1 Úvod

Výuka statistických metod má na lesnické a dřevařské fakultě Mendelovy zemědělské a lesnické univerzity v Brně (FLD MZLU) dlouholetou tradici, což také dokládá řada studijních textů, vydaných pro výuku tohoto oboru (z nejvýznamnějších jmenujme např. LEPORSKÝ 1953, ŠMELKO-WOLF 1977, ZACH 1994). K vydání dalšího učebního textu vedly především následující důvody:

- stále vzrůstající význam uplatnění matematicko-statistických metod ve výzkumné i provozní praxi, včetně rozvoje nových oborů (např. statistická kontrola kvality).
- rozvoj výpočetní techniky a všeobecně dostupného softwaru pro osobní počítače, což umožňuje využití i teoreticky a výpočetně náročnějších metod ve výzkumné a provozní praxi
- další vývoj metodologické základny oboru – do běžné praxe (a do používaného softwaru) jsou zaváděny nové metody
- výuka dalších oborů na LDF – kromě lesnického oboru, pro který byly publikovány všechny předchozí studijní texty, se v současné době na LDF rozvíjí i obory dřevařského a krajinného a nábytkového inženýrství.

Při vypracování tohoto studijního textu byl kladen důraz na podrobné vysvětlení používaných pojmů, přičemž na nezbytné minimum bylo omezeno použití matematického aparátu, důkazů tvrzení apod. (v důležitých případech jsou v textu odkazy na další rozšiřující literaturu, kde čtenář může potřebné důkazy a teoretickou podstatu metod najít). Hlavní těžiště výkladu spočívá ve vysvětlení účelu, základní podstaty a užití jednotlivých metod, včetně interpretace výsledků. Výběr metod je volen tak, aby obsáhly všeobecné základy statistiky, které jsou použitelné v mnoha oborech, nejen v těch, pro které je primárně určen tento učební text.

Téměř všechny používané metody a postupy jsou dokumentovány na řešených příkladech, kterých studijní text obsahuje několik desítek. U každého příkladu je uvedeno zadání (vycházející z praktické úlohy), všechna data potřebná k vyřešení (včetně měřených dat) a podrobná interpretace výsledků.

Značná pozornost byla také věnována obrazovému a tabulkovému doprovodu textu. Skripta obsahují několik desítek obrázků a tabulek, které názorně vysvětlují zvláště ty úseky textu, které jsou náročnější na představivost (např. podstata chyb v testování statistických hypotéz).

Vzhledem ke značné šíři používaných statistických metod a předepsanému rozsahu studijního textu, jsme byli nuceni celou problematiku rozdělit do dvou dílů.

Tento, první díl, obsahuje základní statistické pojmy, statistické charakteristiky a základy matematické statistiky, včetně odhadů parametrů základního souboru a testování statistických hypotéz.

Autoři děkují paní D. Harazimové za přepis podstatné části textu.

2 Základní pojmy

2.1 Pojem statistiky

Vymezení pojmu „statistika“ není tak jednoznačné a jednoduché, jak by se na první pohled mohlo zdát a jak by se „slušelo“ na obor, se kterým je obvykle spojována blízkost matematických metod, využití vzorců, výpočetní techniky a přesného, jednoznačného vyjadřování. Vždyť v kolika různých souvislostech můžeme zaslechnout nebo číst slovo statistika (a jeho odvozeniny):

(1) „Vyplňte, prosím, tento statistický výkaz o ….“

(2) „Na základě statistického šetření na reprezentativním souboru respondentů je možné konstatovat, že …“

(3) „Český statistický úřad odhaduje, že HDP v příštím roce vzroste o …“

(4) „ Na základě poznatků matematické statistiky se velikost náhodného výběru stanoví podle vzorce…“

Na těchto několika příkladech je možné si uvědomit, že v běžné komunikační praxi tímto slovem označujeme nejrůznější jevy a činnosti. Jednou se tak nazývá vyplněný statistický výkaz nebo dotazník (1), jindy organizace nebo realizace statistických zjišťování (2), v jiných případech organizace, které jsou pověřeny statistickou činností (3) a konečně je tak nazývána soustava poznatků o statistických metodách, tj. vědní obor (4).

Ovšem pokud se na výše uvedenými výroky zamyslíme, je zřejmé, že něco mají společné – všechny se zabývají zkoumáním jevů, které jsou výsledkem různých vlivů a příčin, a které je možné hodnotit jen v souborech, kde znalost izolovaných jednotlivých výskytů nebo hodnot jevu nám o zkoumané skutečnosti mnoho neřekne (např. je těžké hodnotit názor veřejnosti na určitou otázku, pokud se zeptáme jen jednoho člověka nebo ČSÚ těžko může odhadovat vývoj HDP na základě znalosti údajů jen o jednom podniku apod.). Tyto jevy se nazývají **jevy hromadné**.

Hromadné jevy tedy můžeme charakterizovat jako jevy (např. předměty, události, procesy apod.), které

- zpravidla vznikají jako výsledek působení mnoha náhodných i nenáhodných vlivů,
- jejichž podstatné vlastnosti se neprojevují v jednotlivých výskytech těchto jevů, ale pouze ve větším množství – v souborech.

Pokud za jev označíme např. „průměrnou výšku porostu“ (tj. výšku, která s nejmenší chybou reprezentuje všechny výšky prostu), pak tuto veličinu nemůžeme správně určit, pokud známe údaje o výšce jen u jednoho stromu. Musíme znát tyto údaje buď o všech stromech porostu nebo alespoň o takovém počtu, který je podle statistických zásad dostatečný k tomu, aby bylo možné požadovaný jev s potřebnou přesností odhadnout. Výška stromu patří mezi tzv. náhodné veličiny, kdy na výslednou podobu jevu (číselně vyjádřenou výšku v m) má vliv mnoho faktorů (např. dřevina, věk, bonita, charakter stanoviště, klimatické vlivy, původ porostu, genetické dispozice jednotlivých stromů, způsob výchovy, apod., a také způsob shromáždění dat a výpočtu). Podobně na klíčivost semene určitého druhu stromu není možné usu-

zovat na základě vyklíčení nebo nevyklíčení jednoho semene, ale na základě pokusů, kdy hodnotíme větší počet (např. 100) náhodně vybraných semen, a tyto statisticky vyhodnotíme.

Z výše uvedených tvrzení je možné učinit závěr, že statistika je vědní obor zkoumající jevy,

- které mají hromadný charakter (např. měření výšky stromu, vlhkosti dřeva apod.),
- u kterých existuje možnost opakovat soubor podmínek, za nichž může uvažovaný jev nastat (např. opakované měření výšky jednoho stromu nebo vlhkosti dřeva jednoho vzorku stejnou metodou apod.). Metody zpracování výsledků tohoto typu měření studuje tzv. teorie chyb.

Smyslem statistiky je tedy zobecňovat vlastnosti jevů a vztahy mezi nimi na základě studia velkého počtu údajů o těchto jevech – hromadných jevů.

Je nutné zdůraznit, že statistika se zabývá především proměnnými (variabilními) jevy, tj. takovými jejichž vlastnosti (např. hodnoty) se mění s každým novým nastoupením daného jevu. Tak např. těžko můžeme statisticky vyhodnotit fakt, že v porostu 123 B3 roste strom (to je stálá, neměnná vlastnost, pokud ten strom není vytěžen nebo se nezmění označení porostu), ale můžeme statisticky vyhodnocovat jeho proměnné vlastnosti – např. výšku, výčetní tloušťku, délku zelené koruny, objem apod. a soubor údajů o všech stromech v porostu – hromadný jev – nám umožní statisticky charakterizovat celý porost (např. pomocí střední tloušťky, střední výšky apod.).

2.2 Popisná a matematická statistika, statistický soubor

Z hlediska charakteru studované množiny hromadných jevů - statistického souboru - je možné statistiku rozdělit na:

- **popisnou** statistiku – popisuje stav jevů nebo jejich vývoj na základním souboru
- **matematickou** statistiku – jevy na vyšetřovaných souborech zkoumá prostřednictvím výběrů

Popisná statistika popisuje stav nebo vývoj jevů na **základním statistickém souboru**. Vymezí se soubor jednotek, na nichž se jev zkoumá, a všechny jednotky se vyšetří z hlediska studovaného jevu, čímž se získá úplná statistická informace o studovaném jevu v příslušném souboru.

Matematická statistika vznikla propojením poznatků popisné statistiky a matematiky (především teorie pravděpodobnosti). Používá se tehdy, jestliže je základní soubor buď neznámého (nebo nekonečně velkého) rozsahu nebo jestliže je tak velký, že není technicky a/nebo ekonomicky možné všechny jeho jednotky vyšetřit z hlediska studovaného jevu. Potom se podle stanovených zásad vymezí určitá podmnožina jednotek – **výběrový soubor** – na kterém se vyšetření zkoumaného jevu provede. Na základě takto získaných údajů se potom s určitou pravděpodobností odhadují příslušné vlastnosti základního souboru. Prvek pravděpodobnosti se do matematické statistiky dostal právě implementací teorie pravděpodobnosti.

Vymezení základního a výběrového souboru vychází z cíle statistické analýzy. Jestliže je naším cílem statisticky charakterizovat jeden porost, potom je základním statistickým souborem množina všech stromů daného porostu a výběrovým souborem by byla zkusná plocha založená uvnitř porostu. Pokud nás ovšem zajímá tatáž veličina (např. tloušťka nebo výška stromu) pro lesní oblast nebo dokonce pro celé území státu (např. pro konstrukci oblastních nebo všeobecných růstových tabulek), potom uvažovaný porost se může stát "zkusnou plochou", tj. výběrovým souborem, zatímco základní soubor by byla množina všech stromů dané dřeviny pro studovanou oblast nebo stát. V případě jednoho porostu se můžeme rozhodnout (podle cíle a požadované přesnosti analýzy), zda využijeme analýzy základního

nebo výběrového souboru, v druhém případě je volba jasná – v žádném případě není technicky a ekonomicky možné (a v podstatě ani potřebné, neboť metody matematické statistiky dávají při správném použití spolehlivé výsledky) změřit všechny stromy v rámci rozsáhlé oblasti nebo dokonce státu – zde musíme použít výběrový soubor.

Na tomto příkladu je možné také vysvětlit jiný způsob dělení statistických souborů: na **soubory přirozené a umělé**. Jednotlivý smrkoý porost je možné považovat soubor přirozený, protože všechny jednotky tohoto souboru – stromy – patří přirozeně k sobě, navzájem se reálně ovlivňují (např. konkurenčními vztahy) a všechny společně závisí na růstových podmínkách daného stanoviště.

Naproti tomu soubor smrků z různých oblastí státu použitý jako podklad pro tvorbu růstových tabulek je umělý, protože tyto stromy nemají spolu žádný reálný vztah a byly sloučeny do jednoho souboru pouze účelově.

2.3 Statistická jednotka, statistický znak

V předchozí kapitole jsme se seznámili s pojmem statistického souboru – tj. s množinou objektů, jejichž určitá vlastnost je předmětem statistického šetření. Zkoumané vlastnosti se však nemohou zjišťovat u všech jevů najednou, ale je nutné danou množinu – soubor – rozdělit na základní prvky, na nichž se konkrétní zkoumaná vlastnost zjišťuje. Jednotlivé prvky této množiny se nazývají **statistické jednotky** (prvky statistického souboru). Tyto jednotky musí být jednoznačně vymezeny už před začátkem statistického šetření, a to ze tří hledisek:

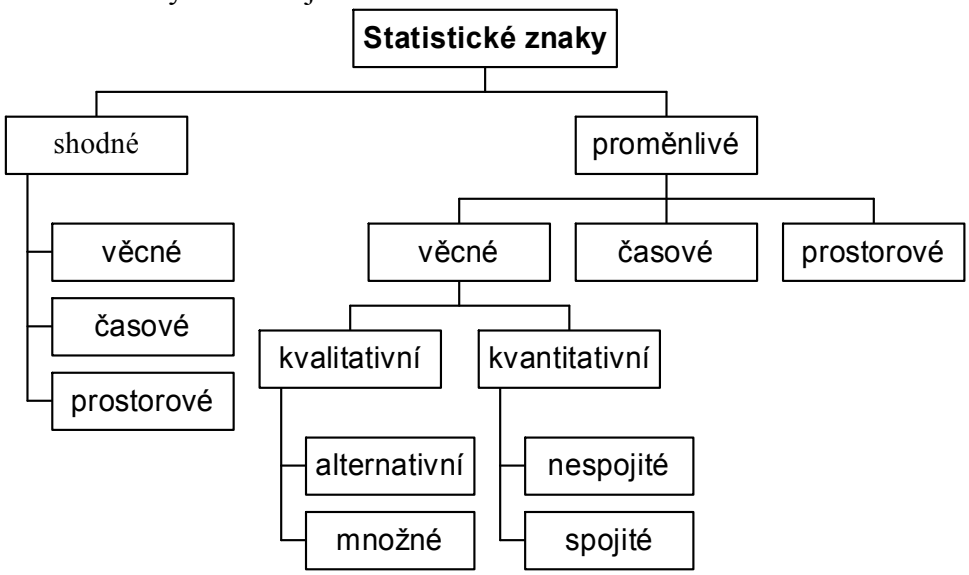
- věcného – vymezení, co se za statistickou jednotku bude považovat (strom, vzorek dřeva, budova, …),
- časového – vymezení, kdy bude statistická jednotka zařazena do zkoumání (přesnost záleží na typu zkoumané veličiny, např. pro zásobu porostu stačí určit rok, neboť se mění velmi pomalu, naopak při měření teploty vzduchu je nutné zaznamenat čas s přesností až na hodiny),
- prostorového – vymezení území, na které se bude statistické zkoumání vztahovat.

Statistické jednotky tedy vymezujeme ze tří hledisek – CO (věcně), KDY (časově), KDE (prostorově). Příkladem může být:

- dřevinu smrk (CO) v porostu 145 D8 (KDE) v roce 1998 (KDY)
- měření teploty vzduchu (CO) na měřící stanici Hůrka (KDE) 25.6.1999 v 14,00 hod. (KDY)
- student(ka) II. ročníku (CO) FLD MZLU v Brně (KDE) ve školním roce 1997/98 (KDY)

Každá šetřená vlastnost statistické jednotky je při statistickém zkoumání charakterizována **statistickým znakem**. Pokud hodnotu statistického znaku vyjádříme konkrétní hodnotou, nazýváme tuto hodnotu statistickou veličinou. V mnoha případech je určení vhodného statistického znaku jednoznačné (např. výška v m a výčetní tloušťka v cm pro vyjádření základních taxačních veličin stromu pro výpočet objemu pomocí objemových tabulek), jindy je volba vhodných znaků obtížnější, neboť dotyčnou vlastnost je možné popsat různými znaky. Např. velikost podniku může být popsána obratem, počtem zaměstnanců, u lesních nebo zemědělských podniků výměrou obhospodařované půdy, velikostí majetku, apod. Podobně kvalita stanoviště může být charakterizována absolutní nebo relativní bonitou, slovním popisem nebo jiným vyjádřením půdních, orografických a dalších vlastností apod. V těchto případech musíme vhodné znaky volit podle účelu analýzy.

Statistickými jednotkami určitého souboru mohou být jen takové jevy a procesy, které mají na jedné straně vlastnosti shodné, na druhé straně se jinými vlastnostmi od sebe liší. Základní dělení statistických znaků je na obr.2.1 .



Obrázek 2.1 - Druhy statistických znaků

Statistické znaky shodné slouží pro vymezení statistických jednotek (a tím pro vymezení statistického souboru).

Pro vlastní statistickou analýzu jsou rozhodné jiné vlastnosti statistických jednotek (obměny nebo varianty znaku), které se od jednotky k jednotce liší. Nazývají se znaky variabilní (proměnlivé) a jsou vlastním předmětem statistického zkoumání. Např. pro již uvedený příklad statistické jednotky: dřevina smrk v porostu 145 D8 v roce 1998 (toto jsou statistické znaky shodné, vymezují statistickou jednotku) můžeme podle cíle analýzy definovat různé proměnlivé znaky, které se mohou měnit pro každou – shodnými znaky definovanou - statistickou jednotku souboru – např. výšku stromu v m, výčetní tloušťku v cm, apod.

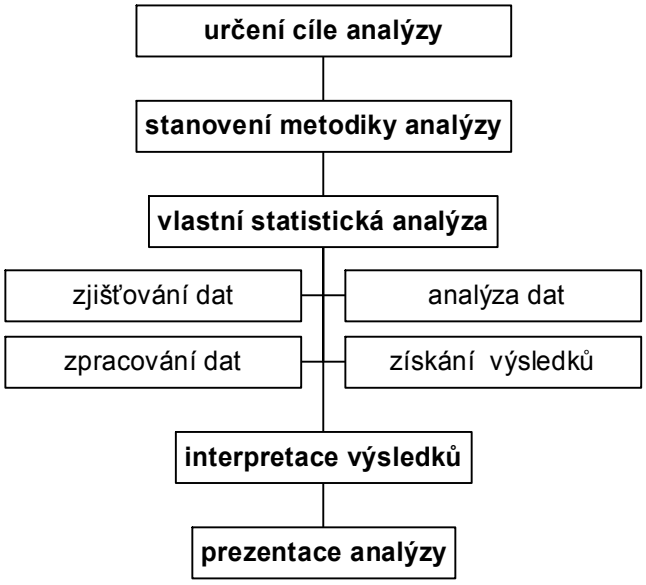
Nejpoužívanější variabilní znaky jsou znaky věcné. Dělíme je do dvou skupin:

- znaky kvantitativní (též znaky číselné) – charakterizují určité množství nebo velikost variabilního znaku, přičemž hodnota tohoto znaku se získá zpravidla měřením (nebo odvozením z přímo měřených nebo jinak zjišťovaných veličin). Příkladem mohou být taxační veličiny stromů a porostů (výška, tloušťka, kruhová výčetní základna jako přímo měřené, zásoba porostu jako odvozená – získaná výpočtem z přímo měřených veličin). Kvantitativní znaky se dělí na dvě skupiny:
  - znaky nespojité (diskrétní) – znaky vyjádřitelné jen celými čísly (např. počet stromů/ha, počet zaměstnanců v podniku, počet mechanizačních prostředků apod.),
  - znaky spojité – mohou nabývat v daném intervalu libovolných hodnot (např. taxační veličiny stromů a porostů, rychlost automobilu apod.)
- znaky kvalitativní (též znaky slovní) – užívají se pro zachycení takových variabilních vlastností statistického souboru, které jsou neměřitelné, na statistických jednotkách se buď vyskytují nebo nevyskytují a musí se vyjádřit slovním popisem. Rozdělují se také do dvou skupin:
  - znaky alternativní – vyskytují se jen ve dvou obměnách (ano/ne, žena/muž, přítomnost škůdce/nepřítomnost škůdce apod.),

- znaky množné – vyskytují se ve více obměnách (např. dřevina, národnost, barva květů, lesní typ, apod.).

Prostorové a časové variabilní znaky jsou méně časté (tyto typy se více využívají pro vymezování statistických jednotek. Používají se tehdy, jestliže u statistických jednotek se mění místo nebo čas jejich vzniku nebo existence. Např. pokud statistickou jednotku definujeme jako smrk v lesní oblasti Drahanská vrchovina, jedním z variabilních znaků může být porost, ve kterém konkrétní měřený strom roste. Podobně je tomu s časovými údaji. Časové a prostorové znaky se často používají ke třídění statistických souborů.

2.4 Základní etapy statistické analýzy



Obrázek 2.3-Základní etapy statistické analýzy

Základním úkolem statistické analýzy dat je získat potřebné informace ze zpracovávaného souboru dat. Aby tohoto cíle bylo co nejefektivněji dosaženo, je nutné celý postup statistické analýzy pečlivě promyslet a naplánovat – rozčlenit do jednotlivých etap, které na sebe logicky navazují.

Základní etapy statistické analýzy ukazuje obr. 2.3.

První, zcela základní a nepominutelnou etapou, je určení **cíle analýzy**. Vždy musíme dopředu vědět, k čemu analýza bude sloužit, na jaké otázky má odpovědět. Zásadně nesprávný je „postup“ opačný, kdy nejprve „něco“ spočítáme a pak uvažujeme, k čemu by to mohlo být dobré, jak by se dala získaná čísla a grafy interpretovat a co vlastně znamenají nebo používat statistickou analýzu pouze k vyšperkování díla, „aby to lépe – ‘vědečtěji‘ - vypadalo“. Určení cíle vychází z dobré znalosti řešeného problému a cíl by měl určovat odborník v daném oboru. Pokud zároveň zná i statistické metody, může ihned určit, zdali je daný problém řešitelný pomocí statistické analýzy a jakými konkrétními postupy. Pokud tyto znalosti nemá nebo je problém složitější, je nutná konzultace se specialistou ve statistice.Tento postup může v budoucnu ušetřit mnohá zklamání a také zbytečně vynaložené prostředky na zjišťování a analýzu dat.

Na základě stanoveného cíle je možné vybrat odpovídající statistické metody a stanovit **metodiku** celé analýzy. Ta se zpravidla skládá z těchto etap:

- zjišťování dat – samotný sběr prvotních dat podle vypracovaného plánu – měření, anketa apod.,
- zpracování dat – příprava prvotních dat do podoby vhodné pro počítačové zpracování – uspořádání, třídění, vytváření tabulek dat vhodných pro analýzu apod.
- analýza dat – vlastní číselné a grafické zpracování statistické analýzy
- stanovení potřebných výstupů a výsledků

Stanovená metodika by měla směřovat k získání co největšího počtu využitelných informací s vynaložením minimálních nákladů a objemu práce. Z tohoto hlediska se téměř vždy vyplatí konzultace se statistikem.

Poté proběhne **vlastní statistická analýza**, dnes prováděná téměř výhradně na počítačích s použitím vhodných statistických programů. Je nutné důrazně upozornit na to, že není možné celou statistickou analýzu zužovat pouze na tuto etapu, i když se tak často, bohužel, děje – ve stylu „nevím, co s tím, tak to proženu počítačem“ nebo „výsledky z počítače vypadají hned líp“. Bez předcházejících a navazujících etap nemají samotné výpočty, grafy a jiné výstupy valnou cenu.

Etapa samotných výpočtů musí být následována etapou **interpretace výsledků**. Ta by měla odpovědět především na otázky kladené v zadání (cíli) analýzy, popřípadě rozebrat veškeré okolnosti, které mohou interpretaci ovlivnit. Zvláště v případě, že závěry neodpovídají předpokladům, je nutný podrobný rozbor příčin, proč tomu tak je. Je nutno zdůraznit, že kvalitu celé analýzy vytváří z velké části tato etapa analýzy, neboť teprve zde se odpovídá na otázky kladené zadáním. Předchozí etapy jsou sice důležitou, nezbytnou, ale v podstatě technickou částí.

Závěrečnou částí bývá prezentace analýzy, zpravidla zadavateli, ať již je formě předložené zprávy nebo živou prezentací spojenou s výkladem.

2.5 Statistické vyjadřovací prostředky

Vzhledem k tomu, že se statistika zabývá zpracováním hromadných dat, tj. mnohdy velmi rozsáhlých datových souborů, musela si vypracovat specifické vyjadřovací prostředky, zaměřené především na co nejnázornější a nejvýstižnější vyjádření vlastností dat a jejich vztahů. Je zřejmé, že slovní popis datových souborů není vždy dostatečně názorný a výstižný, stejně jako tabulky hodnot prvotních dat (tj. dat v tom pořadí a podobě, jak byly získávány měřením nebo jiným zjišťováním). Proto nejdůležitějšími statistickými vyjadřovacími prostředky se staly **tabulky** a **grafy**, zpravidla speciálně uzpůsobené pro vyjádření potřebné statistické informace. Mezi hlavní výhody statistických tabulek a grafů patří:

- názornost a přehlednost zobrazení,
- možnost rychlého vystižení hlavních vlastností, vztahů, trendů apod.,
- odpadá nutnost zdlouhavého (a mnohdy nevýstižného) popisování jevů (tím se ovšem nezbavujeme povinnosti údaje tabulek a grafů interpretovat, uvádět do souvislostí, upozorňovat na příčiny, které vedly k takovému charakteru jevů, jaké vidíme v údajích tabulky).
- v současné době je tvorba tabulek a grafů velmi snadná a rychlá pomocí grafických a textových editorů nebo tabulkových kalkulátorů.

Velmi vhodná je kombinace tabulky a grafu (v grafu zpravidla neuvádíme kvůli přehlednosti přesné hodnoty jednotlivých veličin, je tedy vhodné doplnění grafu vhodně konstruovanou tabulkou s hodnotami).

Mezi společné zásady, platné pro tvorbu jak tabulek, tak i grafů, patří:

- základním požadavkem je vždy obsahová a věcná správnost, s určitostí je nutné vyloučit chyby vzniklé přehlédnutím, nepřesností výpočtů, nedostatkem kontroly,
- výstižný a stručný popis (měl by shrnovat to, o čem příslušná tabulka nebo graf má vypovídat), mělo by být dosaženo takové srozumitelnosti, aby nebylo nutno k výkladu obsahu užít zvláštního vysvětlení (kromě zcela speciálních případů, např. u grafů a tabulek obsahujících pojmy, o kterých se předpokládá, že je uživatel nemůže znát, např. v učebních textech),

- přehlednost - „přeplácanost“, jak obsahová, tak i grafická, tabulkám i grafům škodí, je nutné si vždy rozmyslet, popřípadě i prakticky vyzkoušet, zdali množství údajů není nadbytečné a zdali použité grafické prvky přispívají k přehlednosti tabulky nebo grafu a nejsou pouze samoúčelnou demonstrací grafických možností příslušného programu a (ne)vkusu zhotovitele,
- pokud to není zřejmé z doprovodného textu, je nutné uvést zdroj údajů, odkud bylo čerpáno,
- vždy je nutné uvádět jednotky měřených veličin (jak v popisech tabulky, tak i na osách a v legendě grafu) a hodnoty uvádět s přesností odpovídající měřené veličině (je např. nesmyslné uvádět zásoby porostů, které jsou zjišťovány s přesností ± 10 %, tj. s možným rozpětím hodnot i několika desítek m^3/ha, na několik desetinných míst nebo jestliže uvádíme experimentální třídní četnosti na celá čísla, uvádět modelové třídní četnosti např. na 5 desetinných míst, apod.),

V případě tabulky je nutné upozornit ještě na další zásady:

- hlavička (záhlaví) vyjadřuje obsah sloupců, legenda obsah řádků, text hlavičky i legendy musí být stručný, jasný a výstižný.
- každé políčko tabulky musí být vyplněné – jestliže příslušný údaj chybí, potom je nutné použít smluvenou značku:
  - jestliže údaj není, vyplní se políčko vodorovnou čárkou (-),
  - je-li hodnota malá tak, že nedosahuje hodnoty zvolené přesnosti, užije se nuly s odpovídajícím počtem desetinných míst (např. jestliže měříme s přesností na desetinu jednotky, pak zápis 0,0 znamená, že příslušné hodnoty jsou menší než 0,05 jednotek),
  - v případě, že není hodnota tabulkového znaku známá, užívá se tečky (.),
  - jestliže by zápis číslem odporoval logice věci, zapíše se do políčka tabulky ležatý křížek (x) – např. tak označíme políčko v řádku označeném „součet“, jestliže by součet pro daný sloupec neměl logický smysl
- je vhodné využít různých typů čar a jejich síly ke grafické hierarchizaci údajů tabulky – mělo by být patné, jaké údaje spolu souvisí, které údaje jsou na stejné úrovni, apod. - nebojte se využít této možnosti, ale pozor na výslednou přehlednost a nebezpečí „přeplácanosti“.
- v případě hodně rozsáhlých tabulek je vhodné sloupce a řádky číslovat.

V případě grafů je nutné dodržovat následující pravidla:

- grafické zobrazení vztahu mezi dvěma proměnnými se běžně děje v rovnoběžkové souřadnicové soustavě,
- nejobvyklejší je soustava pravoúhlá se stupnicemi umístěnými na osách:
  - vodorovná osa je nositelkou stupnice pro veličinu nezávisle proměnnou,
  - svislá osa je určena pro veličinu závisle proměnnou,
- stupnice pro čas při časových závislostech vynášíme na vodorovnou osu,
- volba stupnice je závislá na rozpětí dat, která se mají zobrazovat, její délka je určena požadovanou velikostí grafu,
- hodnoty zobrazovaných veličin, které se připisují k bodům stupnice, se nazývají kóty - okrouhlá čísla rovnoměrně rozmístěná na stupnici podle hodnot vhodné funkce (dostaneme tzv. funkční stupnice), mají přispívat k přehlednosti grafu a mají umožnit snadné zjištění hledaných hodnot,
- vždy se kótou označuje nulová hodnota (pokud ji graf obsahuje),

- vzdálenost mezi dvěma celými sousedními hodnotami vynášené veličiny se nazývá modul stupnice, moduly počítáme z rovnic, které matematicky vystihují reálnou situaci poměru délky intervalu funkčních hodnot vynášené veličiny,
- zvláště při počítačové tvorbě grafů musíme dbát na to, aby i při černobílém tisku bylo možné rozlišit všechny proměnné – tj. čáry musí rozlišeny svým stylem (plné, čárkované, čerchované apod.), popř. tloušťkou, není tedy možné se spoléhat na automatické návrhy grafů tabulkových procesorů nebo grafických programů, které většinou „navrhují“ stejné čáry odlišené pouze barvou, nikoli typem (počítají s tiskem na barevných tiskárnách). Na tyto zásady musíme pamatovat i v případě, že máme k dispozici barevnou tiskárnu, ale víme, že zpráva bude dále kopírována černobíle.

2.6 Statistický software

Jak již bylo řečeno, statistika se zabývá zpracováním hromadných dat. Často tedy musíme při statistické analýze zpracovat velké objemy číselných dat nebo tato data získávat z rozsáhlých databází. Proto je statistika jedním z oborů, které jsou „předurčeny“ pro využití počítačových programů a skutečně použití výpočetní techniky má ve statistice dlouhou tradici..

Zásadním problémem první fáze využití statistického softwaru byla „odtrženost“ experimentátorů od výpočetní techniky – experimentátor (odborník v určitém oboru) zjistil (buď sám nebo po konzultaci se statistikem), že pro vyřešení daného úkolu potřebuje využít výpočetně náročných statistických postupů. Zpravidla musel sehnat programátora, který mu napsal (většinou jednoúčelový) program, operátoři ve výpočetním centru programem a daty „nakrmili“ sálový počítač a ten vydal výsledky. Pokud se ukázalo, že výsledky nejsou uspokojující, musel se přepracovat nebo upravit program a celý koloběh experimentátor → programátor → operátor počítače → experimentátor se opakoval. Je jasné, že komunikace byla pomalá a mnohdy neefektivní, výzkumné úkoly se tím opožďovaly a prodražovaly (strojový čas sálových počítačů byl velmi drahý).

Prvním krokem ke zpřístupnění statistických výpočtů bylo vytváření knihoven funkcí a metod, a to jak pro matematiku, tak i pro statistiku. To odstranilo nutnost vytvářet stále znovu jednoúčelové programy. V 60. letech vznikla v USA (Houston) knihovna statistických a matematických programů IMSL (International Mathematical and Statistical Library) a v Oxfordu NAG (Numerical Algorithmus Group), což byly původně soustavy programů určených pro sálové počítače a psané ve strojovém kódu nebo v programovacích jazycích určených pro vědeckotechnické výpočty (FORTRAN). Tyto knihovny ve zmodernizované podobě existují dodnes a používají se především pro UNIX. Ovšem skutečný průlom v používání počítačově orientovaných statistických metod znamenal až nástup osobních počítačů. Především proto, že rozšíření osobních počítačů umožňuje interaktivní práci se statistickými programy, tj. příslušný odborník může používat širokou paletu statistických metod, vidí okamžité numerické i grafické výsledky, může na ně bezprostředně reagovat (např. měnit metodiku, určovat rozsah a druh výstupů apod.). Znamená to ovšem, že kromě své odbornosti musí zvládnout alespoň základní statistické metody a ovládání příslušných programů.

Cílem této kapitoly není popis jednotlivých statistických programů – je jich celá řada, většina je v procesu neustálého vývoje a popis jednotlivých verzí by byl pravděpodobně zastaralý již v okamžiku vydání tohoto textu. Chtěli bychom zde upozornit pouze na základní

principy práce se statistickými programy a na některá úskalí, které jejich používání může přinést.

Programy, které se používají ve statistice, můžeme rozdělit na několik skupin:

- vlastní statistické programy,
- programy primárně určené pro jiné účely, ale využitelné pro statistickou analýzu,
- statistické a matematické jazyky.

Vlastní statistické programy jsou určeny primárně pro statistické výpočty, čemuž je uzpůsoben vstup dat, jejich zpracování i výstup.

Vstup dat je řešen zpravidla formou tabulkového procesoru, tj. tabulky, kam je možné zadat data manuálně nebo je nahrát ze souboru, z Internetu nebo – v případě vyspělejších programů – získat dotazem z databáze. Většina programů také umožňuje import dat z různých jiných rozšířených formátů. V tabulce je obvykle možné i provádět potřebné úpravy dat (např. transformace, třídění, uspořádání podle velikosti apod.), aby byly optimálně připraveny pro následnou analýzu.

Rozsah funkcí je určen posláním a rozsahem programu. V zásadě je možné statistické programy rozdělit na všeobecné a specializované.

Všeobecné programy zahrnují celou škálu statistických metod, které jsou obecně využitelné (např. výpočet základních charakteristik, práci se statistickými rozděleními, testování, odhady parametrů základního souboru, ANOVA, regrese a korelace apod.), i když v různém rozsahu a propracovanosti jednotlivých metod. Patří mezi ně nejpoužívanější a nejznámější statistické „balíky“, např. SPSS, SAS, STATISTICA, S PLUS a další.

Specializované programy se zaměřují na různé statistické techniky, které zpravidla nejsou do hloubky propracovány ve všeobecných programech (např. výpočet síly testů, plánování pokusů, analýza kvality apod.). Tyto programy se vyznačují menším rozsahem funkcí, ale na druhé straně ty funkce, na které se program specializuje, jsou mimořádně podrobně rozpracovány. Tyto programy jsou funkční samostatně (mají tedy svůj vstup, analýzu i výstup), ale mnohé z nich jsou schopny spolupracovat s některým všeobecným balíkem, zvláště jestliže pocházejí z dílny jednoho výrobce. Např. velmi rozsáhlou skupinu specializovaných programů má firma SPSS, přičemž tyto přídavné programy je možné spouštět a ovládat přímo ze základního modulu SPSS.

Mnohé programy mají modulární strukturu, to znamená, že „povinný“ je pouze základní modul (obsahující zpravidla nejpoužívanější funkce) a je možné k němu dokoupit další moduly (které ovšem samostatně nefungují) s dalšími potřebnými funkcemi.

Důležitou vlastností programů je jejich otevřenost. Tím rozumíme možnost úpravy připravených metod, popř. možnost jejich kombinace a přidávání nových funkcí. Některé programy toto umožňují pomocí makrojazyka nebo dokonce specializovaného matematického nebo statistického jazyka (např. S PLUS, který je založen na jazyku S nebo open source statistické programovací prostředí R). Uzavřené programy mají pevně definovanou škálu metod, ze kterých můžeme vybírat, v nejlepším případě umožňují jejich dávkové zpracování.

Všechny programy jsou vybaveny rozsáhlými možnostmi výstupů, zvláště grafických, a v současné době prakticky všechny v tomto směru využívají grafických možností Windows. Samozřejmostí je také možnost exportu výsledků do jiných rozšířených programů.

Další skupinou jsou programy, které jsou určeny pro jiné účely než je statistika, ale jsou v případě potřeby pro tento účel využitelné. Typickým příkladem jsou tabulkové procesory – spreadsheety (např. EXCEL, QUATTRO PRO a další), což jsou primárně ekonomické programy, ale vzhledem ke svým výpočetním a grafickým možnostem jsou pro jednodušší statistickou analýzu použitelné. V současné době také disponují speciálními statistickými moduly, které rutinně zvládají jednoduché statistické analýzy (např. u Excelu je to 1 – a 2 – faktorová

ANOVA, popisná statistika, třídění souboru, základní testy, korelační a regresní analýza, generování náhodných čísel několika základních rozdělení apod.). Kromě toho jsou zpravidla vybaveny množstvím dalších statistických a matematických funkcí, jejichž vhodnou kombinací je možné dosáhnout překvapivě kvalitních výsledků (je to ovšem podmíněno znalostí podstaty používaných statistických metod). Další výhodou tabulkových procesorů je rozsáhlá možnost konfigurace a vývoje nových aplikací v makrojazyku (např. VBA v Excelu). Na druhé straně je nutné počítat s omezenými statistickými možnostmi tabulkových kalkulátorů a nechtít po nich „nemožné“, pro účely skutečně hloubkové a seriózní analýzy je nezbytně nutné použít profesionální statistický program.

V této souvislosti je vhodné zmínit zvláštní skupinu statistických programů, které rozšiřují možnosti tabulkových procesorů o rozsáhlé statistické možnosti. V současné době jsou nejznámější zástupci této kategorie UNISTAT a xlSTAT, které umí spolupracovat s Excelem. Oba mohou být nainstalovány do Excelu jako doplňky a rozšířit menu Excelu o širokou nabídku statistických funkcí (UNISTAT může fungovat i jako samostatný program). Hlavní výhodou tohoto přístupu je fakt, že data ukládaná v Excelu se přímo v Excelu zpracovávají a výsledky se také ukládají jako listy excelovského sešitu. Tyto programy tedy umožní zachování všech možností Excelu jako tabulkového procesoru a navíc ho doplní o statistické funkce.

Při výběru vhodného programu je nutné si odpovědět na několik otázek:

- jaké spektrum metod budu v obvyklých případech potřebovat a do jaké hloubky (je např. rozdíl mezi občasným spočítáním regresní rovnice přímky, což dnes umí každá „lepší“ kalkulačka a mezi komplexní regresní analýzou nelineárního modelu včetně regresní diagnostiky, analýzou vhodnosti modelu apod., což bývá problém i pro mnohý statistický program),
- jak často budu statistickou analýzu potřebovat (čím častěji, tím více potřebuji program, který co nejvíce zrychluje práci, který umožňuje automatizaci činnosti apod.),
- jaké mám finanční a hardwarové možnosti (statistické programy jsou velmi drahé – řádově desítky až stovky tisíc korun – a ty rozsáhlé, s vyspělými grafickými možnostmi, vyžadují příslušně „vyspělý“ počítač, zvláště pro zpracování rozsáhlých datových souborů).

Druhým krokem je shromáždění dostatečných informací o statistických programech. Pravděpodobně nejrozsáhlejším a nejaktuálnějším zdrojem informací je Internet, protože všechny významné softwarové firmy mají své webové stránky. Na nich je možné získat informace o rozsahu funkcí programu, kompatibilitě s jinými formáty (nutno brát s určitou rezervou), hardwarových nárocích (ty je zpravidla dobré „násobit“ nejméně dvěma pro bezproblémový chod programu v případě větších souborů a grafiky). Není dobré se ale na tyto informace „slepě“ spoléhat, protože pochopitelně každá firma vychvaluje svůj produkt jako nejlepší. Po získání základních informací je vhodné - pokud je to možné - se poradit s kolegy o jejich zkušenostech s těmito programy, popř. navštívit některé z webových diskusních fór věnovaných příslušným programům (jejich adresy najdete obvykle na webových stránkách příslušného programu). Vynikající možností k ocenění jednotlivých programů jsou demoverze (zpravidla omezené verze programů, které buď neumožňují některé základní funkce – např. ukládání výsledků nebo tisk, event. mají omezený vstup dat) nebo trial verze (obvykle plné verze programů s časovým omezením, nejobvyklejší doba je 1 měsíc). Tyto zkušební programy je možné si nechat poslat na disketě nebo CD (zpravidla zdarma nebo jen za manipulační poplatek) nebo si je stáhnout z Internetu. Nejlepší je, když zkušební verze umožňuje vyzkoušet program s vlastními daty (u některých metod je to zásadní – např. u nelineární regrese, kde různé algoritmy pracují různě úspěšně s různými rovnicemi i daty, a je jasné, že jako ukázkový bude u programu přidán takový soubor dat, kde program pracuje bezproblémově). Velmi

důležité také je, zda umí uvažovaný program bez problémů načíst data, která jsou již uložena v jiných formátech (např. v tabulce EXCELU, QUATTRA apod. – ruční přepisování rozsáhlých tabulek určitě není ničím příjemným, navíc je to „oblíbený“ zdroj hrubých chyb) a také vyexportovat výsledky do výstupní zprávy. Webové adresy jednotlivých programů se dají najít obvyklými vyhledávacími stroji Mezi další kritéria výběru je možné zařadit:

- podporu v ČR – důležitější programy mají zastoupení v ČR, což je výhodné jak z hlediska bezproblémové koupě za Kč, tak i z hlediska další podpory (školení, upgrady, rady ohledně používání programu apod.),
- jazykové hledisko – většina programů je v angličtině (v současné době jediné rozšířenější statistické programy v češtině jsou ADSTAT nebo jeho rozšířená Windows verze QC EXPERT (www.trilobyte.cz), částečně lokalizovanými programy (přeložené je uživatelské prostředí a výstupy, v angličtině zůstala nápověda) jsou UNISTAT (www.unistat.cz) a STATISTICA(www.statsoft.cz)).
- zda je možné k programu získat doplňkovou literaturu, především z hlediska statistické teorie používaných metod – některé firmy dodávají přímo k programu přehled používaných metod (nebo alespoň odkazy na literaturu) – protože používat „neznámé“ metody (kdy nic nevím o její teorii, podmínkách použití, interpretaci výsledků) nelze v žádném případě doporučit.

V souvislosti se skutečností, že statistické programy jsou dnes již běžně dostupné, je nutné uvést několik poznámek k jejich postavení v rámci celého procesu statistické analýzy.

Uživatel programu si musí být vždy vědom, že statistický program je pouze technický prostředek, který napomáhá k vyřešení určitého odborného problému. Pokud si připomeneme obr. 2.3, na kterém jsou uvedeny základní etapy statistické analýzy, tak statistické programy jsou použitelné v podstatě pouze v etapě zpracování dat a jejich analýzy. Vždy musí předcházet rozbor řešené úlohy (a v rámci stanovení metodiky též posouzení možností využití specializovaného softwaru) a následovat komplexní interpretace výsledků z hlediska řešeného problému. Nikdy nepoužívejme tyto programy samoúčelně – „aby to vypadalo vědecky…“, ale pouze s jasným cílem.

Další důležitá podmínka použití statistických programů již byla výše stručně zmíněna – tou je dobrá znalost používaných metod. Nikdy není možné použít statistickou metodu bez znalosti její teorie, bez znalosti interpretace jejích výsledků. Musíme si uvědomit, že program zpravidla vypíše číselné (popř. grafické) výsledky a tím jeho úloha končí. Pokud neznáme používanou metodu teoreticky, sedíme nad „hromadou“ čísel a čar a nevíme, co s nimi. Proto s koupí programu vždy je dobré si opatřit doporučenou literaturu, kde jsou jednotlivé metody popsány.

Pokud budeme dodržovat alespoň ty nejdůležitější zásady používání statistického softwaru, které zde byly uvedeny, stane se nám dobrým pomocníkem. Výpočetní technika a statistický software jsou již dnes nepostradatelným vybavením statistika. Je nutné si uvědomit, že právě rozšíření výpočetní techniky zpřístupnilo široké obci uživatelů výpočetně náročné a statisticky velmi účinné metody, které byly dříve nepoužitelné především pro „nepřekonatelný“ objem výpočtů.

3 Základní zpracování jednorozměrného statistického souboru

3.1 Prvotní zápis

Jednorozměrný základní statistický soubor je tvořen všemi uvažovanými hodnotami sledované statistické veličiny. Znamená to, že o zkoumané veličině máme k dispozici **úplnou statistickou informaci**. Počet hodnot značíme N a nazýváme **rozsah souboru**.

Prvotní zápis statistického souboru je neuspořádaná forma zápisu zjištěných hodnot. Např. při měření taxačních veličin stromů v porostu zapisujeme údaje podle toho, jak procházíme porostem a měříme jednotlivé stromy, při měření v laboratoři jsou zpravidla údaje zapisovány podle pořadí měřených vzorků, při anketách podle pořadí respondentů apod. Prvotní (tj. laboratorní, terénní) zápis není vhodný k přímému statistickému zpracování, protože informace jsou v něm nepřehledné. Hodnoty se obvykle přenášejí na různá média, která umožní snadnější a rychlejší zpracování zjištěných dat.

