nezávislé* náhodné veličiny s *identickým* (tj. naprosto stejným) *rozdělením*. V anglické literatuře se užívá označení *i.i.d. sample*, kde zkratka *i.i.d* znamená *independent identically distributed*. Náhodný výběr v matematické statistice je tedy abstrakce výběru jedinců z fyzicky existující populace a změření hodnot jedné veličiny na těchto jedincích.

Příklad: Z dospělé mužské populace obyvatel Ostravy vybereme náhodně n mužů a změříme jejich výšku. Získáme hodnoty 176, 168, 191, 179, ... Na tyto hodnoty pohlížíme tak, že jsou to hodnoty nezávislých náhodných veličin téhož rozdělení. Nechť toto rozdělení má střední hodnotu μ a rozptyl σ^2. Pozorované hodnoty jsou výsledkem náhody, ale tato náhoda se řídí daným rozdělením pravděpodobnosti, pozorované hodnoty jsou rozházeny okolo střední hodnoty. Pozorovanou hodnotu náhodné veličiny X_i (výsledek měření na i-tém jedinci) můžeme vyjádřit jako $X_i = \mu + \varepsilon_i$, kde μ je střední hodnota a ε_i je náhodná složka, jejíž rozdělení je totožné pro všechny jedince výběru, tj. pro $i = 1, 2, \ldots, n$.

Z pozorovaných hodnot výběru (sloupec datové matice, realizace abstraktního náhodného výběru) můžeme počítat různé výběrové charakteristiky podle formulí, se kterými jsme se seznámili už v kapitole o deskriptivní statistice.

Výběrovým charakteristikám, které můžeme takto spočítat, se říká *statistiky*. Obecně můžeme statistiku T vyjádřit jako funkci náhodného výběru, tedy $T = T(X_1, X_2, \ldots, X_n)$. Tím máme statistiku vyjádřenou obecněji a můžeme se pak i obecněji vyslovit o jejích vlastnostech.

Příklady statistik jsou

- výběrový průměr $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$,
- výběrový rozptyl $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$.

Jelikož statistiky jsou funkcemi náhodných veličin $X_1, X_2, \ldots, X_n$, jsou i statistiky náhodnými veličinami, které mají nějaké pravděpodobnostní rozdělení, střední hodnotu, rozptyl atd. Pravděpodobnostní rozdělení statistik se nazývají *výběrová rozdělení*.

Předpokládejme, že všechny náhodné veličiny ve výběru mají střední hodnotu μ a rozptyl σ^2. Pak pro střední hodnotu výběrového průměru platí

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{n}{n}\mu = \mu, \tag{3}$$

tedy vidíme, že střední hodnota výběrového průměru je rovna střední hodnotě rozdělení populace.

Podobně pro rozptyl výběrového průměru snadno ukážeme

$$\mathrm{var}(\bar{X}) = \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}. \tag{4}$$

Vidíme, že rozptyl výběrového průměru se zmenšuje s rostoucím rozsahem výběru.

Mnoho metod matematické statistiky bylo navrženo a používá se pro analýzu výběrů z normálně rozdělené populace $N(\mu, \sigma^2)$. Proto uvedeme rozdělení některých výběrových charakteristik výběrů z normálního rozdělení. Výběrový průměr z normálního rozdělení $N(\mu, \sigma^2)$ má opět normální rozdělení $\bar{X} \sim N(\mu, \sigma^2/n)$. Pak standardizovaná náhodná veličina $U = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ má normované normální rozdělení $N(0,1)$.

Dále lze ukázat (viz např. Anděl, 1978), že

$$\frac{(n-1)s^2}{\sigma^2} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{\sigma^2} \sim \chi^2_{n-1}, \tag{5}$$

$$T = \frac{\bar{X} - \mu}{s / \sqrt{n}} \sim t_{n-1}. \qquad (6)$$

Vidíme, že veličiny U a T jsou definovány podobně, pouze na rozdíl od veličiny *U*, která má ve jmenovateli populační směrodatnou odchylku σ (v aplikacích její hodnotu zpravidla neznáme), má veličina T ve jmenovateli výběrovou směrodatnou odchylku *s*. To je sice náhodná veličina, ale její hodnotu umíme spočítat z výběru.

Dalšími ve statistice často užívanými výběrovými rozděleními jsou rozdělení náhodných veličin, ve kterých vystupuje rozdíl dvou výběrových průměrů. Předpokládejme, že máme dva nezávislé výběry (nemají žádné jedince, kteří jsou v obou výběrech) o rozsahu n_1, resp. n_2, ze dvou normálně rozdělených populací, první populace má rozdělení $N(\mu_1, \sigma_1^2)$, druhá $N(\mu_2, \sigma_2^2)$. Pak výběrové průměry mají rozdělení

$$\bar{X}_1 \sim N(\mu_1, \frac{\sigma_1^2}{n_1}), \qquad \bar{X}_2 \sim N(\mu_2, \frac{\sigma_2^2}{n_2}).$$

Potom i rozdíl těchto průměrů má opět normální rozdělení

$$\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}). \qquad (7)$$

Po standardizaci standardizovaná náhodná veličina *U* má normované normální rozdělení:

$$U = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1). \qquad (8)$$

V aplikacích však většinou neznáme hodnoty parametrů σ_1^2, σ_2^2. Pak lze využít toho, že platí (viz např. Anděl, 1978)

$$\frac{(n_1 - 1)\, s_1^2}{\sigma_1^2} + \frac{(n_2 - 1)\, s_2^2}{\sigma_2^2} \sim \chi^2_{n_1 + n_2 - 2} \qquad (9)$$

Pokud neznámé parametry σ_1^2, σ_2^2 můžeme považovat za shodné, tedy $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (rozptyl v obou populacích je shodný), pak náhodná veličina

$$T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1 + n_2 - 2} \qquad (10)$$

tedy má Studentovo *t* - rozdělení s $n_1 + n_2 - 2$ stupni volnosti. Tato statistika má klíčový význam v mnoha aplikacích.

Po tomto seznámení se základními pojmy můžeme říci, že základními úkoly induktivní statistiky jsou:

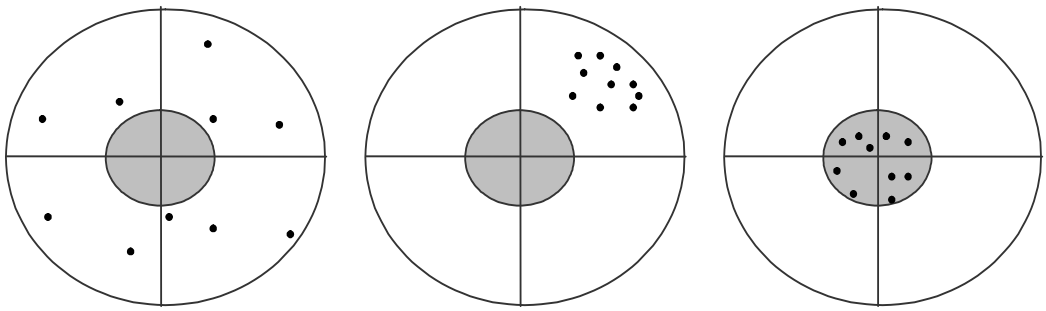
- *odhady parametrů* rozdělení populace,
- *testy hypotéz* o parametrech rozdělení populace.

Oba tyto typy základních úloh matematické statistiky vysvětlíme podrobněji v následujících odstavcích.

4.2 Statistický odhad

Cílem statistického odhadu je zjistit z výběru charakteristiky rozdělení, případně parametry rozdělení populace. Je to jedna z úloh statistické indukce, kdy z informací o části populace chceme dospět k tvrzení, které se týká celé populace. Víme, že zde existuje riziko nesprávného úsudku a úkolem matematické statistiky je toto riziko minimalizovat nebo alespoň poskytnout informace o jeho velikosti.

Podívejme se nejdříve na zdánlivě odtažitý příklad - terče s výsledky tří střelců:



a) velký rozptyl kolem středu

b) malý rozptyl, ale vychýlená muška

c) dobrý střelec

Všichni tři se snažili trefit střed terče, z různých důvodů (vítr, třes ruky, špatný nástroj) se jim to nepodařilo. Přesto snadno usoudíme, že nejlepšího výsledku dosáhl střelec (c), který má malý rozptyl a nevychýlenou mušku.

Při statistickém odhadu jsme v situaci velmi podobné střelcům. Rádi bychom z výběrových dat „trefili“ neznámou hodnotu populační charakteristiky. Je jasné, že se budeme snažit užít takový nástroj a postup, který bude dávat výsledky podobné střelci (c), totiž užívat metody odhadu, které „míří na střed“ a mají co nejmenší rozptyl. Nyní se pokusíme tyto pojmy vyjádřit přesněji.

4.2.1 Bodové odhady

Nechť náhodná veličina X má hustotu $f(x,\theta_1,\ldots,\theta_p) = f(x,\boldsymbol{\theta})$. Říkáme, že $\boldsymbol{\theta} = \theta_1,\ldots,\theta_p$ je vektor parametrů (bod v p-rozměrném prostoru). $\boldsymbol{\theta} \in \Omega$, pak Ω je *parametrický prostor*.

Funkcí $f(x,\boldsymbol{\theta})$ je specifikován systém rozdělení, třeba systém všech normálních rozdělení $N(\mu,\sigma^2)$ s různými hodnotami parametrů μ,σ^2. Každému $\boldsymbol{\theta} \in \Omega$ odpovídá jedno rozdělení z tohoto systému. Úkolem bodového odhadu je nalézt co nejlépe hodnotu vektoru parametrů $\boldsymbol{\theta}$, tzn. nalézt hodnoty jednotlivých složek $\theta_1,\ldots,\theta_p$. O složce tohoto vektoru budeme v dalším textu mluvit jako o parametru a budeme ji označovat θ, tj. bez indexu.

Hodnotu θ budeme odhadovat nějakou statistikou $T = T(X_1, X_2, \ldots, X_n)$, spočítanou z výběrových hodnot.

Říkáme, že statistika T je *nestranný* odhad parametru θ, když platí $E(T) = \theta$, tj. střední hodnota odhadu je rovna odhadovanému parametru.

Příklad 1: Předpokládejme, že všechny náhodné veličiny ve výběru mají střední hodnotu μ. Pak pro střední hodnotu výběrového průměru platí

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{n}{n}\mu = \mu .$$

Výběrový průměr je nestranným odhadem střední hodnoty populace.

Příklad 2: V tomto příkladu ukážeme, že výběrový rozptyl

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

je nestranným odhadem populačního rozptylu σ^2. Tím bude vysvětleno, proč ve jmenovateli výrazu pro výpočet s^2 je *n-1* - viz odst. 2.3. Předpokládejme, že všechny náhodné veličiny ve výběru mají střední hodnotu μ a rozptyl σ^2.

$$E(s^2) = \frac{1}{n-1} E\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \frac{1}{n-1} E\left[\sum_{i=1}^{n}(X_i - \mu + \mu - \bar{X})^2\right] =$$

$$= \frac{1}{n-1} E\left[\sum_{i=1}^{n}\left((X_i - \mu) - (\bar{X} - \mu)\right)^2\right] =$$

$$= \frac{1}{n-1} E\left[\sum_{i=1}^{n}\left((X_i - \mu)^2 - 2(X_i - \mu)(\bar{X} - \mu) + (\bar{X} - \mu)^2\right)\right] =$$

$$= \frac{1}{n-1} E\left[\sum_{i=1}^{n}(X_i - \mu)^2 - 2(\bar{X} - \mu)\sum_{i=1}^{n}(X_i - \mu) + \sum_{i=1}^{n}(\bar{X} - \mu)^2\right] =$$

$$= \frac{1}{n-1} E\left[\sum_{i=1}^{n}(X_i - \mu)^2 - 2(\bar{X} - \mu)\, n(\bar{X} - \mu) + n(\bar{X} - \mu)^2\right] =$$

$$= \frac{1}{n-1} E\left[\sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2\right] = \frac{1}{n-1}\left[n\sigma^2 - n\, E(\bar{X} - \mu)^2\right] =$$

$$= \frac{1}{n-1}\left[n\sigma^2 - n\frac{\sigma^2}{n}\right] = \frac{\sigma^2(n-1)}{n-1} = \sigma^2$$

Jestliže T_n je statistika z výběru s rozsahem n, pak odhad je *asymptoticky nestranný*, když platí $\lim_{n\to\infty} E(T_n) = \theta$, tzn. s rostoucím výběrem se střední hodnota odhadu přibližuje odhadovanému parametru.

Ukázali jsme, že s^2 je nestranný odhad. Pokud populační rozptyl σ^2 odhadujeme druhým výběrovým centrálním momentem (ve jmenovateli n)

$$M_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n-1}{n}s^2,$$

pak tento odhad není nestranný (je *vychýlený*), neboť jeho střední hodnota

$$E(M_2) = \frac{n-1}{n} E(s^2) = \frac{n-1}{n} \sigma^2$$

se nerovná hodnotě odhadovaného parametru σ^2. Je však *asymptoticky nestranný*, neboť

$$\lim_{n\to\infty} E(M_2) = \lim_{n\to\infty}\left[\frac{n-1}{n} E(s^2)\right] = \lim_{n\to\infty}\left(\frac{n-1}{n}\sigma^2\right) = \sigma^2 .$$

Odhad je *konzistentní*, když pro něj platí

$$\lim_{n\to\infty} P\left(\left|T_n - \theta\right| < \varepsilon\right) = 1, \quad \varepsilon > 0,$$

tzn., že s rostoucím rozsahem výběru roste pravděpodobnost, že hodnota statistiky T_n se nalézá v blízkosti hodnoty parametru.

Vydatný (*eficientní, nejlepší*) *odhad* je ten odhad, který má nejmenší rozptyl. *Nejlepší nestranný odhad* je takový odhad T, který má nejmenší rozptyl, tj. pro který platí: Nechť T, T' jsou nestranné odhady a pro každé T' platí, $\mathrm{var}(T) \le \mathrm{var}(T')$, pak T je *nejlepší nestranný odhad*.

V matematické statistice se metody odhadu dělí do dvou skupin - momentová metoda a *metoda maximální věrohodnosti*. S principy těchto metod se lze seznámit např. v knize Cyhelský et al. (1996), kde jsou tyto metody vysvětleny přístupnou formou. Odhady získané metodou maximální věrohodnosti, tzv. *ML-odhady*, mají řadu dobrých vlastností, např. jsou *konzistentní*, *asymptoticky nestranné* a *asymptoticky vydatné*. Metoda maximální věrohodnosti je využívána v mnoha statistických programech v odhadu parametrů statistických modelů.