Nejčastěji se používá:

- kartotéka (různé formy) - každá jednotka souboru má vlastní kartotéční lístek. Vhodné pro ruční zpracování, v současnosti se používá jen velmi vzácně,
- disketa, disk (dříve děrný štítek, děrná páska) – vzhledem k tomu, že většina statistických analýz se provádí s pomocí výpočetní techniky, je to nejčastější způsob ukládání dat, zde má každá jednotka souboru svůj kód (označení, identifikaci).

Příklad prvotního zápisu je v tabulce 3.1 . Jedná se o zápis z měření na výzkumné ploše, kdy zápis byl prováděn při postupu od stromu ke stromu. Měřenou veličinou je výčetní tloušťka (tj. tloušťka ve výši 1,3 m nad patou kmene) v cm. Je zřejmé, že z takového zápisu nejsme na první pohled ani přibližně schopni odhadnout základní charakteristiky měřených stromů (např. nejsilnější strom, nejslabší strom, průměrnou tloušťku apod.), je zde zřetelná chaotičnost. Grafické znázornění prvotního zápisu je na obrázku 3.1 . Ihned vidíme, že grafické znázornění je přehlednější, je zde možné určit alespoň extrémní hodnoty a zhruba odhadnout určitý interval, ve kterém se může pohybovat střední hodnota. Ovšem pro podrobné a přesnější posouzení souboru je nutné ho určitým způsobem připravit. nejjednodušší a nejčastěji používané metody zpracování prvotních dat jsou:

- uspořádání souboru
- třídění souboru

3.2 Uspořádání souboru

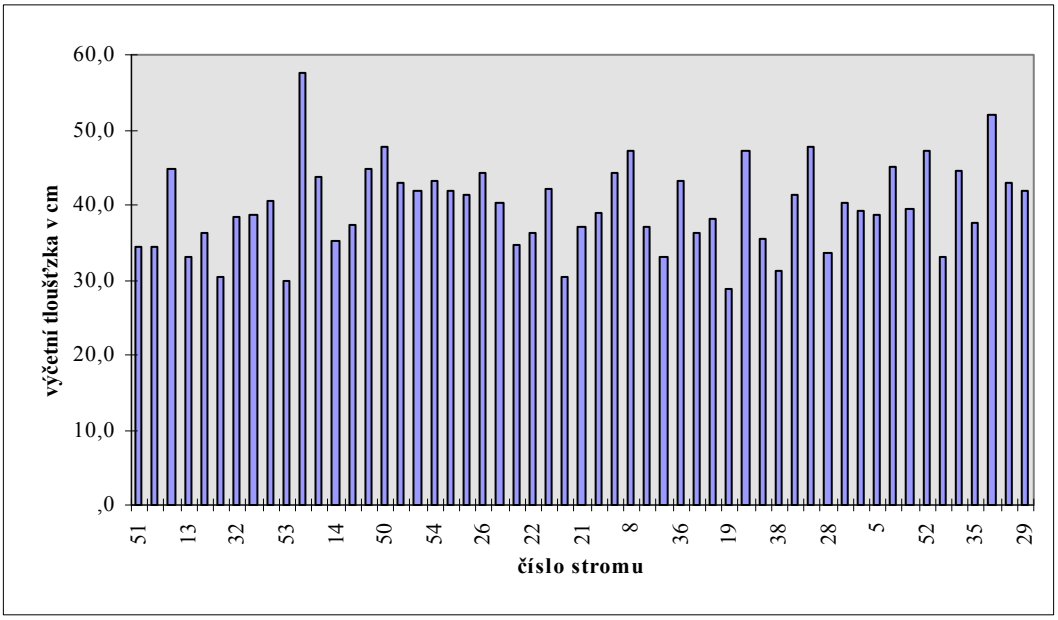
Je to zpravidla první (ale nikoliv nezbytně nutný) krok zpracování souboru. Hodnoty se seřadí podle velikosti vzestupně nebo sestupně. Takto uspořádanému souboru se říká **pořádková statistika**. Tím se zpřehlední soubor, zpřístupní informace a zjednoduší další zpracování souboru. Dále je toto uspořádání nezbytné např. pro práci s tzv. kvantily (podrobněji viz kapitolu 4). Grafické znázornění vzestupně uspořádaného souboru je na obrázku 3.2 . Z graficky znázorněného vzestupně uspořádaného souboru můžeme lehce zjistit např. tyto informace:

- minimální hodnota (x_{min}),
- maximální hodnota (x_{max}) souboru,

- počet hodnot souboru, které patří do libovolně určeného intervalu,
- interval, ve kterém leží všechny hodnoty souboru $< x_{min}, x_{max}>$ a jeho šířka v jednotkách veličiny, tzv. variační rozpětí, $R = x_{max} - x_{min}$.

Pořadové číslo měření	Číslo stromu	Výčetní tloušťka v cm	Pořadové číslo měření	Číslo stromu	Výčetní tloušťka v cm	Pořadové číslo měření	Číslo stromu	Výčetní tloušťka v cm
1	51	34,3	21	37	41,3	41	40	41,3
2	7	34,5	22	26	44,2	42	55	47,7
3	16	44,9	23	18	40,3	43	28	33,5
4	13	33,1	24	39	34,6	44	34	40,3
5	49	36,2	25	22	36,4	45	9	39,1
6	12	30,3	26	20	42,1	46	5	38,7
7	32	38,5	27	44	30,3	47	4	45,0
8	43	38,6	28	21	37,1	48	23	39,5
9	41	40,6	29	47	39,0	49	52	47,2
10	53	29,9	30	1	44,2	50	31	33,0
11	2	57,7	31	8	47,2	51	6	44,5
12	10	43,8	32	25	37,2	52	35	37,7
13	14	35,2	33	24	33,1	53	27	52,0
14	42	37,4	34	36	43,1	54	48	43,0
15	15	44,7	35	45	36,2	55	29	41,8
16	50	47,7	36	46	38,2			
17	11	42,9	37	19	28,8			
18	30	42,0	38	3	47,3			
19	54	43,3	39	33	35,5			
20	17	41,8	40	38	31,2			

Tabulka 3.1 - Prvotní zápis statistického souboru na základě terénního měření



Obrázek 3.1 - Grafické znázornění neuspořádaného statistického souboru (prvotního zápisu)

Na obr 3.3 jsou schematické grafy tří uspořádaných souborů:

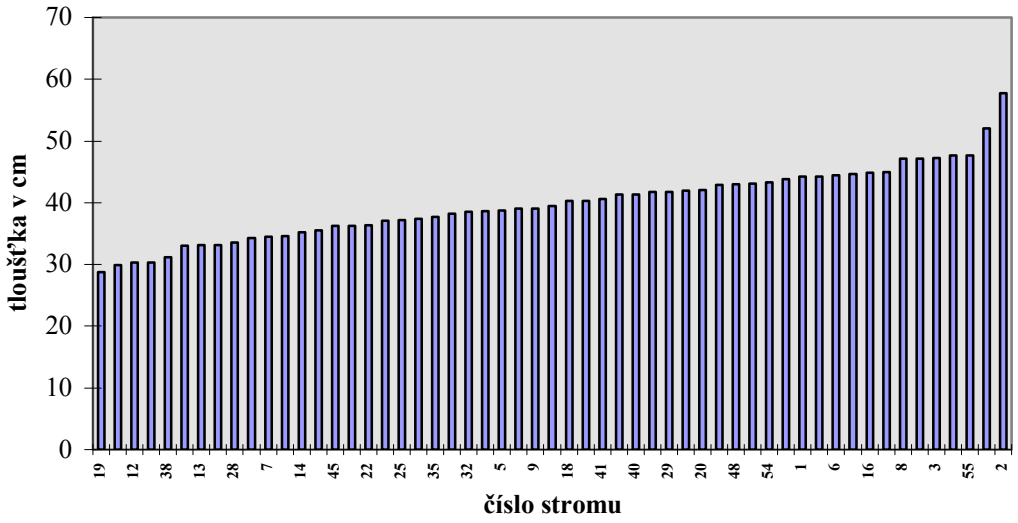
- I. soubor bez výrazné úrovně, tj. hodnoty jsou „každá jiná“,

- II. soubor s významnou úrovní tj. s velkým počtem stejných, resp. málo odlišných hodnot, mluvíme zde o lokální koncentraci dat,
- III. soubor se dvěma výraznými úrovněmi. Může to znamenat, že soubor je složen ze dvou samostatných souborů.

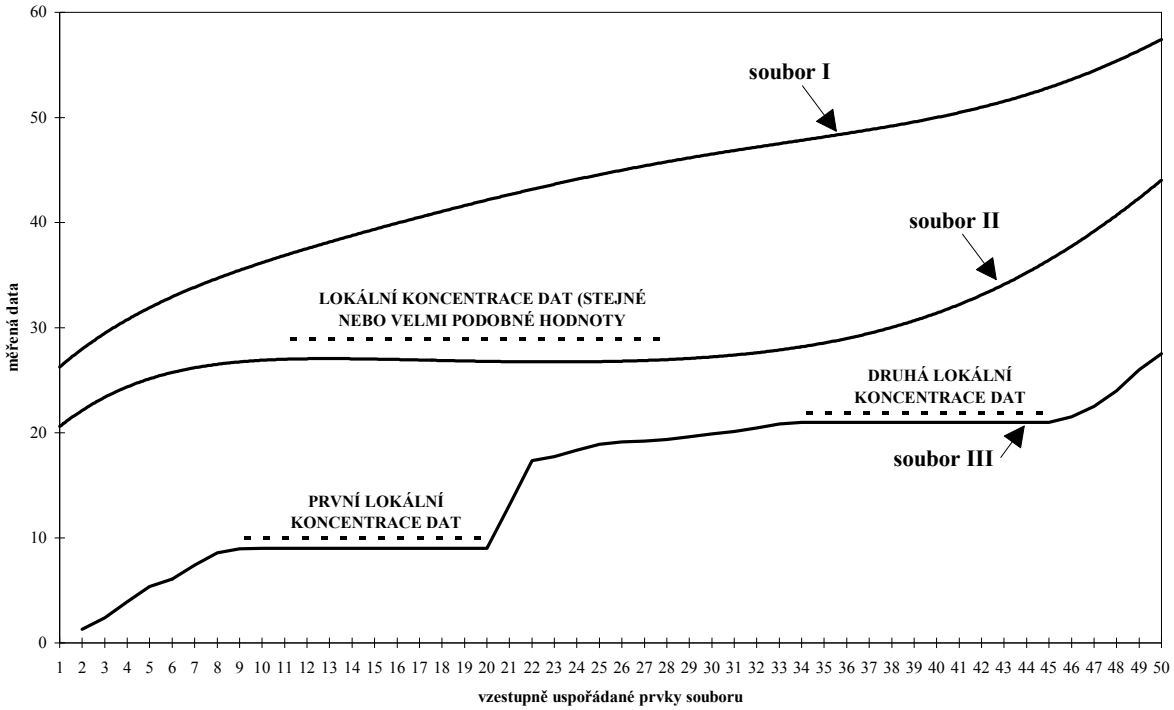
Je zřejmé, že již prvotní zpracování získaných dat může podstatně přispět ke zvolení dalšího správného postupu statistické analýzy. Je vhodné především u menších souborů. Dokonalejším způsobem prvotního zpracování statistického souboru – zvláště rozsáhlejšího (řádově stovky, tisíce ,… dat) - je třídění souboru.

Vzestupné pořadí	Pořadové číslo měření	Číslo stromu	Výčetní tloušťka v cm	Vzestupné pořadí	Pořadové číslo měření	Číslo stromu	Výčetní tloušťka v cm	Vzestupné pořadí	Pořadové číslo měření	Číslo stromu	Výčetní tloušťka v cm
1	37	19	28,8	**21**	52	35	37,7	**41**	19	54	43,3
2	10	53	29,9	**22**	36	46	38,2	**42**	12	10	43,8
3	6	12	30,3	**23**	7	32	38,5	**43**	22	26	44,2
4	27	44	30,3	**24**	8	43	38,6	**44**	30	1	44,2
5	40	38	31,2	**25**	46	5	38,7	**45**	51	6	44,5
6	50	31	33,0	**26**	29	47	39,0	**46**	15	15	44,7
7	4	13	33,1	**27**	45	9	39,1	**47**	3	16	44,9
8	33	24	33,1	**28**	48	23	39,5	**48**	47	4	45,0
9	43	28	33,5	**29**	23	18	40,3	**49**	31	8	47,2
10	1	51	34,3	**30**	44	34	40,3	**50**	49	52	47,2
11	2	7	34,5	**31**	9	41	40,6	**51**	38	3	47,3
12	24	39	34,6	**32**	21	37	41,3	**52**	16	50	47,7
13	13	14	35,2	**33**	41	40	41,3	**53**	42	55	47,7
14	39	33	35,5	**34**	20	17	41,8	**54**	53	27	52,0
15	5	49	36,2	**35**	55	29	41,8	**55**	11	2	57,7
16	35	45	36,2	**36**	18	30	42,0				
17	25	22	36,4	**37**	26	20	42,1				
18	28	21	37,1	**38**	17	11	42,9				
19	32	25	37,2	**39**	54	48	43,0				
20	14	42	37,4	**40**	34	36	43,1				

Tabulka 3.2 - Vzestupně uspořádaný soubor



Obrázek 3.2 - Grafické znázornění vzestupně uspořádaného statistického souboru



Obrázek 3.3 - Příklady možných průběhů vzestupně uspořádaných souborů

3.3 Třídění souboru

Zpracování rozsáhlých datových souborů není technicky jednoduché. Zvláště v „předpočítačové“ éře to znamenalo úmornou rutinní práci, při které zpravidla vznikala spousta chyb (což znamenalo stejně úmorné přepočítávání a hledání chyb). Proto byly hledány způsoby, jak zjednodušit alespoň základní statistické výpočty (např. statistické charakteristiky). Vhodným řešením byla technika třídění souboru.

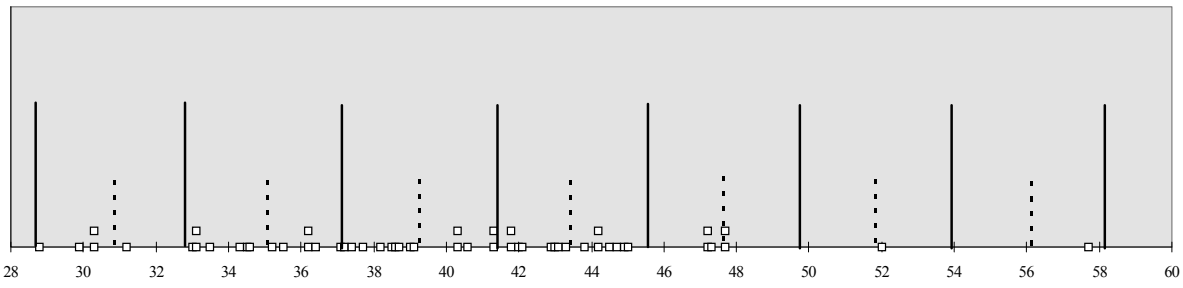
Je nutné podotknout, že v současné době již nemá třídění souborů takový význam jako v minulosti, protože při počítačovém zpracování dat už příliš nezáleží na velikosti zpracovávaných souborů a také nebezpečí výpočetních numerických chyb je minimalizováno. Přesto se v některých oborech tříděná data stále používají (jak z důvodů historických – pro tříděná data existují vyzkoušené a dlouhodobě používané metody zpracování, tak i z praktických – např. třídění do velikostních nebo kvalitativních tříd, apod.) a mezi ně patří i lesnictví. Je tomu tak především z prvního uvedeného důvodu (např. pro tříděná data – tloušťkové třídy nebo stupně - jsou k dispozici prakticky používané tabulky ke zjišťování zásob, např. objemové nebo JOK) i praktických (rychlejší měření a přitom prakticky dostatečná přesnost výsledku). Typickým příkladem tříděného statistického souboru (který jako tříděný soubor již vzniká při měření) je běžně používaný „svěrkovací zápisník (manuál)“, tj. výsledek měření výčetních tlouštěk stromů v porostu. Měřené tloušťky nejsou do dalších výpočtů postoupeny ve své skutečné hodnotě, ale jsou zařazovány do „tloušťkových stupňů“ (které mají podle účelu a přesnosti metody zpravidla rozsah 2 – 4 cm). Tato metoda má několik výhod: zrychlí se tím měření (není nutné odečítat přesné hodnoty, ale pouze zařazovat do daných intervalů) a výpočet (např. při výpočtu průměrné tloušťky není nutné sčítat několik set čísel, ale pouze násobit středy tloušťkových stupňů jejich četnostmi, tj. počty stromů zařazených do jednotlivých stupňů), přičemž přesnost tohoto měření a výpočtu je dostatečná pro uvažované praktické vy-

užití (např. výpočet zásob porostů). Z hlediska statistiky jsou potom střední hodnoty tloušťkových stupňů třídní reprezentanti, hraniční hodnoty tloušťkových stupňů hranicemi tříd a počty stromů v jednotlivých tloušťkových stupních absolutní třídní četnosti (podrobné vysvětlení těchto pojmů následuje níže).

3.3.1 Podstata a účel třídění

Třídění souboru spočívá v rozdělení celého souboru do skupin – tříd. **Třída** je vymezena dolní a horní hranicí, tj. nejmenší a nejvyšší hodnotou, která bude považována za zařazenou do uvažované třídy. V rámci každé třídě se vhodně zvolí jedna hodnota, která zastupuje (reprezentuje) všechny hodnoty patřící do dané třídy (to je třídní reprezentant) a spočítá se počet hodnot zařazených do dané třídy (to je absolutní třídní četnost). Hlavní význam třídění již nyní začíná vystupovat. Místo toho, abychom v dalších výpočtech počítali se všemi hodnotami souboru (např. několika sty nebo tisíci), budeme používat jen třídní reprezentanty a třídní četnosti (nejvýše několik desítek hodnot), přičemž při správně provedeném třídění získáme stejnou statistickou informaci jako pro netříděný soubor (např. statistické charakteristiky). Postup, při kterém z N hodnot statistického souboru, z nichž každá nese určitou informaci, se stejných informací dosáhne z minimálního počtu vhodně stanovených hodnot (v případě třídění jsou to třídní reprezentanti a třídní četnosti), se nazývá zhušťování informací.

Schématické znázornění třídění je na obrázku 3.4 . Jedná se o soubor měřených dat, která jsou uvedena v tabulce 3.1 Je nutné upozornit, že soubor této velikosti není zpravidla potřebné třídit, je použit pouze pro ilustraci podstaty třídění. Předpokládejme, že nepoužijeme obvyklé tloušťkové stupně, ale vytvoříme si vlastní třídění. Bylo vytvořeno celkem 7 tříd se stejným třídním intervalem. Jednotlivé měřené tloušťky (uspořádané podle velikosti) jsou znázorněny bílými čtverečky, hranice tříd dlouhými plnými čarami a třídní reprezentanti kratšími čárkovanými čarami. (Jak se určí počet tříd, jejich velikost apod. se dozvíme v následující části). Vidíme tedy, že do první třídy (dolní hranice 28,75; horní hranice 32,95) patří pět prvků souboru (konkrétně hodnoty 28,8; 29,9; 30,3; 30,3; 31,2), do další 13 prvků, atd. Všechny tyto hodnoty budou v dalších výpočtech nahrazeny jednou hodnotou – třídním reprezentantem (což může být hodnota, která v souboru vůbec fyzicky neexistuje) – v případě první třídy je to hodnota 30,85 a třídní četností (pro první třídu 5). Tedy dále už nebudeme používat pro výpočty všech 55 prvků, ale pouze 7 třídních reprezentantů a 7 třídních četností. V případě správně provedeného třídění získáme prakticky stejné výsledky jako při počítání se všemi původními hodnotami. V případě velkých souborů tato výhoda vynikne daleko více.



Obrázek 3.4 - Schématické znázornění třídění. Bližší vysvětlení viz v textu.

Účelem a podstatou třídění je tedy:

- vyloučit všechny nepodstatné rozdíly v hodnotách znaku souboru,
- zachovat podstatné, charakterizující rozdíly.

3.3.2 Technika třídění

Třídění sestává z následujících kroků:

- **volba třídního intervalu** a **počtu tříd**
- **seřazení tříd do stupnice**
- stanovení **třídního reprezentanta**
- **zjištění četnosti ve třídách**

3.3.2.1 Volba třídního intervalu a počtu tříd

Podmínky volby třídního intervalu vycházejí z účelu a podstaty třídění:

- reprezentant třídy zastupuje bez velké chyby všechny hodnoty do třídy zařazené,
- interval je co největší.

Zpravidla se volí pravidelný třídní interval, to znamená, že všechny třídy mají stejnou vzdálenost mezi dolní a horní hranicí (jako na obr. 3.4). Pouze ve speciálních případech se používají nepravidelné třídní intervaly. V následujícím textu se budeme věnovat jen základnímu případu pravidelných intervalů.

Nejčastěji se délka třídního intervalu (h) odvozuje až z určeného počtu tříd (m) a rozsahu souboru (n). Z teoretických úvah a praktických zkušeností bylo navrženo několik způsobů pro první, orientační určení počtu tříd.

Nejčastěji se používá tzv. **Sturgesovo pravidlo**:

$$m = 1 + 3{,}3 \cdot \log(n) \tag{3.1}$$

které udá přibližný doporučený počet tříd (vypočítané číslo se zpravidla zaokrouhluje nahoru).

Kromě tohoto vzorce se mohou orientačně použít počty tříd uvedené v tabulce 3.3

n	m
25 – 40	5 – 6
41 – 60	6 – 8
61 – 100	7 – 10
101 – 200	8 – 12
201 +	10 – 15

Tabulka 3.3 - Orientační počty tříd v závislosti na velikosti souboru

V praxi obvykle vycházíme z počtu tříd navržených Sturgesovým pravidlem nebo tabulkou 3.3 (a ve většině případů se to dá doporučit), ale v odůvodněných případech je možné zvolit jiné třídění. Konečné rozhodnutí závisí na statistikovi.

Pokud známe počet tříd *m*, odpovídající délka intervalu (také orientační) se určí podle vzorce

$$h = \frac{x_{max} - x_{min}}{m} \tag{3.2}$$

Délku intervalu lze také navrhnout (při obráceném postupu, tj. bez znalosti počtu tříd) podle vzorců

$$h = 0{,}08(x_{max} - x_{min}) \tag{3.3}$$

nebo

$$h = \langle \frac{x_{max} - x_{min}}{12} \langle 2h\,. \tag{3.4}$$

Známe-li h, určí se počet tříd podle vzorce

$$m = \frac{x_{max} - x_{min}}{h}\,. \tag{3.5}$$

Z praktického hlediska je vhodné velikost třídního intervalu určit tak, aby to nebylo číslo s příliš mnoha desetinnými místy (např. vyjde-li h podle vzorce (3.2) 3,243535, je vhodné použít podle potřeby hodnot 3,25 nebo 3,3, popř. i mírně vyšší). Tím je zaručeno, že hranice tříd i třídní reprezentanti vyjdou jako celá čísla nebo čísla s málo desetinnými čísly a bude snadné s nimi dále počítat. Při určení počtu tříd a velikosti třídního intervalu je nutné respektovat vztah

$$m \cdot h \geq x_{max} - x_{min}\,. \tag{3.6}$$

Rovnost v tomto vztahu je minimálním požadavkem (potom je dolní hranice první třídy rovna x_{min} a horní hranice poslední třídy rovna x_{max}), ale je lépe jestliže součin počtu tříd a třídního intervalu (můžeme jej nazvat variačním rozpětím třídící stupnice) je vyšší než skutečné variační rozpětí souboru ($x_{max} - x_{min}$). Umožní to v případě potřeby pohybovat s hranicemi tříd tak, aby bylo dosaženo optimálního výsledku. Na druhé straně nesmí být rozdíl mezi variačním rozpětím třídící stupnice a variačním rozpětím souboru příliš velký, aby na okrajích souboru nevznikly „prázdné“ třídy (tj. třídy s nulovou četností).

Při návrhu počtu tříd a délky třídního intervalu je třeba počítat i s nutností respektovat následující zásady:

- zásada **úplnosti** - každá jednotka souboru musí být zařazena do některé třídy (pokud by tomu tak nebylo, ztrácela by se část statistické informace původního souboru a tříděný soubor by už nemohl dostatečně přesně vypovídat o statistických vlastnostech původního souboru),
- zásada **jednoznačnosti** - každá jednotka je zařazena jednoznačně (tj. nesmí vzniknout pochybnost, do které třídy je nutno určitou jednotku zařadit).

Oběma zásadám je nutno přizpůsobit hranice tříd. Zásadu úplnosti splníme respektováním vztahu (3.6). Přísné dodržení druhé zásady je poněkud komplikovanější. Volíme-li hranice se stejnou nebo menší přesností, než mají hodnoty souboru, může se stát, že některá hodnota je rovna některé hranici. Pro takové případy musí být předem jednoznačně stanoveno, do které třídy (předcházející nebo následující) bude hodnota zařazena. Lze také postupovat tak, že s přihlédnutím k pravidlům zaokrouhlování a k přesnosti hodnot veličiny se hranice určí na 5 jednotek prvního zaokrouhleného místa. Tím je možná shoda třídění hodnoty a hranice třídy vyloučena. Např. pro data z tabulky 3.1 , která jsou měřena na jedno desetinné místo, je možné volit hranice tříd na dvě desetinná místa (např. první třída 27,85 – 32,95; druhá třída 32,95 – 37,15; atd.). Tím nemůže nikdy dojít ke shodě mezi měřenou hodnotou a hranicí třídy a zásada jednoznačnosti třídění je bezezbytku dodržena.

3.3.1.1 Stanovení třídního reprezentanta

Třída obsahuje ty hodnoty, jejichž rozdíly jsou nepodstatné. Může je tedy zastupovat hodnota vhodná, tzv. **reprezentant třídy**. Třídního reprezentanta je nutno volit tak, aby co nejlépe zastupoval hodnoty celé třídy. Zpravidla se tomuto požadavku dostatečně vyhoví, volíme-li za reprezentanta střed třídy, tj. hodnotu aritmetického průměru hranic třídy. Je tomu tak proto, že při dostatečně velkých souborech, vhodných ke třídění, mají i jednotlivé třídy (zvláště středové, obvykle nejpočetnější) dostatečný počet prvků, aby bylo možné jejich hodnoty s rizikem nejmenší chyby nahradit střední hodnotou. Částečně je to možné vidět na obrázku 3.4 (i když tento soubor je pro účely třídění hodně malý). U druhé, třetí, čtvrté a částečně páté třídy jsou třídní reprezentanti skutečně přibližně středními hodnotami měřených jednotek svých tříd. U krajních tříd tomu zpravidla tak nebývá, ale jejich vliv („váha“) na souhrnné charakteristiky souboru (o nich viz kapitola 4) je dána jejich četností a je tedy velmi malá.

Při stejně širokých třídách dostaneme tak rovnoměrnou stupnici třídních reprezentantů, výhodnou při dalším početním zpracování.

Třídní reprezentant se označuje $\bar{x}_i$ (střední hodnota i-té třídy). Rozdíly mezi třídními reprezentanty jsou podstatné, pro soubor charakteristické rozdíly.

Informace, které nesou hodnoty ve třídě, nesou po třídění reprezentant třídy $\bar{x}_i$ a počet hodnot ve třídě n_i .

Informace, které nesou hodnoty celého souboru, nese po třídění m dvojic $\bar{x}_i$, n_i.

Třídění musí být provedeno tak, aby požadované informace byly v roztříděném souboru stejné nebo jen nepodstatně odlišné od informací, které obsahuje původní netříděný soubor (např. souhrnné statistické charakteristiky, jako je aritmetický průměr, směrodatná odchylka a další, vypočítané z roztříděného souboru, by se měly jen velmi málo lišit od týchž hodnot platných pro původní soubor).

3.3.2 Typy četností tříděného souboru

Důležitou úlohu mají v procesu zjišťování informací počty hodnot příslušné třídám – třídní četnosti.

Absolutní třídní četnost n_i je počet hodnot v *i*-té třídě. Je to základní charakteristika každé třídy a spolu s třídním reprezentantem se podílí na výpočtu statistických charakteristik. Je mírou vlivu dané třídy na všechny studované vlastnosti tříděného souboru.

Relativní třídní četnost w_i je poměr absolutní třídní četnosti k rozsahu souboru, tedy

$$w_i = \frac{n_i}{N} \tag{3.7}$$

Relativní četnost se vyjadřuje buď jako desetinné číslo nebo, po vynásobení 100, v procentech. Uvážíme-li, že každá hodnota souboru je v souboru „jistě“ obsažena, určuje w_i díl počtu prvků, který v souboru zaujímá *i*-tá třída. Má charakter pravděpodobnosti, s jakou se v souboru vyskytuje hodnota zařazená do i-té třídy. Relativní četnosti rozšiřují (jako všechny relativní veličiny) možnosti srovnávání souborů mezi sebou, i když tyto soubory nemají stejný rozsah nebi nejsou ve stejných jednotkách. Mají velký význam ve všech statistických úvahách využívajících pravděpodobnosti. Výpočet relativních četností se může provést tak, že se

vyčíslí podíl jednotky souboru $p = \frac{1}{n}$ (nebo $p = \frac{100}{n}\%$) a tím se postupně vynásobí absolutní četnosti tříd.

Absolutní součtová (kumulativní) četnost ν_i je součet absolutních četností od první až po uvažovanou *i*-tou třídu, tedy

$$\nu_i = \sum_{j=1}^{i} n_j \tag{3.8}$$

Relativní součtová (kumulativní) četnost (ω_i) je součet relativních četností od první až po i-tou třídu, tedy

$$\omega_i = \sum_{j=1}^{i} = w_j \tag{3.9}$$

Kumulativní četnosti nesou informace založené na údajích o počtu hodnot znaků, které nepřesahují hodnotu horní hranice třídy a tedy udávají nám, jaká část jednotek souboru je obsažena do určité třídy včetně. Je to užitečné při úlohách typu: kolik hodnot, resp. kolik % hodnot, přesahuje určitou hodnotu (rovnou příslušné hranici třídy)? kolik hodnot (% hodnot) leží mezi určitými hodnotami, shodnými s hranicemi tříd – např. daných určitou normou, předpisem? apod. Relativní verze kumulativní četnosti slouží, stejně jako absolutní, hlavně ke srovnávání mezi různými soubory.

Pro řadu četností za sebou jdoucích tříd se používá název rozdělení četností.

3.3.3 Grafické vyjádření rozdělení četností

Při statistickém popisu souboru je výhodné graficky znázornit rozdělení četností.

Graf rozdělení absolutních třídních četností se nazývá frekvenční graf (též frekvenční funkce).

Graf rozdělení absolutních kumulativních četností se nazývá distribuční graf (též distribuční funkce).

Ve statistice se nejčastěji používá bodového grafu nebo sloupcového grafu (histogramu).

Grafy rozdělení absolutních třídních četností jsou plošné. Body bodového grafu se spojují úsečkami. Vytvoří se tzv. frekvenční polygon. Plocha omezená osou hodnot a polygonem resp. sloupci je v příslušných jednotkách číselně rovna rozsahu souboru.

Ze součtových četností můžeme odvodit rozdělení četností pro jiný třídní interval nebo posunutou třídní stupnici. Při použití statistických metod je často nutné analyzovat různě získané a zpracované soubory. Přitom je třeba pracovat s rozdělením o stejném počtu tříd. Nemáme-li k dispozici původní neroztříděný soubor, můžeme roztříděný soubor určitého rozdělení převést na jiné rozdělené následujícím postupem: vykreslíme ve vhodném měřítku graf součtových četností, určíme nové horní meze tříd a z grafu odečteme součtové četnosti nového třídění. Početně se celý postup provede obdobně příslušnou interpolací. Polygon součtové četnosti předpokládá lineární rozložení počtu hodnot znaku ve třídě. Tento ideální předpoklad ve skutečnosti splněn zpravidla není. Takto získané nové rozdělení nemusí být zcela totožné s rozdělením, které by bylo získáno z původního souboru. Pokud tyto rozdíly vzniknou, bývají při dodržení logiky věcného zkoumání zanedbatelné. Jejich existenci musíme zahrnout do všech úvah založených na výsledcích získaných ze součtových četností.

V běžné statistické praxi zpravidla vystačíme s popsaným jednoduchým přístupem ke třídění, protože soubory, se kterými se běžně setkáváme, jsou jednoduché, nekomplikované. Komplikovaných souborů není však tak málo, jak se všeobecně předpokládá. Třídění takových souborů musí být podrobněji analyzováno, zejména z hlediska vhodnosti reprezentantů a šířky intervalu. I tyto podrobnější analýzy mohou být provedeny na základě zásad a pouček o postupu třídění uvedených v této kapitole.

Příklad 3.1:

Na skladě pily byly měřeny střední tloušťky navážených smrkových kmenů. Po skončení navážení byl k dispozici soubor hodnot středních tlouštěk, který bylo třeba statisticky zpracovat. Hodnoty tlouštěk jsou měřeny s přesností na 0,1 cm. Z prvotního zápisu vyplynulo:

Rozsah souboru $N = 224$ *ks*

Krajní hodnoty $d_{min} = 16{,}1$ *cm;* $d_{max} = 26{,}6$ *cm*

Počet tříd odhadneme z tabulky 3.3 - m´ = 12. Dále vypočítáme první odhad třídního intervalu h´:

$$h' = \frac{26{,}6 - 16{,}1}{12} = 0{,}875 \cong 1{,}0;$$

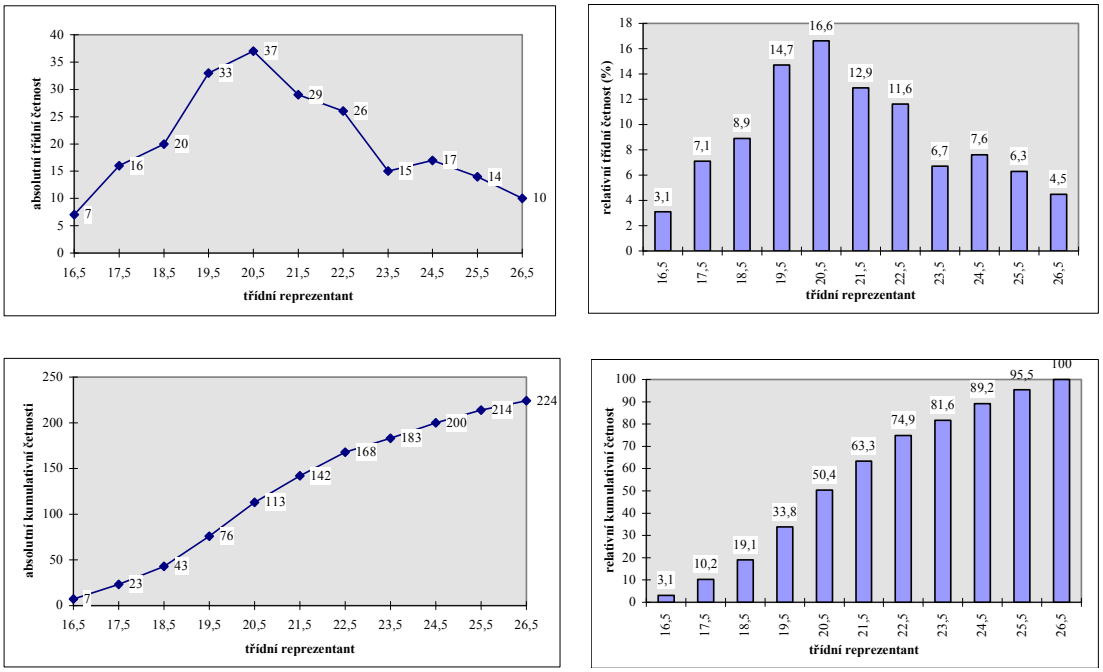
tedy šíře třídního intervalu byla stanovena na 1 cm. Dále se provede prvotní odhad dolní hranice první třídy (počátku třídící stupnice) a´, a horní hranice poslední třídy b´ takto: a´ = 16,0 cm, b´ = 28,0 cm (16 + 12 tříd po 1 cm). Vzhledem k tomu, že nejvyšší měřená hodnota je 26,6 cm, poslední třída s hranicemi 27,0 – 28,0 cm by byla prázdná, upravíme definitivně třídící stupnici takto: a = 16,0 cm = 11 b = 27,0 cm, h = 1 cm. Toto třídění vyhovuje podmínce úplnosti. Abychom splnili i podmínku jednoznačnosti třídění, musíme určit předpis, do které třídy budeme zařazovat hraniční hodnoty: budou zahrnuty do předcházející třídy. Reprezentanty třídy jsou středy tříd, tj. první třída zahrnuje hodnoty 16,1 –17,0 cm, druhá 17,1 – 18,0 cm, atd.

Výsledky třídění jsou v tabulce 3.3 , grafy četností rozdělení na obrázku 3.5 . Absolutní četnosti jsou zde znázorněny pomocí polygonů, relativní četnosti pomocí histogramu.

Číslo třídy	Třídní reprezentant	Absolutní třídní četnost	Relativní třídní četnost (%)	Absolutní kumulativní četnost	Relativní kumulativní četnost (%)
1	16,5	7	3,1	7	3,1
2	17,5	16	7,1	23	10,2
3	18,5	20	8,9	43	19,1
4	19,5	33	14,7	76	33,8
5	20,5	37	16,6	113	50,4
6	21,5	29	12,9	142	63,3
7	22,5	26	11,6	168	74,9
8	23,5	15	6,7	183	81,6
9	24,5	17	7,6	200	89,2
10	25,5	14	6,3	214	95,5
11	26,5	10	4,5	224	100,0

Tabulka 3.3 - Výsledky třídění

Statistické charakteristiky se pro roztříděný soubor počítají pomocí speciálních postupů, které zohledňují rozdělení do tříd a využívají pouze třídní reprezentanty a absolutní třídní četnosti. Třídní četnost působí jako váha dané třídy na výslednou hodnotu charakteristiky (početnější třída má vyšší váhu než méně početná). Těmto charakteristikám se proto říká vážené. Při správně provedeném třídění by se vážené a prosté charakteristiky měly sobě rovnat (roztříděný soubor musí vydat stejnou statistickou informaci jako neroztříděný). Tento požadavek není možno naplnit zcela přesně, vždy mezi váženými a prostými charakteristikami budou určité rozdíly, ale neměly by být podstatné a měly by umožňovat shodnou statistickou interpretaci. Vzorce (včetně příkladu) na výpočet vážených statistických charakteristik budou uvedeny v následující kapitole.



Obrázek 3.5 - Grafické znázornění třídních četností

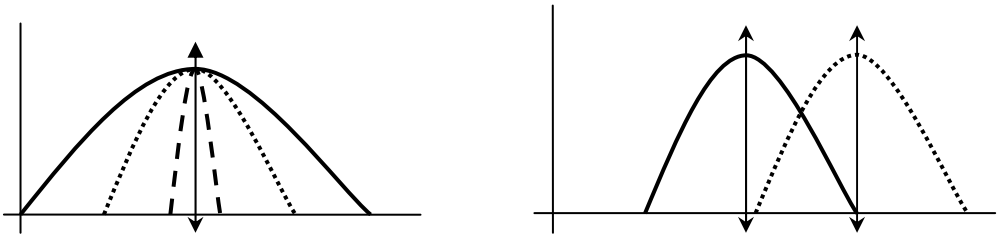
4 Statistické charakteristiky souboru

Statistické charakteristiky (přesně **souhrnné statistické charakteristiky**) jsou veličiny, které **podávají koncentrovanou formou informaci o podstatných statistických vlastnostech** analyzovaného souboru. Jsou výsledkem procesu **zhušťování informací**. Určitým stupněm zhuštění informací o statistickém souboru je již rozdělení třídních četností statistického souboru (tabulka, graf). Statistické charakteristiky právě vystihují ty skutečnosti, ve kterých se mohu rozdělení (a tedy soubory) lišit.

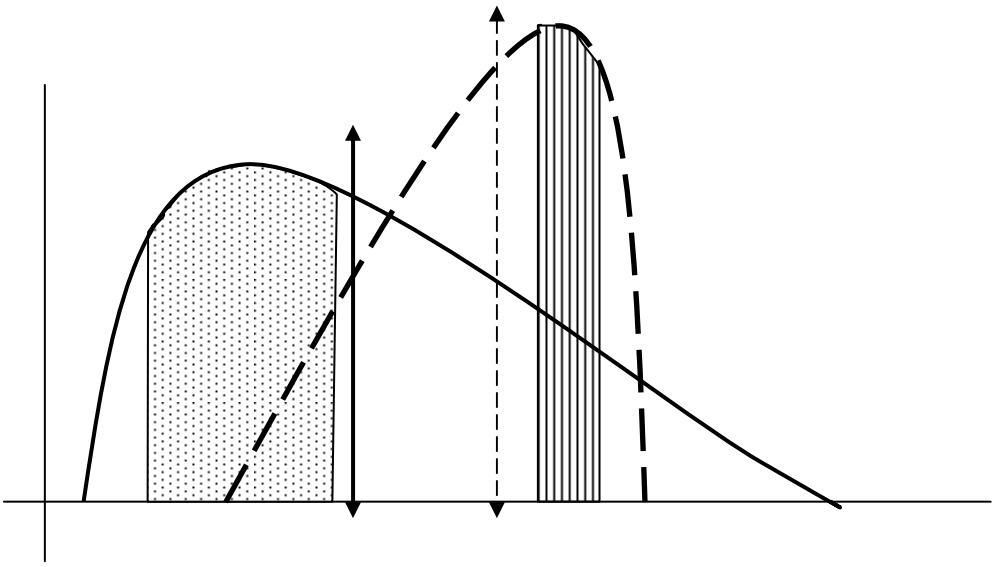
4.1 Typy statistických charakteristik

Důležitou statistickou charakteristikou je rozsah souboru (N). Dále používáme charakteristiky:

- polohy,
- variability,
- tvaru
  - souměrnosti
  - špičatosti



Obrázek 4.1 - Příklady souborů lišících se variabilitou (vlevo) a polohou (vpravo)



Obrázek 4.2 - Příklad souborů lišících se polohou, variabilitou a tvarem

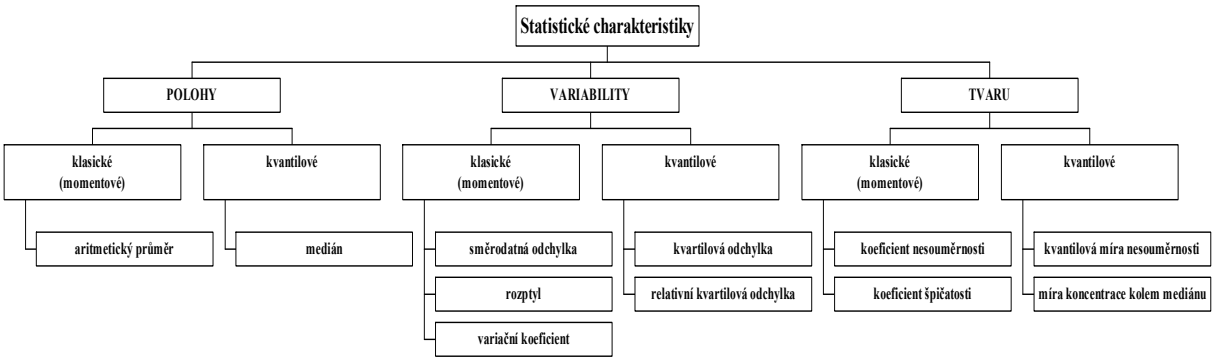
Názornější představu, co tyto pojmy znamenají, je možné si utvořit na základě studia schematických obrázků 4.1 a 4.2 . U všech těchto obrázků jsou na ose x vynášeny jednotlivé hodnoty, na ose y jejich četnosti. Vznik těchto křivek si můžeme představit tak, že polygony nebo histogramy četností z příkladu 3.1 ze třetí kapitoly proložíme určitou funkcí, která nám může popsat četnost výskytu každé hodnoty (s tím, že bychom dále zjemňovali třídní dělení až na „nekonečně malé“ intervaly). Střední („průměrná“) poloha souborů je na číselné ose (ose *x*) znázorněna oboustrannou šipkou.