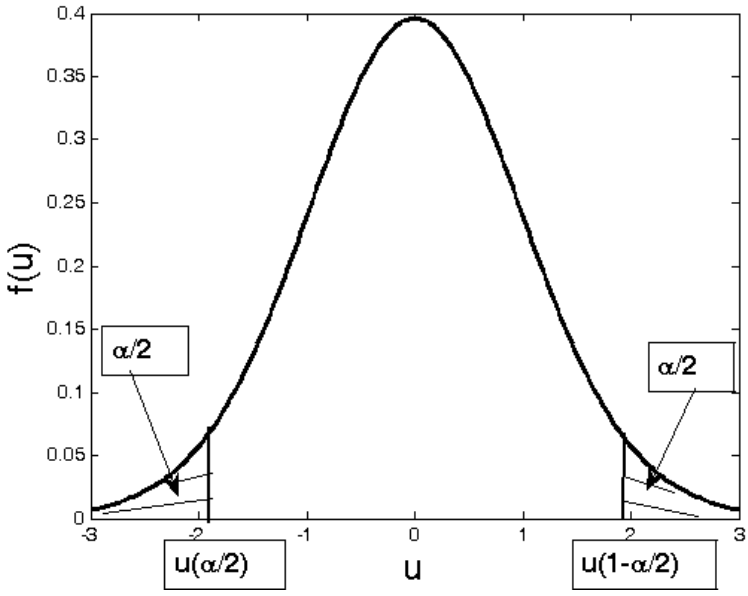
4.2.2 Intervalové odhady

Úkolem intervalového odhadu je určit interval $\left[\theta_1, \theta_2\right]$, v němž leží odhadovaný parametr θ se zadanou pravděpodobností $(1-\alpha)$. Dvojici hodnot θ_1, θ_2 nazýváme intervalovým odhadem (mezemi spolehlivosti) a interval $\left[\theta_1, \theta_2\right]$ pak $100(1-\alpha)$*-procentním intervalem spolehlivosti*, jestliže platí

$$P\left(\theta_1 \le \theta \le \theta_2\right) = 1-\alpha \qquad (1)$$

Hodnota $(1-\alpha)$ se nazývá stupeň spolehlivosti, hodnota α hladina významnosti.

Ukážeme si nyní postup při intervalovém odhadu parametrů normálního rozdělení. Z kapitoly 4.1 víme, že výběrový průměr z normálně rozdělené populace $N(\mu, \sigma^2)$ má rozdělení $\bar{X} \sim N(\mu, \sigma^2/n)$. Po standardizaci dostaneme náhodnou veličinu s normovaným normálním rozdělením: $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$. Jak ukazuje obrázek, platí $P\left(u(\alpha/2) \le \dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le u(1-\alpha/2)\right) = 1-\alpha$.



Pravděpodobnost $(1-\alpha)$ je přesně ta hodnota, kterou požaduje rov. (1) definující intervalový odhad. Proto postupnými úpravami výrazu v závorce jej převedeme na tvar odpovídající rov. (1).

$$P\left(u(\alpha/2)\cdot\sigma/\sqrt{n} \le \bar{X}-\mu \le u(1-\alpha/2)\cdot\sigma/\sqrt{n}\right)=1-\alpha,$$

$$P\left(-\bar{X}+u(\alpha/2)\cdot\sigma/\sqrt{n} \le -\mu \le -\bar{X}+u(1-\alpha/2)\cdot\sigma/\sqrt{n}\right)=1-\alpha.$$

Jelikož normované normální rozdělení je symetrické kolem nulové střední hodnoty, platí $u(\alpha/2)=-u(1-\alpha/2)$, takže po dosazení a vynásobení nerovností hodnotou -1 dostaneme

$$P\left(\bar{X}-u(1-\alpha/2)\cdot\sigma/\sqrt{n} \le \mu \le \bar{X}+u(1-\alpha/2)\cdot\sigma/\sqrt{n}\right)=1-\alpha. \quad (2)$$

Tvar rov. (2) odpovídá tvaru definice (1), takže interval

$$\left[\bar{X}-u(1-\alpha/2)\cdot\frac{\sigma}{\sqrt{n}},\ \bar{X}+u(1-\alpha/2)\cdot\frac{\sigma}{\sqrt{n}}\right] \quad (3)$$

je *dvoustranným* $100(1-\alpha)$*-procentním intervalem spolehlivosti* pro parametr μ, tj. střední hodnotu normálního rozdělení. Pokud bychom znali hodnotu druhého parametru σ^2, mohli bychom meze spolehlivosti pro zvolené α spočítat z výběrových dat. V praktických úlohách však většinou hodnotu tohoto parametru σ^2 neznáme a musíme ji odhadovat výběrovým rozptylem. Pak zcela analogickým postupem dojdeme ke *dvoustrannému* $100(1-\alpha)$*-procentnímu intervalu spolehlivosti* pro parametr μ

$$\left[\bar{X}-t_{n-1}(1-\alpha/2)\cdot\frac{s}{\sqrt{n}},\ \bar{X}+t_{n-1}(1-\alpha/2)\cdot\frac{s}{\sqrt{n}}\right], \quad (4)$$

kde s je výběrová směrodatná odchylka a $t_{n-1}(1-\alpha/2)$ je kvantil t - rozdělení s *n-1* stupni volnosti.

Podobně jako jsme zavedli dvoustranný interval spolehlivosti, mohli bychom i zavést jednostranné intervaly spolehlivosti:

Levým jednostranným $100(1-\alpha)$ *-procentním intervalem spolehlivosti* pro parametr μ:

$$\left(-\infty, \bar{X} + t_{n-1}(1-\alpha)\cdot\frac{s}{\sqrt{n}}\right]. \qquad (5)$$

Pravým jednostranným $100(1-\alpha)$ *-procentním intervalem spolehlivosti* pro parametr μ:

$$\left[\bar{X} - t_{n-1}(1-\alpha)\cdot\frac{s}{\sqrt{n}}, +\infty\right). \qquad (6)$$

Při odvození dvoustranného intervalu spolehlivosti pro parametr σ^2 normálního rozdělení vyjdeme z toho, že

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}, \qquad \text{viz kapitola 4.1, rov.(5).}$$

Pak platí, že

$$P\left(\chi^2_{n-1}(\alpha/2) \le \frac{(n-1)s^2}{\sigma^2} \le \chi^2_{n-1}(1-\alpha/2)\right) = 1-\alpha$$

Po úpravě pak dostaneme

$$P\left(\frac{(n-1)s^2}{\chi^2_{n-1}(1-\alpha/2)} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{n-1}(\alpha/2)}\right) = 1-\alpha$$

a interval

$$\left[\frac{(n-1)s^2}{\chi^2_{n-1}(1-\alpha/2)}, \frac{(n-1)s^2}{\chi^2_{n-1}(\alpha/2)}\right] \qquad (7)$$

je *dvoustranným* $100(1-\alpha)$ *-procentním intervalem spolehlivosti* pro parametr σ^2, tj. pro rozptyl normálního rozdělení.

Příklad 1: Předpokládejme, že populace má normální rozdělení s neznámými parametry μ a σ^2, zkráceně zapsáno $N(\mu, \sigma^2)$. Z výběru o rozsahu 26 jsme spočetli průměr 105 a výběrový rozptyl 25. Naším úkolem je určit oboustranné 95 %-ní intervaly spolehlivosti pro parametry μ a σ^2.

Řešení:

Interval spolehlivosti pro parametr μ určíme dosazením do (4), příslušnou hodnotu kvantilu $t_{25}(0.975) = 2{,}06$ nalezneme v tabulce 6.3, takže oboustranný 95%-ní interval spolehlivosti pro parametr μ je

$\left[105 - 2{,}06\frac{5}{\sqrt{26}}; 105 + 2{,}06\frac{5}{\sqrt{26}}\right]$, po vyčíslení pak přibližně $\left[103{,}0; 107{,}0\right]$.