Na obrázku 4.1 jsou vlevo znázorněny soubory, které se neliší svou polohou (střední hodnoty všech souborů jsou shodné – vyjádřené společnou dvojitou šipkou), ale liší se variabilitou (mírou rozptýlení jednotlivých hodnot kolem hodnoty střední). Nejvyšší variabilitu má soubor znázorněný plnou čarou – křivka četností zabírá největší část číselné osy kolem střední hodnoty, tj. hodnoty tohoto souboru jsou kolem ní nejméně koncentrovány , nejmenší variabilitu vykazuje soubor vykreslený čárkovanou čarou („nejužší“ křivka, hodnoty jsou nejvíce koncentrovány kolem střední hodnoty). Z hlediska tvaru jsou všechny soubory souměrné.

Vpravo na témže obrázku je opačný případ – oba soubory mají shodnou variabilitu (všechny křivky mají stejný tvar), ale liší se polohou (umístěním na číselné ose *x* – jejich střední hodnoty se liší).

Obrázek 4.2 znázorňuje nejobvyklejší případ – kdy se soubory mezi sebou liší ve všech hlavních skupinách charakteristik: v poloze (jejich střední hodnoty jsou posunuty na ose x), variabilitou (soubor znázorněný plnou čarou má zřejmě vyšší variabilitu, protože jeho hodnoty jsou rozptýleny kolem střední hodnoty v širším intervalu) i tvarem: žádný soubor není souměrný – „plný“ soubor má nejvyšší koncentraci hodnot nalevo od střední hodnoty (znázorněno tečkovanou oblastí), „čárkovaný“ soubor napravo (znázorněno podélným šrafováním), přičemž jeho míra koncentrace hodnot je vyšší (čárkovaná oblast je „vyšší“, tj. četnosti hodnot zobrazené na ose *y* jsou v této oblasti větší, souvisí to i s nižší variabilitou).

V této souvislosti je nutné zdůraznit, a z výše uvedených obrázků to také vyplývá, že pro úplný popis statistického souboru je nutné používat charakteristik všech tří kategorií, protože jen tak lze vyjádřit všechny důležité vlastnosti souboru. Nejčastější chybou je uvádění samotného aritmetického průměru (základní charakteristika polohy) bez údajů o variabilitě a tvaru souboru.



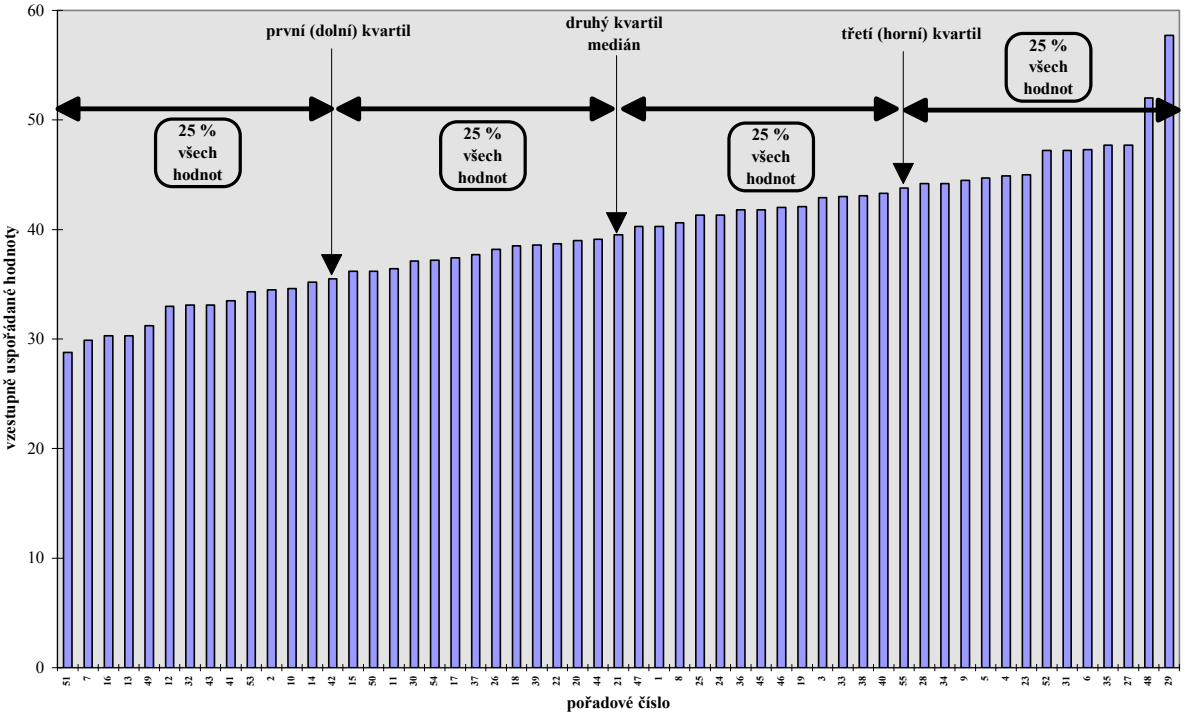
Obrázek 4.3 - Přehled základních statistických charakteristik

Obrázek 4.3 obsahuje přehled nejdůležitějších charakteristik používaných ke statistickému popisu souborů (kromě nich existuje ještě řada dalších, které zde nejsou uvedeny, protože mají buď speciálnější použití nebo složitější teoretickou podstatu a výpočet).

Statistické charakteristiky se obecně dělí na dvě základní skupiny – momentové a kvantilové.

Momentové charakteristiky vycházejí z teorie tzv. statistických momentů. Podrobněji o tomto tématu (včetně odvození jednotlivých typů momentů a jejich vztah k jednotlivým charakteristikám) pojednává např. ZACH (1994) a ŠMELKO, WOLF (1978). Z praktického hlediska je důležité, že jejich výpočet vychází ze všech hodnot souboru, což na jedné straně znamená, že všechny statistické jednotky mají stejný vliv (váhu) na výslednou charakteristiku, ale na druhé straně existují pro jejich použití určitá omezení (normalita souboru, ovlivnění extrémními hodnotami apod. – tyto pojmy budou vysvětleny později). Momentovým charakteristikám se také říká klasické, protože jsou standardně používány vždy, kdy je to přípustné. Patří mezi ně nejobvyklejší charakteristiky, např. aritmetický průměr nebo směrodatná odchylka.

Kvantilové charakteristiky vycházejí z pořádkové statistiky, tj. ze vzestupně uspořádaného souboru. **Kvantil** je hodnota, kterou nese v uspořádaném souboru **prvek určitým způsobem umístěný**. Kvantil je speciálním případem tzv. **percentilu**. Většinou hovoříme o $100 \cdot p$-% percentilu, kde platí $0 < p < 1$. Hodnota p znamená konstantu, která oddělí p-tý díl (100p procent) nejmenších hodnot. V případě percentilu může hodnota p nabývat jakékoli hodnoty mezi 0 a 1. Pojem kvantil se používá pro určité percentily, které dělí soubor na určité „zaokrouhlené“ části a takový kvantil má zpravidla také svůj speciální název. Např. je-li $p = 0,5$, potom příslušný kvantil dělí soubor na poloviny (100.0,5 = 50 %) a nazývá se medián (podrobněji viz dále), podobně pro $p = 0,25$ je příslušným kvantilem dolní kvartil, který odděluje 25 % nejmenších hodnot (podobně existuje horní kvartil pro $p = 0,75$). Kvartily



Obrázek 4.4 - Grafické znázornění kvartilů

(spolu s mediánem) dělí počet hodnot na čtyři stejné díly. Decily dělí počet hodnot na deset stejných dílů atd. Lze tedy zobecnit, že kvantily jsou čísla, která rozdělují řadu vzestupně uspořádaných hodnot znaku na určitý počet skupin o stejném počtu prvků. Grafické znázornění kvartilů jako typického příkladu kvantilů je na obrázku 4.4 . Použita jsou data z tabulky 3.1.

Z vlastností kvantilů je nutno vyzvednout jejich nezávislost na extrémních hodnotách a na rozdělení souboru. Na druhé straně s nimi nelze provádět počtářské operace v takové míře jako s momentovými charakteristikami. Používají se tam, kde z určitých důvodů nelze použít momentové charakteristiky, a také v těch případech, kde je nutno vzít do úvahy spíše rozdělení hodnot na části než nahrazení všech hodnot hodnotou střední (např. při porovnání naměřených hodnot s určitou normou, konstantou).

Pořadové číslo (*i*) prvku příslušného kvantilu se určí podle vzorce

$$i = \frac{N+1}{p} \cdot r \qquad (4.1)$$

kde je

N rozsah souboru
p počet skupin dělení.
r pořadí kvantilu.

Příklad 4.1:
Určete pořadové číslo kvartilů a mediánu podle dat z tabulky 3.1 (obrázek 4.4).

Použijeme rovnice 4.1. V našem případě je N = 55, p = 4 (kvartily dělí soubor na 4 díly) a r = 1, 2, 3 (1 pro dolní kvartil, 2 pro medián – „druhý kvartil“, 3 pro horní kvartil):

$$i = \frac{55+1}{4} \cdot 1 = 14$$

tedy dolní kvartil nese 14. prvek vzestupně uspořádaného souboru. Obdobně dostaneme dosazením r = 2 hodnotu 28 (medián je 28. prvek) a pro r = 3 je to hodnota 42 (horní kvartil je 42. prvek).

Pro výpočet kvantilů existuje celá řada dalších postupů. Některé z nich uvádí např. ZVÁRA (1998).

Je nutné zdůraznit, že statistické charakteristiky uvedené v následujícím textu neobsahují jejich úplný výčet. Statistická teorie zná celou řadu dalších různým způsobem konstruovaných měr. Rovněž výčet vlastností a použití je pouze strohým uvedením nejdůležitějších.

Kterou charakteristiku zvolíme k popisu vlastnosti souboru, záleží na požadavcích a cílech úkolu, který se statickým popisem řeší. Volbu charakteristiky ovlivňuje potřebný charakter informací, které úkol vyžaduje. Při praktickém řešení úkolu určíme nejdříve rozsah a hloubku informací, které potřebujeme a na základě tohoto zjištění volíme ty míry, které jsou schopny požadované informace podat.

4.2 Charakteristiky polohy

Charakteristiky polohy informují o **poloze souboru v uspořádané číselné množině možných hodnot zkoumané veličiny**. Jsou to hodnoty s rozměrem dané veličiny, okolo kterých se hodnoty znaku jednotek souboru soustřeďují.

Pro „ideální“ charakteristiku polohy byly stanoveny následující požadavky:

- má být přesně určena, ne odhadnuta
- má se zakládat se na všech hodnotách souboru

- má mít zřetelnou vlastnost, podle které jsou podávané informace snadnou zjistitelné a pochopitelné
- má být co nejméně dotčeny „kolísáním“ náhodného výběru.

O charakteristikách polohy se někdy hovoří jako o středních hodnotách. Je nutné rozlišovat od matematicky definované střední hodnoty (matematické naděje), která se jako základní pojem definuje a používá v matematické statistice.

Základními charakteristikami polohy jsou **průměry**. Jsou vypočítány ze všech hodnot souboru. Způsob výpočtu odpovídá charakteru zkoumané veličiny a požadavkům dalšího zpracování. Nejčastěji používaným průměrem je **aritmetický průměr**. Ostatní průměry (např. geometrický, harmonický, chronologický, apod.) se používají pouze ve speciálních případech. Podrobněji o nich pojednává např. ZACH (1994).

4.2.1 Aritmetický průměr ($\bar{x}$)

Aritmetickým průměrem charakterizujeme hodnotu, okolo níž kolísají jednotlivé prvky souboru. Fyzikálně odpovídá aritmetický průměr těžišti N stejně hmotných bodů umístěných na přímce se souřadnicemi x_i (to jsou jednotlivé hodnoty souboru).

Pro neroztříděný soubor používáme jednoduchý aritmetický průměr

$$\bar{x}_1 = \frac{\sum_{i=1}^{N} x_i}{N} \tag{4.2}$$

Pro soubor roztříděný do tříd se používá vážený aritmetický průměr

$$\bar{x}_2 = \frac{\sum_{i=1}^{m} n_i \cdot \bar{x}_i}{N} \tag{4.3}$$

kde je *m* počet tříd, n_i absolutní třídní četnost a $\bar{x}_i$ třídní reprezentant, nebo též

$$\bar{x}_2 = \sum_{i=1}^{m} w_i \cdot \bar{x}_i \tag{4.4}$$

kde je w_i jinak než třídní četností určená váha *i*-té třídy

Dolní hranice	Horní hranice	Třídní reprezentant	Třídní četnost	Součtová četnost
28,75	32,95	30,85	5	5
32,95	37,15	35,05	13	18
37,15	41,35	39,25	15	33
41,35	45,55	43,45	15	48
45,55	49,75	47,65	5	53
49,75	53,95	51,85	1	54
53,95	58,15	56,05	1	55

Tabulka 4.1 - Třídění dat z tabulky 3.1

Hovořili jsme již dříve o tom, že při třídění musíme zachovat podstatné informace o statistickém souboru. Informace o poloze je podstatná. Musí tedy být zaručeno, že rozdíly mezi jednoduchým a váženým aritmetickým průměrem jsou nepodstatné. Tedy teoreticky platí $\bar{x}_1 = \bar{x}_2$, tedy by mělo platit $\sum_{i=1}^{N} x_i = \sum_{i=1}^{m} n_i \cdot \bar{x}_i$. Nejvyšší možná odchylka mezi prostým a váženým aritmetickým průměrem je rovna polovině třídního intervalu a vznikla by tehdy, jestliže

by všechny hodnoty zahrnuté do tříd ležely na jejich dolní nebo horní hranici. V praxi jsou ovšem hodnoty ve třídách rozděleny náhodně kolem třídního reprezentanta, a proto je výsledná chyba zpravidla minimální. Tuto vlastnost si ukážeme na následujícím příkladu:

Příklad 4.2:

Vypočítejte jednoduchý aritmetický průměr pro data z tabulky 3.1. Poté měřené hodnoty vhodně setřiďte a vypočítejte vážený aritmetický průměr. Oba průměry porovnejte.

Výpočet jednoduchého aritmetického průměru je zřejmý – sečteme všechny měřené tloušťky (získáme součet 2189), podělíme počtem hodnot (55) a získáme hodnotu 39,8 cm.

Třídění tohoto souboru je uvedeno v tabulce 4.1 Zde postupujeme tak, že vždy vynásobíme třídního reprezentanta dané třídy s příslušnou třídní četností, tj. např. 30,85*5, 35,05*13 atd. a výsledky těchto součinů sečteme. Získáme hodnotu 2196,55, kterou analogicky jako u jednoduchého průměru vydělíme počtem prvků (55) a získáme hodnotu 39,9 cm. Vzhledem k měřené veličině – tloušťce stromů – je výsledný rozdíl 0,1 cm zanedbatelný a vidíme, že z hlediska polohy získáme z tříděného souboru prakticky stejnou informaci jako z netříděného.

Důležité vlastnosti aritmetického průměru:

- aritmetický průměr je využitelný u znaků, ve kterých má logický smysl operace $\sum x_i$.
- určující vlastnosti aritmetického průměru je stálost součtu hodnot $\sum_{i=1}^{N} x_i$. Proto musí být $\sum_{i=1}^{N} x_i = \sum_{i=1}^{m} n_i \cdot \bar{x}_i$, aby chyby při určení váženého aritmetického průměru byly zanedbatelné. Toho lze dosáhnout, když třídní reprezentant co nejlépe zastupuje hodnoty ve třídě. Uvedená vlastnost také znamená $\sum_{i=1}^{N} x_i = N \cdot \bar{x}$.
- platí, že $\sum_{i=1}^{N} (x_i - \bar{x}) = 0$,
- platí, že $\sum_{i=1}^{N} (x_i - \bar{x})^2$ je minimální, tedy platí že $\sum_{i=1}^{N} (x_i - \bar{x})^2 < \sum_{i=1}^{N} (x_i - a)^2$ pro, $a \in S, a \neq \bar{x}$.

Důkazy těchto tvrzení viz např. ZACH (1994).

Z uvedených vlastností plyne:

- aritmetický průměr reprezentuje jednoduchý vztah mezi celkem a jeho částmi - všude tam, kde bude vystupovat hodnota souboru, můžeme ji nahradit aritmetickým průměrem souboru s nejmenším rizikem chyby,
- aritmetický průměr splňuje všechny podmínky kladené na dobrou charakteristiku polohy:
  - je vypočítán,
  - do výpočtu vstupují všechny hodnoty souboru,
  - jeho úloha je snadno pochopitelná,
  - způsob výpočtu je snadný,

- teoreticky je známo rozdělení aritmetických průměrů výběrových souborů z daného základního statistického souboru (blíže bude objasněno v části věnované matematické statistice).

Aritmetický průměr není vhodné použít tehdy, když současně rozsah souboru je malý (obvykle méně než 10 prvků), soubor obsahuje extrémní hodnoty, rozdělení souboru je výrazně nenormální a další úvahy nejsou založené na přímém početním odvození charakteristiky polohy.

4.2.2 Medián ($\tilde{x}$)

Medián je hodnota, kterou nese prostřední prvek v statistickém souboru uspořádaném podle velikosti. Rozděluje počet hodnot uspořádaného souboru na dvě poloviny.

Neroztříděný soubor je tedy pro určení mediánu vždy nutné uspořádat podle velikosti.

Hodnotu mediánu v neroztříděném, pouze vzestupně uspořádaném souboru, určíme takto:

$$\tilde{x} = \begin{cases} x_{\left(\frac{N+1}{2}\right)} & \text{pro N liché} \\ \frac{1}{2}\cdot\left(x_{\left(\frac{N}{2}\right)} + x_{\left(\frac{N}{2}+1\right)}\right) & \text{pro N sudé} \end{cases} \tag{4.5}$$

Znamená to tedy, že pro soubor s lichým počtem hodnot je medián roven přímo prostřednímu prvku, v souboru se sudým počtem hodnot (zde žádný prvek není „prostřední“) se stanoví jako průměrná hodnota dvou „prostředních“ prvků.

Z roztříděného zápisu se medián určí postupem:

- určí se číslo prostředního prvku $\frac{N+1}{2} = n_{(\tilde{x})}$
- z tabulky součtových četností se určí mediánová třída $i(\tilde{x})$, platí

  $$\nu_{i-1} < \frac{N+1}{2} \leq \nu_{i(\tilde{x})}$$

- podle následujícího vzorce (za předpokladu lineárního rozložení počtu hodnot ve třídě)

$$\tilde{x} = \left(\overline{x}_{i(\tilde{x})} - \frac{h}{2}\right) + \frac{h}{n_{i(\tilde{x})}}\left(\frac{N+1}{2} - \sum_{i=1}^{i(\tilde{x})-1} n_i\right) \tag{4.6}$$

kde je

$\overline{x}_{i(\tilde{x})}$ třídní reprezentant mediánové třídy

h třídní interval

$n_{i(\tilde{x})}$ absolutní třídní četnost mediánové třídy

$\sum_{i=1}^{i(\tilde{x})-1} n_i$ součtová četnost po třídu předcházející mediánové třídě (např. je-li mediánová třída šestá, je to součet třídních četností tříd 1. –5.)

N rozsah souboru

Výpočet mediánu roztříděného souboru je pouze orientační a může se od skutečné hodnoty mediánu lišit. Ilustruje to následující příklad.

Příklad 4.3:

Určete medián souboru dat z tabulky 3.1 a poté pro roztříděný soubor podle tabulky 4.1

Zadaný soubor má 55 prvků, tedy lichý počet. Potom podle vztahu 4.5 určíme pořadové číslo mediánu podle „horního“ vzorce $\tilde{x} = x_{\left(\frac{55+1}{2}\right)} = x_{(28)}$, tedy medián je 28. prvek vzestupně uspořádaného souboru. Podle tabulky 3.2 je to hodnota 39,5.

Nyní si budeme postupovat podle třídění v tabulce 4.1 a podle návodu v této kapitole.

Pořadové číslo mediánu určíme stejně – 28. prvek. Podle sloupce součtových četností si určíme mediánovou třídu – tedy tu, kam patří pořadové číslo mediánu. V našem případě to je třetí třída, kam patří prvky s pořadovými čísly 19 – 33. Třídní interval v této tabulce je 4,2, reprezentant mediánové třídy je 39,25 a její četnost je 15 prvků. Součtová četnost tříd předcházejících mediánové třídě je 18.

Na základě těchto údajů dosadíme do vzorce 4.6:

$$\tilde{x} = \left(35{,}05 - \frac{4{,}2}{2}\right) + \frac{4{,}2}{15}\left(\frac{55+1}{2} - 18\right) = 39{,}95\ cm$$

V tomto konkrétním případě vyšel odhad mediánu poměrně blízko skutečné hodnoty, ale v jiných případech (v závislosti na použitém třídění) může být rozdíl mezi mediánem netříděného a tříděného souboru značný.

Z vlastností mediánu je nutno vyzvednout jeho **nezávislost na extrémních hodnotách v souboru**. Nelze s ním provádět počtářské operace v takové míře jako s aritmetickým průměrem. Medián nesplňuje většinu požadavků na dobrou charakteristiku, dále necitlivost mediánu k velikosti jednotlivých hodnot je nevýhodou při srovnávání úrovně v několika souborech, proto se používá pouze tehdy, jestliže nelze použít aritmetický průměr nebo je použití mediánu vhodnější vzhledem k cíli analýzy. Používá se tam, kde je nutno vzít do úvahy spíše rozdělení hodnot na (dvě) části než nahrazení všech hodnot hodnotou střední (např. pro účely sortimentace dříví). Jestliže víme, že medián je vysoký, máme jistotu, že druhá polovina hodnot je vysokých. Je-li medián nízký, máme jistotu, že první polovina hodnot je nízkých. Tak jako ostatní kvantily je vhodný pro porovnávání hodnot souboru s danou normou, hranicí (typ otázky: kolik % hodnot souboru odpovídá normě?).

4.2.3 Modus ($\hat{x}$)

Modus je hodnota, která se ve vyšetřovaném souboru nejčastěji vyskytuje.

Z neroztříděného a uspořádaného zápisu se určí modus odečtem nejčastější hodnoty. V tomto případě rozlišujeme soubory:

- amodální – nemají modus (všechny hodnoty se vyskytují stejně často, např. jen jednou),
- unimodální – soubor má jeden modus (pouze jedna hodnota je nejčastější),
- polymodální – soubor má více modů (více hodnot je nejčastějších).

Z roztříděného souboru určíme modus výpočtem jako souřadnice x maxima křivky, která co nejlépe přiléhá polygonu rozdělení třídních četností. Ve většině praktických případů stačí za křivku volit parabolu druhého stupně. Při této volbě platí vzorec

$$\hat{x} = \bar{x}_0 + \frac{h}{2} \cdot \frac{p-1}{2s-p-1} \qquad (4.7)$$

kde je

$\bar{x}_0$ střed třídy s největší četností

s četnost třídy s největší četností

p četnost třídy vpravo od třídy s největší četností

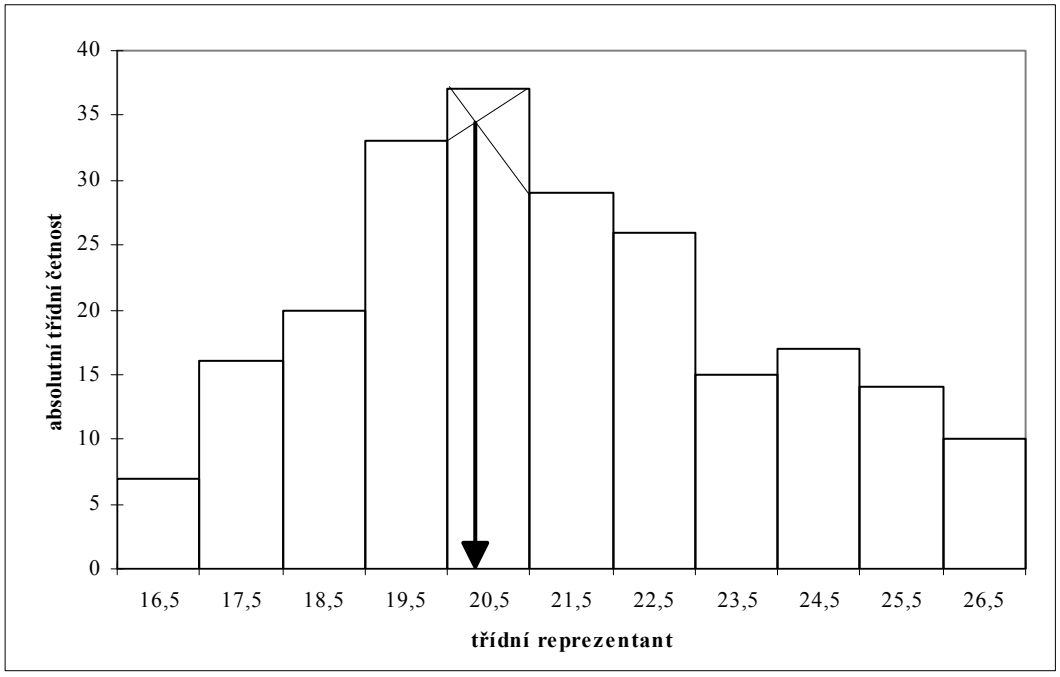
l četnost třídy vlevo od třídy s největší četností

Určení modu z roztříděného souboru je jen velmi orientační (např. se zde nedá zohlednit fakt možné polymodality souboru) a používá se jen ojediněle.

Graficky se modus určuje z histogramu rozdělení třídních četností v souladu s úvahou početního zjištění modu konstrukcí, která je zřejmá z obrázku 4.5 .

Z vlastností modu nutno vědět, že modus nezáleží na krajních hodnotách, k výpočtu není nutné znát všechny hodnoty souboru, nezávisí na celkovém počtu prvků. Nelze s ním provádět početářské operace možné u aritmetického průměru. Používá se tam, kde je významný právě největší počet určité hodnoty. To se stane např. u nesouměrného rozdělení četností (mj. dále), kdy aritmetický průměr a modus nevystihují plně celkovou charakteristiku souboru. Potom modus má důležitou úlohu typické hodnoty. Při porovnání souborů podle typických hodnot je nezastupitelný.

U roztříděného souboru mluvíme často o modálním intervalu. Přitom je třeba předpokládat, že třídní interval je konstantní. Modální interval je závislý na třídění.



Obrázek 4.5 - Grafický způsob určení modu

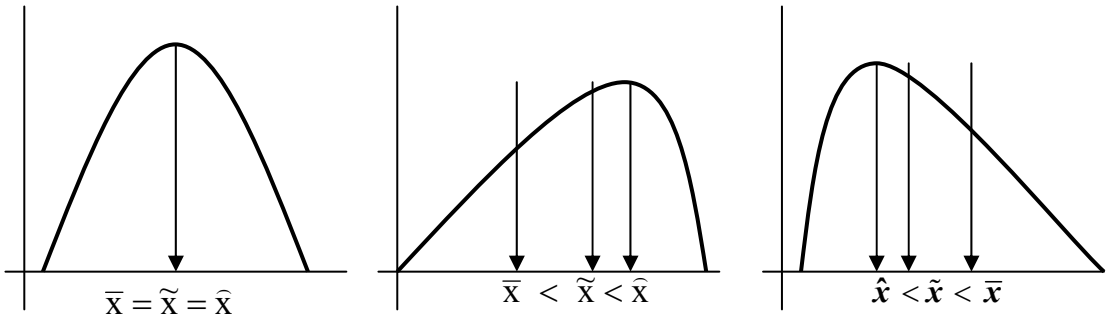
4.2.4 Vztahy mezi charakteristikami polohy

Mezi hodnotami charakteristik polohy existují vztahy, které mají logický smysl a význam vyplývající z vlastností souboru a z definicí charakteristik. V prvním pohledu, ze kterého se dají logicky odvozovat pro konkrétní soubor další tvrzení, jde o přesnou závislost seřazení průměru, mediánu a modu pro souměrná či nesouměrná rozdělení četností.

Platí:

1. V dokonale souměrném rozdělení je $\overline{x} = \tilde{x} = \hat{x}$, tj. střední hodnota se rovná prostřední i nejčastější hodnotě
2. V mírně nesouměrných rozděleních je $\overline{x} - \hat{x} \cong 3(\overline{x} - \tilde{x})$
3. V silně nesouměrných rozděleních je $\tilde{x} \in <\overline{x}, \hat{x}>$
4. Nastane-li případ 2. nebo 3. potom, když
   a) $\hat{x} < x < \overline{x}$ jde o levostranně nesouměrné rozdělení
   b) $\overline{x} < x < \hat{x}$ jde o pravostranně nesouměrné rozdělení

Uvedené vztahy ilustruje obrázek 4.6 , kde levý obrázek představuje souměrný soubor (všechny základní charakteristiky polohy se sobě rovnají), prostřední obrázek představuje pravostranný soubor (s převahou hodnot vyšších než je aritmetický průměr, tedy koncentrovaných napravo od průměru, odtud název), pravý obrázek představuje druhý možný případ nesouměrnosti – levostranný soubor.



Obrázek 4.6 - Vztahy mezi charakteristikami polohy

4.3 Charakteristiky variability

Charakteristiky variability informují o tom, jak jsou jednotlivé hodnoty souboru rozptýleny, tj. jak se jednotlivé hodnoty znaku liší vzhledem k sobě navzájem nebo vzhledem ke střední hodnotě.

Při konstrukci míry variability řady hodnot znaku se postupuje dvěma způsoby:

- pokud se chce vyjádřit variace pouze ve smyslu vzájemné odlišnosti jednotlivých hodnot znaku, vychází se ze vzájemných odchylek od každé hodnoty jednotlivě v souhrnu,
- pokud se chce měřit variace ve smyslu odlišnosti od střední hodnoty, vychází se z odchylek všech hodnot znaku od této střední hodnoty.

Charakteristiky variability se používají v podobě

- absolutní - mají rozměr studované veličiny
- relativní (poměrné) - bez rozměru nebo v procentech. Jsou vhodné pro porovnání variability různých souborů.

Požadavky na dobrou míru variability:

- být přesně určena, ne odhadnuta,
- zakládat se na všech hodnotách souboru,

- mít zřetelnou vlastnost, podle které lze zaručit dobrou schopnost vystihnout variabilitu jevu, tj. dosáhnout, aby vyšší číselná hodnota charakteristiky skutečně znamenala větší variabilitu souboru a naopak,
- snadný způsob výpočtu a možnost početních operací.

Variabilita souboru je v teorii i praxi statistiky velmi důležitá a využívaná vlastnost. Proto je jí v našem kursu věnována větší pozornost než ostatním vlastnostem.

4.3.1 Variační rozpětí

Variační rozpětí je rozdíl mezi maximální a minimální hodnotou souboru vyjádřený buď absolutně v jednotkách měřené veličiny nebo relativně.

$$R = x_{max} - x_{min} \tag{4.8}$$

$$R\% = \frac{x_{max} - x_{min}}{\bar{x}} \cdot 100 \tag{4.9}$$

Ve speciálních případech (kdy jako míru polohy nepoužíváme aritmetický průměr) je možné ve vzorci 4.9 nahradit aritmetický průměr jinou mírou polohy (mediánem, modem apod.).

Variační rozpětí vyjadřuje pouze šířku intervalu možných hodnot. Relativní míry se použijí souhlasně k odpovídající míře polohy. Méně variabilní je soubor s menším R%. Slouží k rychlému posouzení variability. Není vhodné pro nehomogenní soubory (tj. pro soubory se zřetelně odlehlými krajními hodnotami, které vzhledem k ostatním hodnotám velmi zvyšují hodnotu variačního rozpětí.

4.3.2 Průměrná odchylka

Vzorec pro netříděný soubor

$$\bar{d}_1 = \frac{\sum_{i=1}^{N} |x_i - \bar{x}|}{N} \tag{4.10}$$

Vzorec pro roztříděný soubor

$$\bar{d}_2 = \frac{\sum_{i=1}^{m} n_i |\bar{x}_i - \bar{x}|}{N} \tag{4.11}$$

Relativní průměrná odchylka

$$\bar{d}\% = \frac{\bar{d}}{\bar{x}} .100\ (\%) \tag{4.12}$$

Stejně jako v případě variačního rozpětí lze počítat odchylky nejen k aritmetickému průměru, ale v případě potřeby i k ostatním mírám polohy, a to jak v absolutním, tak i v relativním vyjádření.

Průměrná odchylka závisí na všech hodnotách znaku. Udává průměrnou velikost odchylek všech hodnot od zvolené střední hodnoty. Není možné provádět početní operace s průměrnými odchylkami při operacích se soubory. Nepodává informace o rozdělení hodnot v souboru. Nejčastěji se počítá průměrná odchylka od aritmetického průměru, i když logicky je využitelnější průměrná odchylka od mediánu. Platí totiž, že průměrná odchylka od mediánu je nejmenší ze všech průměrných odchylek počítaných od středních hodnot.

Průměrná odchylka se používá jako doplněk ostatních středních hodnot (nepodává informace o odlišnosti hodnot mezi sebou).

4.3.3 Rozptyl

Je definován vzorcem

$$S^2 = \operatorname{var} X = \frac{\sum_{j=1}^{N} \left(x_j - \bar{x}\right)^2}{N} \tag{4.13}$$

Rozměr rozptylu je kvadrát jednotky veličiny.

Pro roztříděný soubor se počítá podle vzorce

$$S^2 = \frac{\sum_{i=1}^{m} n_i \left(\bar{x}_i - \bar{x}\right)^2}{N} \tag{4.14}$$

Analogicky jako v případě jednoduchého a váženého aritmetického průměru se zde předpokládá, že platí $\sum_{j=1}^{N} \left(x_j - \bar{x}\right)^2 = \sum_{i=1}^{m} n_i \left(\bar{x}_i - \bar{x}\right)^2$. Protože tomu tak zpravidla není, dopouštíme se určité systematické chyby, která, jak bylo zjištěno, je přímo závislá na šířce třídy (třídním intervalu - h). Tuto chybu lze minimalizovat Shepardovou korekcí podle vztahu

$$S_{kor}^2 = S^2 - \frac{1}{12} h^2 \tag{4.15}$$

Rozptyl je tedy aritmetický průměr čtverců odchylek od $\bar{x}$ a je tedy konstruován k vyjádření variability hodnot kolem aritmetického průměru, ale vyjadřuje i vzájemnou odlišnost hodnot znaku. Měří tedy obě uvažované možnosti variability souboru. Důkaz tohoto tvrzení viz např. ZACH (1994).

Vlastnosti rozptylu:

1. Součet čtverců hodnot znaku od aritmetického průměru je minimální.
2. Hodnota rozptylu souboru $S_1 = \{x_i + k; N\}$

   $S_1^2 = \frac{1}{N} \sum_i \left[\left(x_i + k\right) - \left(\bar{x} + k\right)\right]^2 = \frac{1}{N} \sum \left[x_i - \bar{x}\right]^2$, tj. rozptyl souboru o rozsahu N, jehož všechny prvky byly zvětšeny o konstantu k, je stejný jako rozptyl původního souboru.
3. Hodnota rozptylu souboru $S_2 = \{x_i \cdot k; N\}$

$S_2^2 = \frac{1}{N}\sum(x_1 \cdot k - \bar{x} \cdot k)^2 = \frac{1}{N}k^2\sum_i(x_i - \bar{x})^2$, tj,. rozptyl souboru o rozsahu N, jehož všechny prvky byly vynásobeny konstantou k, se zvětší k^2-krát.

4. Hodnota rozptylu souboru $S_3 = \{x_i + y_i; N\}$

$$S_3^2 = \frac{1}{N}\sum_i[(x_i + y_i) - (\bar{x} + \bar{y})]^2 = \frac{1}{N}\sum_i[(x_i - \bar{x}) + (y_i - \bar{y})]^2 =$$

$$= \frac{1}{N}\sum_i\left[(x_i - \bar{x})^2 + (y_i - \bar{y})^2 + 2(x_i - \bar{x})(y_i - \bar{y})\right] = Sx^2 + Sy^2 + 2.\mathrm{cov}_{xy}$$

Výraz $\frac{1}{N_i}\sum(x_i - \bar{x})(y_i - \bar{y}) = \mathrm{cov}_{xy}$ se nazývá kovariance. Často se s ní budeme při dalším studiu setkávat.

Rozptyl splňuje všechny požadavky na dobrou charakteristiku variability. Jeho použití je mnohostranné (zvláště v matematické statistice, v popisné statistice se často nahrazuje směrodatnou odchylkou, která je vyjadřována v jednotce měřené veličiny) a bude v hlavních aplikacích vysvětleno v příslušných kapitolách.

4.3.4 Směrodatná odchylka

Je definována odmocninou rozptylu

$$S = \sqrt{\mathrm{var}\,X} = \sqrt{\frac{\sum_{j=1}^{N}(x_j - \bar{x})^2}{N}} \tag{4.16}$$

Definiční vzorec lze jednoduchým postupem upravit na vzorec vhodnější k výpočtu. Platí

$$S = \sqrt{\frac{\sum_{j=1}^{N}(x_j - \bar{x})^2}{N}} = \sqrt{\frac{\sum_{j=1}^{N}x_j^2}{N} - \bar{x}^2} \tag{4.17}$$

Pro roztříděný soubor platí obdobně

$$S = \sqrt{\frac{\sum_{j=1}^{k}n_i(\bar{x}_i - \bar{x})^2}{N}} = \sqrt{\frac{\sum_{j=1}^{k}n_i\bar{x}_i}{N} - \bar{x}^2} \tag{4.18}$$

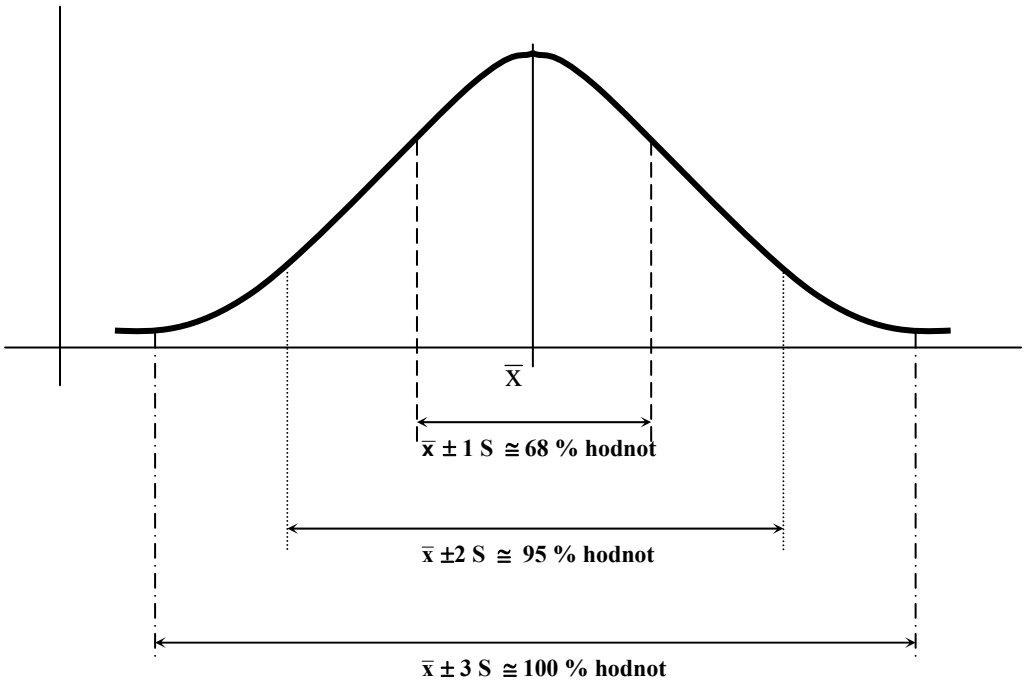
Směrodatná odchylka je nejlepší a nejpoužívanější charakteristikou variability. Splňuje všechny požadavky na dobrou charakteristiku variability. Rozměr směrodatné odchylky je stejný jako rozměr veličiny, což je její hlavní výhodou oproti rozptylu pro účely popisné statistiky. Požadavek dobré a snadno pochopitelné schopností podávat informace o variabilitě je u směrodatné odchylky výborně naplněn. U souboru s normálním rozdělením četností totiž platí, že v určitých intervalech daných násobky směrodatné odchylky kolem aritmetického průměru je určitá část počtu hodnot:

- v rozmezí $\bar{x} \pm 1S$ je to asi 68 % všech hodnot,

- v rozmezí $\bar{x} \pm 2S$ je to asi 95 % všech hodnot,
- v rozmezí $\bar{x} \pm 3S$ je to asi 100 % všech hodnot.

Názorně je tato skutečnost vidět na obrázku 4.7 . Silná plná čára představuje normální rozdělení (o tomto i jiných rozděleních podrobněji viz kapitola o spojitých rozděleních náhodných veličin), které se vyznačuje mimo jiné souměrností a tím, že aritmetický průměr je nejčastější hodnotou. Tak jako i u jiných grafů rozdělení četností jsou i zde na ose *x* vynášeny hodnoty (vlastně číselná osa) a na ose *y* jejich četnosti. Celá plocha pod křivkou představuje 100 %-ní pravděpodobnost výskytu všech hodnot. První interval (znázorněný čárkovanou čárou) představuje (vzhledem k vysokým četnostem hodnot zahrnutým v tomto intervalu) 68 % všech hodnot souboru. Dále směrem k okrajům se četnosti hodnot prudce snižují, takže v intervalu $\pm$ 2 S je 95 % hodnot a konečně v intervalu $\pm$ 3 S (čerchovaný interval) leží přibližně 100 % hodnot souboru. Tato skutečnost (také nazývaná pravidlo tří nebo šesti směrodatných odchylek) může pomoci např. při odhadu možného rozpětí hodnot jejich vzájemných početních proporcí. Např. jestliže v porostu očekáváme normální rozdělení výčetních tlouštěk stromů a víme, že průměrná tloušťka je 38,0 cm a směrodatná odchylka tlouštěk 5,5 cm, můžeme usuzovat, že přibližně 68 % stromů bude mí výčetní tloušťku v intervalu 32,5 – 42,5 cm, 95 % všech tlouštěk padne do rozmezí 27 – 49 cm a prakticky žádná tloušťka nebude menší než 21,5 cm a větší než 54,5 cm.

Podrobněji bude o této problematice pojednáno v příslušných partiích matematické statistiky.



Obrázek 4.7 – Grafické znázornění pravidla 3 směrodatných odchylek

4.3.5 Variační koeficient

Variační koeficient je relativní mírou variability a používá se k vzájemnému porovnávání variability různých souborů. Vychází z průměru a směrodatné odchylky, má tedy vlastnosti obdobné jako S a S^2. Čím je S % menší, tím je menší variabilita souboru.

Stanoví se podle vzorce

$$S\% = \frac{S}{\bar{x}} \cdot 100 \tag{4.19}$$

4.3.6 Kvantilové odchylky

Kvantilové odchylky se používají spolu s mediánem. Výhodou kvantilů a z nich odvozených charakteristik je, že k jejich určení nepotřebujeme znát všechny hodnoty souboru. Stačí seřadit prvky (pokud to lze) podle velikosti, určit ty, které jsou nositeli kvantilů, a tyto proměřit. Obecně lze říci, že kvantilové odchylky jsou horší mírou variability než klasické charakteristiky (nesplňují většinu požadavků kladených na míry variability). Vypovídají, ve shodě s definicí kvantilu, o průměrné rozdílnosti stejně početných skupin hodnot a používají se tam, kde nelze použít klasické charakteristiky, především rozptyl a směrodatnou odchylku. Jako příklad může uvést kvartilovou odchylku:

$$Q = \frac{(\tilde{x}_{75} - \tilde{x}) + (\tilde{x} - \tilde{x}_{25})}{2} = \frac{\tilde{x}_{75} - \tilde{x}_{25}}{2} \tag{4.20}$$

kde je

$\tilde{x}_{25}$ dolní kvartil (odděluje dolních 25 % hodnot)

$\tilde{x}$ medián

$\tilde{x}_{75}$ horní kvartil

Jedná se tedy o polovinu rozpětí horního a dolního kvartilu. Záleží tedy na velikosti rozpětí prostředí poloviny prvků.

Podobně lze konstruovat např. decilovou odchylku (a další):

$$D = \frac{(\tilde{x}_{90} - \tilde{x}_{80}) + (\tilde{x}_{80} - \tilde{x}_{70}) + \ldots\ldots + (\tilde{x}_{20} - \tilde{x}_{10})}{8} = \frac{\tilde{x}_{90} - \tilde{x}_{10}}{8} \tag{4.21}$$

4.4 Charakteristiky tvaru

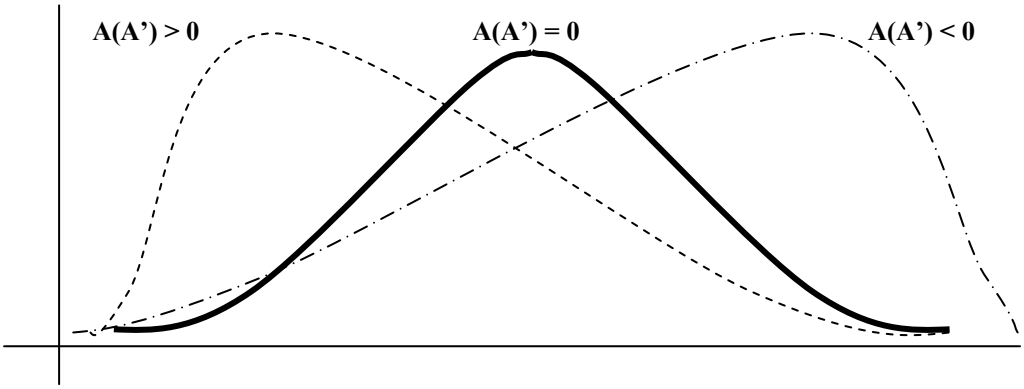
4.4.1 Míry nesouměrnosti

Nesouměrnost (též asymetrie, nebo šikmost) se projevuje tím, že v souboru je více hodnot menších než větších ve srovnání se střední hodnotou (levostranná nesouměrnost) nebo více hodnot větších než menších ve srovnání se střední hodnotou (pravostranná nesouměrnost).

Požadavky na míru nesouměrnosti:

- dobře měřit nesouměrnost, tj. pro souměrný soubor by měla být rovna 0, u levostranné, resp. pravostranné, nesouměrnosti mít opačná znaménka,
- aby byla bezrozměrným číslem, což umožní měření nesouměrnosti souborů nezávisle na jednotkách,

- aby ji bylo možno snadno určit.



Obrázek 4.8 – Grafické znázornění levostranného (čárkovaně), souměrného (plně) a pravostranného (čerchovaně) rozdělení četností

4.4.1.1 Pearsonova míra nesouměrnosti

Je nejméně výstižnou mírou nesouměrnosti a používá se k rychlému posouzení nesouměrnosti. Stanoví se podle vztahu

$$A' = \frac{\bar{x} - \hat{x}}{S} \cong \frac{3 \cdot (\bar{x} - \tilde{x})}{S}$$

Pokud je $A' = 0$ je rozdělení četností souměrné
$A' > 0$ levostranně nesouměrné
$A' < 0$ pravostranně nesouměrné

4.4.1.2 Kvantilové míry nesouměrnosti

Využívají rozdílu rozptýlenosti mezi nižší polovinou a vyšší polovinou hodnot souboru, tedy na podobném principu jako kvantilové odchylky (viz kapitola 4.3.6). Používá se vzorec

$$a_i = \frac{(\tilde{x}_{100-i} - \tilde{x})(\tilde{x} - \tilde{x}_i)}{(\tilde{x}_{100-i} - \tilde{x}) + (\tilde{x} - \tilde{x}_i)} = \frac{\tilde{x}_{100-i} + \tilde{x}_i - 2\tilde{x}}{\tilde{x}_{100-i} - \tilde{x}_i} \tag{4.22}$$

kde je

$\tilde{x}_i$ dolní kvantil
$\tilde{x}_{100-i}$ horní kvantil
$\tilde{x}$ medián

Místo horního kvantilu můžeme vzít x_{max} dolní kvantil nahradit x_{min}. Potom

$$a = \frac{x_{max} + x_{min} - 2\tilde{x}}{x_{max} - x_{min}} \tag{4.23}$$

Interpretace hodnot a_i je stejná jako u Pearsonovy míry nesouměrnosti. Hodnoty a_i, a se pohybují v rozmezí -1 až +1. Je tedy možné přesněji posuzovat velikost nesouměrnosti. Podstatnou nevýhodou je závislost na extrémních hodnotách. Tento nedostatek lze odstranit vhodně volenou mírou a_i (tj. volit kvantil x_i tak, aby vyloženě extrémní krajní hodnoty nebyly do výpočtu zahrnuty).

Používají se v úlohách, kde využíváme mediánu a kvantilových odchylek.

4.4.1.3 Koeficient nesouměrnosti

Je nejlepší mírou nesouměrnosti, která je vypočítána ze všech hodnot souboru za použití nejlepších charakteristik polohy a variability. Používá se přednostně všude, kde je vyžadována plnohodnotná informace o nesouměrnosti a je možné pracovat s aritmetickým průměrem a směrodatnou odchylkou.

Pro neroztříděný soubor

$$A = \frac{\sum_{j=1}^{N}\left(x_j - \bar{x}\right)^3}{N \cdot S^3} \tag{4.24}$$

Pro roztříděný soubor

$$A = \frac{\sum_{i=1}^{m} n_i \left(\bar{x}_i - \bar{x}\right)^3}{N \cdot S^3} \tag{4.25}$$

Interpretace hodnot koeficientu nesouměrnosti je následující:

Je-li $A = 0$ je rozdělení souměrné,
$A < 0$ pravostranné,
$A > 0$ levostranné.

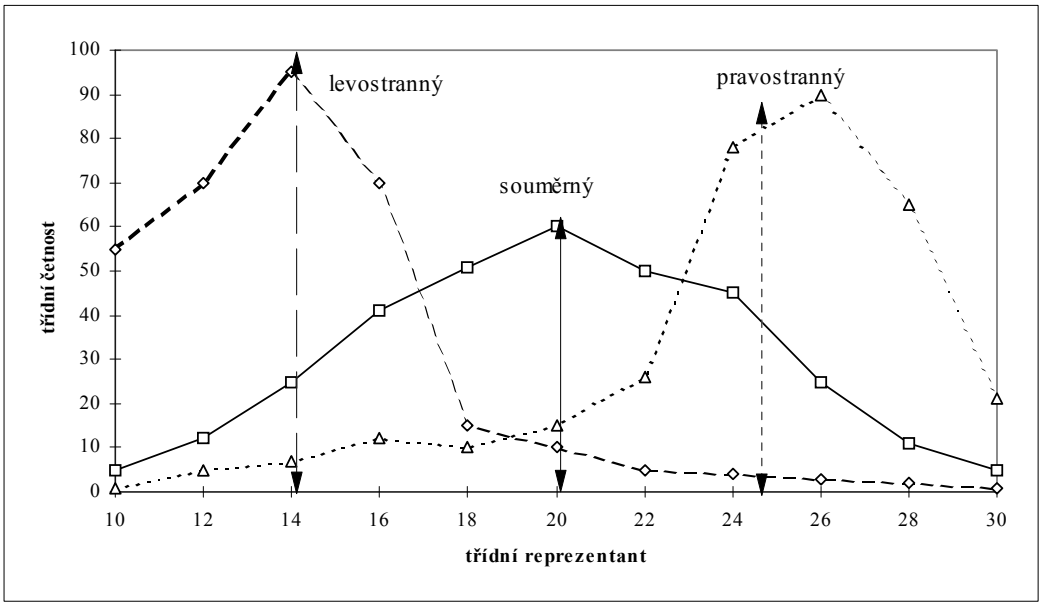
Nesouměrnost je tím větší, čím více se A odlišuje od nuly. Pravidla interpretace vycházejí z toho, že koeficient nesouměrnosti je založen na třetích mocninách odchylek hodnot od aritmetického průměru, takže odchylky si ponechávají svoje znaménko (+ nebo -) a v celkovém součtu převáží ty, které jsou větší. Tedy je možné zobecnit, že pravostranně nesouměrné rozdělení znamená, že hodnot vyšších něž aritmetický průměr je více než hodnot nižších. Levostranně nesouměrné rozdělení znamená, že hodnot nižších než aritmetický průměr je více než hodnot vyšších. Tento princip je ilustrován v tabulce 4.2 a na obrázku 4.9 .
V tabulce 4.2 jsou generovány tři roztříděné soubory – jeden velmi výrazně levostranný (L), jeden přibližně souměrný (S) a jeden velmi výrazně pravostranný (P). Pro tyto soubory byly vypočítány příslušné vážené statistické charakteristiky $\bar{x}$ a S. Je zřejmé, že nesouměrnost souboru velmi ovlivňuje polohu aritmetického průměru (vyznačeno na obrázku 4.9 oboustrannými šipkami příslušného typu čáry). Z tohoto obrázku je zřejmé, že např. u levostranného rozdělení bude kladný součet součinů odchylek třídních reprezentantů od aritmetického průměru a třídních četností výrazně vyšší než součet záporných odchylek, u pravostranného rozdělení tomu bude naopak a u souměrného rozdělení se oba typy odchylek budou vyrovnávat. Je nutno podotknout, že zde byly pro ilustraci generovány skutečně extrémně nesouměrné soubory. V běžné praxi se za výraznou nesouměrnost považují již hodnoty kolem 1 a více.

A je bezrozměrná veličina (resp. vyjádřená v jednotkách směrodatné odchylky), což umožňuje bezproblémové porovnávání různých souborů.

TR	Třídní četnost (NI)			Levostranný		Souměrný		Pravostranný	
	L	S	P	OD	Suma	OD	Suma	OD	Suma
10	55	5	1	-4022,149		-5027,322		-3027,648	
12	70	12	5	- 727,032	-4749,752	-6185,986		-9687,708	
14	95	25	7	- 0,571		-5449,240	-19741,722	-8026,445	
16	70	41	12	420,736		-2659,945		-7283,140	-32464,024
18	15	51	10	834,951		- 419,229		-2704,216	
20	10	60	15	1969,527		0,000		-1336,724	
22	5	50	26	2389,391	23587,174	389,190		- 390,216	
24	4	45	78	3785,761		2840,906		- 7,927	
26	3	25	90	4951,916		5351,058	19147,617	324,453	
28	2	11	65	5276,947		5593,687		2867,261	6749,500
30	1	5	21	3957,944		4972,777		3557,786	
N	330	330	330		18837,421		- 594,105		-25714,524
X	14,18	20,02	24,47						
S	3,41	4,28	3,96						
A	16,76	- 0,42	-19,68						

VYSVĚTLIVKY: TR - třídní reprezentant, L - levostranný, S-souměrný, P-pravostranný, N -počet prvků, X- aritm. průměr, S-směrodatná odch., A - koef. nesouměrnosti, OD - $NI*(TR-X)^3$, Suma - součet záporných odchylek (tečkovaně), kladných (šedě), celkově (černě)

Tabulka 4.2 - Příklad výpočtu koeficientu zahrocenosti pro výrazně levostranný, přibližně souměrný a výrazně pravostranný soubor. Bližší vysvětlení viz v textu



Obrázek 4.9 – Grafické znázornění nesouměrných a přibližně souměrného souboru s vyznačením jejich vážených aritmetických průměrů (oboustranné šipky). Bližší vysvětlení viz v textu.

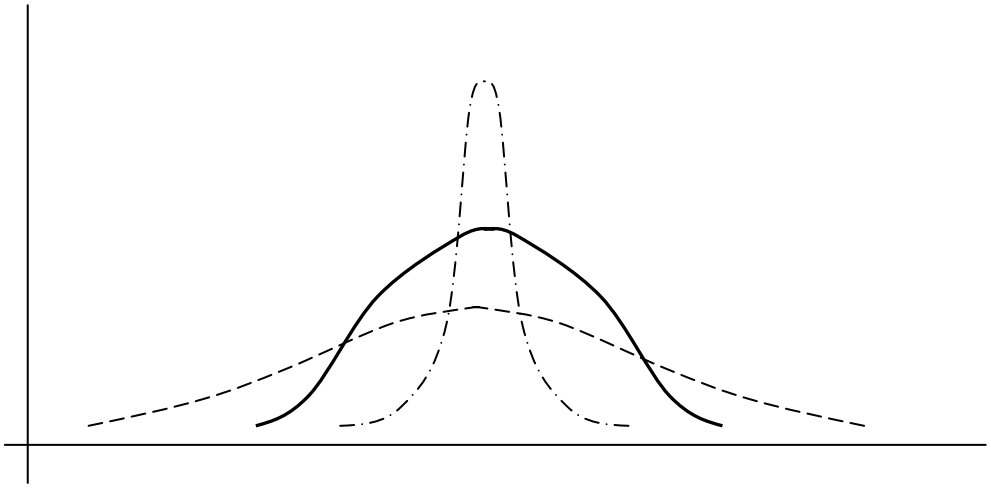
4.4.2 Míry zahrocenosti

Zahrocenost (též špičatost, koncentrace, exces) je druhá základní tvarová vlastnost rozdělení četností souboru. Jedná se o srovnání „výšky a strmosti kopce“ polygonu rozdělení četností (statisticky řečeno o srovnání koncentrace dat kolem určité skupiny hodnot) se zá-

kladním vzorem rozložení hodnot daným normálním rozdělením. Základní situaci ilustruje obrázek 4.10 .Četnosti jednotlivých hodnot souboru daných normálním rozdělením jsou znázorněny plnou silnou čarou. Soubor, který má četnosti jednotlivých hodnot nižší, než odpovídá normálnímu rozdělení (a zpravidla vyšší variabilitu – hodnoty jsou kolem střední hodnoty rozptýleny v širším intervalu), se nazývá plochý („kopec je nižší, širší a plošší“). Naopak soubor, jehož nejvyšší četnosti jsou koncentrovány u několika hodnot (na schématickém obrázku 4.10 kolem střední hodnoty) a jehož variabilita je zpravidla menší („kopec je vyšší, užší a špičatější“), se nazývá špičatý.

Hlavní požadavky na dobrou míru zahrocenosti:

- dobře měřit zahrocenost.- vyšší hodnota má znamenat špičatější rozdělení,
- aby byla bezrozměrným číslem, což umožní měření zahrocenosti nezávisle na jednotkách,
- aby ji bylo možno snadno určit.



Obrázek 4.10 - Schématické znázornění normálně zahroceného souboru (plná čára), plochého (čárkovaná) a špičatého (čerchovaná)

4.4.2.1 Míra koncentrace kolem mediánu

Využívá rozdílu rozptýlenosti všech hodnot znaku proti rozptýlenosti hodnot blízkých mediánu. Větší špičatost se projeví v poklesu rozpětí kvantilů a tím hodnota *e* vzroste. Vlastnosti odpovídají kvantilovému charakteru. Používá se společně s mediánem.

Stanoví se podle vzorce

$$e = \frac{x_{max} - x_{min}}{\tilde{x}_{75} - \tilde{x}_{25}} \qquad (4.26)$$

4.4.2.2 Koeficient zahrocenosti

Jeho výpočet je velmi podobný koeficientu nesouměrnosti (využívá čtvrtou mocninu odchylek hodnot od aritmetického průměru).

Pro neroztříděný soubor se stanoví podle vzorce

$$E = \frac{\sum_{j=1}^{N}\left(x_j - \bar{x}\right)^4}{N \cdot S^4} - 3 \qquad (4.27)$$

Pro roztříděný soubor

$$E = \frac{\sum_{i=1}^{m} n_i \left(\bar{x}_i - \bar{x}\right)^4}{N \cdot S^4} - 3 \qquad (4.28)$$

Je to nejlepší charakteristika zahrocenosti. Je vypočítán ze všech hodnot souboru, využívá nejlepších charakteristik polohy a variability.

Interpretace je také podobná koeficientu nesouměrnosti:

Je-li E = 0 je rozdělení normálně zahrocené
E < 0 ploché
E > 0 špičaté

V této souvislosti je nutné upozornit, že v literatuře se mnohdy uvádí velmi podobný vzorec, který ale nemá na konci odečtení hodnoty 3. Potom je interpretace posunuta tak, že normální soubor má E = 3, plochý soubor hodnotu menší než 3 a špičatý vyšší než 3. Zvláště při použití počítačových programů je nutné se ujistit, podle kterého vzorce je hodnota E počítána.

4.5 Příklad použití a interpretace statistických charakteristik

V předchozích kapitolách byl podán obecný výklad způsobu výpočtu, významu a interpretace statistických charakteristik. V této kapitole bude jejich použití ukázáno na jednoduchých příkladech a bude poukázáno na základní fakta, která lze z nich lze vypozorovat.

V tabulce 4.3 jsou uvedena data z měření výčetních tlouštěk (tloušťka měřená ve výšce 1,3 m nad zemí) ve dvou porostech. Naším úkolem je tyto soubory analyzovat pomocí základních charakteristik.

Pro Porost 1 je kompletní výpočet uveden v tabulce 4.4 .

V prvních dvou sloupcích je uvedeno zadání - měřené hodnoty (x_i). V následujících sloupcích (A – D) jsou uvedeny mezivýpočty pro jednotlivé statistické charakteristiky:

A pro aritmetický průměr $\bar{x}$ x_i

B pro rozptyl S^2 $\left(x_i - \bar{x}\right)^2$,

C pro koef. nesouměrnosti A $\left(x_i - \bar{x}\right)^3$,

D pro koef. zahrocenosti E $\left(x_i - \bar{x}\right)^4$.

V řádku „Suma“ jsou součty pro sloupce A – D a v následujícím řádku „Vypočítané charakteristiky“ jsou již vypočítané hodnoty jednotlivých statistických charakteristik. Pro výpočty A a E byla použita hodnota směrodatné odchylky 5,1072.

Tři sloupce na pravé straně zobrazující vzestupně uspořádaný soubor a prvky, ze kterých byl vypočítán medián (prvky 30 a 31 uspořádaného souboru) a také modus (hodnota 23,9, která se vyskytuje třikrát).

POROST 1						POROST 2					
Číslo stromu	$d_{1.3}$(cm)	Číslo stromu	$d_{1.3}$(cm)	Číslo stromu	$d_{1.3}$(cm)	Číslo stromu	$d_{1.3}$(cm)	Číslo stromu	$d_{1.3}$(cm)	Číslo stromu	$d_{1.3}$(cm)
1	22,0	21	25,7	41	37,6	1	12,1	21	20,7	41	13,9
2	26,5	22	27,9	42	37,7	2	12,8	22	19,0	42	56,3
3	26,7	23	31,8	43	23,9	3	14,0	23	21,1	43	
4	29,6	24	21,7	44	35,4	4	14,2	24	19,0	44	
5	27,3	25	23,9	45	32,4	5	14,5	25	21,3	45	
6	26,7	26	16,0	46	27,1	6	14,6	26	17,0	46	
7	37,3	27	29,5	47	30,3	7	14,8	27	22,1	47	
8	27,5	28	16,6	48	27,7	8	16,6	28	16,8	48	
9	27,0	29	22,3	49	29,1	9	18,2	29	22,7	49	
10	24,8	30	23,9	50	36,5	10	18,3	30	15,4	50	
11	38,4	31	29,6	51	28,2	11	18,4	31	50,2	51	
12	32,7	32	26,5	52	25,4	12	18,7	32	19,0	52	
13	27,0	33	32,2	53	23,8	13	18,8	33	59,5	53	
14	33,1	34	36,0	54	31,4	14	19,1	34	15,0	54	
15	27,3	35	27,6	55	27,4	15	19,2	35	24,4	55	
16	23,0	36	37,7	56	31,1	16	19,7	36	17,0	56	
17	27,8	37	30,4	57	33,6	17	19,7	37	24,6	57	
18	31,4	38	20,8	58	36,4	18	19,7	38	14,0	58	
19	24,6	39	32,3	59	26,4	19	56,5	39	13,8	59	
20	20,9	40	30,8	60	32,1	20	18,0	40	26,3	60	

Tabulka 4.3 - Měřené výčetní tloušťky ve dvou porostech

Pokud chceme interpretovat statistické charakteristiky, měli bychom si všimnout následujících hodnot a vztahů:

- aritmetického průměru a mediánu – především jejich vzájemné polohy, pokud jsou tyto hodnoty blízko sebe, svědčí to o nepřítomnosti extrémních hodnot a indikuje normalitu souboru; naopak velké rozdíly (vzhledem k měřítku a jednotkám hodnot) svědčí o opaku),
- hodnot koeficientů tvaru – A a E – a provést jejich interpretaci (přitom musíme dbát na to, abychom pouze formálně nehodnotili odchylku od nuly – podstatné jsou pouze značné odchylky, pokud jde pouze o několik desetin, můžeme takový soubor považovat na přibližně normální. V kapitole věnované testování statistických hypotéz budou uvedeny testy, které toto rozhodnutí objektivizují),
- hodnot variability – extrémní variabilita signalizuje odchylky od normality a přítomnost extrémních hodnot.

Porost 1 tedy můžeme hodnotit následujícím způsobem:

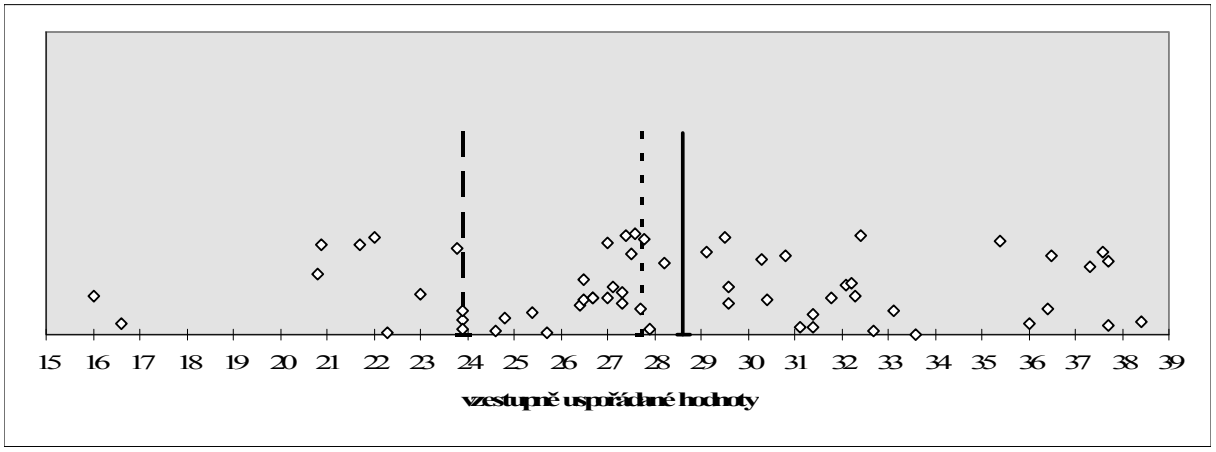
- hodnoty mediánu a aritmetického průměru jsou si poměrně blízké, svědčí to o nepřítomnosti typických extrémních hodnot (dvě nejnižší hodnoty jsou poněkud vzdálené od ostatních, ale nelze je ještě považovat za hodnoty, které by hodnotu průměru zásadním způsobem ovlivňovaly),
- hodnoty A a E svědčí o mírné plochosti a pravostrannosti souboru (znamená to, že v souboru je mírně více tlouštěk vyšších oproti aritmetickému průměru než by odpovídalo normálnímu rozdělení, přičemž variabilita je mírně větší, tj. četnosti jednotlivých tlouštěk jsou menší než odpovídá normalitě), ale odchylky od nuly nejsou podstatné, proto se tento soubor může považovat za velmi blízký normálnímu.

Výpočet statistických charakteristik						Vzestupně uspořádaný soubor		
Číslo stromu	$d_{1.3}$ (cm)	A	B	C	D	Vzestupné pořadí	Číslo stromu	$d_{1.3}$ (cm)
1	22,0	22,0	44,0675	- 292,5346	1941,9419	1	26	16,0
2	26,5	26,5	4,5725	- 9,7775	20,9075	2	28	16,6
3	26,7	26,7	3,7571	- 7,2826	14,1161	3	38	20,8
4	29,6	29,6	0,9248	0,8894	0,8553	4	20	20,9
5	27,3	27,3	1,7911	- 2,3971	3,2082	5	24	21,7
6	26,7	26,7	3,7571	- 7,2826	14,1161	6	1	22,0
7	37,3	37,3	75,0245	649,8369	5628,6710	7	29	22,3
8	27,5	27,5	1,2958	- 1,4751	1,6791	8	16	23,0
9	27,0	27,0	2,6841	- 4,3975	7,2046	9	53	23,8
10	24,8	24,8	14,7328	- 56,5494	217,0555	10	25	**23,9**
11	38,4	38,4	95,2901	930,1905	9080,2100	11	30	**23,9**
12	32,7	32,7	16,4971	67,0059	272,1555	12	43	**23,9**
13	27,0	27,0	2,6841	- 4,3975	7,2046	13	19	24,6
14	33,1	33,1	19,9065	88,8160	396,2675	14	10	24,8
15	27,3	27,3	1,7911	- 2,3971	3,2082	15	52	25,4
16	23,0	23,0	31,7908	- 179,2471	1010,6551	16	21	25,7
17	27,8	27,8	0,7028	- 0,5892	0,4939	17	59	26,4
18	31,4	31,4	7,6268	21,0627	58,1681	18	2	26,5
19	24,6	24,6	16,3081	- 65,8577	265,9553	19	32	26,5
20	20,9	20,9	59,8818	- 463,3854	3585,8303	20	3	26,7
21	25,7	25,7	8,6338	- 25,3690	74,5426	21	6	26,7
22	27,9	27,9	0,5451	- 0,4025	0,2972	22	9	27,0
23	31,8	31,8	9,9961	31,6045	99,9227	23	13	27,0
24	21,7	21,7	48,1405	- 334,0146	2317,5048	24	46	27,1
25	23,9	23,9	22,4518	- 106,3841	504,0834	25	5	27,3
26	16,0	16,0	159,7275	-2018,6890	25512,8645	26	15	27,3
27	29,5	29,5	0,7425	0,6398	0,5513	27	55	27,4
28	16,6	16,6	144,9215	-1744,6130	21002,2323	28	8	27,5
29	22,3	22,3	40,1745	- 254,6392	1613,9880	29	35	27,6
30	23,9	23,9	22,4518	- 106,3841	504,0834	**30**	48	**27,7**
31	29,6	29,6	0,9248	0,8894	0,8553	**31**	17	**27,8**
32	26,5	26,5	4,5725	- 9,7775	20,9075	32	22	27,9
33	32,2	32,2	12,6855	45,1814	160,9211	33	51	28,2
34	36,0	36,0	54,1941	398,9592	2937,0044	34	49	29,1
35	27,6	27,6	1,0781	- 1,1195	1,1624	35	27	29,5
36	37,7	37,7	82,1138	744,0879	6742,6766	36	4	29,6
37	30,4	30,4	3,1035	5,4673	9,6315	37	31	29,6
38	20,8	20,8	61,4395	- 481,5830	3774,8084	38	47	30,3
39	32,3	32,3	13,4078	49,0949	179,7692	39	37	30,4
40	30,8	30,8	4,6728	10,1010	21,8351	40	40	30,8
41	37,6	37,6	80,3115	719,7246	6449,9321	41	56	31,1
42	37,7	37,7	82,1138	744,0879	6742,6766	42	18	31,4
43	23,9	23,9	22,4518	- 106,3841	504,0834	43	54	31,4
44	35,4	35,4	45,7201	309,1443	2090,3308	44	23	31,8
45	32,4	32,4	14,1501	53,2281	200,2264	45	60	32,1
46	27,1	27,1	2,3665	- 3,6404	5,6002	46	33	32,2
47	30,3	30,3	2,7611	4,5881	7,6239	47	39	32,3
48	27,7	27,7	0,8805	- 0,8262	0,7752	48	45	32,4
49	29,1	29,1	0,2131	0,0984	0,0454	49	12	32,7
50	36,5	36,5	61,8058	485,8966	3819,9573	50	14	33,1
51	28,2	28,2	0,1921	- 0,0842	0,0369	51	57	33,6
52	25,4	25,4	10,4868	- 33,9598	109,9730	52	44	35,4
53	23,8	23,8	23,4095	- 113,2628	548,0033	53	34	36,0
54	31,4	31,4	7,6268	21,0627	58,1681	54	58	36,4
55	27,4	27,4	1,5335	- 1,8989	2,3515	55	50	36,5
56	31,1	31,1	6,0598	14,9172	36,7212	56	7	37,3
57	33,6	33,6	24,6181	122,1470	606,0526	57	41	37,6
58	36,4	36,4	60,2435	467,5897	3629,2756	58	36	37,7
59	26,4	26,4	5,0101	- 11,2144	25,1015	59	42	37,7
60	32,1	32,1	11,9831	41,4816	143,5956	60	11	38,4
Suma		**1718,3**	**1565,0018**	**- 424,0236**	**112990,0760**			
Vypočítaná charakteristika		**28,63833333**	**26,0834**	**- 0,0531**	**- 0,2320**		Medián	27,75
Výsledná (zaokrouhlená) charakteristika		**28,6**	**26,1**	**-0,05**	**-0,23**		Modus	23,9

Tabulka 4.4 - Výpočet statistických charakteristik pro Porost 1. Bližší vysvětlení viz v textu.

Rozložení hodnot, polohu aritmetického průměru (plná čára), mediánu (tečkovaná čára) a modu (čárkovaná čára) ukazuje obrázek 4.11. Jednotlivé měřené hodnoty (bílé kosočtverce) jsou náhodně „rozházeny“ ve směru osy Y proto, aby i v případě koncentrace dat na jednom místě byly vidět pokud možno všechny hodnoty.

Nyní provedeme stejné hodnocení pro Porost 2. Tabulka 4.5 ukazuje vypočítané hodnoty základních charakteristik pro tento soubor ve druhém sloupci (v prvním jsou pro srovnání uvedeny odpovídající charakteristiky Porostu 1). Ihned jsou vidět značné rozdíly. Odchylka aritmetického průměru od mediánu je extrémně velká (2,85 cm oproti 0,95 cm), také variabilita (tu je lépe porovnávat podle relativní míry – variačního koeficientu – 53,2 % oproti 17,8 %) podstatně vzrostla. Nejnápadnější změna je u charakteristik tvaru, které vykazují extrémní levostrannost a špičatost. Když budeme pátrat po příčinách, zjistíme, že soubor obsahuje 4 extrémní hodnoty 50,2;56,3;56,5 a 59,5. Grafické znázornění tohoto souboru ukazuje obrázek 4.12 . Tyto hodnoty jsou tak vzdáleny od ostatních, že je nutno je považovat za silné extrémy, které ovlivňují statistické hodnocení celého souboru, přičemž statistická informace obsažená v ostatních hodnotách se zcela ztrácí. V této souvislosti je nutné se zmínit o „zacházení“ s takovými hodnotami.

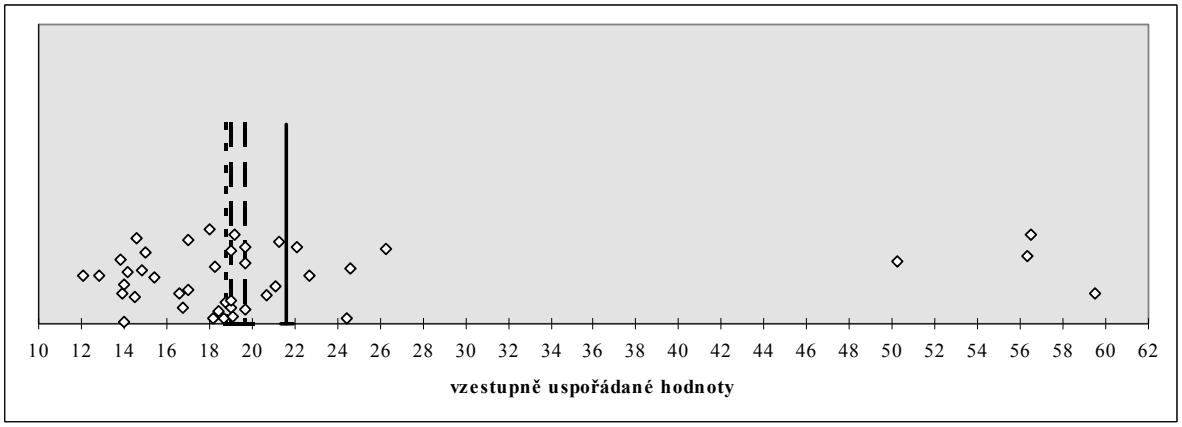


Obrázek 4.11- Grafické znázornění hodnot pro Porost 1 s vyznačením polohy průměru (plná čára), mediánu (tečkovaná) a modu (čárkovaná)

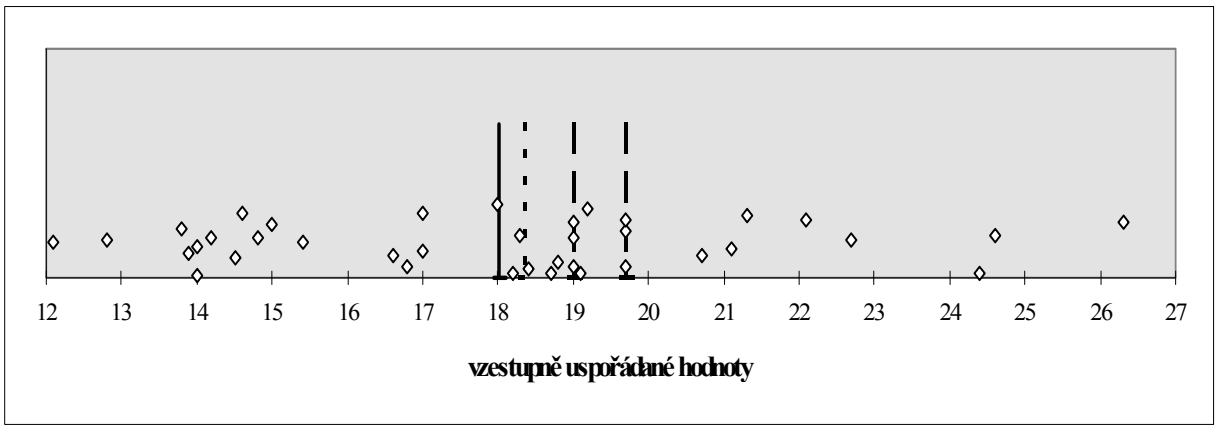
Charakteristika	Porost 1	Porost 2 podle zadání	Porost 2 bez extrémů
aritmetický průměr (cm)	28,60	21,60	18,00
medián (cm)	27,75	18,75	18,35
modus (cm)	23,90	19,00;19,70	19,00;19,70
rozptyl (cm^2)	26,10	136,60	11,40
směrodatná odchylka (cm)	5,10	11,50	3,30
variační koeficient (%)	17,80	53,20	18,30
koef. nesouměrnosti	- 0,05	2,42	0,39
koef. zahrocenosti	- 0,23	4,65	- 0,20
minimum (cm)	16,00	12,10	12,10
maximum (cm)	38,40	59,50	26,30
variační rozpětí (cm)	22,40	47,40	14,20
počet hodnot	60	42	38

Tabulka 4.5 – Statistické charakteristiky analyzovaných souborů

V první řadě je nutné zjistit, jakým způsobem tyto extrémní hodnoty vznikly. Pokud vznikly jako výsledek hrubé chyby měření nebo zápisu, je nutné je vyřadit (a v případě velmi malého výběru je nutné se pokusit – pokud je to technicky možné – získat nová správná měření). Pokud by se ukázalo, že měření je správné, potom je nutné odhalit příčinu prudké změny měřených hodnot. Zde znovu nastávají dvě možnosti. Jestliže se ukáže, že daná měření nesouvisejí s ostatními (vznikly v důsledku prudké změny podmínek, sloučením dvou ve skutečnosti rozdílných souborů apod.), je možné je také vyřadit z dalšího zpracování. Za předpokladu, že je z určitého důvodu není možné vyřadit (příliš malý výběr, který není možné doplnit, nebo se jedná o skutečně možné extrémní hodnoty dané veličiny), je nutné s tímto souborem zacházet jako se souborem s nenormálním rozdělením četností a přizpůsobit tomu statistickou analýzu (používat robustních – kvantilových – charakteristik, pokusit se o vhodnou transformaci pro přiblížení se normalitě apod.). Pro tyto případy není možné dát všeobecně platný návod a při jejich řešení se v plné míře projeví odborná i statistická erudice analytika.



Obrázek 4.12 - Grafické znázornění hodnot pro Porost 2 (včetně extrémních hodnot) s vyznačením polohy průměru (plná čára), mediánu (tečkovaná) a modu (čárkovaná)



Obrázek 4.13 - Grafické znázornění hodnot pro Porost 2 (s vyloučením extrémních hodnot) s vyznačením polohy průměru (plná čára), mediánu (tečkovaná) a modu (čárkovaná)

V našem případě se ukázalo, že extrémní hodnoty měřených tlouštěk zřejmě vznikly změřením tlouštěk výstavků (stromů z předchozího porostu. které byly na holině ponechány pro zabezpečení alespoň částečné přirozené obnovy následného porostu). Je zřejmé, že tyto stromy do měřeného souboru nepatří (mají zcela jiný věk), a proto je možné je z měřeného

souboru vyloučit. Vzhledem k dostatečnému rozsahu souboru není nezbytně nutné měření doplňovat.

Hodnoty statistických charakteristik po vyloučení měřených hodnot uvádí poslední sloupec tabulky 4.5 a graficky je situace znázorněna na obrázku 4.13 .

Z obrázku i z tabulkových hodnot je zřejmé, že tento soubor lze také hodnotit jako přibližně normální.

Pro ilustraci faktu, že správně roztříděný soubor podává stejnou statistickou informaci jako soubor neroztříděný, je v tabulce 4.6 uvedeno třídění souboru Porost 1 a jsou vyčísleny základní statistické charakteristiky. Odhad mediánu z roztříděného souboru je velmi přibližný, odhad modu nebyl prováděn vůbec (z roztříděného souboru se dělá pouze výjimečně).

Roztřídění bylo provedeno podle zásad uvedených v kapitole 3.3 (m = 7, h = 3,3) a na jejich základě byly stanoveny hranice tříd, třídní reprezentanti ($\overline{x}_i$) a třídní četnosti (n_i). Ve sloupcích A – D jsou pomocné výpočty pro stanovení aritmetického průměru (sloupec A - $\overline{x}_i \cdot n_i$), rozptylu (sloupec B - $n_i \cdot (\overline{x}_i - \overline{x})^2$), koeficientu nesouměrnosti (sloupec C - $n_i \cdot (\overline{x}_i - \overline{x})^3$) a koeficientu zahrocenosti (sloupec D - $n_i \cdot (\overline{x}_i - \overline{x})^4$).

Z výsledných hodnot plyne, že rozdíly mezi prostými a váženými charakteristikami jsou nevýznamné a umožňují stejnou statistickou interpretaci. Tento příklad ukazuje, že při správně provedeném třídění nesou vážené charakteristiky stejnou informaci jako prosté, které jsou vypočítány ze všech hodnot souboru. To lze s výhodou využít zvláště tehdy, jsme-li nuceni zpracovávat rozsáhlý datový soubor bez možnosti použít vhodné programové vybavení.

Číslo třídy	Hranice třídy		Třídní reprezentant	Absolutní třídní četnost	A	B	C	D
	dolní	horní						
1	15,85	19,15	17,50	2	35,0000	249,3145	-2783,5958	31078,8475
2	19,15	22,45	20,80	5	104,0000	309,2911	-2432,5747	19132,2000
3	22,45	25,75	24,10	9	216,9000	187,5530	- 856,1796	3908,4597
4	25,75	29,05	27,40	17	465,8000	27,2038	- 34,4128	43,5322
5	29,05	32,35	30,70	14	429,8000	57,9772	117,9835	240,0964
6	32,35	35,65	34,00	5	170,0000	142,3111	759,2299	4050,4913
7	35,65	38,95	37,30	8	298,4000	596,5058	5150,8276	44477,3962
Součet					1719,9000	1570,1565	- 78,7220	102931,0233

Přehled vážených statistických charakteristik pro soubor Porost 1	Charakteristika	Vypočítané vážené charakteristiky	Výsledné (zokrouhlené) vážené charakteristiky
	aritmetický průměr	28,6650	**28,70**
	medián	28,5647	**28,60**
	rozptyl	26,1693	**26,20**
	směrodatná odchylka	5,1156	**5,10**
	variační koeficient	17,8461	**17,80**
	koeficient nesouměrnosti	- 0,0098	**- 0,01**
	koeficient zahrocenosti	- 0,4949	**- 0,49**

Tabulka 4.6 – Výpočet vážených statistických charakteristik pro Porost 1

5 Úvod do matematické statistiky

V předchozích kapitolách jsme se zabývali popisem základního souboru, tj. statistickým zkoumáním takového souboru, o němž máme k dispozici úplnou statistickou informaci, tj. známe měřené (nebo jinak zjišťované) hodnoty příslušného statistického znaku u všech jednotek zkoumaného souboru. Ovšem s tímto přístupem vystačíme jen v relativně malém počtu případů, většinou jen u malých, dobře definovaných souborů s veličinami, jejichž měření není obtížné ani drahé. Ve valné většině případů musíme zvolit jiný přístup – cestu stanovení výběrového souboru a zobecnění jeho výsledků pro soubor základní. A tady se dostáváme do sféry matematické statistiky.

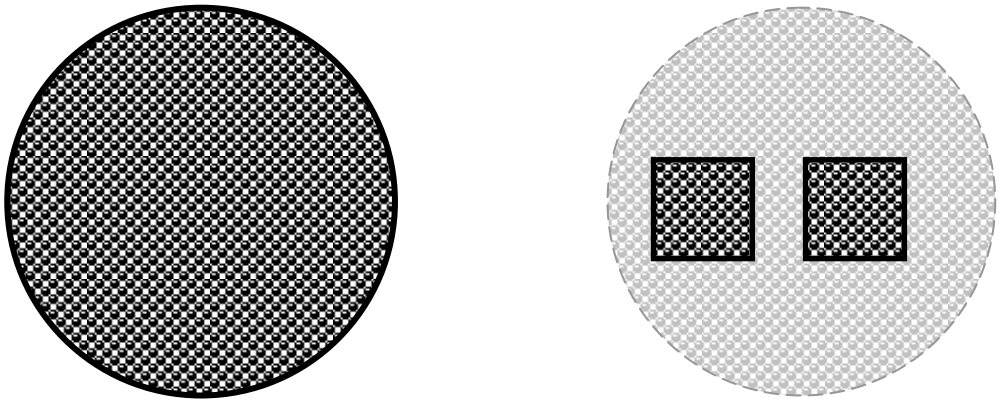
5.1 Podstata výběrového šetření

Rozdíl mezi koncepcí popisné a matematické statistiky ukazuje obrázek 5.1 . Vlevo je základní soubor, kdy je jasně vymezen jeho rozsah (znázorněný tučnou hraniční čárou) a jsou známy všechny měřené jednotky (jednotlivé „kuličky“). Statistické charakteristiky budou výslednicí znalosti vyšetřovaného jevu na všech jednotkách („kuličkách“) souboru. Vpravo je schématicky zachycen výběrový soubor. V tomto případě velikost (počet jednotek) základního souboru není známa (znázorněno tenkou přerušovanou čarou kolem základního souboru a šedou barvou „kuliček“) nebo je „nezměřitelně“ velká. V tomto případě měření provedeme na tzv. **výběrovém souboru**, který je schématicky znázorněn na pravé části obrázku. V tomto případě měříme příslušné veličiny jen na jednotkách spadajících do výběrového souboru (viditelné „kuličky“ na tučně orámovaných „zkusných plochách“), z jejichž statistických vlastností potom odhadujeme statistické vlastnosti celého základního souboru.

Je nutné zdůraznit, že v jak v popisné statistice, tak i v matematické statistice je **objektem zájmu statistické analýzy základní soubor. Výběrový soubor je pouze prostředkem k jeho poznání**. Proto statistická analýza používající matematickou statistiku nesmí končit výpočtem statistických charakteristik výběru, ale musí pokračovat ke stanovení statistických vlastností základního souboru (např. odhady parametrů, testování hypotéz apod.). Tak nás např. při určování zásoby porostu pomocí kruhových zkusných ploch nezajímá zásoba určité dřeviny na zkusných plochách, ale celková zásoba dřeviny v porostu (nebo přepočítaná zásoba na 1 ha).

Výběrová šetření se zpravidla skládají z řešení tří dílčích etap:

- **plánování výběrového šetření** - jedná se především o dvě důležitá rozhodnutí:
  - stanovení potřebného rozsahu výběru (tj. z kolika hodnot se nejméně musí skládat výběrový soubor, abychom získali potřebné údaje o základním souboru s požadovanou přesností),
  - stanovení nejvhodnějšího způsobu výběru (tj. jakým způsobem nejefektivněji potřebný rozsah výběru získat)
- **statistický odhad parametrů základního souboru** ze známých hodnot statistických charakteristik výběru (zde musíme vždy uvažovat pouze o odhadu, protože máme k dispozici pouze údaje o výběru, tj. velmi malé části základního souboru)
- **statistické zkoumání vyslovených hypotéz** o určitých vlastnostech základního souboru na základě vědomostí o výběrovém souboru (vlastně odpovídáme na otázku: mohu s předem stanovenou pravděpodobností na základě znalostí o výběru, který mám k dispozici, tvrdit, že základní soubor má určitou vlastnost?).