Oboustranný 95%-ní interval spolehlivosti pro parametr σ^2 určíme dosazením do (7), potřebné hodnoty kvantilů nalezneme v tabulce 6.2, $\chi^2_{25}(0{,}025) = 13{,}12$, $\chi^2_{25}(0{,}975) = 40{,}65$. Tedy oboustranný interval spolehlivosti je

$\left[\frac{25,25}{40,65}; \frac{25,25}{13,12}\right]$, po vyčíslení je tento interval přibližně $\left[15,4;\ 47,6\right]$.

Vztah (2) lze využít pro výběry velkého rozsahu i v situaci, kdy parametr σ^2 neznáme. Z kapitoly 3.6 o centrální limitní větě víme, že průměr většího počtu nezávislých stejně rozdělených náhodných veličin má přibližně normální rozdělení a že veličina $\frac{f_n - p}{\sqrt{p\,(1-p)/n}}$ má přibližně normované normální rozdělení $N(0,1)$, statistika f_n znamená relativní četnost hodnot 1 ve výběru o rozsahu n z populace, která má alternativní rozdělení s parametrem p. Pro velké výběry tedy z rov. (2) přibližně platí, že

$$P\left(f_n - u(1-\alpha/2)\sqrt{\frac{p(1-p)}{n}} \le p \le f_n + u(1-\alpha/2)\sqrt{\frac{p(1-p)}{n}} \right) = 1-\alpha\,. \qquad (8)$$

Nestranným odhadem rozptylu $\frac{p(1-p)}{n}$ je $\frac{f_n(1-f_n)}{n-1}$ a tedy *dvoustranný* $100(1-\alpha)$*-procentní interval spolehlivosti* pro parametr p je

$$\left\langle f_n - u(1-\alpha/2)\sqrt{\frac{f_n(1-f_n)}{n-1}}\ ,\ \ f_n + u(1-\alpha/2)\sqrt{\frac{f_n(1-f_n)}{n-1}} \right\rangle. \qquad (9)$$

Příklad 2: V průzkumu volebních preferencí dotazem na 900 náhodně vybraných potenciální voličů bylo zjištěno, že politickou stranu ABC by volilo 25 % dotazovaných voličů. Určete oboustranný 95%-ní interval spolehlivosti pro parametr *p*, tj. voličskou preferenci této strany v populaci.

Kvantil normovaného normálního rozdělení *u(0,975)= 1,96.* Dosazením do (9) získáme

$$\left\langle 0{,}25 - 1{,}96\sqrt{\frac{0{,}25\,.\,(1-0{,}25)}{899}}\ ;\ \ 0{,}25 + 1{,}96\sqrt{\frac{0{,}25\,.\,(1-0.{,}25)}{899}} \right\rangle,$$

což po vyčíslení dá $\langle 0{,}222;\ \ 0{,}278\rangle$. V tomto intervalu leží parametr p s pravděpodobností 0,95.

Souhrn:

- *Náhodný výběr* jedinců y populace se realizuje tak, že o zařazení jedince do výběru rozhoduje *náhoda* a každý jedinec z populace má *stejnou pravděpodobnost* zařazení do výběru.
- V matematické statistice je náhodný výběr náhodný vektor o n složkách $(X_1, X_2, \ldots, X_n)$, kde složky tohoto vektoru jsou *nezávislé* náhodné veličiny s *identickým rozdělením*.
- Statistiku T můžeme vyjádřit jako funkci náhodného výběru, tedy $T = T(X_1, X_2, \ldots, X_n)$.
- Statistika je náhodná veličina.
- Rozdělení statistik nazýváme výběrovými rozděleními.
- Úkolem bodového odhadu je nalézt co nejlépe hodnotu parametru θ, tuto hodnotu odhadujeme nějakou statistikou $T = T(X_1, X_2, \ldots, X_n)$, spočítanou z výběrových hodnot.
- Úkolem intervalového odhadu je určit interval $\langle \theta_1, \theta_2 \rangle$, v němž leží odhadovaný parametr θ se zadanou pravděpodobností $(1-\alpha)$, hodnota $(1-\alpha)$ se nazývá stupeň spolehlivosti.

Kontrolní otázky:

1. *Co je náhodný výběr v matematické statistice?*
2. *Co je to statistika? Je to deterministická nebo náhodná veličina?*
3. *Jaké je rozdělení výběrových průměrů z normálního rozdělení? Jaké má parametry?*
4. *Co znamená, že bodový odhad je nestranný? Proč při výpočtu výběrového rozptylu se ve jmenovateli užívá výraz (n-1)?*
5. *Kdy je odhad asymptoticky nestranný?*
6. *Co platí o konsistentním odhadu?*
7. *Co je to nejlepší nestranný odhad?*
8. *Co nám říká intervalový odhad?*

Pojmy k zapamatování:

- náhodný výběr
- statistický odhad
- bodové odhady parametrů
- nestranný odhad, konsistentní odhad, nejlepší nestranný odhad
- intervalový odhad, interval spolehlivosti

Následující část této kapitoly je věnována základům testování statistických hypotéz. Studium této části vám zabere asi tři až čtyři hodiny.

4.3 Testování hypotéz

V mnoha aplikacích statistiky se užívá postup, kterému se říká statistické testování hypotéz. Základní myšlenku se pokusíme vysvětlit na poměrně jednoduchém příkladu. Nejčastěji testovanými hypotézami jsou hypotézy o parametru rozdělení populace. Uvažujme tedy příklad, jehož realizaci si dovedeme snadno představit. Naším úkolem je posoudit (ve statistice se říká testovat) hypotézu, že střední hodnota tělesné výšky studentů-mužů z Ostravské university je 175 cm. Asi vás napadá, že nejjednodušší by bylo všechny tyto studenty změřit, vypočítat průměr a porovnat vypočtený průměr s hypotetickou hodnotou 175 cm a jsme s úlohou hotovi. Vystačili bychom s jednoduchými postupy popisné statistiky a žádným testováním hypotéz se nemusíme zatěžovat. Bohužel ne vždy je takové jednoduché řešení přijatelné. I ve výše uvedené situaci bychom se asi těžko rozhodovali, kdyby populační průměr vyšel velmi blízko hodnotě 175 cm, řekněme 175,01 cm. Tato hodnota je sice různá od 175 cm, ale je tento rozdíl podstatný? V jiných úlohách zkoumaná populace může být početnější než těch zhruba 3000 studentů-mužů na OU, takže změřit všechny jedince je nemožné. Zkrátka řečeno, často jsou k dispozici data jen o výběru jedinců z populace, nikoliv o celé populaci. Pak použití metod statistické indukce je nezbytné a testování hypotéz se nevyhneme.

Vraťme se k uvedenému příkladu. Tělesná výška dospělé mužské populace je spojitá náhodná veličina. Ze zkušenosti několika generací badatelů víme, že tělesná výška má normální rozdělení, $N(\mu,\sigma^2)$. Pokud bychom tuto zkušenost předchozích generací neměli, museli bychom tvar rozdělení zjišťovat sami studiem empirických rozdělení mnoha výběrů nebo na tvar rozdělení usoudit ze zákonitostí procesu vytvářejícího data.

Předpokládejme, že jsme pořídili náhodný výběr n jedinců ze sledované populace a změřili jejich výšku. Jak víme, na náhodný výběr pohlížíme ve statistice jako na vektor $(X_1, X_2, \ldots, X_n)$ nezávislých náhodných veličin stejného rozdělení, v našem případě $X_i \sim N(\mu,\sigma^2), \quad i = 1, 2, \ldots, n$. Problém ovšem je v tom, že hodnoty parametrů μ, σ^2 neznáme. Kdybychom je znali, nepotřebujeme žádná výběrová data, protože pravděpodobnostní rozdělení populace včetně parametrů (charakteristik populace) by bylo známo a žádná data bychom nepotřebovali.