Obrázek 5.1 -Grafické znázornění základního (vlevo) a výběrového (vpravo) souboru

Jako příklad si můžeme uvést zkoumání tloušťkové struktury několika rozsáhlých smrkových porostů (tj. zkoumání rozložení četností tlouštěk v těchto porostech) srovnatelného věku a růstových podmínek, přičemž máme odpovědět na otázku, zdali jsou tyto porosty stejně tloušťkově vyspělé a zda mají shodnou strukturu rozdělení tlouštěk (např. pro účely produkční a sortimentační analýzy).

Kdybychom chtěli tento úkol řešit pomocí popisné statistiky, tj. základního souboru, znamenalo by to v rozsáhlých porostech změřit tloušťky několika tisíc nebo i desítek tisíc stromů, což je prakticky nemožné (hlavně časově a ekonomicky), a navíc to není ani potřebné, protože při správné aplikaci postupů matematické statistiky tento úkol splníme s potřebnou přesností daleko dříve a racionálněji.

Nejprve si musíme určit počet stromů měřených v každém porostu a způsob jejich výběru. Tím splníme první bod postupu výběrových šetření. To stanovíme na základě známé nebo odhadnuté variability základního souboru a požadované přesnosti. Poté provedeme potřebná měření a vypočítáme potřebné statistické charakteristiky výběru (tj. určitého počtu skutečně změřených tlouštěk stromů) podle postupů uvedených v kapitole 4. Ale tyto údaje nás vlastně vůbec nezajímají a jsou pouze prostředkem ke zjištění těchto údajů (např. průměrné tloušťky, rozdělení četností tlouštěk apod.) pro základní soubory, tj. celé porosty. To provedeme pomocí postupů druhé etapy, tj. provedeme kvalifikovaný statistický odhad těchto parametrů (např. aritmetického průměru, směrodatné odchylky, apod.) pro základní soubory. Potom můžeme např. tvrdit, že s pravděpodobností 95 % je průměrná tloušťka analyzovaného porostu v rozmezí 38,6 - 40,2 cm. Dále máme rozhodnout, zdali jsou porosty stejně tloušťkově vyspělé a mají stejnou tloušťkovou strukturu. Stejně tloušťkově vyspělé budou porosty tehdy, budou-li mít stejnou průměrnou tloušťku. Víme, že jednotlivé výběrové soubory mají číselně rozdílné průměry, např. 39,4 cm, 40,6 cm a 42,7 cm. Musíme si ale uvědomit, žer kdybychom z jednotlivých porostů naměřili tloušťky na jiných stromech (udělali jiný výběr), byl by výsledek pro výběrové soubory číselně zase jiný. Musíme tedy odpovědět na otázku: jsou rozdíly mezi výběrovými průměry tak velké, že ani v základním souboru se nedá s rozumnou pravděpodobností očekávat jejich shoda? Vyslovíme tedy hypotézu: všechny porosty mají stejnou střední tloušťku a budeme ji statisticky testovat. Test nám dá objektivní podklad pro naše rozhodnutí. Stejně budeme postupovat i v případě tloušťkové struktury.

Vidíme tedy, že metody matematické statistiky nám dávají do rukou mocnou statistickou zbraň, která nám pomůže kvalifikovaně odpovědět na otázky, jejichž řešení by prostředky popisné statistiky nebylo možné.

Výběrová šetření jsou tedy založena na tom, že poznatky o relativně malém výběrovém souboru zevšeobecníme na základní soubor. Tento princip se nazývá **statistická induk-**

ce, jejímž základem je matematická statistika. Statistická indukce nepodává zobecněné výsledky s naprostou jistotou. Jedním z hlavních úkolů statistiky je, aby zhodnotila nejistotu jednotlivých indukčních závěrů, tj. stanovila pravděpodobnost, s jako daný závěr přijímáme.

Matematická statistika vznikala spojením statistickým metod a teorie pravděpodobnosti. Proto je nutno si nyní uvést některé pojmy teorie pravděpodobnosti, ovšem pouze v rozsahu a úrovni nezbytně nutné pro pochopení dalšího výkladu.

5.2 Základní pojmy teorie pravděpodobnosti

5.2.1 Náhodný experiment

Experiment (pokus) je **realizace neměně vymezeného komplexu podmínek**, **kterou lze** (alespoň teoreticky) **mnohonásobně nezávisle opakovat**. Za experiment můžeme považovat např. měření nebo jiné zjišťování určité veličiny. Realizací je potom naměřená konkrétní hodnota. Experimenty, kdy vymezený komplex podmínek bezezbytku určuje výsledek (tj. při zachování stejných podmínek dostaneme vždy stejný výsledek) nazýváme **experimenty jisté** (deterministické). Příkladem může být určení obsahu kruhu, jestliže známe jeho průměr. Za předpokladu, že použijeme hodnotu π se stejnou přesností, dostaneme vždy stejnou hodnotu výsledku, ať „pokus“ opakujeme libovolně často.

Naproti tomu u **náhodného experimentu** mohou navíc působit na výsledky realizací chaoticky se měnící náhodné vlivy. Do náhodných vlivů zahrnujeme všechny působící podmínky, které nemůžeme přesně popsat, neznáme je nebo je do vymezeného komplexu podmínek nezahrnujeme. Výsledkem jsou hodnoty, jejichž velikost se pokus od pokusu mění a není je možné předem přesně určit.

Příkladem náhodného experimentu mohou být hazardní hry, laboratorní nebo terénní pokusy, hromadná pozorování, hromadná měření, hromadná aplikace technologických postupů, opakované operace ve vojenství, ekonomice, lékařství atd. Např. u karetní hry jsou komplexem podmínek pravidla hry, náhodnými vlivy je to, jaké karty dostane hráč do ruky, jak silné má protihráče, zda neudělá hrubou chybu (a zda ji naopak udělá někdo jiný) a mnoho dalších.

V reálném světě jsou náhodné experimenty daleko častější než jisté a jsou také předmětem statistického zkoumání. Proto se v dalším výkladu budeme věnovat výhradně jim.

5.2.2 Jev a jeho vlastnosti

K popisu a práci s náhodným experimentem se používá pojmů a symbolů teorie množin. Vychází se z toho, že náhodný experiment může mít řadu možných výsledků:

ω	je možný výsledek náhodného experimentu,
$\Omega = \{\omega\}$	množina všech možných výsledků náhodného experimentu,
$A \subset \Omega$	množina některých možných výsledků náhodného experimentu - podmnožina všech možných výsledků náhodného experimentu.

Množina A se nazývá **jev**. Rozlišujeme tři základní skupiny jevů:

- jev jistý (Ω) - jev, kterému je příznivý každý výsledek náhodného experimentu, tj. při opakování daného experimentu vždy nastane,
- jev nemožný ($\varnothing$) - jev, kterému není příznivý žádný výsledek náhodného experimentu, tj. při opakování daného experimentu nikdy nenastane,

- jev náhodný (označuje se velkým písmenem, např. A) – jev, jemuž jsou příznivé některé výsledky náhodného experimentu, které při opakování náhodného experimentu mohou, ale nemusí nastat (vlivem působení náhodných vlivů).

Příkladem může být náhodný experiment „měření výšky stromu“. Jedná se o typický náhodný experiment, protože jeho výsledek (tj. konkrétní výška stromu) je ovlivněn mnoha náhodnými vlivy (jak společnými pro všechny stromy v porostu, např. bonitou stanoviště, klimatickými vlivy, antropogenní vlivy, apod., tak i jedinečnými pro daný strom, např. genetická predispozice, konkurenční vztahy atd.). Jevem jistým zde může být naměření určité výšky, tj. že číselným vyjádřením výšky stromu bude kladné číslo. Naopak jevem nemožným je naměření záporné výšky. Náhodným jevem je pak konkrétní výška stromu, která je výslednicí všech uvažovaných i neuvažovaných náhodných vlivů.

Ze známých formálně zavedených množinových pojmů a množinových operací jsou pro teorii pravděpodobnosti a její aplikace důležité :

Buď A, B, C jevy, potom

1. $A \subset B$ značí: z nastoupení jevu A plyne nastoupení jevu B.
   Je-li $A \subset B \wedge B \subset C$ potom $A \subset C$ - tranzitivnost jevů .
2. $A \subset C \wedge C \subset A \Leftrightarrow A = C$; A , C jsou jevy ekvivalentní .
3. $A \cap B$ značí: nastoupí jev A a zároveň jev B - průnik jevů .
   Platí vztahy: $A \cap B \subset A$; $A \cap B \subset B$
   $A \cap B = B \cap A$
   $A \cap \Omega = A$; $A \cap \emptyset = \emptyset$
   Je-li $A \cap B = \emptyset$ nazývají se A , B neslučitelné .
4. $A \cup B$ značí : nastoupení jevu A nebo nastoupení jevu B - sjednocení jevů .
   Platí vztahy : $A \subset (A \cup B)$; $B \subset (A \cup B)$; $A \cup A = A$; $A \cup B = B \cup A$;
   $A \cup \Omega = \Omega$; $A \cup \emptyset = A$.
5. $\overline{A}$ značí: nenastoupení jevu A; jev opačný k jevu A .
   Platí vztahy: $A \cup \overline{A} = \Omega$; $A \cap \overline{A} = \emptyset$; $\overline{\overline{A}} = A$; $\overline{A \cap B} = \overline{A} \cup \overline{B}$;
   $\overline{A \cup B} = \overline{A} \cap \overline{B}$
6. A/B značí: rozdíl jevů - nastoupení jevu A a nenastoupení jevu B .
7. Jestliže $A \cap B = \emptyset$ potom A, B jsou jevy neslučitelné .
8. Buď $B = \bigcup_{i=1}^{k} A_i$, kde A_i jsou jevy neslučitelné. Potom říkáme, že B se rozpadá na dílčí jevy .
9. Jestliže $\bigcup_{i=1}^{k} A_i = \Omega$ potom A_i tvoří úplnou skupinu jevů .

Příklad 5.1:

Analyzujme s použitím předchozích pojmů a tvrzení hod hrací kostkou na rovnou desku stolu. Komplex podmínek je dán šestistěnným tvarem kostky, na každé stěně vzájemně odlišné označení, a dostatečně velkou rovnou deskou stolu. Formalizace jevů je zde velmi jednoduchá

a názorná. Náhodný experiment je jeden hod. Při jednom hodu se na horní stěně objeví určité označení; označme je ω_i *, když* i *je číslice označující počet bodů.*

1. Jev jistý $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$; na rovné desce stolu „padne“ vždy nějaký počet „bodů“.

2. Jev nemožný $\emptyset = \{\ \}$; jev „nepadne žádný počet bodů“ (na rovné desce stolu kostka nezaujme jinou polohu než stěnou nahoru, a tedy ukáže nějaký počet bodů)..

3. $B = \{\omega_1, \omega_3, \omega_5\}$; jev „padne lichý počet bodů“,
   $A = \{\omega_3\}$; jev „padne trojice bodů“,
   $A \subset B$; jev „když padne trojka, padne lichý počet bodů“.

4. $B = \{\omega_1, \omega_3, \omega_5\}$; jev „padne lichý počet bodů“,
   $D = \{\omega_4, \omega_5, \omega_6\}$; jev „padne počet bodů větší než tři“,
   $B \cap D = \{\omega_5\}$; jev „padne lichý počet bodů a zároveň vyšší než tři “,.
   $B \cup D = \{\omega_1, \omega_3, \omega_5, \omega_4, \omega_6\}$; jev „ padne lichý počet nebo počet vyšší než tři “.

5. $\overline{B} = \{\omega_2, \omega_4, \omega_6\}$; jev „ padne sudý počet bodů “ .

6. B , $\overline{B}$ jsou jevy neslučitelné, dílčí a tvoří úplnou skupinu jevů.

Formalizace jevů s využitím množinových operací pro určitý náhodný experiment je velmi účinná a užitečná při přípravě metodiky řešení rozličných úloh, zejména při přípravě získávání informací např. z měření a jeho následného cíleného zpracování.

Příklad 5.2:

Připravujeme šetření o rozsahu poškození stromů v daném porostu dvěma rozdílnými druhy - houbou a mechanicky. Poškození houbou označme V*, poškození mechanické* D*, žádné poškození označme* Z*. Náhodný experiment je zjištění poškození jednoho stromu.*

Formálně sestavené jevy jsou

1. $\Omega = \{V, D, Z\}$;
2. V ; D ; Z ; $\overline{V}$; $\overline{D}$.
3. $V \cup D$, $V \cap D$.
4. V/D ; D/V .
5. $(V \cup D) \cap Z = \emptyset$

(možno pokračovat)

Skupina formálně sestavených jevů předkládá základní skupinu otázek, na které může experiment odpovědět. Zároveň předkládá potřebné statistické soubory hodnot sestavené ze šetření v porostu potřebné k vyřešení určené úlohy.

5.2.3 Náhodná veličina, náhodný vektor

Náhodná veličina je definována jako **libovolná reálná funkce** X **definovaná na množině** Ω. Množina Ω je množina všech možných výsledků náhodného experimentu, náhodná veličina je taková veličina, **jejíž hodnota je pokus od pokusu mění působením náhodných vlivů.**

Zavedená náhodná veličina je matematicky zobrazením množiny Ω do množiny reálných čísel R.

Matematickým zavedením náhodné veličiny se získala možnost jevy měřit, tj. vyjadřovat číselně, a hodnoty měření matematicky zpracovávat.

Jako příklad náhodné veličiny si můžeme uvést již zmíněnou výšku stromu v porostu. Výška stromu je přesně definovaná, tzn. je přesně určen předpis, jak stromu přiřadit reálné číslo znamenající jeho výšku ve stanovených délkových jednotkách. Výška určitého stromu je výslednicí působení komplexu vnějších podmínek prostředí společného všem stromům porostu a náhodným vlivům.

Realizace náhodné veličiny je hodnota $X(\omega)$, tj. např. měření konkrétní výšky stromu. Zápis $X(\omega) = x$ znamená, že náhodná veličina X nabývá hodnoty x (tj. nějaké konkrétní výšky, např. 25 m).

Náhodný vektor je libovolná uspořádaná n-tice $(X_1, X_2, \ldots, X_n)$ náhodných veličin definovaných na téže množině Ω. Realizace náhodného vektoru je vektor $(X_1(\omega), X_2(\omega), \ldots, X_n(\omega))$. Jako příklad si můžeme uvést trojici náhodných veličin výška, výčetní tloušťka, výtvarnice. Jeho realizací je uspořádaná trojice hodnot výšky, výčetní tloušťky a výtvarnice naměřená na určitém stromě.

5.2.4 Pravděpodobnost

Pravděpodobnost je objektivní vlastnost náhodného jevu. Je to **reálné číslo, které charakterizuje** (poměřuje) **možnost nastoupení určitého jevu při působení vymezeného komplexu podmínek.**

5.2.4.1 Axiomatická definice pravděpodobnosti

Libovolnou reálnou funkci P definovanou na jevovém poli A nazveme pravděpodobnost, splňuje-li pro libovolné jevy $A_1, A_2, \ldots \in A$ následujících 8 vlastností.

1. $P(\Omega) = 1 \ \wedge \ P(\emptyset) = 0$ (axiom)
2. $0 \le P(A) \le 1$ (axiom)
3. Jsou-li $A_1, A_2, \ldots$ neslučitelné jevy, potom
   $P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \ldots$ (axiom)
4. $P(\overline{A}) = 1 - P(A)$
5. $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$
6. $P(A_1 / A_2) = P(A_1) - P(A_1 \cap A_2)$
7. $A_2 \subseteq A_1 \Rightarrow P(A_1) = P(A_1 / A_2) + P(A_2)$
8. $A_2 \subseteq A_1 \Rightarrow P(A_2) \le P(A_1)$

Uvedená definice pravděpodobnosti se nazývá axiomatická. Udává podmínky, za kterých se reálná funkce nazývá pravděpodobnost. Volba určitých hodnot pravděpodobnosti nebo konkrétních vzorců pro jejich výpočet závisí na věcné podstatě řešených úloh.

Pravděpodobnostní prostor je trojice (Ω, A, P).

Každou funkci, uvažovanou jako pravděpodobnost v řešení určitého úkolu, je nutné posoudit axiomatickou definicí pravděpodobnosti. Použít ji můžeme, pokud požadavkům axiomatické definice vyhoví.

5.2.4.2 Empirický zákon velkých čísel (statistická definice pravděpodobnosti)

Při opětovné nezávislé realizaci téhož náhodného experimentu se podíl výskytu sledovaného jevu mezi všemi realizacemi ustaluje kolem nějakého čísla. To znamená, že jestliže vícenásobně opakujeme určitý pokus, např. určité měření nebo zjišťování, je nejprve podíl výskytu studovaného jevu (n_A) ku celkovému počtu pokusů (n) - $n_A : n$ - velmi rozkolísaný a se zvyšujícím se počtem pokusů se ustaluje kolem konstanty nazývané stabilní relativní četnost. Tato zákonitost se nazývá zákon velkých čísel (tj. velkého počtu provedených pokusů):

$$P(A) = \frac{n_A}{n} \tag{5.1}$$

kde je

P(A) pravděpodobnost jevu A

n_A je počet realizací, při kterých nastal jev A ,

n je počet všech realizací (vlastně velikost výběru)

Takto zavedená (pro dostatečně velká n) relativní četnost $P(A)$ se nazývá **statistická pravděpodobnost**.

Ze statistické pravděpodobnosti (jejím zobecněním na základní soubor) vycházela klasická definice pravděpodobnosti nastoupení jevu A:

$$P(A) = \frac{N_A}{N} \tag{5.2}$$

kde je

N_A počet všech možných případů příznivých jevu A

N počet případů v experimentu možných (tedy vlastně základní soubor).

Statistická (klasická) pravděpodobnost vyhovuje axiomatické definici pravděpodobnosti. Toto tvrzení se dokáže tak, že se postupně dokazuje platnost podmínek 1. až 8. pro statistickou pravděpodobnost. Tyto důkazy přesahují ovšem rozsah a poslání tohoto učebního textu.

Příklad 5.3:

Při těžbě zaznamenáváme celkový počet skácených stromů n *a počet stromů s hnilobou běle* n_B *a hnilobou jádra* n_J *. Následující tabulka 5.1 obsahuje sledované počty a vypočítané podíly. Určete pravděpodobnost výskytu stromu s hnilobou běle a hnilobou jádra podle zákona velkých čísel.*

n	5	10	15	20	25	30	35	40	45	50	55	60	65	70
n_B	0	4	4	5	5	5	5	6	7	8	9	9	10	11
n_B:n	0,00	0,40	0,27	0,25	0,20	0,17	0,14	0,15	0,16	0,16	0,16	0,15	0,15	0,16
n_J	3	3	3	10	10	10	10	13	18	21	22	23	26	28
n_J:n	0,60	0,30	0,20	0,50	0,40	0,33	0,29	0,33	0,40	0,42	0,40	0,38	0,40	0,40

Vysvětlivky:

n	počet pokácených stromů	**n_J**	počet stromů napadených hnilobou jádra
n_B	počet stromů napadených hnilobou běle	**n_J:n**	relativní četnost pro hnilobu jádra
n_B:n	relativní četnost pro hnilobu běle		

Tabulka 5.1 - Zadání příkladu 5.3

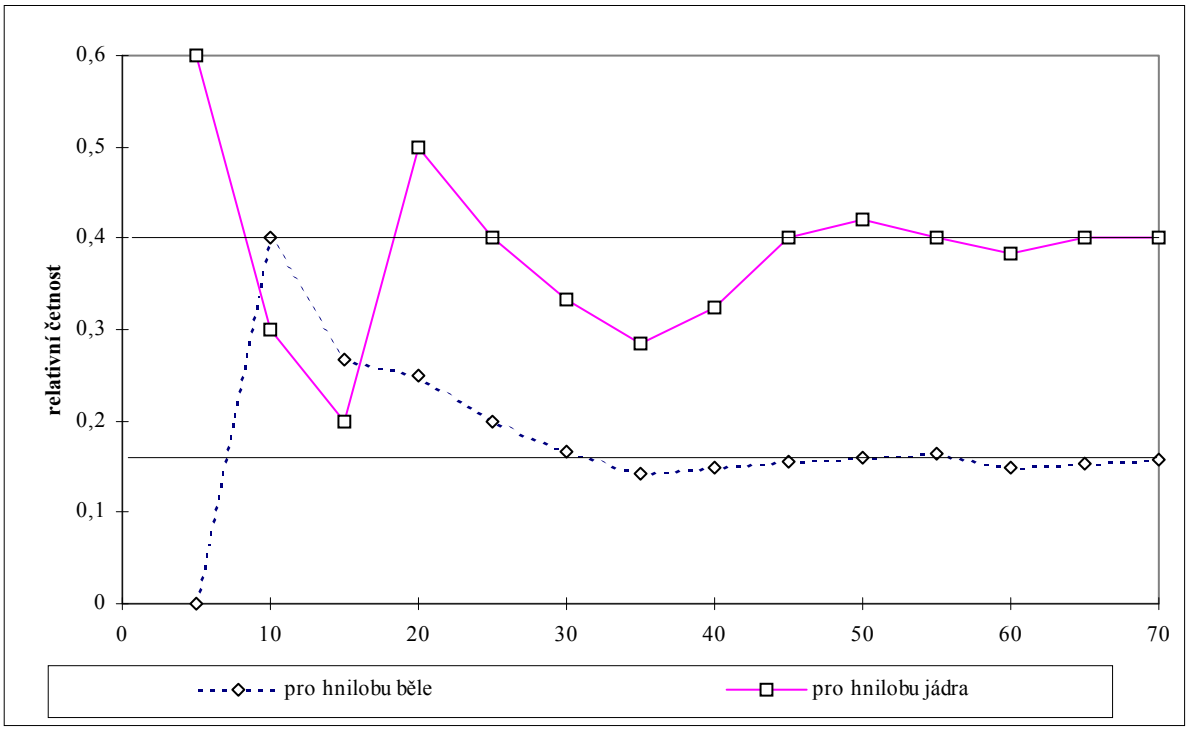
Z tabulky 5.1 a grafu na obrázku 5.2 je zřejmé, že skutečně relativní četnosti pro malé výběry značně kolísají a až pro velké výběry se ustalují kolem určitých konstantních hodnot. Z číselných hodnot tabulky se dá usoudit, že je přibližně $n_B : n = 0{,}16$, $n_J : n = 0{,}40$. Podle statistické definice pravděpodobnosti je pravděpodobnost skácení stromu s hnilobou běle $P(n_B) \cong 0{,}16$, skácení stromu s hnilobou jádra $P(n_J) \cong 0{,}40$.

Můžeme ze zaznamenaných hodnot určit pravděpodobnost skácení zdravého stromu?

Nemůžeme ! Není totiž zaznamenáno kolik stromů je poškozeno hnilobou běle a zároveň hnilobou jádra, tzn. nevíme, zda je $B \cap J = \emptyset$. Nemůžeme použít vzorce

$$P(Z) = \frac{n - (n_B + n_J)}{n} = 0{,}50 \ ,$$

ani množinových operací k vyjádření jevu Z - zdravý strom - a vzorců z axiomatické definice pravděpodobnosti. Můžeme pouze určit, že pravděpodobnost skácení zdravého stromu není menší než 0,50.



Obrázek 5.2 - Relativní četnosti pro data z příkladu 5.3. Na ose X jsou počty pokusů (skácených stromů), vodorovné čáry udávají úroveň stabilní relativní četnosti.

Uvedený příklad dokumentuje nutnost a prospěšnost formalizace jevů s využitím množinových operací pro určitý náhodný experiment při přípravě metodiky řešení úlohy získávání informací např. měření a jejich následném cíleném zpracování. Je vidět, že malým rozšířením zápisu zjišťování o počet $n_{B \cap J}$ získáme možnost mnohem širšího zpracování dat a mnohonásobně širší okruh informací o studovaném problému.

5.2.4.3 Podmíněná pravděpodobnost

Dosud jsme hovořili pouze o případu, kdy náhodný experiment má komplex podmínek K a kdy nastoupení jevu A není závislé na žádném jiném jevu. Jev A, výsledek náhodného experimentu, má potom pravděpodobnost P(A), která se jmenuje nepodmíněná (též prostá).

Jestliže ke komplexu podmínek K připojíme podmínku nastoupení jevu H, potom jev A není pouze výsledkem působení podmínek K, ale zároveň nastoupením jevu H. Bez nastoupení jevu H jev A nastat nemůže. Pravděpodobnost jevu A se tím zmenší a je závislá - podmíněná - na pravděpodobnosti jevu H . Vzorec podmíněné pravděpodobnosti má tvar

$$P\left(A/H\right) = \frac{P(A \cap H)}{P(H)} \tag{5.3}$$

Z uvedeného vztahu lze odvodit, že pravděpodobnost nastoupení průniku jevů je rovna součinu pravděpodobnosti nastoupení jevu H a podmíněné pravděpodobnosti jevu A. Platí

$$P(A \cap H) = P(H) \cdot P\left(A/H\right) \tag{5.4}$$

Tento vzorec se nazývá pravidlo o násobení pravděpodobnosti a v obecné formě má tvar

$$P\left(A_1 \cap A_2 \cap \ldots \cap A_n\right) = P\left(A_1\right) \cdot P\left(A_2/A_1\right) \cdot P\left(A_3/A_1 \cap A_2\right) \cdot \ldots \cdot P\left(A_n/A_1 \cap A_2 \cap \ldots \cap A_{n-1}\right)$$

Uvedené vzorce o násobení pravděpodobnosti se používají v praxi pro výpočet pravděpodobnosti průniku jevů. Jev H se nazývá hypotéza a jeho pravděpodobnost je předem známá - říká se jí pravděpodobnost a priori.

V souvislosti s podmíněnou pravděpodobností je nutné vysvětlit rozdíl mezi dvěma druhy výběrů:

- **výběr s opakováním** (s vracením) je takový typ výběru, kdy jednotlivé prvky výběru před dalším výběrem vracíme do základního souboru. Znamená to, že pravděpodobnost výběru určitého prvku v jednotlivých „kolech“ výběru je opravdu nezávislá, tj. není nijak ovlivněna výsledky předchozího výběru. Je to způsobeno tím, že základní soubor je vždy pro každé výběrové „kolo“ stejný, se všemi prvky.
- **výběr bez opakování** (bez vracení) je takový typ výběru, kdy vybrané prvky již do základního souboru nevracíme. To způsobuje, že pravděpodobnost výběru určitého prvku je ovlivněna výsledky předchozích pokusů (výběrů), protože se tak mění velikost základního souboru i zastoupení určitých prvků v něm. Proto zde musíme pracovat s podmíněnou pravděpodobností, a tedy zahrnout i výsledky předchozích pokusů.

Rozdíl mezi výběrem s opakováním a bez opakování je nejmarkantnější buď u malých základních souborů nebo v těch případech, kdy výběrový soubor tvoří podstatnou část základního souboru. Pokud je základní soubor velmi velký nebo výběrový soubor nepřesahuje 5 % (někteří autoři uvádí 10 %) rozsahu základního souboru, potom je rozdíl mezi oběma typy výběrů jen teoretický a v praxi jsou jejich pravděpodobnosti stejné. Proto, máme-li velký základní soubor, ze kterého provádíme relativně výběr, nemusíme se zpravidla na rozdíl mezi výběrem s opakováním a bez opakování ohlížet.

Příklad 5.4:

Jaká je pravděpodobnost, že potomek laně v daném roce bude korunový jelen?

Hypotéza H: kolouch bude jelen; $P(H) = 0{,}5$ - pravděpodobnost a priori (předpokládáme vyrovnaný poměr pohlaví).

Jev A: jelen bude korunový; $P(A) = 0,4$ - pravděpodobnost získaná pozorováním a platná pro danou populaci.

Pravděpodobnost, že potomek laně bude korunový jelen je

$$P(H \cap A) = 0,5 \cdot 0,4 = 0,2$$

5.3 Spojité a diskrétní náhodné veličiny a jejich zákony rozdělení pravděpodobnosti

Náhodná veličina (náhodná proměnná) **je taková veličina, jejíž hodnota se pokus od pokusu mění v závislosti na náhodných vlivech** (podrobněji viz kapitola 5.2.3).

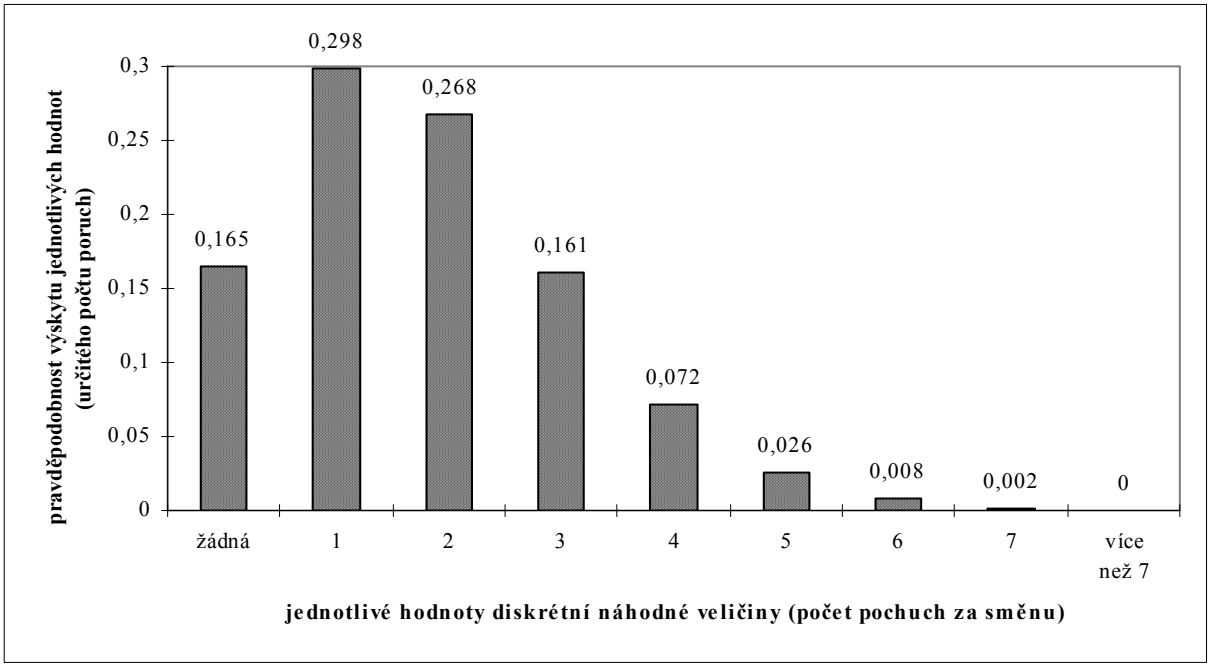
5.3.1 Spojitá a diskrétní náhodná veličina

Náhodné veličiny mohou být

- spojité (mohou nabýt jakékoli hodnoty z určitého intervalu),
- diskrétní (nabývají jen konečného nebo spočetného počtu hodnot, v podstatě jen celočíselné hodnoty).

Zákon rozdělení pravděpodobnosti vyjadřuje pravděpodobnosti výskytu jednotlivých hodnot náhodné veličiny. Může být vyjádřen dvěma různými způsoby:

- frekvenční funkcí
- distribuční funkcí



Obrázek 5.3 - Příklad frekvenční funkce diskrétní náhodné veličiny

5.3.2 Frekvenční funkce

Frekvenční funkce f(x) udává pravděpodobnost, že určitá náhodná veličina X nabude právě konkrétní hodnoty x. Symbolicky se frekvenční funkce dá zapsat takto:

$$f(x) = P(X = x_i) \tag{5.5}$$

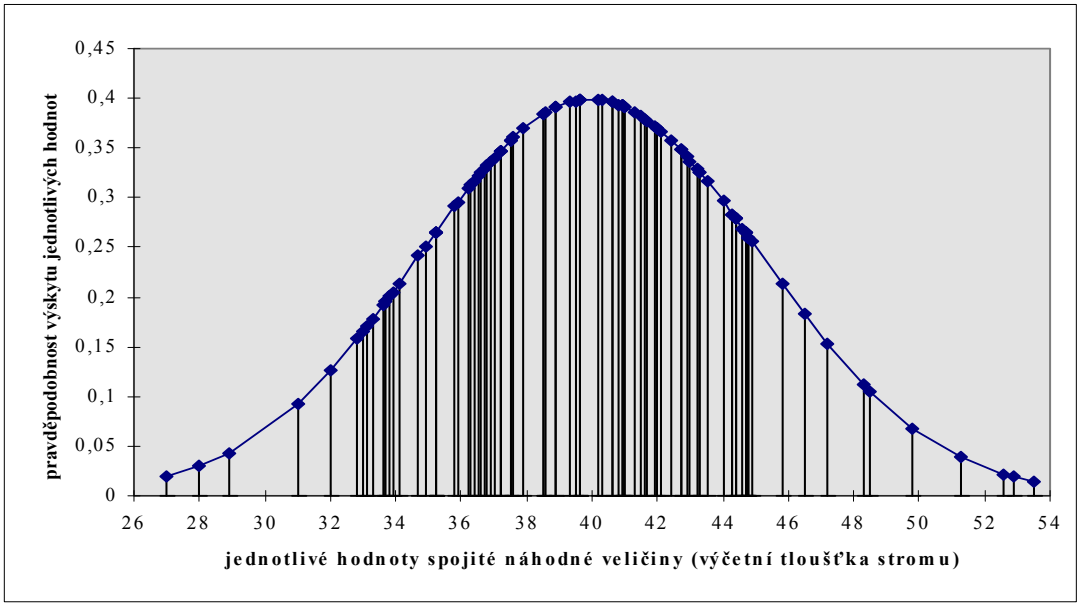
Mějme určitou náhodnou veličinu, např. výšku stromu. Frekvenční funkce nám určí např. pravděpodobnost, s jakou náhodná veličina výška stromu (to je X) nabude hodnotu právě 25 m (to je x). Za výraz f(x) musíme dosadit konkrétní vzorec frekvenční funkce toho rozdělení náhodné veličiny, kterou se příslušná náhodná veličina řídí - předpokládejme, že rozdělení náhodné veličiny výška je normální, potom za f(x) použijeme frekvenční funkci normálního rozdělení (Gauss-Laplaceovu funkci) - podrobněji viz kapitola o spojitých funkcích.

Frekvenční funkce má rozdílnou grafickou interpretaci pro diskrétní a spojitou náhodnou veličinu.

Pro diskrétní veličiny je to soustava kolmic vztyčených v bodech x_1, x_2, ...x_n (každá hodnota je izolovaná, mezi nimi „nic není“), pro spojité veličiny je to plynulá křivka mezi krajními body určitého intervalu (zde může náhodná veličina nabýt jakékoli hodnoty). Frekvenční funkce pro spojité náhodné veličiny se nazývá hustota pravděpodobnosti.

Grafické znázornění frekvenčních funkcí ukazují obrázky 5.3 a 5.4 . Na obrázku 5.3 je příklad ukazující pravděpodobnosti výskytu poruch strojů za směnu v určitém závodě. Počet poruch je typická diskrétní veličina (může nastat pouze 1 porucha, 2 poruchy, … není možné, aby nastalo 3,67854 poruchy apod. – samozřejmě teď neuvažujeme závažnost poruch). Způsob výpočtu těchto pravděpodobností bude ukázán v kapitole věnované diskrétnímu Poissonovu rozdělení. Vidíme, že nejvyšší pravděpodobnost má výskyt 1 a 2 poruch, zatímco výskyt více než 7 poruch za směnu je jev prakticky nemožný.

Obrázek 5.4 ukazuje frekvenční funkci (zde obvykle nazývanou hustota pravděpodobnosti) pro spojitou náhodnou veličinu, v tomto případě pro výčetní tloušťku stromů. Je zřejmé, že tloušťka může nabývat v daném intervalu jakýchkoliv hodnot (jejich přesnost je dána pouze přesností měření). Toto konkrétní rozdělení bylo generováno jako rozdělení normální a způsob výpočtu bude komentován v příslušné kapitole věnované normálnímu rozdělení. Svislé čáry ukazují polohu jednotlivých hodnot náhodné veličiny (tj. konkrétních tlouštěk, ty jsou vyneseny na ose x) a zároveň velikost pravděpodobnosti výskytu (spolu s body na křivce) jednotlivých hodnot náhodné proměnné (čím je čára delší, tím je pravděpodobnost výskytu vyšší).



Obrázek 5.4 - Příklad frekvenční funkce (hustoty pravděpodobnosti) spojité náhodné veličiny

5.3.3 Distribuční funkce

Distribuční funkce udává pravděpodobnost, že náhodná veličina X nabude nejvýše hodnotu x (tj. udává pravděpodobnost, že konkrétní hodnoty náhodné veličiny nepřekročí předem danou horní hranici, která je dána hodnotou x). Symbolický zápis vypadá takto:

$$F(x) = P(X \leq x) \tag{5.6}$$

Pokud bychom aplikovali tento vztah na předchozí příklad týkající se výšky stromů, potom distribuční funkce odpoví na otázku, jaká je pravděpodobnost, že náhodná veličina výška stromu dosáhne nejvýše výšku 25 m (tj. jaká je pravděpodobnost, že jednotlivé výšky dosáhnou jakékoli nižší výšky nebo přímo hodnotu 25 m).

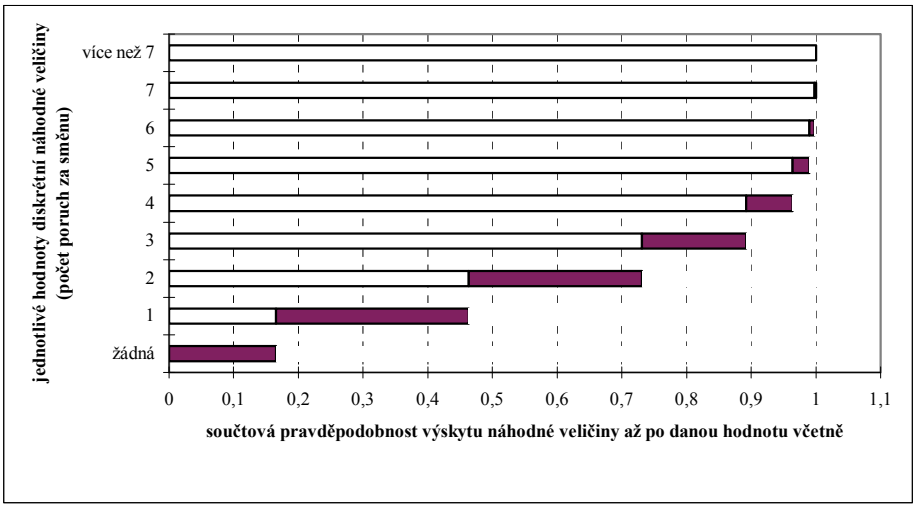
Je zřejmé, že distribuční funkce se dá odvodit z funkcí frekvenčních. Pro diskrétní veličinu platí

$$F(x) = \sum_{x_i \leq x} f(x) \tag{5.7}$$

což znamená, že distribuční funkce diskrétní náhodné veličiny je součtem frekvenčních funkcí všech konkrétních hodnot x_i menších nebo stejně velkých jako je hodnota x.

V kapitole věnované frekvenční funkci jsme se ptali, jaká je pravděpodobnost, že se za směnu vyskytnou právě 3 poruchy? Odpověď zní - 0,161. Jestliže se ale ptáme, jaká je pravděpodobnost, že se vyskytnou nejvýše tři poruchy (tj. žádná, jedna, dvě nebo tři), potom nám odpověď dá distribuční funkce - musíme sečíst hodnoty frekvenční funkce pro x = 0,1,2,3, tj. F(x)= 0,165 + 0,298 + 0,268 + 0,161 = 0,892, tj. téměř 90 %. Grafické znázornění distribuční funkce pro všechny hodnoty z tohoto příkladu je na obrázku 5.5 . Tato stupňovitá čára je nespojitá zprava.

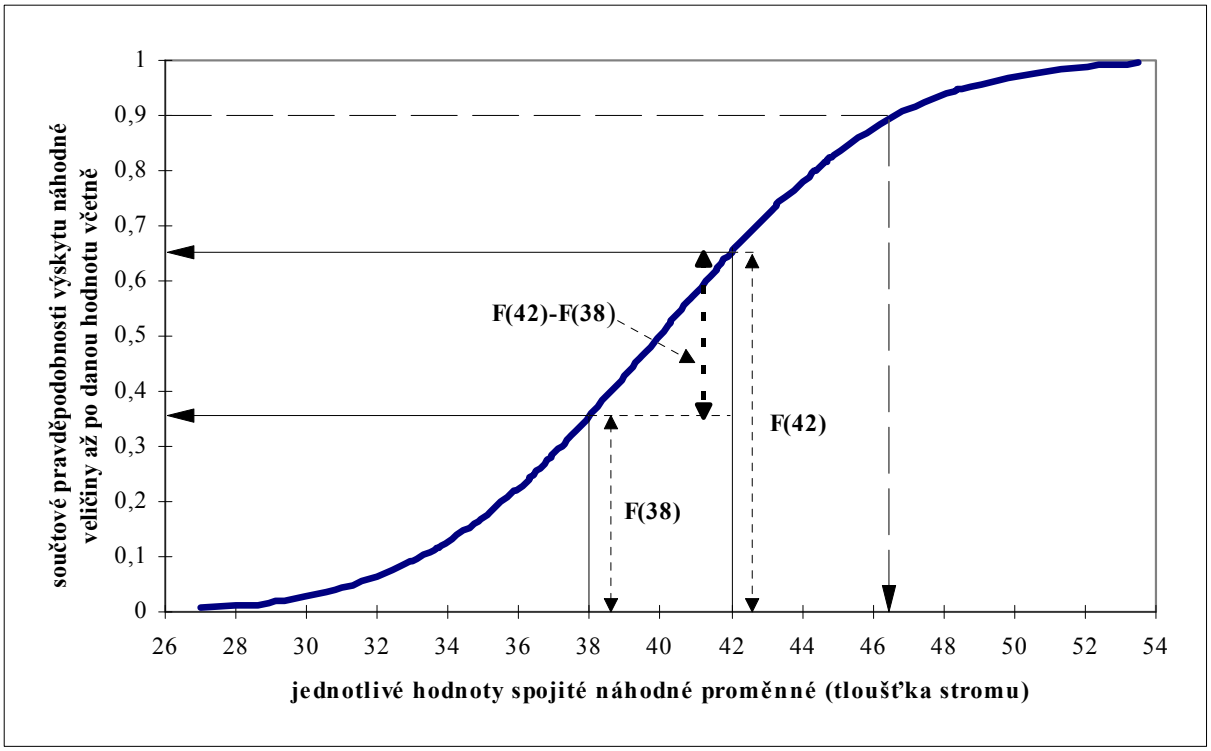
x_i	$F(x_i)$
více než 7	1,000
7	1,000
6	0,998
5	0,990
4	0,964
3	0,892
2	0,731
1	0,463
žádná	0,165



Obrázek 5.5 – Příklad distribuční funkce diskrétní náhodné veličiny

U spojitých funkcí se situace obdobná, jen musíme mít na paměti, že se jedná o spojitý interval, tedy sumace nemůže být provedena jako součet, ale integrálem

$$F(x) = \int_{-\infty}^{x} f(x) \cdot d(x) \tag{5.8}$$



Obrázek 5.6 – Distribuční funkce spojité náhodné veličiny. Bližší vysvětlení viz v textu.

Grafické znázornění je na obrázku 5.6 . Distribuční funkce je rostoucí (nebo minimálně neklesající) funkce, kde na ose x jsou jednotlivé hodnoty náhodné veličiny, na ose y součtové pravděpodobnosti.

Vzhledem k tomu, že se dá distribuční funkce určit z frekvenční funkce a naopak, stačí proto pro úplný popis určité funkce náhodné veličiny znát jednu z nich a druhou odvodit.

Využití distribuční funkce je zřejmé z obrázku 5.6 Jestliže např. potřebujeme vědět, jaká je pravděpodobnost výskytu hodnot výčetních tlouštěk menších nebo rovných 38 cm (zapisujeme $P(X = x \leq 38)$), potom si vyhledáme na ose x hodnotu 38, vztyčíme pořadnici na křivku distribuční funkce a pro průsečík odečteme na ose y příslušnou hodnotu. Pokud chceme znát tuto hodnotu přesně, musíme ji buď vypočítat nebo, pro nejběžnější rozdělení náhodných veličin, ji nalezneme ve statistických tabulkách (zde z grafu normálního rozdělení odečteme zhruba hodnotu 0,36 (přesná hodnota je 0,35498).

Podobně můžeme použít opačný postup (např. která hodnota odpovídá pravděpodobnosti 0,9? V grafu je tento případ naznačen čárkovanou čarou a na ose x odečteme hodnotu 46,6 (přesná hodnota pro normální rozdělení je 46,67). Pokud danou pravděpodobnost nazveme P, potom platí

$$F(x_P) = P \tag{5.9}$$

a hodnota x_P se nazývá P(.100 (%))-kvantil daného rozdělení spojité náhodné veličiny.

Tato vlastnost distribuční funkce je zvláště důležitá pro řešení praktických úkolů matematické statistiky (např. odhady parametrů základního souboru, statistické testování apod.). Významná je také možnost určit pravděpodobnost, že hodnota náhodné veličiny X se nachází v určitém intervalu hodnot $\langle x, x + \Delta x \rangle$, protože platí (viz obrázek 5.6)

$$P[x < X < x + \Delta x] = F(x + \Delta x) - F(x) \tag{5.10}$$

Např. potřebujeme znát, s jakou pravděpodobností se vyskytuje výčetní tloušťka v intervalu $\langle 38,42 \rangle$ cm. Zjistíme hodnoty (výpočtem nebo z tabulek pro příslušné rozdělení) F(42) a F(38) a jejich odečtením získáme potřebnou hodnotu pravděpodobnosti. V případě normálního rozdělení to budou hodnoty F(42) = 0,651951, F(38) = 0,35498, tedy P(38 < X < 42) = 0,651951 - 0,35498 = 0,296971, tj. necelých 30 %. Celý výpočet ze zřejmý z obrázku 5.6 .

Stejně snadno se dá vypočítat, s jakou pravděpodobností hodnota náhodné veličiny překročí danou mez

$$P(X > x) = 1 - F(x) . \qquad (5.11)$$

5.4 Teoretická rozdělení náhodných veličin

Všechny vzorce uvedené v této kapitole vznikly na základě určitých pravděpodobnostních úvah odpovídajícími výpočty. Tyto úvahy a výpočty pouze naznačíme bez použití náročného matematickém aparátu. Umožňuje nám to věta z teorie matematické statistiky, která říká, že libovolná funkce definovaná na číselné ose a splňující podmínku $\sum_{x \in (-\infty,\infty)} \pi(x) = 1$ pro diskrétní náhodnou veličinu je pravděpodobnostní funkce a funkce s podmínkou $\int_{-\infty}^{\infty} f(x)dx = 1$ pro spojitou náhodnou veličinu je hustota pravděpodobnosti.

Náhodné veličiny, jejichž rozdělení zkoumáme, se určitým způsobem zkonstruují - sestaví. Vypočítají se hodnoty a k těmto hodnotám se určí pravděpodobnosti jako hodnoty příslušné funkce pravděpodobnosti. U každého rozdělení pravděpodobnosti uvedeme střed a rozptyl rozdělení.

Uvedeme pouze ta rozdělení, která mají poměrně častou aplikaci v biometrii, ekonometrii nebo v technických vědách.

5.4.1 Teoretická rozdělení diskrétních náhodných veličin

V této kapitole budeme označovat symbolem

$\pi(x)$ frekvenční (pravděpodobnostní) funkci diskrétní náhodné veličiny,

μ střední hodnotu rozdělení,

σ^2 rozptyl rozdělení.

Každé zde uváděné rozdělení má své označení (symboly) a parametry. Parametry jsou ty hodnoty, které musíme znát, abychom mohli libovolné hodnotě x přiřadit její pravděpodobnost na základě daného rozdělení. Jsou uváděny v závorce za označením rozdělení – např. binomické rozdělení je označováno Bi a má parametry n (počet nezávislých pokusů) a p (apriorní pravděpodobnost výskytu jevu). Proto, když se napíše např. Bi(20;0,40), znamená to, že se jedná o binomické rozdělení pro 20 výběrů a pro apriorní pravděpodobnost výskytu studovaného jevu 40 % (0,40).

5.4.1.1 Alternativní rozdělení - A(p)

Alternativní náhodná veličina nabývá pouze dvou hodnot. Při experimentu jev buď nastane, a to s pravděpodobností p, nebo nenastane s pravděpodobností 1 - p. Označíme-li hodnoty veličiny X symboly 1 a 0, je pravděpodobnostní funkce dána formou

$\pi(x) = p$ pro $x = 1$
$= 1 - p$ $x = 0$
$= 0$ všechna ostatní x.

Dále platí, že střední hodnota rozdělení je $\mu = p$ a rozptyl $\sigma^2 = p \cdot (1 - p)$

Příklad 5.5:

V zásilce je ze 100 výrobků 5 poškozených a 95 nepoškozených. Určete pravděpodobnost výskytu nepoškozeného výrobku.

Je-li v zásilce ze 100 výrobků 5 poškozených a 95 nepoškozených, potom platí, že
$\pi(x) = 0,95$ výrobek je nepoškozený
$\pi(x) = 0,05$ výrobek je poškozený
Jev je nepoškozený výrobek.

5.4.1.2 Binomické rozdělení - Bi (n,p)

Předpokládejme, že určitý pokus opakujeme n-krát za stejných podmínek. V každém pokusu může nastat náhodný jev A se stejnou a **předem známou pravděpodobností** *p* a nenastat s pravděpodobností *1 - p*. Podstatné je, že pokusy jsou na sobě **nezávislé**, tj. pravděpodobnost nastoupení jevu A v jednom pokusu není nijak ovlivněna výsledky ostatních pokusů. Typickým příkladem tohoto typu experimentu je např. hod mincí. Za sledovaný jev prohlásíme padnutí jedné strany mince (např. líce). Je známá předem daná pravděpodobnost, že padne jedna strana mince (p = 0,5, tedy 1 – p je také 0,5), jedná se o nezávislé pokusy (předchozí hody mincí nemohou ovlivnit, která strana padne v dalším hodu) a pokusy (hody) můžeme za stejných podmínek mnohonásobně opakovat. Obecně toto schéma platí pro jakýkoli výběr s opakováním (s vracením vybraných prvků). Za těchto podmínek hledáme takové rozdělení, které nám popíše pravděpodobnost nastoupení různého počtu výskytů jevu A (např. pravděpodobnost, že padne jedna strany mince 0-krát, 1-krát, 2-krát , …).

Binomická náhodná veličina tedy udává **celkový počet nastoupení jevu v $n \geq 1$ nezávislých realizacích náhodného experimentu. Při každé realizaci může nastat jev s neměnnou pravděpodobností *p* a neúspěch s pravděpodobností *1 - p*.**

Frekvenční funkce je dána

$$\pi(x) = \begin{cases} \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x} & \text{pro } x = 0,1,2,3,.... \\ \\ 0 & \text{pro jiná x} \end{cases} \qquad (5.12)$$

Střední hodnota a rozptyl se vyčíslí podle vzorců

$$\mu = n \cdot p \qquad (5.13)$$

$$\sigma^2 = n \cdot p \cdot (1 - p) \tag{5.14}$$

Parametry Bi rozdělení jsou *n*, *p*. Vzhledem k výrazu $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ se hodnota frekvenční funkce pro vyšší hodnoty *n* a *x* obtížně numericky vyčísluje (důvodem je nutnost pracovat s faktoriály, takže při výpočtech s vysokými čísly mohou selhat i počítače). K výpočtu se v těchto případech obvykle používá různých aproximací. Jedna z nich je založena na použití lokální Moivreovy-Laplaceovy věty, která využívá aproximaci normálním rozdělením. O tomto způsobu výpočtu si podrobněji povíme v kapitole věnované normálnímu rozdělení.

Tvar binomického rozdělení určují parametry (*n, p*). Při určitém (stejném) *n* je pro

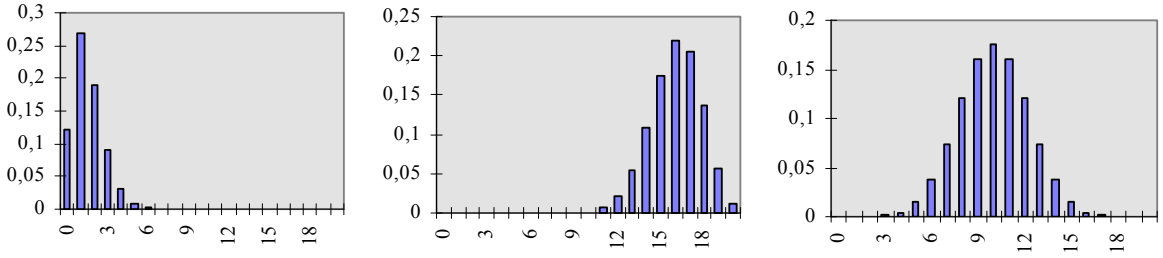
p = 1 - p = 0,5 souměrné

p < 1 - p kladně (levostranně) nesouměrné

p > 1 - p záporně (pravostranně) nesouměrné.

Tuto situaci ilustruje obrázek 5.7 . Ve všech případech je znázorněno binomické rozdělení pro n =20 lišící se hodnotami *p*: na levém obrázku je p = 0,1 (znamená to, že vysokou pravděpodobnost mají nízké hodnoty výskytu studovaného jevu, prakticky od 0 – 4 z 20 pokusů, výskyt více než 5 jevů z 20 pokusů je při této hodnotě *p* prakticky nemožný), na prostředním obrázku je p = 0,8 (znamená to, že nejčastější jsou časté výskyty studovaného jevu, prakticky od 11 do 20 výskytů z 20 pokusů, méně výskytů je prakticky nemožných) a na pravém obrázku je případ p = 0,5, tedy souměrné rozdělení, kdy nejvíce výskytů je pro střední hodnotu (10) a nejméně časté (prakticky nemožné) jsou buď ojedinělé výskyty nebo naopak velmi časté výskyty jevu.

Jestliže platí nerovnost, že n . p. (1-p) > 9, lze prakticky binomické rozdělení nahradit rozdělením normálním N s parametry $\mu = np$, $\sigma^2 = np(1-p)$, jestliže je *n* velké a p < 0,1 rozdělením Poissonovým Po s parametrem $\lambda = np$.



Obrázek 5.7 - Příklady různých binomických rozdělení pro n = 20 a pro hodnoty p = 0,1 (vlevo); p = 0,8 (uprostřed) a 0,5 (vpravo). Bližší vysvětlení viz v textu.

Příklad 5.6 (podle HEBÁK-KAHOUNOVÁ 1988):

Je křížen bělokvětý hrách s fialovokvětým, přičemž předpokládáme, že rostliny, na nichž je pokus prováděn, dosud nebyly kříženy. Podle pravidel dědičnosti je možno očekávat, že 75 % nově vzniklých potomků pokvete fialově a 25 % bíle. Zatím vzklíčilo 10 nových rostlin. Jaká je pravděpodobnost, že

a) žádná nepokvete bíle?

b) fialově pokvetou alespoň 3?

c) fialově pokvete alespoň 6 a nejvýše 8 rostlin?

Za předpokladu, že nové rostliny mohou nezávisle na sobě kvést bíle nebo fialově a že pravděpodobnost fialového květu je pro každou rostlinu stejná, má náhodná veličina „počet fialově kvetoucích rostlin“ binomické rozdělení Bi(10;0,75).

a) pravděpodobnost, že všechny rostliny pokvetou fialově (a tedy žádná bíle), tj.bude 10 fialových rostlin z 10 vyklíčených, je

$$\pi(10) = \binom{10}{10} \cdot 0{,}75^{10} \cdot 0{,}25^{0} = 0{,}0563\,,$$ tj. tato pravděpodobnost je necelých 6 %.

b) pravděpodobnost, že alespoň 3 květy budou fialové, je dána součtem pravděpodobností pro hodnoty 4, 5, 6, 7, 8, 9, 10 (protože chceme znát pravděpodobnost toho, že fialové budou květy 3 nebo 4 nebo 5 ... nebo 10, ale nejméně 3). Na této úloze se dá dobře ukázat význam doplňkové pravděpodobnosti.

Základní vztah je

$$\pi(X \geq 3) = \pi(3) + \pi(4) + \pi(5) + \pi(6) + \pi(7) + \pi(8) + \pi(9) + \pi(10).$$

Abychom nemuseli počítat tolik frekvenčních funkcí, je možné výhodně využít pravděpodobnosti opačného jevu, tj. vypočítat pravděpodobnost, že nejvýše 3 květy budou fialové a pak ji odečíst od jedné a požadovaný výsledek získat jako doplňkovou pravděpodobnost, tedy

$$\pi(X \geq 3) = 1 - \pi(X < 3) = 1 - [\pi(0) + \pi(1) + \pi(2)] = 1 - (9{,}53674.10^{-7} + 2{,}86102.10^{-5} +$$
$$+ 0{,}000386238) = 1 - 0{,}000416 = 0{,}999584,$$ jedná se tedy o jev prakticky jistý.

c) pravděpodobnost, že sledovaný jev „počet fialových květů“ nabude hodnoty v intervalu od 6 do 8, je roven analogicky součtu frekvenčních funkcí pro hodnoty X = 6,7,8, tedy hodnoty 0,146, 0,25 a 0,282, tedy s výsledkem 0,678.

Příklad 5.7 (podle HEBÁK-KAHOUNOVÁ 1988):

Výrobní závod dodává výrobky balené na paletách po 10 kusech. Výrobce předpokládá, že každá paleta, kde bude zjištěn alespoň jeden vadný výrobek, bude reklamována a výrobce vrátí odběrateli peníze. Je známo, že pravděpodobnost vyrobení kvalitního výrobku je 0,95 a náklady na výrobky na jedné paletě jsou 2000 Kč. Jakou cenu má výrobce stanovit, chce-li očekávat zisk 25 %?

Předpokládáme, že výrobky jsou vybírány z velkého množství náhodně a baleny do palet. Potom náhodná veličina „počet zmetků na paletě“ je dána binomickým rozdělením Bi(10; 0,05) Pravděpodobnost toho, že výrobce utrží za dodanou paletu částku (označíme ji jako *c*) se rovná pravděpodobnosti, že na paletě nebude žádný výrobek vadný, tj.náhodná veličina musí nabýt hodnotu 0:

$$\pi(0) = \binom{10}{0} \cdot 0{,}05^{0} \cdot 0{,}95^{10} = 0{,}5987$$, tedy tato pravděpodobnost dosahuje asi 60 %

Výrobce předpokládá z jedné palety zisk 500 Kč (25 % ze 2000 Kč výrobních nákladů) a hledaná cena se vypočítá z rovnice

$$0{,}5987\,c - 2000 = 500 \Rightarrow c = 4\,175,\,70 \cong 4200 \text{ Kč}$$

Příklad 5.8:

Na výrobní lince je pravidelně kontrolována kvalita výrobku během operací určité skupiny. Z každých 100 výrobků je kontrolováno 10. Výrobek je po kontrole ihned vrácen na linku. Jaká je pravděpodobnost, že při kontrole v následující skupině operací budou kontrolovány stejné výrobky jako v předchozí?

Pravděpodobnost výběru při první kontrole je $p = \frac{10}{100} = 0{,}1$; potom je 1 - p =0,9.

Hodnota $\pi(10) = \binom{10}{10}.0{,}1^{10}.0{,}9^{0} = 1{,}0 \cdot .10^{-10}$. Tento výsledek znamená, že uvažovaný jev má mizivou pravděpodobnost - je prakticky nemožný.

5.4.1.3 Hypergeometrické rozdělení - H (n, N, M)

Hypergeometrické rozdělení je zevšeobecněním binomického rozdělení **pro závislé pokusy** (výběry bez opakování). Používá se za následujících podmínek:

- známe velikost základního souboru N (počet všech realizací náhodného experimentu),
- v rámci základního souboru známe počet prvků M, které jsou nositelem zkoumaného jevu (buď tuto hodnotu známe přesně nebo můžeme použít kvalifikovaný odhad),
- jedná se o výběr bez opakování (bez vracení), kdy pravděpodobnost výběru prvku se znakem A (zkoumaným jevem) není při všech pokusech stejná, ale mění se v závislosti na výsledcích předchozích pokusů.

Příkladem může být pravděpodobnost výhry ve Sportce (uvažujeme pouze jeden vyplněný sloupec a neuvažujeme žádná prémiová čísla apod.):

- N je velikost základního souboru, tedy počet všech čísel - 49
- M je nositelem zkoumaného jevu (to je tažené číslo) - 6
- n je počet závislých pokusů bez vracení (to je počet vsazených čísel, výběr bez opakování je to proto, že v jednom tahu nemůžeme žádné číslo zaškrtnout dvakrát
- x je počet vyhrávajících čísel - 6 pro 1. cenu, 5 pro druhou atd.

Hypergeometrická náhodná veličina tedy udává **počet realizací náhodného experimentu, které vykazují zkoumaný jev (x), je-li N ≥ 1 počet všech realizací, M z nich (1 ≤ M ≤ N) je nositelem zkoumaného jevu a ze všech možných N provedeme n (1 ≤ n ≤ N) realizací.**

Potom

$$\pi(x) = \begin{cases} \pi(x) = \dfrac{\binom{M}{x} \cdot \binom{N-M}{n-x}}{\binom{N}{n}} & \text{pro } x = \max(0, M - N + n), \ldots, \min(M, n) \\ 0 & \text{pro jiná } x \end{cases} \tag{5.15}$$

Střední hodnota a rozptyl se vypočítají

$$\mu = n.\frac{M}{N}; \tag{5.16}$$

$$\sigma^2 = n.\frac{M}{N}.\left(1 - \frac{M}{N}\right).\frac{N-n}{N-1} \tag{5.17}$$

Střední hodnota hypergeometrického rozdělení je totožná se střední hodnotou binomického rozdělení Bi(n,p), kde $p = M/N$. Rozptyl hypergeometrického rozdělení je roven součinu rozptylu binomického rozdělení a zlomku $(N-n)/(N-1)$. Protože hodnota tohoto zlomku je menší nebo nejvýše rovna (pro $n = 1$) jedné, je rozptyl hypergeometrického rozdělení většinou menší než binomického. Z toho plyne, že úsudky vytvořené na základě výběru bez vracení jsou obvykle přesnější než na základě výběru s vracením.

Vzhledem k tomu, že vyčíslení hypergeometrického rozdělení je numericky velmi obtížné, používá se různých aproximací, kdy pravděpodobnost $\pi(x)$ se počítá podle vzorců jiných rozdělení pomocí transformovaných parametrů, např. hodnota pravděpodobnostní funkce H(n, N, M) rozdělení daná vzorcem

$$\pi(2) = \frac{\binom{50}{2}\binom{450}{48}}{\binom{500}{50}} \qquad \text{pro N = 500;} \quad \text{m = 50; n = 50; x = 2}$$

se aproximuje binomickým rozdělením Bi (M, p*)

kde $p^* = \frac{2n - x}{2N - N + 1}$, tj. $p^* = \frac{100 - 2}{1000 - 50 + 1} = 0{,}10305$

a $\pi(2) = \binom{50}{2}.0{,}10305^2.0{,}89695^{48} = 0{,}07$

Jestliže je N vzhledem n dostatečně velké (za spolehlivou hranici se považuje podíl $n/N \leq 0{,}05$), potom se hypergeometrické rozdělení blíží rozdělení binomickému a můžeme postupovat při výpočtu pravděpodobnosti u výběru bez opakování stejně jako výběru s opakováním. Tato možnost se často využívá, protože zpřesnění závěrů snížením rozptylu při použití hypergeometrického rozdělení není obvykle úměrné numerické náročnosti výpočtů (HEBÁK-KAHOUNOVÁ 1988).

Hypergeometrické rozdělení má široké uplatnění v praxi. Používá se ve statistické kontrole jakosti, vyskytuje se jako pravděpodobnostní model některých sázkových her, např. Sportky, používá se při zjišťování počtu jedinců v populaci, atd.

Příklad 5.9 (podle HEBÁK-KAHOUNOVÁ 1988):

Odběratel stanovil následující podmínky výběrové kontroly přejímaného zboží. Z každé zásilky je náhodně (bez vracení) vybráno 30 % výrobků. Je-li ve výběru méně než 1 % zmetků, je zásilka automaticky přijata, je-li ve výběru více než 40 % zmetků, je automaticky odmítnuta. V případě, že ve výběru je procento zmetků v rozsahu 1 – 40 %, je proveden nový náhodný výběr dalších 20 % výrobků (z původního počtu). Pak se zásilka přijme jen tehdy, jestliže ve druhém výběru není více než 1 % zmetků, jinak je zásilka vrácena jako nekvalitní. V zásilce 10 výrobků je 20 % zmetků. Jaká je pravděpodobnost, že odběratel zásilku odmítne, bude-li provádět kontrolu podle výše uvedeného postupu?

Zkoumaný jev je „počet zmetků v zásilce“. Ze zadání vyplývá, že v zásilce jsou z celkem 10 výrobků (velikost základního souboru N) 2 zmetky (20 % z 10, to je hodnota M, tj. počet nositelů zkoumaného jevu)

Zásilka bude zamítnuta ve dvou případech:

1) jestliže v kontrolním výběru tří výrobků (30 % z 10, to je hodnota n – velikost výběru) budou oba zmetky (protože 2/3 = 0,67, tj. podíl zmetků v kontrolovaném výběru by byl

67 %, tj. nad normou 40%). Musíme tedy vypočítat, jaká je pravděpodobnost, že ve výběru o velikosti 3 výrobky budou oba zmetky.

2) jestliže v kontrolním výběru tří výrobků bude jeden zmetek (1/3 je asi 33%, tj. v rozmezí 1 - 40 %), provede se další výběr dvou výrobků (20% z původních 10) a zásilka se odmítne, jestliže zde bude nalezen druhý zmetek (potom 1/2 je 50 %, tj. nad normou 1 % zmetků pro druhý výběr). Musíme tedy nalézt pravděpodobnost, že v prvním výběru bude 1 zmetek a zároveň ve druhém výběru také jeden zmetek.

Protože zásilka může být odmítnuta oběma způsoby, musíme sečíst výsledné pravděpodobnosti obou možností.

Tabulka 5.2 udává jednotlivé veličiny pro obě varianty odmítnutí zásilky. Z výsledného výpočtu je možné učinit tyto závěry:

- celková pravděpodobnost odmítnutí zásilky je 20 %, tj. za zadaných podmínek bude odmítnuta každá pátá zásilka se zadanými parametry
- pravděpodobnost, že při kontrole budou „odhaleny“ oba zmetky, je velmi nízká (necelých 7%),
- pravděpodobnost odmítnutí zakázky 2. variantou je pravděpodobnější – více než 13 % .

$$\pi = \underbrace{\frac{\binom{M_1}{x_1}\cdot\binom{N_1-M_1}{n_1-x_1}}{\binom{N_1}{n_1}}}_{\text{pravděpodobnost odmítnutí podle 1.varianty}} + \underbrace{\underbrace{\frac{\binom{M_2}{x_2}\cdot\binom{N_2-M_2}{n_2-x_2}}{\binom{N_2}{n_2}}}_{\text{1.výběr}}\cdot\underbrace{\frac{\binom{M_3}{x_3}\cdot\binom{N_3-M_3}{n_3-x_3}}{\binom{N_3}{n_3}}}_{\text{2.výběr}}}_{\text{pravděpodobnost odmítnutí podle 2.varianty}}$$

Obecný význam veličiny	Symbol	Význam veličiny v příkladu	1. varianta		2.varianta			
					1. výběr		2. výběr	
			Symbol	Hodnota	Symbol	Hodnota	Symbol	Hodnota
velikost základního souboru	N	celkový počet výrobků	N_1	10	N_2	10	N_3	7
počet nositelů jevu v základním souboru	M	celkový počet zmetků	M_1	2	M_2	2	M_3	1
velikost výběru	n	počet kontrolovaných výrobků	n_1	3	n_2	3	n_3	2
hodnota náhodné veličiny	x	počet zmetků ve výběru	x_1	2	x_2	1	x_3	1

Tabulka 5.2 - Vstupní veličiny pro výpočet příkladu 5.9

$$\pi = \frac{\binom{2}{2}\binom{8}{1}}{\binom{10}{3}} + \frac{\binom{2}{1}\binom{8}{2}}{\binom{10}{3}}\cdot\frac{\binom{1}{1}\binom{6}{1}}{\binom{7}{2}} = 0{,}066667 + (0{,}466667\cdot 0{,}285714) = 0{,}20$$

Příklad 5.10:

Při kontrole výrobků před expedicí je vždy ze 100 výrobků najednou /výběr bez vracení) vybrána skupina 5 výrobků, které jsou zkontrolovány. při převzetí zásilky je rovněž ze 100 kusů vybrána skupina 5 výrobků a zkontrolována. Jaká je pravděpodobnost, že při přebírací kontrole budou ke kontrole vybrány 2 výrobky kontrolované před expedicí?

$$\pi(2)=\frac{\binom{2}{5}\binom{95}{3}}{\binom{100}{5}}=0{,}018 \qquad \text{N = 100; M = 5; n = 5; x = 2}$$

5.4.1.4 Poissonovo rozdělení - Po(λ)

S Poissonovým rozdělením se setkáme všude tam, kde jde o **četnost jevu v mnoha pokusech a když výskyt jevu má v jednotlivém pokusu jen malou pravděpodobnost**. Jev se objeví ojediněle ve velkých sériích realizací náhodného experimentu. Využívá se zejména v problematice spolehlivosti strojů a zařízení a v problematice hromadné obsluhy.

Dále se tímto rozdělením řídí náhodná veličina, kterou je počet výskytu sledovaného jevu v určitém časovém intervalu délky *t*. Uvedená veličina má Poissonovo rozdělení, jestliže

- jev může nastat v kterémkoli časovém okamžiku,
- počet výskytů jevu během časového intervalu závisí jen na jeho délce a ne na jeho počátku ani na tom, kolikrát jev nastoupil před jeho počátkem,
- pravděpodobnost, že jev nastoupí více než jednou v intervalu délky *t*, konverguje k nule rychleji než *t,*
- λ je střední hodnota počtu výskytů jevu za časovou jednotku.

Z hlediska praktického použití je Poissonovo rozdělení nejdůležitější rozdělení diskrétní náhodné veličiny. Příkladem jeho použití může být počet zmetků ve velké sérii výrobků (pokud pravděpodobnost výskytu zmetku je velmi malá), výskyt úrazů nebo chorob, úmrtí v určitém časovém intervalu apod.

Poissonovo rozdělení vzniká limitním přechodem z binomického při $n \to \infty$ a $p \to 0$. Říká se mu též zákon malých čísel.

Poissonova náhodná veličina má pravděpodobnostní funkci

$$\pi(x)=\begin{cases}\pi(x)=\dfrac{e^{-\lambda}.\lambda^{x}}{x!} & \text{pro } x = 0, 1, 2,....\\ \\ 0 & \text{pro jiná x}\end{cases} \qquad (5.18)$$

Střední hodnota a rozptyl se rovnají parametru λ, tedy $\mu = \sigma^2 = \lambda$.. Poissonovo rozdělení má parametr λ, přičemž $\lambda = n \,.\, p$, kde n je počet pokusů (roste nade všechny meze, tedy $n \to \infty$) a p je pravděpodobnost výskytu jevu (je velmi malá, tedy $p \to 0$). Hodnota λ je vždy kladná.

Příklad 5.11 (podle HEBÁK-KAHOUNOVÁ 1988):
Podle tabulek pravděpodobnosti dožití je pravděpodobnost toho, že 25-tiletý muž přežije další rok, rovna 0,998. Pojišťovna nabízí mužům tohoto věku, že při ročním pojistném 500 Kč vyplatí pozůstalým v případě úmrtí pojištěnce 100 000 Kč. U pojišťovny je takto pojištěno 1000 mužů. Jaká je pravděpodobnost, že zisk pojišťovny na konci roku z tohoto druhu pojištění bude alespoň 300 000 Kč?

Náhodná veličina „počet zemřelých pojištěnců během roku" má Poissonovo rozdělení, protože *n* je velmi vysoké (1000 pojištěných) a pravděpodobnost výskytu jevu (úmrtí pojištěnce během roku) je velmi nízká (p = 0,002). Parametr λ je 1000. 0,002 = 2. Pokud žádný pojištěnec během roku nezemře, bude zisk pojišťovny 500 000 Kč (500 Kč pojistného od 1000 mužů). Zisk nejméně 300 000 Kč pojišťovna dosáhne, jestliže během roku nezemřou více než 2 pojištěnci.Řešením úlohy je tedy

$$\pi(X \leq 2) = \pi(0) + \pi(1) + \pi(2)$$

Dosadíme do vzorce 5.18 za x, postupně 0,1,2 a za λ hodnotu 2 a získáme následující řešení:

$$\pi(X \leq 2) = 0{,}1353 + 0{,}2707 + 0{,}2707 = 0{,}6767$$

Pojišťovna má za těchto podmínek tedy pravděpodobnost necelých 70 % (67,7 %), že dosáhne očekávaného zisku.

Příklad 5.12:
Z výsledků dlouhodobých pozorování poruchovosti strojů ve výrobním závodě bylo zjištěno, že střední hodnota počtu poruch strojů za směnu v rámci závodu je 1,8. Určete, jaká je pravděpodobnost výskytu počtu poruch v intervalu 1-7.
Pravděpodobnost, že se za směnu vyskytne žádná, jedna, dvě, atd. poruchy, je dána Poissonovým rozdělením a hodnotami danými vztahem

$$\pi(x) = \frac{e^{-1,8}.1,8^{x}}{x!} \qquad \text{pro x = 0, 1, 2, ...}$$

X	0	1	2	3	4	5	6	7
π (x)	0,165	0,298	0,268	0,161	0,072	0,026	0,008	0,002
Σ π (x)	0,165	0,463	0,731	0,892	0,964	0,990	0,998	1,000

Graficky jsou tyto hodnoty zobrazeny v grafech 5.3 a 5.5 . Tabulka obsahuje i kumulativní pravděpodobnost - distribuční funkci. Z tabulky lze např. zjistit, že velmi malou pravděpodobnost má výskyt 4 a více poruch za směnu. Více jak 7 poruch je jev prakticky nemožný.

5.4.2 Teoretická rozdělení spojitých náhodných veličin

Nejprve je třeba zdůraznit, že v teorii matematické statistiky se k popisu hustoty pravděpodobnosti a distribuční funkce používají speciální funkce a to gama funkce Γ (a) a beta funkce B(a,b). Protože teorie těchto funkcí je složitá, nebudeme příslušné vztahy pomocí těchto funkcí odvozovat. Zpravidla pouze uvedeme výsledný vztah popisující zákony pravděpodobnosti příslušného rozdělení.

5.4.2.1 Exponenciální rozdělení - Ex (λ)

Exponenciální náhodná veličina X zpravidla vyjadřuje dobu čekání na nějakou náhodnou událost. Jde o případ, kdy dosud pročekaná doba nijak neovlivňuje šanci na urychlené nastoupení očekávané náhodné události - tzv. čekání bez paměti.

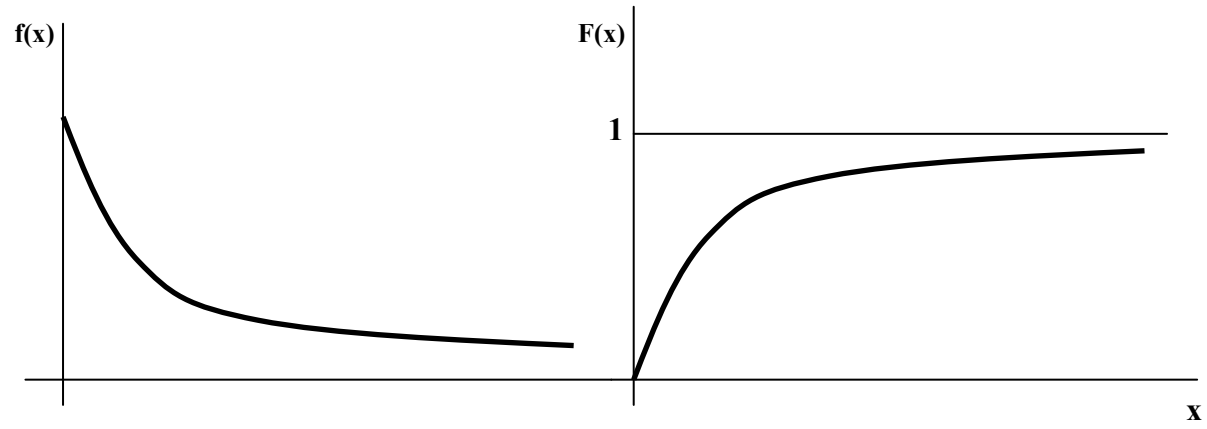
$$f(x) = \begin{cases} \lambda \cdot e^{-\lambda \cdot x} & \text{pro } x \geq 0 \\ 0 & \text{pro jiná } x \end{cases} \qquad (5.19)$$

pro střední hodnotu a rozptyl platí vztahy

$$\mu = \lambda^{-1}; \qquad \sigma^2 = \lambda^{-2}$$

Toto rozdělení se používá v teorii obnovy, v teorii hromadné obsluhy. Popisuje se jím doba životnosti některých výrobků, doba trvání některých akcí, doba čekání na nějakou událost.

Přibližný tvar frekvenční funkce (hustoty pravděpodobnosti) a distribuční funkce znázorňuje obrázek 5.8 .



Obrázek 5.8 - Tvar frekvenční funkce (vlevo) a distribuční funkce (vpravo) exponenciálního rozdělení

Příklad 5.13:

Ukázalo se, že doba opravy poruchy vznětového motoru se řídí exponenciálním rozdělením se střední hodnotou 105 min. Určíme pravděpodobnost, že porucha bude opravena do 8,5 hodiny (jedna pracovní směna).

$$\lambda = \frac{1}{105} = 0{,}010\,\text{min}^{-1}.F(x \leq 510) = \int_0^{510} 0{,}010.e^{-0{,}010x}dx = \left[-e^{-0{,}010x}\right]_0^{500} dx =$$

$$= \left[-e^{-0{,}010x}\right]_0^{500} = 1 - 0{,}006 = 0{,}994 .$$

Pravděpodobnost, že porucha bude opravena právě za 2 hodiny je

$$f(120) = 0{,}010 \,.\, e^{-0{,}010.120} = 0{,}003.$$

Z výsledku řešení je zřejmé, že klást si takové otázky je v praxi nevhodné.

5.4.2.2 Normální rozdělení - N (μ, σ^2)

Normální rozdělení má ve statistické teorii i ve většině aplikací dominantní postavení. Je nejpoužívanějším a nejdůležitějším zákonem rozdělení pravděpodobnosti. Proto mu také budeme věnovat větší pozornost než ostatním rozdělením. Jeho význam spočívá mimo jiné v tom, že pomocí normálního rozdělení je možno aproximovat mnoho jiných – i diskrétních – rozdělení.

Objev normálního rozdělení sahá až do 18. století, kdy v roce 1733 francouzský matematik žijící v Anglii Abraham de Moivre publikoval malý spisek o binomickém rozdělení pro

velká *n*, kde popové použil vztah frekvenční funkce normálního rozdělení. Tato publikace upadla v zapomenutí a přibližně o století později ke stejnému objevu dospěli Carl Friedrich Gauss a Pierre Simon de Laplace. V roce 1924 anglický statistik Karl Pearson (s jeho jménem se ještě mnohokrát setkáme) objevil zapomenutý spisek de Moivrův a v zájmu zamezení mezinárodních sporů o název rozdělení – bylo už zavedeno označení Gaussovo nebo Gauss-Laplaceovo rozdělení - navrhl, aby se toho základní statistické rozdělení nazývalo normální.

Normální rozdělení má mnoho náhodných veličin v reálném světě. V matematické statistice bývá požadavek normálního rozdělení podmínkou pro použití těch neúčinnějších metod, proto se náhodné veličiny jinak rozdělené uměle při matematicko-statistickém zpracování transformují na alespoň přibližně normální náhodnou veličinu.

5.4.2.2.1 Obecné normální rozdělení

Normální rozdělení je zákonem rozdělení součtu libovolných náhodných veličin. Stačí, aby **sčítanců byl dostatečný počet a aby žádný z nich neměl na výslednou náhodnou veličinu rozhodující vliv**.

Vznik normální náhodné veličiny je možno představit si takto. Kdyby nebylo náhodných vlivů, nabyla by zkoumaná veličina X vždy (ve všech pokusech) konstantní hodnoty μ. Na realizaci náhodného experimentu však působí velké množství drobných a nezávislých náhodných vlivů, které působí malé změny w_1, w_2 ..... hodnot veličiny. Předpokládáme, že se tyto malé změny ke konstantě připočítávají, tedy platí

$$X = \mu + w_1 + w_2 \ldots\ldots .$$

Jak rychle výsledná náhodná veličina konverguje k normalitě, závisí na charakteru jednotlivých sčítanců (pro náhodné veličiny, které mají tvar blízký normálnímu rozdělení, stačí několik málo sčítanců, jestliže jednotlivé veličiny mají výrazně nenormální rozdělení, musí jich být více). Nelze tedy jednoznačně stanovit, kolik je „dostatečný počet“ sčítanců.

Hustota pravděpodobnosti normální náhodné veličiny je dána vztahem

$$f(x) = \frac{1}{\sqrt{2\pi}\cdot\sigma}\cdot e^{\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]} \qquad \text{pro } x \in (-\infty,\ \infty) \text{ a pro } \sigma > 0 \tag{5.20}$$

kde je

σ směrodatná odchylka,

μ střední hodnota,

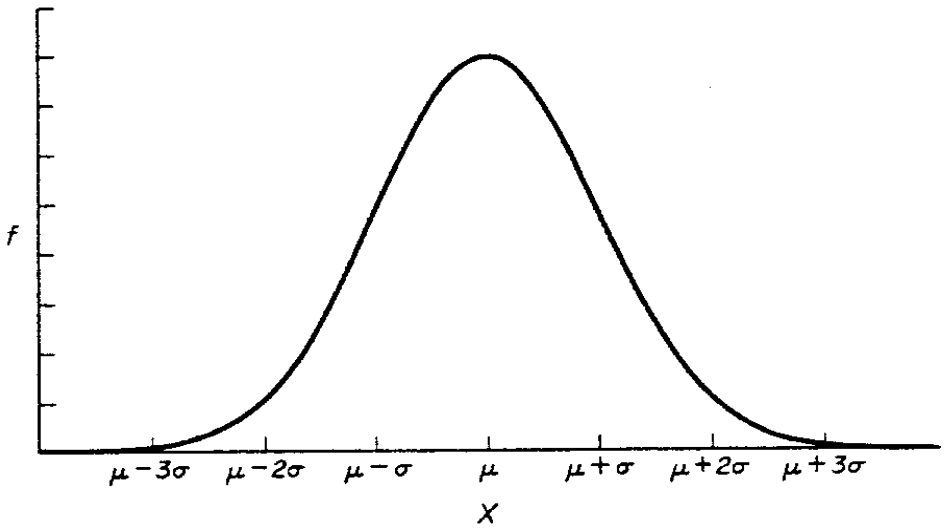
tzv. Gauss - Laplaceovou funkcí. Graf této funkce je běžně znám pod nnázvem Gaussova křivka (viz obrázek 5.9). Gaussova křivka má následující vlastnosti:

- je to symetrická křivka s osou symetrie ve střední hodnotě,
- má inflexní body v s x-ovými souřadnicemi $\mu + \sigma$ a $\mu - \sigma$,
- střední hodnota μ je zároveň i mediánem (prostřední hodnotou) a modem (nejčastější hodnotou)
- v bodě $x = \mu$ dosahuje svého maxima $1/(\sqrt{2\pi}\,\sigma)$
- v intervalu $\langle \mu - \sigma, \mu + \sigma \rangle$ leží necelých 70 % všech hodnot náhodné veličiny X (68,26 %),
- v intervalu $\langle \mu - 2\sigma, \mu + 2\sigma \rangle$ leží asi 95 % všech hodnot náhodné veličiny X (95,44 %), tedy v intervalu mezi jednou a dvěma směrodatnými odchylkami leží asi 27,18

% všech hodnot (13,59 % v intervalu $\langle \mu - \sigma, \mu - 2\sigma \rangle$ a 13,59 % v intervalu $\langle \mu + \sigma, \mu + 2\sigma \rangle$

- v intervalu $\langle \mu - 3\sigma, \mu + 3\sigma \rangle$ leží téměř 100 % všech hodnot náhodné veličiny X (99,724 %), tedy v intervalu mezi dvěma a a třemi směrodatnými odchylkami leží pouze 4,28 % všech hodnot (2,14 % v intervalu $\langle \mu - 2\sigma, \mu - 3\sigma \rangle$ a 2,14 % v intervalu $\langle \mu + 2\sigma, \mu + 3\sigma \rangle$).

Vlastnosti, uvedené v posledních třech bodech, byly již zmiňovány v kapitole věnované směrodatné odchylce (viz kapitola 4. o statistických charakteristikách) včetně příkladu a grafického znázornění na obrázku 4.7.



Obrázek 5.9 - Frekvenční funkce (hustota pravděpodobnosti, Gaussova křivka) normálního rozdělení

Obecné normální rozdělení má dva parametry – střední hodnotu a rozptyl. Kombinací těchto dvou hodnot a jejich dosazením do vztahu 5.20 se mění tvar (změnou rozptylu) a poloha (změnou střední hodnoty) Gaussovy křivky (viz obrázky 5.10 a 5.11).

Tato vlastnost normálního rozdělení je velmi „nepříjemná“, protože to znamená, že při použití obecného normálního rozdělení $N(\mu, \sigma^2)$ musíme hodnoty frekvenční a distribuční funkce počítat pro každé použité obecné normální rozdělení (kterých je vlastně nekonečně mnoho – tolik, kolik je možných kombinací μ a σ^2). Proto byl hledán způsob, jak tuto nevýhodu obejít a zkonstruovat „jednotné“ normální rozdělení, které by bylo lehce použitelné ve statistické praxi. Řešením je tzv. standardizované (normované) normální rozdělení.

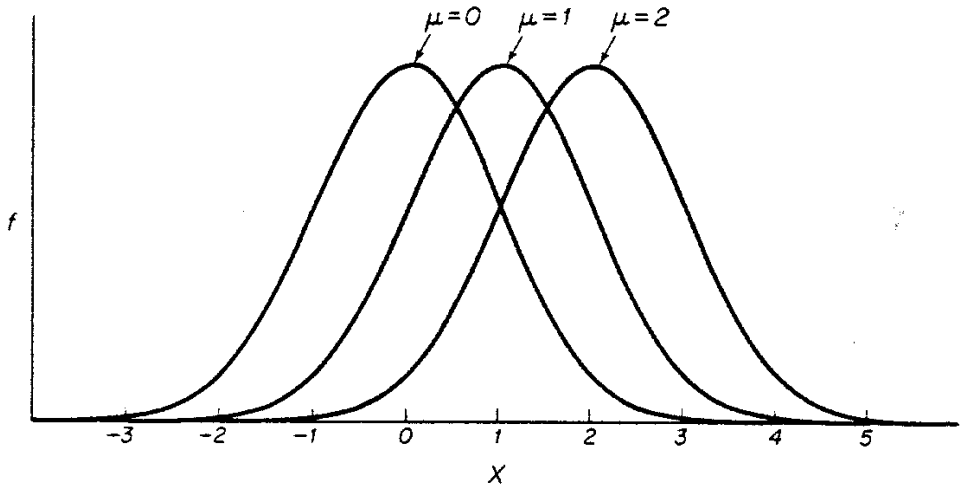
5.4.2.2.2 Standardizované (normované) normální rozdělení

Standardizované normální rozdělení dostaneme z normálního rozdělení substitucí

$$z = \frac{x - \mu}{\sigma} \tag{5.21}$$

Tím vlastně převedeme obecné normální rozdělení, kde jednotlivé hodnoty *x* jsou vyjadřovány v jednotkách měřené veličiny, na takové normální rozdělení, kde jednotlivé hodnoty *x* jsou představovány hodnotami *z* vyjádřenými ve „směrodatných odchylkách“. Mějme např. obecné normální rozdělení $N(50,5^2)$, tj. nějakou měřenou veličinu se střední hodnotou 50 a směro-

datnou odchylkou 5. Pokud chceme vyjádřit např. skutečnou hodnotu 55 pomocí standardizované hodnoty *z*, potom platí z = (55 – 50)/5 = 5/5 = 1 (směrodatná odchylka). Je to pravda protože skutečně hodnota 55 je ve vzdálenosti jedné směrodatné odchylky od střední hodnoty (50).