My však můžeme spočítat jen hodnoty charakteristik výběrových, např.výběrový průměr $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ a výběrový rozptyl $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.
Z kapitol 4.1 a 4.2 už víme, že výběrový průměr má normální rozdělení, $\bar{X} \sim N(\mu, \sigma^2/n)$ a normovaná náhodná veličina $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$. Kdybychom hodnotu parametru σ^2 znali, uměli bychom hodnotu této náhodné veličiny

vyčíslit. Jelikož σ^2 neznáme, musíme jej odhadnout výběrovým rozptylem s^2. Pak náhodná veličina T má t-rozdělení

$$T = \frac{\bar{X} - \mu}{s / \sqrt{n}} \sim t_{n-1}. \qquad (1)$$

Tato náhodná veličina je vyjádřena jen pomocí výběrových statistik ($\bar{X}, s$), rozsahu výběru n a parametru μ, o jehož hodnotě máme testovat nějaké tvrzení, tzv. *nulovou hypotézu*, H_0, v našem případě hypotézu

H_0 : $\mu = 175$ cm

proti tzv. *alternativní hypotéze* nebo krátce *alternativě* H_1, v našem případě

H_1 : $\mu \neq 175$ cm.

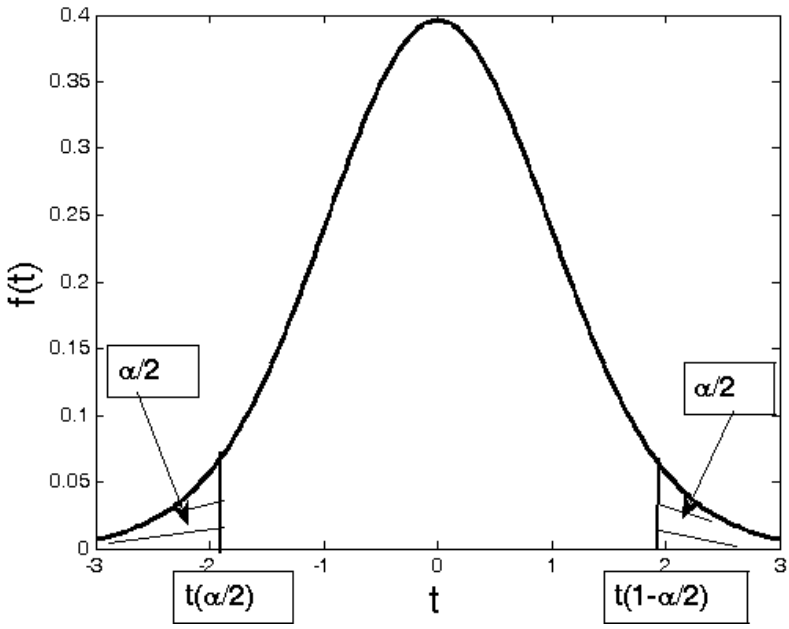
Pokud tvrzení formulované nulovou hypotézou H_0 je pravdivé, v našem příkladu $\mu = 175$ cm, pak pro rozdělení náhodné veličiny, kterou získáme dosazením této hodnoty do (1), platí

$$\frac{\bar{X} - 175}{s / \sqrt{n}} \sim t_{n-1}.$$

Této náhodné veličině říkáme *testová statistika (testové kriterium)*, neboť ji můžeme užít k testu hypotézy H_0.

Testem hypotézy rozhodujeme mezi přijetím či odmítnutím tvrzení formulovaného nulovou hypotézou H_0. Je to situace podobná rozhodování soudu, který rozhoduje o nevině či vině obžalovaného. Rozhodování je zatíženo rizikem nesprávného rozhodnutí, soud může odsoudit nevinného (justiční omyl) nebo propustit viníka bez potrestání, pokud se jeho vinu nepodařilo prokázat. Na rozdíl od soudu lze však pravděpodobnost neoprávněného zamítnutí nulové hypotézy H_0 předem stanovit, neboť známe rozdělení testové statistiky. Tato pravděpodobnost se nazývá *hladina významnosti* testu a většinou se označuje symbolem α.

Za platnosti nulové hypotézy má testová statistika t-rozdělení s $n-1$ stupni volnosti a může teoreticky nabývat jakoukoliv reálnou hodnotu, tj. $(-\infty, +\infty)$.



Otázkou je, kdy nulovou hypotézu zamítnout. Intuitivně je zřejmé, že zamítnout H_0 můžeme tehdy, když $\overline{X}$ se bude podstatně lišit od hodnoty předpokládané v nulové hypotéze, v našem příkladu od 175. Přitom chceme, aby pravděpodobnost chybného, nesprávného zamítnutí byla rovna hladině významnosti testu α. Z obrázku je vidět, že zamítnout H_0 můžeme, když absolutní hodnota rozdílu $\left|\overline{X} - 175\right|$ je velká, přesněji vyjádřeno, když pro hodnotu testového kritéria platí

$$\left|\frac{\overline{X} - 175}{s/\sqrt{n}}\right| \geq t_{n-1}(1-\alpha/2).$$

Tzn., že H_0 zamítneme, když hodnota testového kritéria je z množiny W,

$$W \equiv \left(-\infty, t_{n-1}(\alpha/2)\right] \cup \left[t_{n-1}(1-\alpha/2), +\infty\right)$$

Množině W se říká *kritický obor*. Volně řečeno, „padne-li“ hodnota testového kriteria při hodnocení výběrových dat do kritického oboru, zamítáme nulovou hypotézu.

Vrátíme se k našemu příkladu:

1. Máme zformulovanou nulovou hypotézu i alternativu:
   H_0 : μ = 175 cm,
   H_1 : $\mu \neq$ 175 cm.

2. Zvolíme hladinu významnosti testu α = 0,05.

3. Víme, že vhodným testovým kritériem pro test této hypotézy je statistika
   $$\frac{\overline{X} - 175}{s/\sqrt{n}} \sim t_{n-1}.$$

4. Určíme kritický obor, potřebný kvantil $t_{15}(0{,}975) = 2{,}13$ nalezneme v tab. 6.3 $W \equiv (-\infty, -2{,}13] \cup [2{,}13, +\infty)$.

5. Z výběru o rozsahu $n = 16$ jsme zjistili $\overline{X} = 177{,}2$ cm a $s^2 = 39{,}5$ cm^2.

6. Vypočteme hodnotu testového kritéria
$$\frac{\overline{X} - 175}{s / \sqrt{n}} = \frac{177{,}2 - 175}{\sqrt{39{,}5} / \sqrt{16}} = 1{,}40 .$$

7. Přijmeme rozhodnutí: Jelikož hodnota testového kritéria není v kritickém oboru, nemůžeme zamítnout nulovou hypotézu, že $\mu = 175$ cm.