Obrázek 5.10 - Vliv změny střední hodnoty na polohu normálního rozdělení (při konstantním rozptylu)

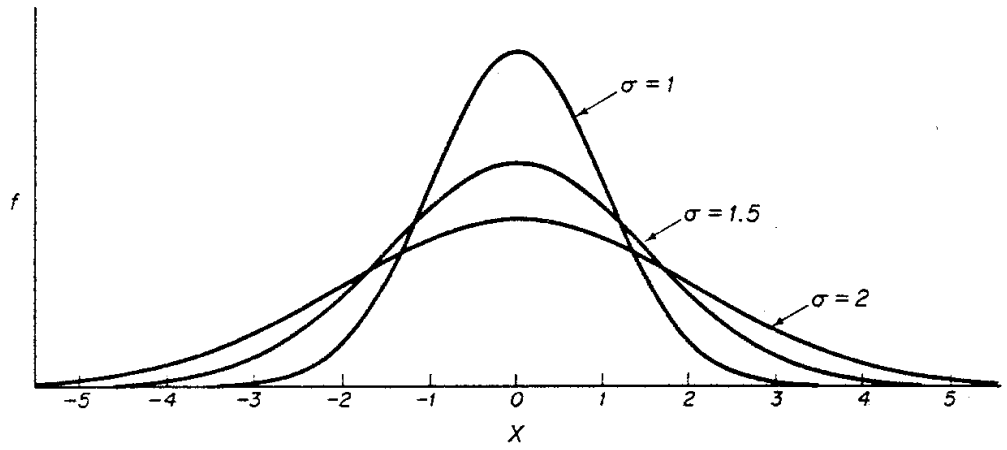
Standardizované normální rozdělení má konstantní střední hodnotu 0 a rozptyl 1, symbolicky to zapisujeme N(0,1).

Tento postup má velkou výhodu v tom, že **jakékoli normální rozdělení je možné převést na standardizované normální rozdělení, které má křivku jednotného tvaru a vlastností a jehož hodnoty mohou být snadno tabelovány a tedy pro výpočet frekvenční a distribuční funkce není nutné používat relativně složité Gauss Laplaceovy funkce**.

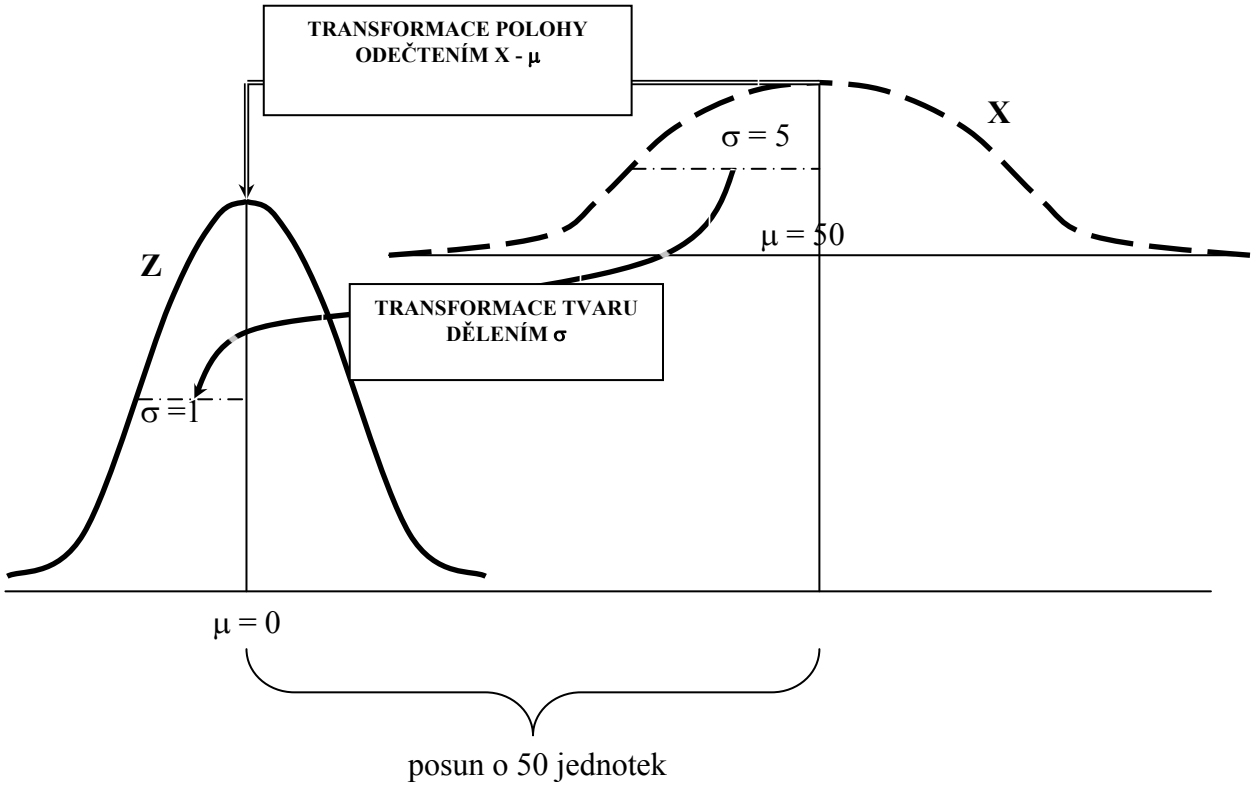
Princip standardizace je schématicky graficky znázorněn na obrázku 5.12 - zde je tučnou čárkovanou čarou znázorněno obecné normální rozdělení N(50,5) náhodné veličiny X. Plnou tučnou čarou je znázorněno standardizované normální rozdělení N(0,1). Princip standardizace spočívá v tom, že **PLOCHA POD OBĚMA KŘIVKAMI – PŮVODNÍ I STANDARDIZOVANÉ - (tj. hodnota jejich distribučních funkcí) JE STEJNÁ (= 1), KŘIVKY MAJÍ POUZE JINÝ TVAR.** V našem případě se od všech hodnot veličiny X odečetla střední hodnota 50 a tím se poloha posunula o 50 jednotek vlevo. Poté je nutné upravit tvar na jednotnou směrodatnou odchylku 1. To se provede vydělením všech posunutých hodnot pěti (směrodatnou odchylkou). Tvar křivky se upraví tak, že V NAŠEM PŘÍPADĚ je standardizovaná křivka užší a vyšší (má menší směrodatnou odchylku a tím nižší variabilitu a aby zachovala stejnou plochu pod křivkou, musí být „vyšší“).

Při $\mu = 0$ a $\sigma = 1$ dostáváme tedy frekvenční funkci standardizovaného normálního rozdělení ve formě následujícího vztahu

$$f(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{\left[-\frac{z^2}{2^2}\right]} \qquad (5.22)$$



Obrázek 5.11 - Vliv změny rozptylu na tvar hustoty pravděpodobnosti normálního rozdělení (při konstantní střední hodnotě)



Obrázek 5.12 - Transformace obecného na standardizované normální rozdělení. Bližší vysvětlení viz v textu.

Distribuční funkce rozdělení N (0, 1) je důležitým podkladem pro mnohá statistická šetření. Její hodnoty jsou proto tabelovány ve vhodných formách, nejčastěji ve formě tzv. Laplaceovy funkce - též pravděpodobnostního integrálu - definovaného vztahem

$$\Phi(z) = P(0 < Z < z) = \frac{1}{\sqrt{2\pi}} \int_0^z e^{-\frac{z^2}{2}} dz. \qquad (5.23)$$

Má-li náhodná veličina X rozdělení N (μ, σ^2), potom pravděpodobnost P (-x<X<x) lze transformací na standardizovanou proměnnou x = z σ + μ a tím přes rozdělení N (0,1) převést na tvar

$$P(\mu - \sigma\ z < X < \mu + z\,\sigma) = P(-z < Z < z) = 2\,\Phi(z).$$

Hodnoty 2 Φ (z) nebo Φ(z) se tabelují. Jestliže např. volíme z = 3 dostaneme P (μ - 3σ < X < μ + 3 σ) = 0,9973. S touto vysokou pravděpodobností - prakticky s jistotou - leží všechny hodnoty normální náhodné veličiny v intervalu < $\mu \pm 3\ \sigma$>. Jestliže naopak volíme pravděpodobnost, určíme interval, ve kterém se zvolenou pravděpodobností leží hodnoty normální náhodné veličiny. Volme např. P = 0,95, potom z = 1,96. V intervalu < $\mu \pm 1{,}96\ \sigma$> leží 95 % všech hodnot normální náhodné veličiny.

Využití standardizovaného normálního rozdělení si ukážeme na následujícím příkladu.

Příklad 5.14:

Předpokládejme, že výčetní tloušťky stromů v určitém porostu mají normální rozdělení. Střední tloušťka je 30 cm, směrodatná odchylka je 5 cm. Celkem bylo měřeno 500 stromů. Určete

a) kolik stromů je silnějších než 36 cm

b) jaká je pravděpodobnost, že náhodným výběrem vybereme strom silnější než 36 cm

c) kolik stromů leží v rozmezí tlouštěk 25 – 36 cm

V našem případě je výčetní tloušťka stromů náhodná veličina s normálním rozdělením D ~ N(30,5^2). Znamená to, že tloušťka 30 cm je nejčastější hodnotou a že zhruba polovina tlouštěk je slabších a polovina silnějších než tato hodnota.

a) Řešení tohoto úkolu je úlohou na distribuční funkci normálního rozdělení. Potřebujeme zjistit pravděpodobnost výskytu všech tlouštěk do hodnoty 36 cm včetně (tedy vypočítat hodnotu distribuční funkce F(D = 36)). Poté odečtením této pravděpodobnosti od jedné a vynásobením získané pravděpodobnosti počtem všech měřených stromů získáme odpověď na danou otázku. Situaci graficky znázorňuje obrázek 5.13 . Jde nám o stanovení pravděpodobnosti znázorněné plochou na obrázku vyznačenou vodorovným šrafováním.

Zde se ihned projeví výhoda standardizace. Kdybychom pracovali s obecným normálním rozdělením, museli bychom vypočítat distribuční funkci normálního rozdělení N(30,5^2), což znamená určitý integrál funkce 5.20, tedy relativně velmi složitý výpočet. V tomto případě mohou pomoci statistické programy nebo EXCEL, kde bychom využili funkce NORMDIST. Jestliže máme k dispozici jen statistické tabulky, nejprve vypočítáme standardizovanou veličinu Z

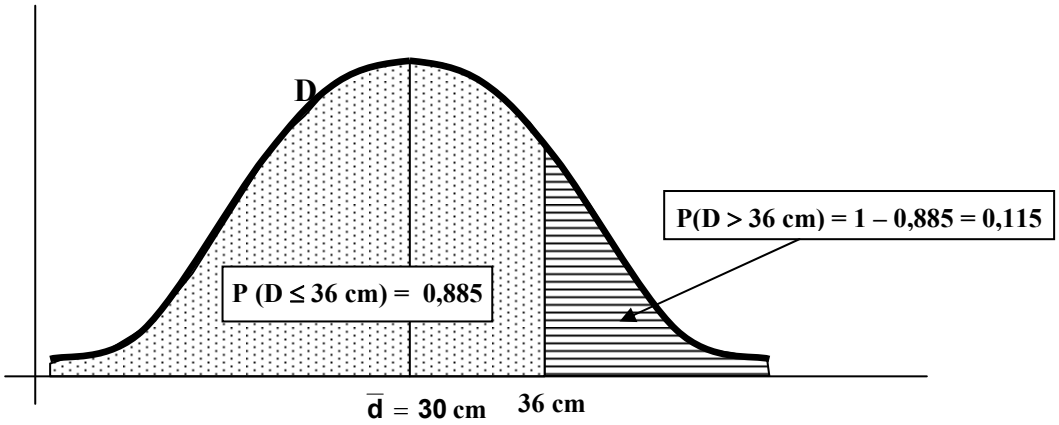
$$Z = \frac{36 - 30}{5} = 1{,}2;$$

čímž jsme si převedli původní hodnotu 36 cm na hodnotu vzdálenou 1,2 směrodatné odchylky od střední hodnoty, tedy na relativní vyjádření nezávislé na původních jednotkách. Teď můžeme použít statistické tabulky (nebo v EXCELu funkci NORMSDIST) a najdeme potřebnou hodnotu, v tomto případě 0,8849. Na obrázku 5.13 je tato pravděpodobnost znázorněna tečkovanou plochou. Hodnota 0,8849 znamená, že asi 88,5 % všech tlouštěk je slabších než 36 cm nebo této hodnotě rovných. Znamená to, že silnějších bude 1 – 0,8849 = 0,1151, tedy asi

11,5 %. Po převedení na absolutní čísla to znamená 500 . 0,1151 = 57,55 ≅ 58 stromů. V porostu tedy můžeme očekávat 58 stromů silnějších než 36 cm.

b) odpověď na tuto otázku je stejná jako v předchozí úloze v bodu a). Jedná se pouze o jinak formulovaný týž problém výpočtu distribuční funkce.

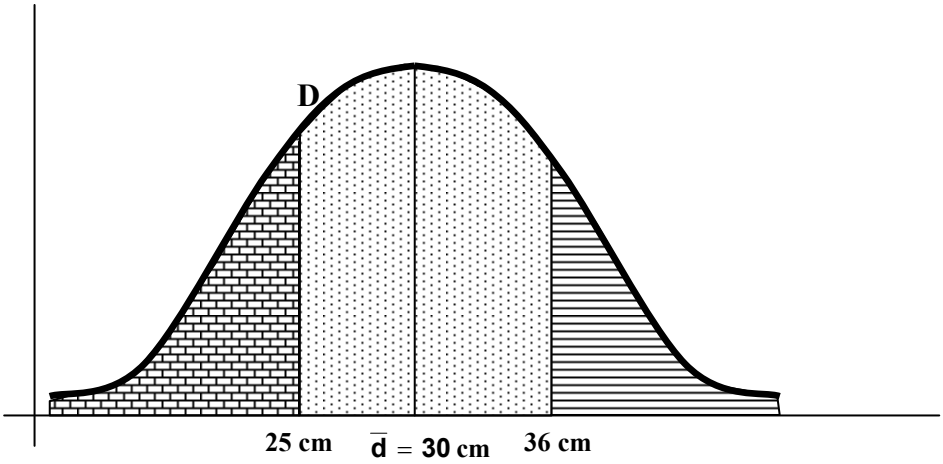
c) zde si ukážeme využití symetričnosti normálního rozdělení. Situaci ukazuje obrázek 5.14 . Jde nám o zjištění pravděpodobnosti vyznačené vytečkovanou plochou. Znamená to, že musíme určit pravděpodobnosti, že tloušťka nepřesáhne 25 cm („cihlová“ plocha), že přesáhne 36 cm (vodorovně šrafovaná plocha) a odečtením jejich součtu od hodnoty 1 (celá plocha pod křivkou) získáme výsledek.



Obrázek 5.13 – Grafické znázornění řešení příkladu 5.14 a)

1. Nejdříve si pro obě hranice intervalu (25 a 36 cm) vypočítáme standardizované hodnoty *Z* a příslušné pravděpodobnosti

- pro hodnotu 36 cm ji máme již vypočítanou – Z = 1,2, tedy P(D > 36 cm) = P(Z > 1,2) = 0,1151.
- pro hodnotu 25 cm $\frac{25-30}{5} = -1$, tedy P(D < 25 cm) = P(Z < -1) = P(Z > 1) = 1 – 0,8413 = 0,1587. Tento výpočet je umožněn právě symetričností normálního rozdělení. Pravděpodobnost, že hodnota *Z* bude menší než –1 je stejná, jako že hodnota *Z* bude větší než 1.



Obrázek 5.14 – Grafické znázornění zadání příkladu 5.14 c)

2. Vypočítáme pravděpodobnost

P(25 cm < D < 36 cm) = P(-1 < Z < 1,2) = 1 – (Z < -1 nebo Z Z > 1) =

= 1 – (0,1587 + 0,1151) = 0,7262. Tedy pravděpodobnost, že tloušťky budou v požadovaném rozmezí, je asi 72,6 %.

5.4.2.2.3 Transformace na normální tvar

Máme-li náhodnou veličinu X, která nemá normální rozdělení a je třeba ji transformovat na normální náhodnou veličinu, musíme najít takovou funkci ψ (x). která náhodnou veličinu H převede na normální náhodnou veličinu ψ (X). Problematika transformací je velmi složitá a přesahuje oblast určení tohoto učebního textu. Uvádíme zde pouze nejjednodušší a nejpoužívanější typy transformací:

$$\Psi(x)=\sqrt{x} \qquad \text{pro} \qquad x \in (0, \infty)$$

$$\Psi(x)=\sqrt{x+\frac{1}{2}} \qquad x \in <0, \infty)$$

$$\Psi(x)=\sqrt{x}+\sqrt{x+1} \qquad x \in (0, \infty)$$

Transformace pro náhodné veličiny X, které vyjadřují výběrovou relativní četnost úspěchů v posloupnosti nezávislých pokusů:

$$\Psi(x)=2\arcsin\sqrt{x}, x \in < 0,1 > .$$

Funkce je tabelovaná. Je součástí statistických tabulek.

Fisherova transformace pro náhodné veličiny X, které vyjadřují výběrové korelační koeficienty:

$$\Psi(x)=\frac{1}{2}\ln\frac{1+x}{1-x}, x \in (-1,1)$$

Funkce je tabelovaná. Je součástí statistických tabulek.

Logaritmická transformace pro náhodné veličiny s logaritmicko normálním rozdělením a podobným:

$$\psi(x) = \ln x, \qquad x \in (0, \infty);$$

$$\psi(x) = \ln(x+1) \qquad x \in <0, \infty)$$

Jednou z nejúčinnějších způsobů přiblížení výběru k normálnímu rozdělení je **Boxova - Coxova transformace**

$$\Psi(x)=\begin{cases}\dfrac{x^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ \ln x & \lambda = 0\end{cases} \qquad (5.24)$$

Tato transformace účinně přibližuje výběr normalitě jak z hlediska šikmosti a špičatosti, tak i z hlediska extrémních hodnot. Určení hodnoty λ, samotná transformace a především následná retransformace parametrů jsou teoreticky i výpočetně velmi náročné postupy a pro jejich realizaci je třeba výkonný statistický program (z těch dostupnějších tuto transformaci provádí- např. ADSTAT). Teorii Boxovy – Coxovy transformace viz např. MELOUN - MILITKÝ 1994.

5.4.2.2.4 Výpočet frekvenční funkce normálního rozdělení pro roztříděný soubor

V praxi je někdy třeba vypočítat normální rozdělení třídních četností roztříděného souboru. Upravený vzorec pro výpočet rozdělení četností má tvar

$$n_i = \frac{N}{\sqrt{2\pi}.S_{x_{(i)}}}.e^{-\frac{\left(\bar{x}_i - \bar{x}\right)^2}{2S_x^2}} \qquad (5.25)$$

kde je

N	rozsah souboru
S_x	směrodatná odchylka souboru
$S_{x_{(i)}} = \frac{S_x}{h}$	směrodatná odchylka v třídních jednotkách (h – třídní interval)
n_i	četnost normálního rozdělení i-té třídy

Příklad 5.15:

Pro zadané parametry μ = 21,22; σ^2 = 5,59; N = 221 a pro třídění m = 11, h = 1 $\bar{x}_1$ = 16,5 vypočítejte třídní četnosti normálního rozdělení.

Výsledné četnosti (n_i) jsou uvedeny v následující tabulce:

$\bar{x}_i$	16,5	17,5	18,5	19,5	20,5	21,5	22,5	23,5	24,5	25,5	26,5
n_i	5,1	10,8	19,2	28,6	35,6	37,0	32,2	23,4	14,2	7,2	3,1

Vypočtenou - očekávanou - četnost uvádíme na počet desetinných míst odpovídajících úloze.

5.4.2.2.5 Užití normálního rozdělení pro výpočet frekvenční funkce binomického rozdělení

V kapitole 5.4.1.2 věnované binomickému rozdělení jsme uvedli, že pro velká *n* se frekvenční funkce binomického rozdělení vyčísluje pomocí vzorce 5.12 obtížně. Proto byly vyvinuty postupy, které umožňují dostatečnou aproximaci pomocí normálního rozdělení. Jedná se o aplikaci tzv. lokální Moivre-Laplaceovy věty, která říká, že hodnotu frekvenční funkce binomického rozdělení pro velká *n* lze přibližně vyjádřit pomocí vztahu

$$\pi(x) \cong \frac{1}{\sqrt{n \cdot p \cdot (1-p)}} \cdot f(z) \qquad (5.26)$$

kde

$$z = \frac{x - n \cdot p}{\sqrt{n \cdot p \cdot (1-p)}} \qquad (5.27)$$

Ověříme tento výpočet na příkladu 5.6 a), který se týkal křížení hrachu. Připomeňme, že v tomto případě bylo x = 10, n =10, p = 0,75. Potom po dosazení do vzorců 5.26 a 5.27 dostaneme hodnotu z = 1,82574 a podle vzorce 5.22 získáme hodnotu f(z) = 0,07535 (tuto hodnotu lze také získat ze statistických tabulek). Výsledná hodnota pravděpodobnostní funkce bino-

mického rozdělení bude potom podle vzorce 5.26 0,055. Pomocí základního vzorce pro frekvenční funkci binomického rozdělení jsme dostali hodnotu 0,056. Vzhledem k poměrně malému výběru je tato shoda velmi dobrá a potvrzuje použitelnost tohoto postupu (v praxi by se tento výpočet používat zpravidla pro vyšší hodnoty *n* než 10).

5.4.2.3 Rozdělení $\chi^2(f)$ – Pearsonovo (chi kvadrát)

Pearsonovu náhodnou veličinu lze konstruovat následovně. Mějme normální náhodnou veličinu X, s rozdělením N (μ, σ^2). Ze souboru hodnot této veličiny provedeme všechny možné nezávislé výběry rozsahu f. Pro každý výběr vypočítáme hodnotu

$$y_i = \sum_{i=1}^{f}\left[\left(x_i - \mu\right).\sigma^{-1}\right]^2 = \sum_{i=1}^{f} z_i^2 \qquad (5.28)$$

Všemi hodnotami y_i je definována Pearsonova náhodná veličina χ^2 (čte se chi kvadrát). Počet prvků ve výběrech f se nazývá počet stupňů volnosti. Hodnota f je parametrem Pearsonova rozdělení.

Hustota pravděpodobnosti se zavádí formálně

$$f_f(y) = p_f(y) \text{ pro } y \in < 0, \infty).$$

Přesně je

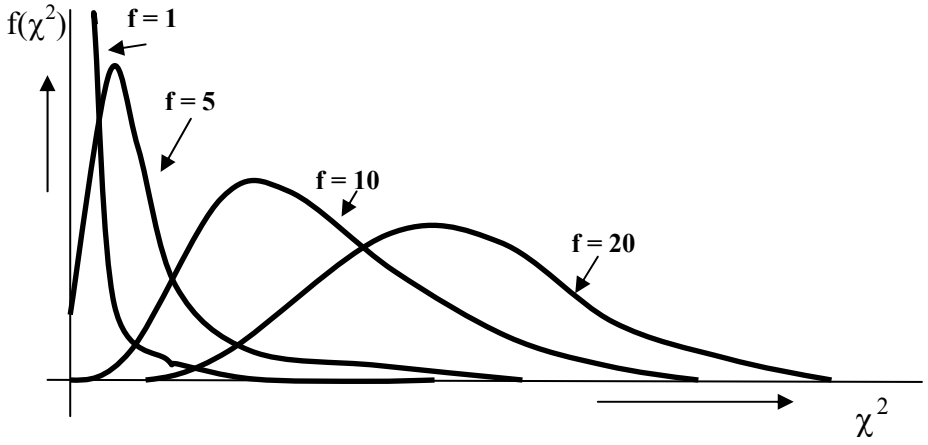
$$f_f\left(y\right) = \frac{1}{2^{\frac{f}{2}}\Gamma\left(\frac{f}{2}\right)}.y^{\frac{f}{2}-1}.e^{-\frac{y}{2}}; y > 0 \qquad \textbf{(5.29)}$$

$\Gamma\left(\frac{f}{2}\right)$ je tzv. gama funkce

Pro střední hodnotu a rozptyl Pearsonova rozdělení platí

$$\mu = f; \qquad \sigma^2 = 2f.$$

.Pearsonovo rozdělení je levostranně nesouměrné (míra nesouměrnosti závisí na počtu stupňů , čím je počet stupňů volnosti menší, tím je frekvenční funkce více levostranná). Schématicky je tato vlastnost zobrazena na obrázku 5.15. Pro $f \to \infty$ přechází Pearsonovo rozdělení v rozdělení normální.



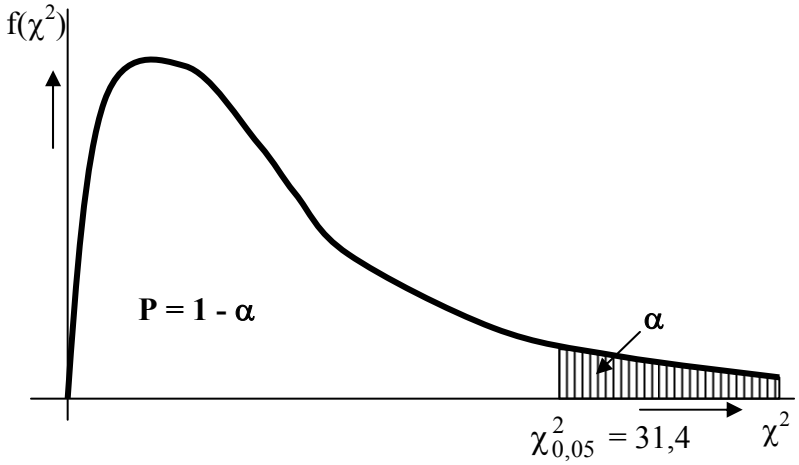
Obrázek 5.15 - Křivky Pearsonova rozdělení pro různé počty stupňů volnosti f

Distribuční funkce náhodné veličiny χ_f^2 je tabelovaná pro určité hodnoty pravděpodobnosti a různá f. Tabulky χ_f^2 rozdělení patří mezi základní statistické tabulky. Nejdůležitější jsou tzv. kritické hodnoty $\chi^2_{\alpha,f}$, tj. takové hodnoty, které náhodná veličina χ^2 překročí s předem stanovenou pravděpodobností α. Situaci ukazuje obrázek 5.15. Máme např. χ^2_{20}, tj. Pearsonovo rozdělení o 20-ti stupních volnosti, potom můžeme z tabulek určit, že hodnota, kterou toto rozdělení překročí s pravděpodobností α = 5 % (= 0,05) – svisle vyčárkovaná plocha pod křivkou, je $\chi^2_{0,05;20}$ = 31,4. Naopak s pravděpodobností P = 1 - α = 1 – 0,05 = 0,95 (= 95 %) – bílá plocha pod křivkou - tuto hodnotu nepřekročí. Tyto kritické hodnoty můžeme zjistit i statistickými programy nebo v EXCELu pomocí funkce CHIINV.

Rozdělení χ_f^2 má velkou použitelnost pro statistických šetřeních. Obecně řečeno jde o rozdělení tzv. kvadratických forem s aplikací v odhadech parametrů, testování hypotéz, testech kontingenčních tabulek atd. χ_f^2 rozdělení má každá náhodná veličina s konstrukcí obdobnou jako Pearsonova veličina. Použití Pearsonova rozdělení bude ukázáno na příkladech v příslušných kapitolách.

V souvislosti s Pearsonovým rozdělením je nutno ještě blíže vysvětlit pojem, se kterým se v matematické statistice budeme často setkávat – **stupně volnosti**. Jejich podstatu je možné vysvětlit následujícím příkladem:

Předpokládejme výběr n = 5 určité náhodné veličiny X: 12,14,15,16,18. Jejich aritmetický průměr je 15. Nyní vytvořme nový výběr o stejné velikosti tak, aby průměr byl stejný. Takovýchto souborů můžeme vytvořit nekonečně mnoho, přičemž *n – 1* hodnot můžeme volit libovolně, ale poslední číslo je určeno danou podmínkou. Součet všech členů nového výběru musí zase dát hodnotu 75, aby se získal aritmetický průměr 15. Např. jestliže zvolíme čísla 5, 10,20,30, musí být poslední číslo 10. Znamená to, že uvedenou podmínkou je jedna hodnota vázaná, ostatní jsou „volné“ – to jsou ony „stupně volnosti“. Podobný případ jsou např. odchylky od průměru $(x_i - \bar{x})$ - vzhledem k tomu, že součet odchylek hodnot od průměru musí být 0, je *n - 1* odchylek volných a jedna vázaná. Tyto příklady se dají shrnout, že počet stupňů volnosti se rovná počtu hodnot (nemusí to být vždy prvotní měřená data, ale i např. odchylky) zmenšeného o počet omezujících podmínek.



Obrázek 5.16 – Grafická interpretace kritické hodnoty a pravděpodobností P a α Pearsonova rozdělení

5.4.2.4 Rozdělení Fisherovo - Snedecorovo - F (f_1, f_2)

Mějme náhodné veličiny X s rozdělením $\chi^2_{f_1}$ a Y s rozdělením $\chi^2_{f_2}$, které jsou nezávislé. Fisher - Snedecorova náhodná veličina je definovaná vztahem

$$F = \frac{\frac{1}{f_1}X}{\frac{1}{f_2}Y} \tag{5.30}$$

Hustota pravděpodobnosti se zavádí formálně

$$f(F) = p_{f_1, f_2}(F) \qquad \text{pro } F > 0$$

$$0 \qquad \text{pro } F \le 0$$

kde f_1, f_2 jsou přirozená čísla. Grafem hustoty pravděpodobnosti F rozdělení je nesouměrná jednovrcholová křivka, jejíž tvar závisí na počtu stupňů volnosti f_1 a f_2. Při konstantní hodnotě f_1 vzrůstající hodnota f_2 mění tvar křivky tak, že se stává souměrnější a zahrocenější. Při $f_2 \to \infty$ přechází v Pearsonovo rozdělení a při $f_1 = 1$ a $f_2 = f$ ve Studentovo t-rozdělení.

Pokud $f_2 > 2$, pak existuje konečná střední hodnota F veličiny a platí

$$\mu = \frac{f_2}{f_2 - 2}.$$

Pokud $f_2 > 4$, pak existuje konečný rozptyl F veličiny, který je roven

$$\sigma^2 = \frac{2f_2^2(f_1 + f_2 - 2)}{f_1(f_2 - 2)^2(f_2 - 4)}$$

Distribuční funkce je dána odpovídajícím integrálem

$$F(F) = \int_0^F p_{f_1, f_2}(F)\,dF$$

Tabelované kritické hodnoty - kvantily - F_{f_1, f_2} rozdělení patří mezi základní statistické tabulky.

Pro práci s kvantily je třeba vědět, že

$$F_{f_1, f_2, \alpha} = 1\left[F_{f_2, f_1, 1-\alpha}\right]^{-1}.$$ (Pozor na převrácení pořadí stupňů volnosti!!)

Příklad 5.16:

Pro F- rozdělení se stupni volnosti $f_1 = 4$ a $f_2 = 10$ hledáme hodnotu, která bude překročena s pravděpodobností 5 % a 95 %.

Jedná se o náhodnou veličinu s rozdělením $F_{0,05;4;10}$. V tabulkách (nebo statistickými programy nebo v EXCELu pomocí funkce FINV) najdeme kritickou hodnotu 3,48. Tato hodnota bude překročena s pravděpodobností $\alpha = 0,05$. Hodnotu F, která nebude překročena s pravděpodobností $\alpha = 0,05$, tj. bude překročena s pravděpodobností $P = 1 - \alpha = 1 - 0,05 = 0,95$ získáme tak, že nejprve vyhledáme hodnotu $F_{0,05;10;4}$ (pozor na výměnu pořadí stupňů volnosti, je to velmi důležité) = 5,96 a z ní spočítáme převrácenou hodnotu 1/5,96 = 0,168. Tedy s 5 % pravděpodobností bude překročena hodnota 3,48, s 95 % pravděpodobností hodnota 0,168.

V praktickém použití je výhodné zavést takové označení, aby platilo $X \geq Y$. Potom totiž v oboustranném testu není třeba počítat převrácené hodnoty kritických hodnot. F_{f_1, f_2} rozdělení má četné aplikace. Nejčastěji se s ním setkáme pro ověřování hypotéz o kvadratických formách, např. rozptylech. F-rozdělení má každá náhodná veličina, která má obdobnou konstrukci jako Fisher - Snedecorova náhodná veličina.

5.4.2.5 T-rozdělení (Studentovo) - t_n

Mějme nezávislé náhodné veličiny X s rozdělením N (0, 1) a Z s rozdělením χ^2_k.
Potom náhodná veličina

$$T = \frac{X}{Z.k} \qquad 5.31)$$

má Studentovo t - rozdělení o *k* stupních volnosti.

Objevitelem tohoto rozdělení byl W.S. Gosset, který publikoval pod pseudonymem Student.

Hustota pravděpodobnosti je dána formálně vztahem

$f(t) = g_k(t)$, kde $x \in (-\infty, \infty)$
k je přirozené číslo

Pokud $k > 1$, existuje střední hodnota a platí $\mu = 0$

Pokud $k > 2$, existuje rozptyl a platí $\sigma^2 = \frac{k}{k-2}$.

Grafem této funkce je křivka souměrná vzhledem k t = 0, probíhá od t = - ∞ do t = +∞ a je tím plošší, čím je menší f (počet stupňů volnosti). Pro $f \to \infty$ přechází v normální rozdělení. Tvar tohoto rozdělení tedy závisí jen na velikosti výběru *n*, nezávisí na rozptylu hodnot, což je jeho výhoda. Pro výběry větší než 30 prvků se dá s dostatečnou přesností nahradit normálním rozdělením.

T-rozdělení má četné aplikace. Pro jeho značné využití uveďme jmenovitě např. interval spolehlivosti pro μ při neznámém σ^2 nebo t- testy (jednovýběrový, dvouvýběrový, párový) .

5.5 Náhodný výběr a výběrové postupy

5.5.1 Teoretická východiska

Výsledky náhodného experimentu pozorujeme prostřednictvím náhodné veličiny nebo náhodného vektoru. Pravděpodobnostní chování náhodné veličiny (náhodného vektoru), ze kterého čerpáme informace o náhodném experimentu, je popsáno

- distribuční funkcí $\Phi(X)$; $\Phi(X_1, X_2, ..., X_n)$
- pravděpodobnostní funkcí (pro diskrétní náhodné veličiny) $\pi(X)$; $\pi(X_1, X_2, ..., X_n)$
- hustotou pravděpodobnosti (pro spojité náh. veličiny) $f(X)$; $f(X_1, X_2, ..., X_n)$

Realizace náhodného experimentu obsahují informace, které při vhodném matematickém zpracování vypovídají o obecných zákonitostech nebo potvrzují či zamítají ty hypotézy,

k jejichž ověření byl náhodný experiment prováděn. V matematické statistice vycházíme z konkrétního a snažíme se vyvodit závěry v obecném. Tento postup nazýváme – jak již bylo uvedeno - statistická indukce.

Podkladem pro statistickou indukci je pravděpodobnostní chování náhodné veličiny či náhodného vektoru. To znamená, že toto charakteristické pravděpodobnostní chování musí mít i soubor realizací, tj. hodnoty náhodné veličiny X, resp. vektoru (X_1, X_2, ..., X_n), které z celé množiny možných realizací nastaly. Uskutečněné realizace dávají tzv. výběrový soubor ze základního souboru.

Podmínkou, kterou musíme splnit, je zajištění reprezentativnosti výběru, tj. toho, aby výběr podal úplnou informaci o základním souboru. Podle teorie pravděpodobnosti je podmínka splněna, když zůstane zachována pravděpodobnost nastoupení realizace, tzn., že každá realizace musí mít zachovanou pravděpodobnost být vybrána. To splňuje tzv. výběr s opakováním. Jestliže není možné „vybranou" realizaci vrátit do základního souboru, je třeba zajistit, aby se pravděpodobnosti nastoupení realizace téměř nezměnily. To lze splnit např. velkým počtem prvků základního souboru.

Většina statistických metod je založena na předpokladu použití náhodného výběru s opakováním (nezávislý výběr).

Nezávislý náhodný výběr rozsahu *n* ze základního souboru, na němž je definována náhodná veličina X s rozdělením $\Phi(x)$, je n - rozměrný náhodný vektor $X = (X_1, ..., X_n)$, který má vlastnosti

1. X_1, X_2, ..., X_n jsou nezávislé náhodné veličiny,
2. všechny veličiny X_1, X_2, ..., X_n mají stejné rozdělení pravděpodobnosti.

Libovolná náhodná veličina T, která vznikne jako funkce náhodného výběru $X = (X_1, X_2, ..., X_n)$ se nazývá statistika. Píšeme

$$T = T(X_1, X_2, ..., X_n).$$

Můžeme tedy shrnout, že **reprezentativní náhodný výběr** je charakterizován následujícími předpoklady:

- **jednotlivé prvky výběru jsou vzájemně nezávislé,**
- **výběr je homogenní**, tj. **všechny veličiny mají stejné rozdělení pravděpodobnosti s konstantním rozptylem** (předpokládá se, že jde o normální rozdělení),
- **všechny prvky základního souboru mají stejnou pravděpodobnost, že budou zařazeny do výběru**.

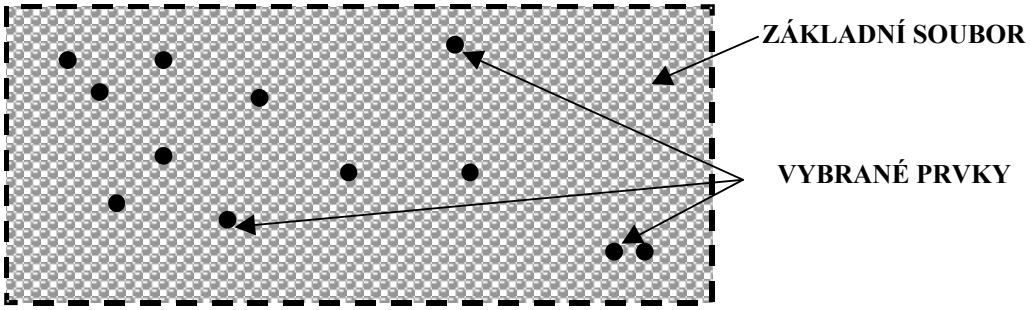
Pokud výběr nesplňuje výše uvedené předpoklady, je jeho statistická analýza komplikovanější a zpravidla se musí používat speciální postupy. Proto je před vlastní analýzou nutné vyšetřit, zda výběry splňují základní předpoklady, tj. potřebný rozsah, nezávislost, homogenitu a normalitu. Touto problematikou se zabývá průzkumová analýza dat (o ní podrobněji ve II. díle tohoto textu).

5.5.2 Druhy výběrů

V praxi se setkáváme při výběrových metodách s různou úrovní známých informací o základním souboru. Máme tedy určitou odstupňovanou možnost posoudit reprezentativnost výběrového souboru. Nejsou-li informace o základním souboru dostatečné, považujeme každý náhodný výběr za reprezentativní. Jestliže známe některé vlastnosti základního souboru, můžeme je využít k upřesnění způsobu výběru, ale pouze tak, aby princip náhodného výběru zůstal zachován.

5.5.2.1 Jednoduchý výběr

Při malém rozsahu základního souboru můžeme prvky očíslovat. Výběr potom uskutečníme „losováním“ nebo pomocí tabulek náhodných čísel. Tabulky náhodných čísel jsou součástí všech statistických tabulek (viz obrázek 5.17).



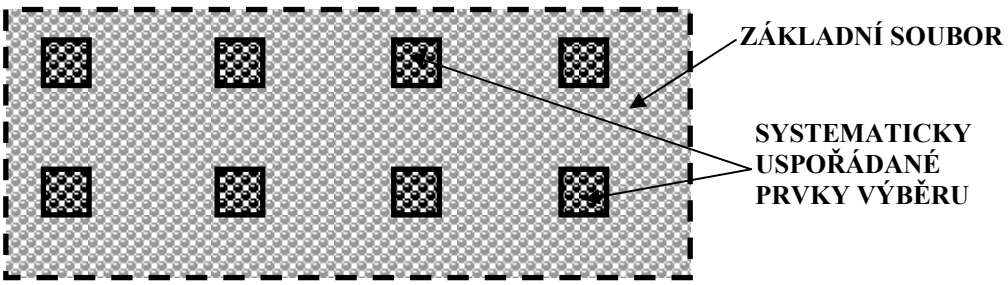
Obrázek 5.17 - Grafická interpretace jednoduchého výběru

5.5.2.2 Systematický výběr

Prvky základního souboru můžeme pokládat za seřazené v určitém pořadí. Jestliže rozsah základního souboru je N a výběrového souboru n můžeme vzít číslo $k = \frac{N}{n}$ za výběrový krok a z řady prvků vybírat postupně b-tý, (b+k)-tý, (b+2k)-tý, přičemž pořadové číslo b náhodně vybereme z množiny (1, 2, ..., k). Pak hovoříme o systematickém výběru s náhodným startem.

Jestliže jsou prvky základního souboru rozmístěny po ploše, lze systematický výběr prvků uskutečnit průsečíky pravoúhlé sítě určitých konstantních vzdáleností. Obvykle se používá čtvercová nebo obdélníková síť. Typickým příkladem tohoto typu výběru jsou zkusné plochy nebo relaskopická stanoviska používaná ke zjišťování zásoby porostů (viz obrázek 5.18). Na základě variability zásoby porostu (tj. měřené nebo častěji odhadnuté veličiny základního souboru) se určí potřebný počet ploch (velikost výběru) a ty se rozmístí do porostu tak, aby rovnoměrně a v pravidelných intervalech pokryly celou plochu porostu. Podstatou prvku náhodnosti je to, že plochy se zakládají v pravidelném sponu – „padni, kam padni“ – kromě extrémních míst (např. porostní okraje).

Systematický výběr se snadno realizuje a zabezpečuje dobře požadavky na náhodný výběr. Pro jeho použití je důležité, aby uspořádání jednotek v základním souboru bylo vzhledem ke zjišťovanému znaku úplně nahodilé. V případě, že zjišťovaný znak vykazuje systematické změny (např. porostní veličiny v porostech, které vznikly pruhovou okrajovou obnovou) a krok systematického výběru by se náhodnou s intervalem těchto změn shodoval, může dojít k systematickému nadhodnocení nebo podhodnocení měřených veličin, a tím i výsledků.



Obrázek 5.18 - Grafická interpretace systematického výběru

5.5.2.3 Oblastní (stratifikovaný) výběr

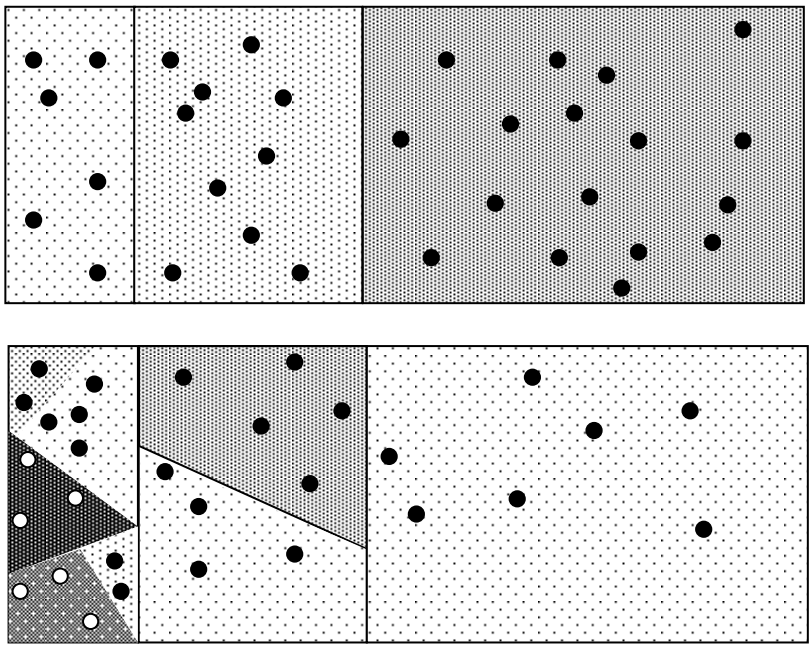
Základní soubor se rozdělí na několik dílčích souborů, tzv. oblastí čili strat. K rozdělení na oblasti mnou vést různé důvody. Velmi často bývá důvodem menší variabilita veličin v oblastech než v celém souboru. Výhodou oblastního výběru je to, že umožňuje statisticky zpracovávat (včetně predikce) jak celý základní soubor, tak jednotlivé oblasti. Podle toho, jak se rozsah celého výběru rozdělí na jednotlivé oblasti, dostaneme (grafická interpretace viz obrázek 5.19):

1. úměrný (proporcionální) náhodný oblastní výběr, při němž je poměr (n_i/N_i) ve všech oblastech stejný,
2. optimálně umístěný oblastní výběr, při němž rozsah výběru v jednotlivých oblastech je úměrný rozsahu a variabilitě (popř. jiným vlastnostem) oblastí.