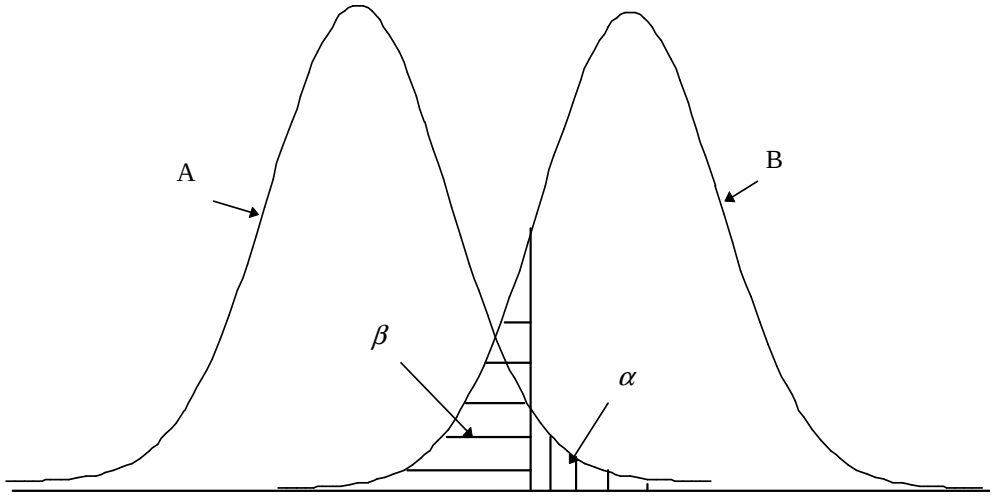
Data z našeho výběru tedy neopravňují k zamítnutí nulové hypotézy. Tím jsme však nedokázali, že tvrzení touto hypotézou formulované je pravdivé. Přijmeme-li analogii s rozhodováním soudu, pouze nemáme dostatečný „důkaz“ o vině obžalovaného a nezbývá, než jej propustit a věřit v jeho nevinu. Je zřejmé, že pravděpodobnost nesprávného odsouzení nevinného záleží na přísnosti soudu. Pokud je soud přísný, tj. stačí málo a obžalovaný jde do vězení, pak je větší pravděpodobnost justičního omylu, ale sníží se pravděpodobnost, že na svobodě zůstanou nepotrestaní viníci. Podobné je to i se statistickým testováním hypotéz. Zvolíme-li hladinu významnosti α velkou („přísný soud“), je větší riziko neoprávněného zamítnutí nulové hypotézy („odsouzení nevinného“). Zvolíme-li hladinu významnosti α nízkou („benevolentní soud“, prohřešek musí být velký, aby odsoudil), je větší riziko neoprávněného nezamítnutí nulové hypotézy („nepotrestaný viník“). Při testování hypotéz se tedy můžeme dopustit nesprávného rozhodnutí dvojího druhu. Situaci ukazuje následující tabulka.

	NAŠE ROZHODNUTÍ:	
SKUTEČNOST (nám neznámá)	H_0 : nezamítáme	H_0 : zamítáme
Tvrzení H_0 je pravdivé	*SPRÁVNÉ*	*CHYBA I. druhu*
Tvrzení H_0 je nepravdivé	*CHYBA II. druhu*	*SPRÁVNÉ*

Při testování hypotéz pravděpodobnost chyby I. druhu stanovujeme předem, je rovna hladině významnosti α. Obvykle se volí $\alpha = 0{,}05$ nebo $\alpha = 0{,}01$ či $\alpha = 0{,}001$ podle závažnosti chyby I. druhu. Pravděpodobnost chyby druhého druhu označujeme obyčejně symbolem β a veličina $(1-\beta)$ se nazývá *síla testu*.

Již jsme vysvětlili, že pravděpodobnosti chyb I. a II. druhu spolu souvisejí. Snižujeme-li při daném rozsahu výběru pravděpodobnost chyby I. druhu α, roste pravděpodobnost chyby II. druhu β. Situaci ilustruje obrázek. Křivka A je hustota, která odpovídá hustotě testové statistiky za platnosti nulové hypotézy a kterou užíváme při testu. Křivka B je hustota odpovídající skutečnosti (nulová hypotéza neplatí). Vidíme, že snižováním hladiny významnosti testu se zvětšuje

pravděpodobnost chyby II. druhu, β. Pokud při pevném α chceme zvýšit sílu testu, tj. snížit pravděpodobnost chyby II. druhu β, je nutné zvětšit rozsah výběru.



Základní myšlenky statistického testování hypotéz jsme si ukázali na testu, kterému se říká *jednovýběrový dvoustranný t-test*. Jednovýběrový proto, že využívá data z jednoho výběru, dvoustranný (někdy se užívá přívlastek *oboustranný*) proto, že kritický obor je na obou koncích rozdělení testové statistiky.

V tomto *jednovýběrovém dvoustranném* t - testu testujeme hypotézu, že střední hodnota normálně rozdělené populace, ze které máme výběr, je rovna nějaké dané hodnotě μ_0, proti alternativě, že tomu tak není:

H_0 : $\mu = \mu_0$,
H_1 : $\mu \neq \mu_0$.

Povšimněme si, že při tomto testu zamítáme nulovou hypotézu tehdy, když *dvoustranný* $100(1-\alpha)$*-procentní interval spolehlivosti* pro parametr μ neobsahuje hodnotu μ_0 předpokládanou nulovou hypotézou, srovnej se vztahem (4) v kap. 4.2.2. Další možnosti využití *t*-testů ukážeme v následujícím semestru v předmětu Analýza dat.

Souhrn:

- Statistický test hypotézy se užívá k rozhodování za nejistoty. Rozhodujeme mezi nulovou hypotézou a alternativou.
- Jsou dva druhy chybného rozhodnutí.
- Pravděpodobnost chyby I. druhu při testu volíme předem (hladina významnosti).
- Test hypotézy je analogický rozhodování soudu, ale rozdíl je v tom, že pravděpodobnost chyby prvního druhu je u statistických testů známa, dokonce ji zvolíme.
- Kritický obor testu závisí na tom, jak je zformulována alternativa.

Kontrolní otázky:

1. *Proč testy o parametrech jsou rozhodování v nejistotě?*
2. *Vysvětlete rozdíl mezi chybou prvního a druhého druhu.*
3. *Proč je zamítnutí nulové hypotézy pro praktické rozhodování užitečnější výsledek než nezamítnutí nulové hypotézy?*

Pojmy k zapamatování:

- statistické testování hypotéz,
- nulová hypotéza, alternativa,
- chyby prvního a druhého druhu,
- hladina významnosti,
- síla testu,
- testová statistika (testové kriterium),
- kritický obor,
- jednovýběrový *t*-test.

Korespondenční úlohy budou zadávány vždy na začátku semestru.
K této kapitole bude korespondenční úloha obsahovat jeden příklad.

5 Literatura - komentovaný seznam

Seznam je pouze zlomkem rozsáhlé statistické literatury dostupné v současné době. Zařazeny jsou především knihy a skripta českých autorů nebo české překlady z posledního období. Při výběru byl brán zřetel na dostupnost pro studenty Ostravské university a také na přístupnost textu začátečníkům ve statistice.

Anděl, J.: Matematická statistika, SNTL Praha, 1978
Klasická učebnice matematické statistiky. Úplné sledování vyžaduje hlubší znalosti matematické analýzy a lineární algebry, ale kniha obsahuje řadu příkladů, které jsou srozumitelné i bez těchto matematických znalostí a pomohou čtenáři orientovat se v aplikaci statistických metod.

Anděl, J.: Statistické metody, Matfyzpress Praha, 1993
Příručka, která pomůže prohloubit porozumění v mnoha oblastech aplikací statistických metod, nepředpokládá hlubší znalosti matematiky. Pokrývá širokou paletu běžně užívaných metod, přístupným způsobem vysvětluje jejich základy a předpoklady pro aplikaci.

Cyhelský, L., Kahounová, J. , Hindls, R.: Elementární statistická analýza, Management Press, Praha, 1996
Kniha přístupným způsobem vysvětluje základy deskriptivní statistiky a počtu pravděpodobnosti nutné pro aplikace statistiky. Zabývá se základy teorie odhadu a testování hypotéz. Neobsahuje analýzu rozptylu a regresi. Knihu je možno doporučit čtenáři se středoškolskými znalostmi matematiky jako první učebnici pro seznámení s problémy statistické analýzy dat. Dostupná v knihovně OU.

Havránek, T. a kol.: Matematika pro biologické a lékařské vědy, Academia, 1981
Obsahuje rozsáhlou kapitolu o počtu pravděpodobnosti.

Havránek, T.: Statistika pro biologické a lékařské vědy, Academia, 1993
Kniha vynikajícího, bohužel předčasně zesnulého českého statistika, která vyšla až dva roky po jeho smrti. Kniha poměrně přístupným způsobem vykládá i obtížné partie matematické statistiky. Aplikace statistických metod je ilustrována na řadě příkladů z autorovy praxe v analýze biomedicínských dat.

Hebák, P., Kahounová, J.: Počet pravděpodobnosti v příkladech, SNTL, 1988
Kniha obsahuje úvod do počtu pravděpodobnosti a ukazuje použití jeho výsledků na řadě řešených příkladů. Vhodná pro cvičení a upevňování znalostí tohoto oboru.

Hebák, P., Hustopecký, J.: Průvodce moderními statistickými metodami, SNTL Praha, 1990
Na více než třiceti příkladech inspirovaných praktickými úlohami je důkladně ilustrována aplikace různých metod induktivní statistiky, včetně formulace úlohy, zdůvodnění různých alternativ řešení a interpretace výsledků

Hradecký, P., Madryová, A., Turčan, M.: Pravděpodobnost, skriptum VŠB-TU Ostrava, 1998
Učební text k jednosemestrálnímu úvodnímu kurzu pravděpodobnosti. Může posloužit jako vhodná doplňková literatura.

Komenda, S.: Biometrie, skriptum PřF UP Olomouc, 1994
Autor do učebního textu promítá dlouholetou zkušenost z oblasti aplikací statistiky v biomedicínském výzkumu. Přístupnou formou jsou vysvětleny základy pravděpodobnosti, statistiky i mnohé metodologické otázky. Čtenářskou zajímavost textu zvyšuje řada původních aforismů. Vhodný úvodní text. Skriptum je dostupné ve více výtiscích v knihovně OU.

Komenda, S.: Politometrie, Vydavatelství UP, Olomouc, 1995
Podtitulek „Základy statistické gramotnosti v politologii s úvodem do metodologie vědy“ vystihuje přesně obsah učebního textu. Jak je u prof. Komendy obvyklé, jeho text potěší i čtivostí a vtipem. Skriptum je dostupné ve více výtiscích v knihovně OU.

Křivý, I.: Úvod do teorie pravděpodobnosti, skriptum PF Ostrava, 1983
Učební text byl určen původně pro studenty učitelství matematiky, takže v některých jeho částech se předpokládá přiměřená matematická znalost. Skriptum je dostupné ve více výtiscích v knihovně OU.

Křivý, I.: Základy matematické statistiky, skriptum PF Ostrava, 1985
Učební text pro studenty učitelství matematiky. Pokrývá základní aplikační oblasti matematické statistiky. K úplnému sledování je potřeba vyšší než středoškolská úroveň matematiky. Skriptum je dostupné ve více výtiscích v knihovně OU.

Křivý, I.: Základy teorie pravděpodobnosti, skriptum PřF OU Ostrava, 2004.

Laga, J., Likeš, J.: Základní statistické tabulky, SNTL, 1978
Obsáhlé „klasické“ statistické tabulky českých autorů, obsahují i důkladné vysvětlení pojmů důležitých pro správné užití tabulek v aplikacích metod matematické statistiky.

Lepš, J.: Biostatistika, skriptum, Jihočeská universita, Čes. Budějovice, 1996
Netradičně napsaný učební text (autor je biolog), ve kterém je čtenář na příkladech veden od základních pojmů až k mnohorozměrným metodám analýzy dat.

Likeš, J., Machek, J.: Matematická statistika, SNTL, Praha, 1983
Přístupně napsaná učebnice statistiky pro vysoké školy technické. Předpokládá znalost základů matematické analýzy v rozsahu vyučovaném na technických školách.

Meloun, M.,Militký, J.: Statistické zpracování experimentálních dat, PLUS, 1994
Rozsáhlá kniha aplikačně orientovaná, zejména na metody regresní analýzy.

NCSS Statistical System for Windows – User ‘s Guide,
Popisuje nejen ovládání programového systému, ale také základy implementovaných metod a doporučení pro interpretaci výsledků. K dispozici je on-line jako součást instalace NCSS.

Potocký R., Kalas J., Komorník J, Lamoš F.: Zbierka úloh z pravděpodobnosti a matematickej štatistiky, 2. vydání, Alfa, Bratislava, 1991
Rozsáhlá sbírka příkladů, určená především studentům matematicko-fyzikálních a přírodovědeckých fakult. Každá část knihy obsahuje i zavedení pojmů a nezbytný výklad teorie, které jsou potřebné pro řešení příkladů.

Swoboda, H.: Moderní statistika, Svoboda, Praha, 1977
Autorem této výjimečné knihy je rakouský žurnalista, snad právě proto dokáže vysvětlovat složité statistické pojmy srozumitelným a poutavým způsobem. I české vydání této knihy je vybaveno spoustou výstižných ilustrací a na dobu vydání až překvapivě dobře typograficky zpracováno. Dostupná v naší univerzitní knihovně.

Řezanková H.: Analýza kategoriálních dat, VŠE, Praha, 2005
Učební text zaměřený na metody analýzy nominálních a ordinálních veličin a jejich vztahů.

Tošenovský, J., Noskievičová D: Statistické metody pro zlepšování jakosti. Monanex, Ostrava, 2000
Úvodní kapitoly obsahují vysvětlení základů pravděpodobnosti a zavedení pojmů nutných pro aplikaci statistických metod. Další kapitoly jsou věnovány statistickým metodám užívaným v hodnocení kvality výrobních procesů.

Tvrdík J.: Základy statistické analýzy dat, Přírodovědecká fakulta Ostravské university, Ostrava 1998
Přístupně napsaný učební text zaměřený na pochopení důležitých pojmů nutných pro aplikaci statistických metod.

Zvára, K.: Biostatistika, Karolinum, Praha, 1998
Velmi zdařilá učebnice statistiky určená především studentům biologie. Důraz je kladen na aplikaci statistických metod, která je ilustrována řadou řešených příkladů z biologického výzkumu.

Zvára K., Štěpán J.,: Pravděpodobnost a matematická statistika, Matfyzpress, Praha, 2001
Vynikající učebnice původně napsaná pro studenty matematiky na pedagogických fakultách. Vhodná doplňující literatura, prohlubující znalosti matematické statistiky.

Interaktivní učebnice pro základní kurs statistiky:

Härdle W. et al., MM*Stat - Základy statistiky,
http://www.quantlet.com/mdstat/scripts/mmcze/java/start.html, 2005

Řezanková, H., Marek, L., Vrabec, M., Kalenský, L., Řezanka, P.,
IASTAT - Interaktivní učebnice statistiky, http://badame.vse.cz/iastat/, 2005

Dear, K. et al., Surf-Stat,
http://www.anu.edu.au/nceph/surfstat/surfstat-home/surfstat.html, 2005

6 Statistické tabulky

Statistické tabulky byly pořízeny s využitím statistických funkcí NORMSDIST, CHIINV, TINV, FINV programu Microsoft Excel pro Windows 95, verse 7.0. Pokud jste u počítače, na kterém je nainstalován Excel nebo některý ze statistických programů (NCSS atd.) statistické tabulky nepotřebujete, neboť potřebné hodnoty distribučních funkcí či kvantilů snadno zjistíte pomocí těchto programových prostředků.

6.1 Distribuční funkce normovaného normálního rozdělení

$$X \sim N(0,1), \quad \Phi(x) = P(X < x)$$

$\Phi(x)$

x	*+0*	*+0,02*	*+0,04*	*+0,06*	*+0,08*
0,0	0,5000	0,5080	0,5160	0,5239	0,5319
0,1	0,5398	0,5478	0,5557	0,5636	0,5714
0,2	0,5793	0,5871	0,5948	0,6026	0,6103
0,3	0,6179	0,6255	0,6331	0,6406	0,6480
0,4	0,6554	0,6628	0,6700	0,6772	0,6844
0,5	0,6915	0,6985	0,7054	0,7123	0,7190
0,6	0,7257	0,7324	0,7389	0,7454	0,7517
0,7	0,7580	0,7642	0,7704	0,7764	0,7823
0,8	0,7881	0,7939	0,7995	0,8051	0,8106
0,9	0,8159	0,8212	0,8264	0,8315	0,8365
1,0	0,8413	0,8461	0,8508	0,8554	0,8599
1,1	0,8643	0,8686	0,8729	0,8770	0,8810
1,2	0,8849	0,8888	0,8925	0,8962	0,8997
1,3	0,9032	0,9066	0,9099	0,9131	0,9162
1,4	0,9192	0,9222	0,9251	0,9279	0,9306
1,5	0,9332	0,9357	0,9382	0,9406	0,9429
1,6	0,9452	0,9474	0,9495	0,9515	0,9535
1,7	0,9554	0,9573	0,9591	0,9608	0,9625
1,8	0,9641	0,9656	0,9671	0,9686	0,9699
1,9	0,9713	0,9726	0,9738	0,9750	0,9761
2,0	0,9772	0,9783	0,9793	0,9803	0,9812
2,1	0,9821	0,9830	0,9838	0,9846	0,9854
2,2	0,9861	0,9868	0,9875	0,9881	0,9887
2,3	0,9893	0,9898	0,9904	0,9909	0,9913
2,4	0,9918	0,9922	0,9927	0,9931	0,9934
2,5	0,9938	0,9941	0,9945	0,9948	0,9951

6.2 Vybrané kvantily rozdělení Chí-kvadrát

$$X \sim \chi_n^2, \quad P\left[X < x(p)\right] = p$$

	x(p)			
n	*p=0,025*	*p=0,95*	*p=0,975*	*p=0,99*
1	0,00	3,84	5,02	6,63
2	0,05	5,99	7,38	9,21
3	0,22	7,81	9,35	11,34
4	0,48	9,49	11,14	13,28
5	0,83	11,07	12,83	15,09
6	1,24	12,59	14,45	16,81
7	1,69	14,07	16,01	18,48
8	2,18	15,51	17,53	20,09
9	2,70	16,92	19,02	21,67
10	3,25	18,31	20,48	23,21
11	3,82	19,68	21,92	24,73
12	4,40	21,03	23,34	26,22
13	5,01	22,36	24,74	27,69
14	5,63	23,68	26,12	29,14
15	6,26	25,00	27,49	30,58
16	6,91	26,30	28,85	32,00
17	7,56	27,59	30,19	33,41
18	8,23	28,87	31,53	34,81
19	8,91	30,14	32,85	36,19
20	9,59	31,41	34,17	37,57
25	13,12	37,65	40,65	44,31
30	16,79	43,77	46,98	50,89
40	24,43	55,76	59,34	63,69
50	32,36	67,50	71,42	76,15
100	74,22	124,34	129,56	135,81

6.3 Vybrané kvantily Studentova t-rozdělení

$$X \sim t_n, \quad P\left[X < x(p)\right] = p$$

	x(p)				
n	*p=0,9*	*p=0,95*	*p=0,975*	*p=0,99*	*p=0,995*
1	3,08	6,31	12,71	31,82	63,66
2	1,89	2,92	4,30	6,96	9,92
3	1,64	2,35	3,18	4,54	5,84
4	1,53	2,13	2,78	3,75	4,60
5	1,48	2,02	2,57	3,36	4,03
6	1,44	1,94	2,45	3,14	3,71
7	1,41	1,89	2,36	3,00	3,50
8	1,40	1,86	2,31	2,90	3,36
9	1,38	1,83	2,26	2,82	3,25
10	1,37	1,81	2,23	2,76	3,17
11	1,36	1,80	2,20	2,72	3,11
12	1,36	1,78	2,18	2,68	3,05
13	1,35	1,77	2,16	2,65	3,01
14	1,35	1,76	2,14	2,62	2,98
15	1,34	1,75	2,13	2,60	2,95
16	1,34	1,75	2,12	2,58	2,92
17	1,33	1,74	2,11	2,57	2,90
18	1,33	1,73	2,10	2,55	2,88
19	1,33	1,73	2,09	2,54	2,86
20	1,33	1,72	2,09	2,53	2,85
25	1,32	1,71	2,06	2,49	2,79
30	1,31	1,70	2,04	2,46	2,75
40	1,30	1,68	2,02	2,42	2,70
50	1,30	1,68	2,01	2,40	2,68
70	1,29	1,67	1,99	2,38	2,65
100	1,29	1,66	1,98	2,36	2,63
500	1,28	1,65	1,96	2,33	2,59

6.4 Vybrané kvantily Fisherova-Snedecorova F-rozdělení

$$X \sim F_{m,n}, \quad P\left[X < x(0{,}95)\right] = 0{,}95$$

x(0,95)

	m							
n	1	2	3	4	5	10	20	40
1	161,45	199,50	215,71	224,58	230,16	241,88	248,02	251,14
2	18,51	19,00	19,16	19,25	19,30	19,40	19,45	19,47
3	10,13	9,55	9,28	9,12	9,01	8,79	8,66	8,59
4	7,71	6,94	6,59	6,39	6,26	5,96	5,80	5,72
5	6,61	5,79	5,41	5,19	5,05	4,74	4,56	4,46
6	5,99	5,14	4,76	4,53	4,39	4,06	3,87	3,77
7	5,59	4,74	4,35	4,12	3,97	3,64	3,44	3,34
8	5,32	4,46	4,07	3,84	3,69	3,35	3,15	3,04
9	5,12	4,26	3,86	3,63	3,48	3,14	2,94	2,83
10	4,96	4,10	3,71	3,48	3,33	2,98	2,77	2,66
11	4,84	3,98	3,59	3,36	3,20	2,85	2,65	2,53
12	4,75	3,89	3,49	3,26	3,11	2,75	2,54	2,43
13	4,67	3,81	3,41	3,18	3,03	2,67	2,46	2,34
14	4,60	3,74	3,34	3,11	2,96	2,60	2,39	2,27
15	4,54	3,68	3,29	3,06	2,90	2,54	2,33	2,20
20	4,35	3,49	3,10	2,87	2,71	2,35	2,12	1,99
30	4,17	3,32	2,92	2,69	2,53	2,16	1,93	1,79
40	4,08	3,23	2,84	2,61	2,45	2,08	1,84	1,69
60	4,00	3,15	2,76	2,53	2,37	1,99	1,75	1,59
120	3,92	3,07	2,68	2,45	2,29	1,91	1,66	1,50
500	3,86	3,01	2,62	2,39	2,23	1,85	1,59	1,42