Typickým příkladem je např. zjišťování zásoby v porostech nebo dílcích, které se skládají z částí, které se liší variabilitou zásoby nebo jinými charakteristikami. Potom se postupuje následujícím způsobem:

- zjistí se plošné podíly (W_j) plochy jednotlivých částí (P_j) z celkové plochy (P), jejichž vlastnosti (především variabilita) se liší, podle vztahu $W_j = P_j/P$,
- pro každou část se zjistí stupeň variability (např. variačním koeficientem, pomocí stupně rozrůzněnosti zásoby, apod.) - R_j,
- vypočítá se průměrná variabilita pro všechny části dohromady $\overline{R}$
- určí se celkový počet zkusných ploch (relaskopických stanovišť) (n) – buď podle vztahů pro velikost náhodného výběru nebo podle speciálních grafikonů
- celkový počet zkusných ploch n se rozdělí úměrně podle jejich výměry a variability sledované veličiny podle vztahu

$$n_j = n \cdot k_j, \text{ kde } k_j = \frac{W_j \cdot R_j}{\overline{R}}$$



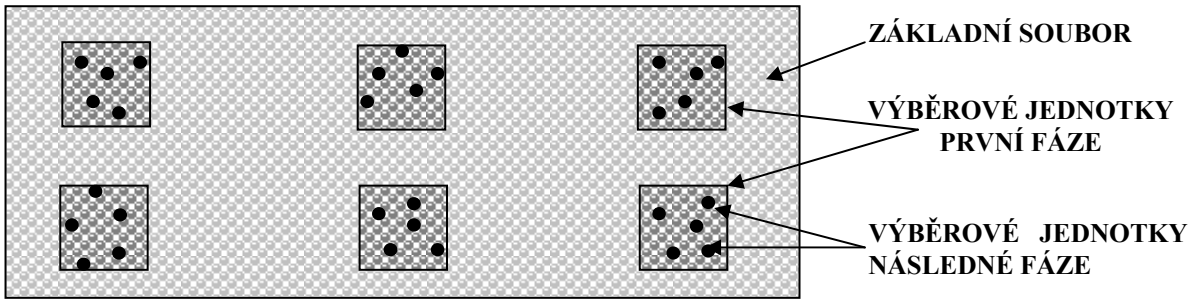
ÚMĚRNÝ VÝBĚR: počet vybraných prvků se řídí pouze velikostí (např. plochou) jednotlivých oblastí

OPTIMÁLNÍ VÝBĚR: počet vybraných prvků se řídí variabilitou a velikostí jednotlivých oblastí (tedy plocha s menší plochou, ale vysokou variabilitou může mít více prvků výběru než větší plocha s menší variabilitou

Obrázek 5.19 - Grafická interpretace oblastního výběru. Různá variabilita je značena různým rastrem.

5.5.2.4 Dvou- (a více-) stupňový výběr

Používá se při velmi rozsáhlých základních souborech, ve kterých technika výběrů podle předchozích postupů by byla prakticky nepoužitelná (např. základní soubor je příliš prostorově nebo časově rozptýlený).



Obrázek 5.20 - Grafická interpretace vícestupňového výběru

V prvním stupni výběru se z této množiny vyberou podle předcházejících metod výběru některé skupiny. Každá skupina má pokud možno dobře reprezentovat celý základní soubor (variabilita ve skupinách by měla být stejná jako z základním souboru, tedy opačně než v případě oblastí!).

Ve druhém stupni výběru se v každé vybrané skupině uskuteční opět vhodnou technikou výběr určitých prvků (buď stejný počet prvků, jestliže jsou skupiny stejně velké, nebo různý počet prvků pro každou skupinu, jestliže skupiny jsou rozdílně velké).

Grafická interpretace tohoto principu je na obrázku 5.20 .

Typickým příkladem je inventarizace lesů nebo ploch pro monitoring zdravotního stavu apod. na velkých plochách (např. území lesních oblastí nebo celého státu). První stupeň představuje výběr oblastí, kde bude měření prováděno, druhý stupeň stanovení (obvykle pravidelné) sítě ploch, další stupeň může představovat výběr určitého počtu stromů pro různé typy měření.

5.5.2.5 Dvou- (a více-) fázový výběr

Od předcházejícího výběrového postupu se liší hlavně v tom, že v jednotlivých fázích se veličiny zjišťují různým způsobem a s různou přesností.

V první fázi se určí jistý počet výběrových jednotek (n_1), na nichž se měřený znak určí relativně jednoduchým a laciným způsobem. Tím se získají hodnoty x_A.

Ve druhé) fázi (a případných dalších) se stanoví menší počet výběrových jednotek n_2, kde se tatáž veličina stanoví přesnějším, ale zpravidla náročnějším a dražším způsobem. Tím se získají hodnoty x_B.

Mezi hodnotami x_A a x_B se najde korelační vztah (blíže ve II. díle tohoto učebního textu) $x_B = f(x_A)$, která se potom použije na korekci méně přesných údajů první fáze výběru, která byla ale provedena na daleko větším rozsahu výběru.

Typickým příkladem je např. kombinované fotogrammetricko – terestrické zjišťování zdravotního stavu lesů. V první fázi se pomocí leteckých nebo kosmických snímků změří (automaticky nebo poloautomaticky) potřebné charakteristiky zdravotního stavu na velkém počtu ploch studovaného území. V další fázi se vybere relativně malý počet ploch, kde se provede pozemní šetření, které je detailnější, ale také náročnější a dražší. Najdou se vhodné korelační vztahy mezi „smímkovými“ a „terénními“ daty, které umožní korekci hodnot získaných ze snímků pro celé studované území.

5.5.3 Určení minimální velikosti výběru

Velikost (rozsah) výběru je velmi důležitou vlastností. Má vliv zejména na:

- přesnost odhadu parametrů polohy a variability
- velikost intervalů spolehlivosti (s rostoucím rozsahem výběru se zužují, protože je k dispozici více informací)
- snižuje se pravděpodobnost chyby II. druhu u statistických testů (pravděpodobnost přijetí nesprávné alternativní hypotézy) a tedy roste síla testu (tyto pojmy budou vysvětleny v příslušných kapitolách)

Nejčastěji se rozsah výběru určí podle vzorce (ZACH 1990 A)

$$n = U_{\alpha}^2 \cdot \frac{\hat{S}^2}{\Delta^2} \tag{5.32}$$

kde je

U_{α} je symbol normované náhodné veličiny (pro $\alpha = 0.05$ je to hodnota 1.96, pro $\alpha = 0.01$ se rovná 2.58)

$\hat{S}^2$ je míra variability, v tomto případě odhad rozptylu (je možné použít i variační koeficient jako relativní míru variability). Problém je v tom, že tato hodnota při plánování výběru není známa (určuje se buď odhadem, na základě předběžného výběru, výsledku předchozích šetření, údajů z literatury apod.)

Δ^2 je hodnota požadované přesnosti výběrového průměru $\bar{x}$, která je definovaná jako polovina intervalu spolehlivosti. Potom s pravděpodobností (1 - α) platí, že $\mu - \Delta \le \bar{x} \le \mu + \Delta$.

Tohoto postupu je možné užít jak pro zjišťování potřebného rozsahu měření při plánování experimentu, tak i pro zpětné ověření dostatečné velikosti výběru.

Příklad 5.17:

Byly zjišťovány výčetní tloušťky modřínu na dvou plochách (ZACH 1990 B) s těmito výsledky:

Plocha	Počet měření	Střední tloušťka (cm)	Rozptyl tlouštěk (cm^2)
I	30	40.8	29.5
II	20	36.0	22.8

Zjistěte, zdali rozsah obou výběrů byl dostatečný pro přesnost určení střední tloušťky v intervalu tloušťkové třídy 4 cm (± 2 cm) s 95 % pravděpodobností.

Jde o zjištění, zda na obou zkusných plochách byl změřen tak velký počet stromů, že můžeme se zadanou pravděpodobností tvrdit, že „neznámá“ střední tloušťka základního souboru (např. pro porost, dílec, apod.), se od výběrové střední hodnoty neliší o víc než ± 2 cm. Použijeme vzorec 5.32).

Pro P = 95 % a plochu I je výpočet následující:

$$n = 1{,}96^2 \cdot \frac{29{,}5 \cdot (30/29)}{2^2} = 29{,}3 \cong 30$$

Hodnota 29,5.(30/29) je tzv. bodový odhad rozptylu základního souboru (podrobněji bude vysvětleno v kapitole 6). Vypočítaná hodnota se zaokrouhluje vždy nahoru (bez ohledu na běžná pravidla zaokrouhlování).

Vidíme, že počet měřených stromů byl na ploše I dostatečný (požadovaný rozsah výběru je 30 stromů, naměřeno bylo 30). Pro druhou plochu obdobným postupem vychází asi 24 stromů (přesně 23,04), tj. na ploše II nebyl změřen dostatečný počet stromů. Je samozřejmé, že se vzrůstající variabilitou (vyšší rozptyl), zvyšující se přesností (menší „povolený" interval, tj. zmenšující se hodnota Δ) a s rostoucí pravděpodobností (např. místo 95 % bude požadována pravděpodobnost 99 %) se nutný rozsah výběru dále zvětšuje. Např. pro pravděpodobnost 99 % by byl nutný rozsah výběru při jinak stejném zadání asi 51 stromů pro plochu I a 40 stromů pro plochu II.

Pomocí tohoto vzorce lze také zpětně vypočítat přesnost, se kterou byl již uskutečněný výběr proveden, a to tak, že vypočítáme hodnotu Δ.

Jiný způsob stanovení minimálního rozsahu výběru (MELOUN - MILITKÝ 1994) je založen na tom, že při vypočítané velikosti výběru nepřesáhne relativní chyba směrodatné odchylky určenou hodnotu $\delta(s)$. Ve vzorci figuruje špičatost výběrového rozdělení ($g_2(x)$), které musíme buď zjistit na základě předvýběru nebo ji pro běžná rozdělení známe (např. pro rovnoměrné 1.8, pro normální 3.0, pro exponenciální 6.0 apod.). Velikost výběru se vypočítá podle vzorce

$$n = \frac{g_2(x) - 1}{4 \cdot \delta^2(s)} + 1 \qquad (5.33)$$

kde relativní chyba např. 10 % se zadává jako 0.1. Podle tohoto vzorce vychází předpokládaná velikost výběru pro normální rozdělení a $\delta(s)$ = 10 % 51 prvků. Z toho vyplývá, že často volené malé rozsahy výběrů (např. 5, 10, ... prvků) jsou v mnoha případech nedostatečné.

6 Odhady parametrů základního souboru

6.1 Teoretická východiska

Parametrem budeme nazývat statistickou charakteristikou základního souboru. **Statistikou budeme nazývat statistickou charakteristiku výběrového souboru.**

Za odhad parametru můžeme považovat a prohlásit věcně odpovídající hodnotu získanou nějakým postupem z výběru X. Důležité ovšem je, abychom uměli posoudit, který postup a hodnota jím získaná je z hlediska pravdivosti informace nejlepší. K tomuto posouzení slouží řada kritérií na posouzení vlastností odhadu. Podle povahy problému se volí to kritérium, které nejlépe vyhovuje podmínkám řešené úlohy.

Uveďme příklad. Výškoměrem chceme zjistit výšku stromu. Protože víme, že při měření se dopouštíme chyb, chceme jejich vliv vyloučit a proto měříme výšku několikrát. Tím dostaneme soubor naměřených výšek $H = (H_1 , H_2 , \ldots , H_n)$, který je výběrovým souborem ze základního souboru všech možných měření výšek (těch je teoreticky nekonečné množství – tedy nekonečně veliký základní soubor). Takovýchto výběrů můžeme také učinit nekonečně mnoho a pokaždé získáme jiné naměřené hodnoty. Proto je naším úkolem z konkrétních naměřených hodnot výběrového souboru provést „co nejlepší“ odhad nám neznámého parametru základního souboru (v našem příkladu je to výška stromu získaná z velmi mnoha – teoreticky nekonečného počtu – měření).Víme, že náhodné chyby mají normální rozdělení s nulovou střední hodnotou a určitým rozptylem. Jestliže to vezmeme v úvahu, můžeme za „správnou“ výšku stromu považovat, tj. ji odhadovat, některou hodnotu „uprostřed“ naměřeného souboru, kde očekáváme chybu velmi blízkou nule, tedy např.

- aritmetický průměr všech naměřených výšek,
- průměr nejnižší a nejvyšší naměřené výšky,
- průměr hodnot s vyloučením nejnižší a nejvyšší naměřené výšky,
- medián souboru naměřených výšek,

atd.

Který z těchto odhadů je nejlepší, tj. který se nejvíce přibližuje skutečné výšce stromu? To musíme poměřit objektivními matematickými kritérii.

Odhad parametru základního souboru z hodnot charakteristik výběru lze provést dvojím způsobem:

1. Z hodnot výběru vypočítáme stanoveným způsobem jedno číslo, které prohlásíme za odhad parametru. Metodě, kterou je toto číslo sestrojeno, se říká **bodový odhad** a číslu **odhad**.
2. Z výběrových hodnot se vypočítají dvě číselné hodnoty stanoveným způsobem tak, aby dotyčný parametr ležel uvnitř číselného intervalu určeného oběma vypočítanými hodnotami. Metodě, kterou se interval sestrojuje, se říká **intervalový odhad**, intervalu **interval spolehlivosti**.

Bodový odhad i krajní hodnoty intervalu spolehlivosti jsou statistiky. Proto metody a kriteria odhadů se odvozují z rozdělení těchto statistik nebo z charakteristik těchto rozdělení.

6.2 Bodové odhady

6.2.1 Základní vlastnosti bodových odhadů

Nechť X = (X_1 , X_2 , ... , X_n) je náhodný výběr rozsahu *n* z rozdělení $\Phi(x)$, P_i je parametr rozdělení. Potom

1. nesporný (konzistentní) odhad parametru P_i je statistika T, pro kterou platí

$$\lim_{n\to\infty} P\left[\left(T-P_i\right) < \varepsilon\right] = 1 \tag{6.1}$$

pro $\varepsilon > 0$ libovolné. Vztah říká, že odhad konverguje ke skutečnému parametru (tj. „neznámé“ charakteristice základního souboru), když rozsah výběru roste nade všechny meze. Prakticky to znamená, že velké chyby odhadu mají při dostatečně velkém rozsahu výběru malou pravděpodobnost výskytu.

2. nestranný (nevychýlený) odhad parametru je statistika T , pro kterou platí

$$E(T) = P_i, \tag{6.2}$$

tj. že se střední hodnota výběrové statistiky se rovná odhadovanému parametru základního souboru. V opačném případě se statistika T nazývá **vychýlený odhad parametru**. Je-li pouze $\lim_{n\to\infty} E(T) = P_i$, nazýváme statistiku T **asymptoticky nevychýlený odhad parametru P_i**. Rozdíl $E(T) - P_i$ se **nazývá vychýlení odhadu.** Jsou-li T_1, T_2 dva nevychýlené odhady parametru P_i, řekneme, že odhad T_1 je lepší než odhad T_2 , když rozptyl $D(T_1)$ je menší než rozptyl $D(T_2)$. **Nejlepší** nestranný odhad ze skupiny nestranných odhadů je odhad, jehož rozptyl je **minimální,** tj. $D(T) \le D(T^*), kde T^*$ značí libovolný nestranný odhad.

3. Vydatný bodový odhad parametru P_i je statistika T, která je nestranným odhadem parametru P_i a platí-li pro dané *n* $D^2(T) \le D^2(T^*), kde T^*$ značí libovolný nestranný odhad. Poměr

$$e(T) = \frac{D^2(T)}{D^2(T^*)} \tag{6.3}$$

se nazývá **vydatnost nestranného odhadu parametru** P_i . Je-li $\lim_{n\to\infty} e(T) = 1$, potom T je odhad asymptoticky vydatný.

Statistická teorie zná různé druhy bodových i intervalových odhadů, které se liší vlastnostmi, jež mají splňovat. V následujících odstavcích uvedeme pouze některé. Základním předpokladem k provedení odhadu je znalost rozdělení důležitých statistik.

6.2.2 Bodový odhad parametrů μ (střední hodnoty) a σ^2 (rozptylu)

Má-li základní soubor normální rozdělení N (μ, σ^2), je statistika $T = \overline{X}$ (tj. výběrový aritmetický průměr) konzistentním, nestranným a vydatným odhadem parametru μ. Platí totiž

$$E(\overline{X}) = \mu \tag{6.4}$$

$$D^2\left(\overline{X}\right) = \frac{\sigma^2}{n} \tag{6.5}$$

a odtud

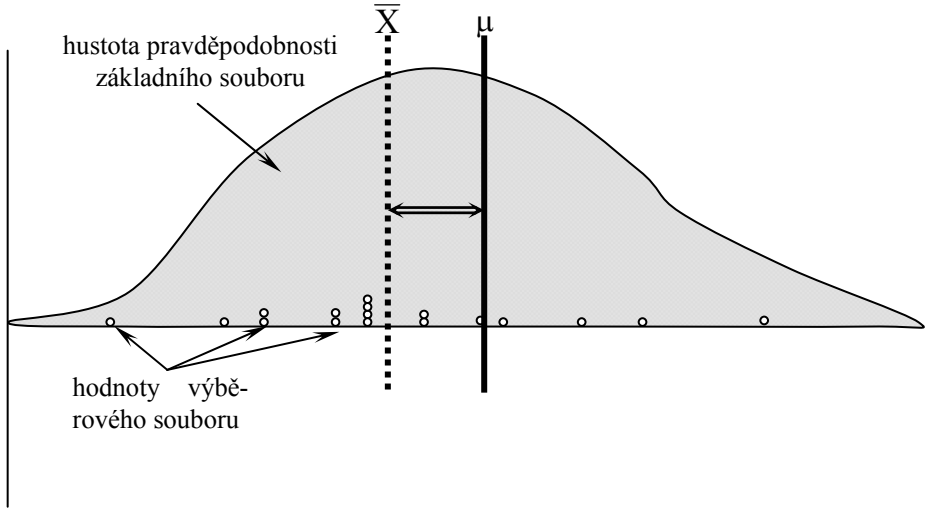
$$\lim_{n\to\infty} \frac{\sigma^2}{n} = 0 \tag{6.6}$$

Má-li základní soubor normální rozdělení N (μ, σ^2), je statistika

$$T = S^2 . \frac{n}{n-1} \tag{6.7}$$

konzistentním a nestranným odhadem parametru σ^2. Pokud bychom použili jako odhad parametru σ^2 pouze hodnotu S^2 (vypočítanou podle vzorce $S^2 = \frac{1}{n}(x_i - \overline{x})^2$), získáme sice konzistentní, ale záporně vychýlený odhad, tj. odhad σ^2 podhodnotíme se systematickou chybou $-(\sigma^2/n)$. Tato chyba je tím menší, čím větší je výběr, takže u velkých výběrů je zanedbatelně malá.

Protože platí Ljapunovova věta, která říká, že statistika $\overline{X}$ je asymptoticky normální, tj. že se nepředpokládá X s rozdělením N (μ, σ^2), můžeme s dobrými výsledky provádět bodové odhady parametrů μ i u souborů, u kterých není zaručeno normální rozdělení.



Obrázek 6.1 - Podstata bodového odhadu

Z vlastností bodových odhadů vyplývá, že na jedné straně mají poměrně jednoduché stanovení (zpravidla je to přímo hodnota výběrové statistiky, někdy upravená o jednoduchou korekci, např. rozptyl), ale druhé straně má jednu velkou nevýhodu: nemůžeme stanovit chybu, s jakou odhad provádíme, tj. jak daleko je náš odhad od „neznámé pravdy“ parametru základního souboru.

Tuto skutečnost graficky znázorňuje obrázek 6.1 . Zde šedá plocha znázorňuje hustotu pravděpodobnosti (rozdělení četností) základního souboru. Průměr (hledaný parametr) základního souboru je znázorněn tučnou plnou čarou (μ) – tuto hodnotu však neznáme a nikdy znát nebudeme. Z tohoto základního souboru vybereme výběrový soubor, jehož konkrétní měřené veličiny jsou znázorněny bílými body. Jejich průměr (výběrová statistika) je znázorněna tučnou tečkovanou čárou ($\overline{x}$). V případě použití bodového odhadu hodnotu $\overline{x}$ prohlásí-

me za nejlepší odhad parametru μ. Ale nejsme schopni žádným způsobem vyjádřit, s jakou přesností je tento odhad proveden (resp. nejsme schopni vyčíslit vzdálenost $\bar{x}$ - μ, která je na obrázku 6.1 vyjádřena oboustrannou dvojitou šipkou, protože neznáme skutečnou hodnotu aritmetického průměru základního souboru μ). Pokud potřebujeme vyjádřit spolehlivost odhadu, použijeme intervalový odhad (viz následující kapitola).

Příklad 6.1:

Odhadneme parametry (střední hodnoty μ a rozptyl σ^2) základního souboru tlouštěk stromů určitého porostu, kde jsme měřením tlouštěk na zkusných plochách získali hodnoty výběrového aritmetického průměru $\bar{x}$ = 21,3 cm a výběrového rozptylu S^2 = 6,62 cm^2, měřený počet stromů (rozsah výběru) n = 224

$\bar{x} = 21{,}3$ cm $\rightarrow \mu = 21{,}3$ cm

$S^2 = 6{,}62$ cm $\rightarrow \sigma^2 = 6{,}62 \cdot \frac{224}{223} = 6{,}65$ cm^2

$\sigma = 2{,}58$ cm (směrodatná odchylka základního souboru)

Znamená to, že odhadovaná průměrná tloušťka celého porostu je 21,3 cm se směrodatnou odchylkou 2,58 cm. Také je zřejmé, že pro velké výběry se odhady parametrů variability základního souboru téměř neliší od výběrových statistik. Kdybychom měli rozsah výběru např. 10, byl by odhad rozptylu 6,62. (10/9) = 7,35, což je již podstatný rozdíl.

6.3 Intervalový odhad

6.3.1 Podstata intervalového odhadu

Při použití intervalového odhadu neurčujeme pouze jednu hodnotu, ale **stanovíme hranice intervalu hodnot, ve kterém leží neznámý parametr základního souboru**, ale tentokrát už **se známou, předem určenou pravděpodobností**.

Statistiky T_1, T_2 příslušné parametru τ, pro které při zvoleném $\alpha \in (0,1)$ platí

$$P(T_1 \leq \tau \leq T_2) = 1 - \alpha \tag{6.8}$$

určují interval spolehlivosti pro parametr τ při α hladině významnosti.

Jestliže jsou t_1 hodnota statistiky T_1 a t_2 hodnota statistiky T_2 v určitém výběru, potom nerovnost

$$t_1 \leq \tau \leq t_2 \tag{6.9}$$

aproximuje parametr τ se spolehlivostí 1 - α.

Pro hranice T_1 a T_2 platí

$$P(\tau > T_2) = \alpha_2 ; \tag{6.10}$$

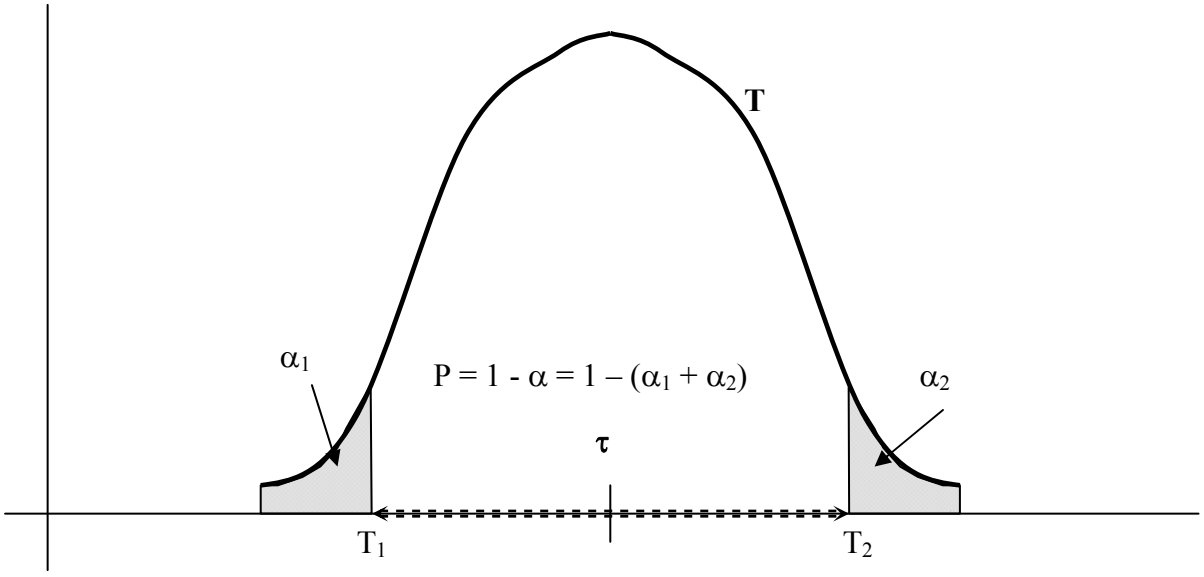
$$P(\tau < T_1) = \alpha_1 \qquad (6.11)$$

přičemž platí, že $\alpha = \alpha_1 + \alpha_2$.

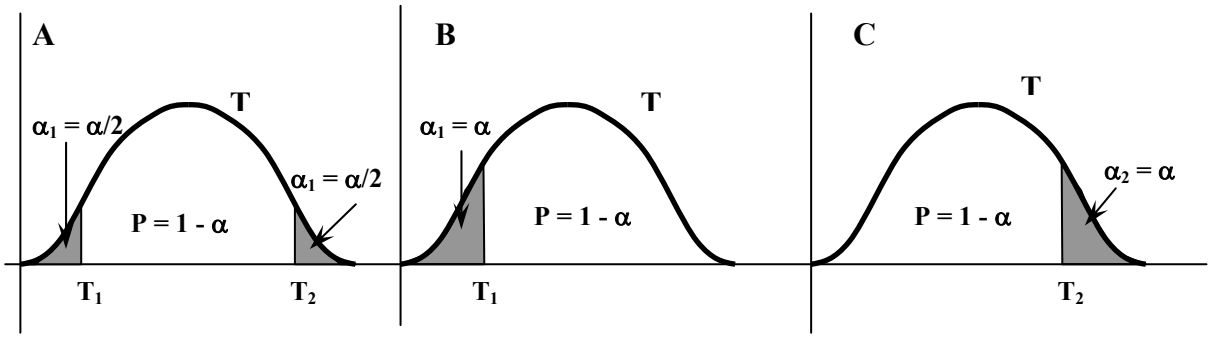
Grafickou interpretaci rovnice 6.8 ukazuje obrázek 6.2 . Křivka znázorňuje rozdělení výběrové statistiky T, kde určíme podle určitého předpisu (vzorce) hodnotu T_1 a T_2, což jsou hranice, ve kterých se s pravděpodobností P = 1 - α nachází nám neznámý parametr základního souboru τ. Plochu P znázorňuje bílá plocha pod křivkou, hodnotu α součet šedých ploch.

Hodnotu α (a tím také hodnotu P) můžeme určit dopředu (a priori) a tím také určit spolehlivost odhadu. Hodnota α se nazývá **hladina významnosti** a představuje **riziko (nejistotu), že neznámý parametr τ ve skutečnosti neleží v intervalu určeném hranicemi T_1 a T_2**, tedy že náš odhad je chybný. Jako hodnotu hladiny významnosti můžeme teoreticky volit jakoukoli hodnotu mezi 0 a 1, obvykle volené hodnoty (pro ně jsou také uzpůsobeny statistické tabulky) jsou α = 0,05 (5 %) a 0,01 (1 %). Znamená to, že neznámá hodnota parametru základního souboru leží v intervalu T_1 a T_2 s pravděpodobností P = 1 - α, což pro α = 0,05 znamená P = 0,95 (95 %) a pro α = 0,01 hodnotu P = 0,99 (99 %), tedy vypočítaný interval spolehlivosti je platný z 95 % (resp. 99 %) pravděpodobností. Tuto pravděpodobnost si můžeme vyložit tak, že kdybychom udělali 100 nezávislých náhodných výběrů z téhož základního souboru, tak v 95 (resp. 99) výběrech určíme interval spolehlivosti tak, že neznámý parametr základního souboru v něm skutečně leží a pouze v 5 (resp. v 1) případě se „zmýlíme“, tj. stanovíme intervalový odhad parametru základního souboru tak, že v něm studovaný parametr neleží.

Interval, kde leží s pravděpodobností P neznámý parametr základního souboru, je na obrázku 6.2 vyznačen oboustrannou šipkou.



Obrázek 6.2 - Schéma intervalu spolehlivosti



Obrázek 6.3 – Oboustranný souměrný (A) a jednostranné (levostranný – B, pravostranný – C) intervaly spolehlivosti

Interval spolehlivosti může být dvojí:

- **oboustranný** – je ohraničený zdola i shora hodnotami T_1 a T_2, přičemž hladina významnosti α se rozdělí na obě strany. Pokud platí, že $\alpha_1 = \alpha_2 = \alpha/2$, potom hovoříme o souměrném oboustranném intervalu spolehlivosti, což je také nejčastěji používaný případ (viz obrázek 6.3 A),
- **jednostranný** – je ohraničený pouze z jedné strany, a to:
  - levostranný – ohraničený zdola a platí $\alpha_1 = \alpha$, $\alpha_2 = 0$ (viz obrázek 6.3 B),
  - pravostranný – ohraničený shora, přičemž platí $\alpha_2 = \alpha$, $\alpha_1 = 0$ – viz obrázek 6.3 C).

6.3.2 Centrální limitní věta

Mějme základní soubor s normálním rozdělení o parametrech μ (aritmetický průměr) a σ^2 (rozptyl) a rozsahu $N \to \infty$. Z tohoto základního souboru provedeme „všechny možné“ náhodné výběry o rozsahu *n* (počet výběrů budeme označovat *k*). Pro každý výběr spočítáme aritmetický průměr $\bar{x}_i$ a rozptyl s_i^2. Získáme tedy *k* dvojic $\bar{x}_i$ a s_i^2. Jednotlivé hodnoty $\bar{x}_i$ a s_i^2 se budou od sebe lišit, protože každá dvojice $\bar{x}_i$ a s_i^2 byla vypočítána z jiných hodnot tj. z jiných výběrů). Musíme se uvědomit, že nejenom původní základní soubor, ale také „soubory“ $\bar{x}$ a s^2 (které jsou výsledkem různých výběrů) jsou náhodnou veličinou, která má svou střední hodnotu, rozptyl a rozdělení. Pro ně platí:

- aritmetický průměr výběrových průměrů se rovná aritmetickému průměru základního souboru

$$E(\bar{x}) = \mu \tag{6.12}$$

- rozptyl výběrových průměrů závisí přímo na rozptylu základního souboru a nepřímo na rozsahu výběru *n* (čím je rozsah výběru větší, tím je rozptyl výběrových průměrů menší) podle vztahu

$$s_{\bar{x}}^2 = \frac{\sigma^2}{n} \tag{6.13}$$

- z čehož se odvodí směrodatná odchylka výběrových průměrů (nazývá se také střední chyba průměru, ve statistických programech v angličtině se často označuje zkratkou MSE – mean standard error – nebo jen SE):

$$s_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{6.14}$$

- **se zvětšujícím se rozsahem výběru se rozdělení výběrových průměrů blíží normálnímu rozdělení $N(\mu, \frac{\sigma}{\sqrt{n}})$. Tato věta se nazývá centrální limitní věta**.

Tyto poznatky jsou velmi důležité a mají praktické využití. Podstatný je fakt, že uvedené vztahy a centrální limitní věta platí i v případě, že původní základní soubor nemá normální rozdělení. Při praktickém použití vztahů 6.13 a 6.14 narážíme na fakt, že rozptyl ani směrodatnou odchylku základního souboru ve valné většině případů neznáme. Proto se nahrazuje bodovým odhadem směrodatné odchylky $\hat{s}_{\bar{x}}$ základního souboru s použitím směrodatné odchylky výběru *s* podle vztahu

$$\hat{s}_{\bar{x}} = \frac{s}{\sqrt{n-1}} \tag{6.15}$$

Potom výraz

$$\Delta_{\bar{x}} = z_{\alpha/2} \cdot \sigma_{\bar{x}} = z_{\alpha/2} \cdot \frac{s}{\sqrt{n-1}}, \tag{6.16}$$

kde je $z_{\alpha/2}$ kvantil normovaného normálního rozdělení pro hladinu významnosti $\alpha/2$, se nazývá chyba výběrového průměru při spolehlivosti $P = 1 - \alpha$ a je to vlastně polovina intervalu spolehlivosti aritmetického průměru.

Pokud se směrodatná odchylka výběrových průměrů vyjádří relativně, dostaneme variační koeficient výběrových průměrů

$$s_{\bar{x}}\% = \frac{s_{\bar{x}}}{\bar{x}} \cdot 100 \tag{6.17}$$

nebo jestliže ve vzorci 6.14 dosadíme variační koeficient $\sigma\%$, resp. výběrový variační koeficient s%

$$\Delta_{\bar{x}}\% = z_{\alpha/2} \cdot \sigma_{\bar{x}}\% = z_{\alpha/2} \cdot \frac{s\%}{\sqrt{n-1}} \tag{6.18}$$

a výraz

$$\Delta_{\bar{x}}\% = z_{\alpha/2} \cdot \sigma_{\bar{x}}\% = z_{\alpha/2} \cdot \frac{s\%}{\sqrt{n-1}} \tag{6.19}$$

se nazývá relativní chybou (přesností) výběrového průměru při spolehlivosti P.

V dalších kapitolách se zaměříme na intervalové odhady nejpoužívanějších statistických charakteristik za předpokladu normálního rozdělení.

6.3.3 Interval spolehlivosti pro parametr μ (střední hodnotu základního souboru)

Výběr ze základního souboru o parametrech μ a σ^2 má rozsah n, aritmetický průměr $\bar{x}$ a rozptyl S^2 (a tedy směrodatnou odchylku S).

1. Je-li rozsah *n* libovolný a neznáme-li parametr σ^2, má průměr $\bar{x}$ při α hladině významnosti interval spolehlivosti

$$\bar{x} - t_{\alpha/2} \cdot \frac{S}{\sqrt{n-1}} \le \mu \le \bar{x} + t_{\alpha/2} \cdot \frac{S}{\sqrt{n-1}} \qquad (6.20)$$

kde $t_{\alpha/2}$ je kritická hodnota Studentova t - rozdělení pro $\alpha/2$ hladinu významnosti pro n - 1 stupňů volnosti. Při jednostranném intervalu spolehlivosti se užívá kvantil t_α.

2. Je-li rozsah n velký ($n \ge 30$) a známe rozptyl σ^2, potom má průměr μ při α stupni významnosti interval spolehlivosti

$$\bar{x} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \qquad (6.21)$$

kde $z_{\alpha/2}$ jsou kritické hodnoty normovaného rozdělení N (0, 1).

Vzorce 6.20 a 6.21 platí za předpokladu nezávislého výběru. Je-li však rozsah základního souboru N konečný a provádíme z něho výběr bez opakování, je třeba hodnotu směrodatné odchylky opravit. Potom nerovnost bude mít tvar

$$\bar{x} - t_{\alpha/2} \cdot \frac{S}{\sqrt{n-1}} \cdot \sqrt{1 - \frac{n}{N}} \le \mu \le \bar{x} + t_{\alpha/2} \cdot \frac{S}{\sqrt{n-1}} \cdot \sqrt{1 - \frac{n}{N}} \qquad (6.22)$$

$\sqrt{1 - \frac{n}{N}}$ se nazývá opravný koeficient pro konečný základní soubor.

Příklad 6.2:

Bodový odhad z příkladu 6.1 doplníme intervalovým odhadem parametru μ.

Podmínky: parametr σ^2 základního souboru neznáme, $\alpha = 0{,}05$. Odhad podle vzorce 6.20:

$$21{,}3 - 1{,}97 \cdot \frac{2{,}57}{\sqrt{223}} \le \mu \le 21{,}3 + 1{,}97 \cdot \frac{2{,}57}{\sqrt{223}}$$

$$21{,}3 - 0{,}34 \le \mu \le 21{,}3 + 0{,}34$$

$$20{,}96 \le \mu \le 21{,}64 \quad \text{(cm)}$$

Hodnoty rozptylu a aritmetického průměru jsou převzaty z příkladu **6.1**, hodnota 1,97 je kvantil Studentova t-rozdělení pro *n – 1* stupňů volnosti.

Znamená to, že průměrná tloušťka zkoumaného porostu se s pravděpodobností 95 % pohybuje v rozmezí 20,96 – 21,64 cm. Pokud bychom udělali další nezávislé výběry (tj. měření tlouštěk na jiných stromech), nejvýše 5 % výběrů by vyústilo v určení chybného intervalu spolehlivosti (tj. takového, ve kterém by skutečná průměrná tloušťka celého porostu neležela).

Kdybychom úkol řešili za předpokladu, že bodový odhad rozptylu z příkladu 6.1 považujeme za rozptyl základního souboru, použijeme statistiku 6.21

$$21{,}3 - 1{,}96.\frac{2{,}58}{\sqrt{224}} \le \mu \le 21{,}3 + 1{,}96.\frac{2{,}58}{\sqrt{224}}$$
$$21{,}3 - 0{,}34 \le \mu \le 21{,}3 \pm 0{,}34$$
$$20{,}9 \le \mu \le 21{,}7 \qquad \text{(cm)}$$

Výsledek při obou předpokladech je prakticky stejný. Je to způsobeno tím, že je použit velký výběr, kdy se kvantily Studentova a normovaného normálního rozdělení sobě hodně přibližují. V případě menšího výběru by byl rozdíl znatelnější.

Pokud bychom znali velikost základního souboru (nebo ji byli schopni korektně odhadnout), můžeme použít vzorec 6.22. V našem případě můžeme odhadnout celkový počet stromů v porostu (tj. rozsah základního souboru N) na základě velikosti porostu v ha a počtu stromů /ha z taxačních tabulek. Uvažujme, že jsme získali hodnotu N = 600. Potom

$$21{,}3 - 1{,}97.\frac{2{,}57}{\sqrt{223}}.\sqrt{1 - \frac{224}{600}} \le \mu \le 21{,}3 + 1{,}97.\frac{2{,}57}{\sqrt{223}}.\sqrt{1 - \frac{224}{600}}$$
$$21{,}3 - 0{,}27 \le \mu \le 21{,}3 + 0{,}27$$
$$21{,}0 \le \mu \le 21{,}6$$

I v tomto případě je výsledek prakticky shodný s výsledky předchozími.

Ukažme si ještě, jak na velikost intervalu působí změna hladiny významnosti a velikosti výběru. Obecně platí:

- se zmenšující se hodnotou hladiny významnosti α (a tedy se zvyšující se hodnotou $P = 1 - \alpha$) se zvětšuje (rozšiřuje) interval spolehlivosti (je to dáno tím, že zvyšující se hodnota P znamená „přísnější“ interval spolehlivosti, - musí vyhovět vyššímu procentu případů),
- se zmenšující se velikostí výběru *n* se také rozšiřuje interval spolehlivosti (to je dáno tím, že při menším výběru máme méně informací o základním souboru, proto musíme být v určení odhadu „opatrnější“).

Tyto skutečnosti si ukážeme na následujícím příkladu.

Příklad 6.3:

Intervalový odhad pro data z příkladu 6.2 vypočítejte pro

- *$\alpha = 0{,}01$ při nezměněné velikosti výběru,*
- *pro výběr n = 35 prvků při hodnotě $\alpha = 0{,}05$,*
- *pro $\alpha = 0{,}01$ a n = 35.*

Využijeme vzorce 6.20, do kterého dosadíme postupně následující hodnoty

n	***224***	***224***	***35***	***35***
α	***0,05***	***0,01***	***0,05***	***0,01***
$t_{\alpha/2,n-1}$	1,97	2,60	2,03	2,73
S	2,57	2,57	2,57	2,57
$\bar{x}$	21,30	21,30	21,30	21,30
dolní hranice	**20,96**	**20,85**	**20,40**	**20,10**
horní hranice	**21,64**	**21,75**	**22,20**	**22,50**

Pro vyšší názornost ukazuje vliv změny velikosti výběru a hladiny významnosti obrázek 6.4.

Všimněme si, že při statistickém šetření máme k dispozici aparát, který umožňuje podrobné rozlišení podmínek. Platí, že každé upřesnění podmínek řešení zpřesňuje výsledek.

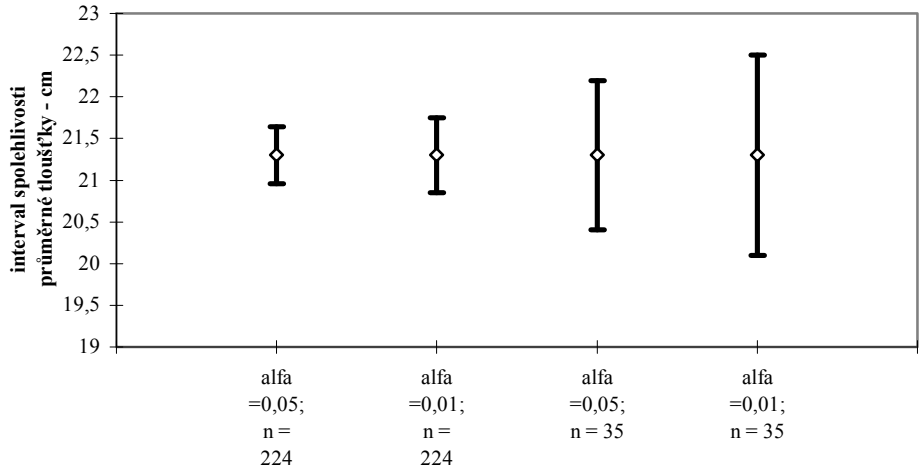
6.3.4 Interval spolehlivosti pro parametr σ^2 (rozptyl)

Výběr ze základního souboru s rozdělením N (μ, σ^2) má rozsah n a rozptyl S^2.

1. Pro n nepříliš velká ($n < 100$) můžeme pro parametr σ^2 použít interval spolehlivosti

$$\frac{n \cdot S^2}{\chi^2_{\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{n \cdot S^2}{\chi^2_{1-\frac{\alpha}{2}}} \tag{6.23}$$

kde $\chi^2_{\frac{\alpha}{2}}; \chi^2_{1-\frac{\alpha}{2}}$ jsou kritické hodnoty chi-kvadrát (Pearsonova) rozdělení pro α stupeň významnosti pro $n-1$ stupních volnosti.



Obrázek 6.4 – Vliv různých hodnot hladiny významnosti a velikosti výběru na intervaly spolehlivosti aritmetického průměru

Na rozdíl od intervalového odhadu průměru je intervalový odhad rozptylu nesouměrný (je to způsobeno tím, že Pearsonovo rozdělení je nesouměrné, zvláště pro menší počet stupňů volnosti).

Pro směrodatnou odchylku σ základního souboru se interval spolehlivosti určí analogicky

$$\sqrt{\frac{n \cdot S^2}{\chi^2_{\frac{\alpha}{2}}}} \le \sigma \le \sqrt{\frac{n \cdot S^2}{\chi^2_{1-\frac{\alpha}{2}}}} \qquad (6.24)$$

2. Pro výběry o velkém rozsahu (prakticky $n > 30$) můžeme pro parametr σ použít interval spolehlivosti

$$\sigma = S \pm z_{1-\alpha/2} \cdot \frac{S}{\sqrt{2n}} \qquad (6.25)$$

kde $z_{\alpha/2}$ se vyhledá v tabulkách distribuční funkce rozdělení N (0,1). Tento intervalový odhad je již souměrný (používá souměrné normální rozdělení).

Příklad 6.4:

Bodový odhad z příkladu 6.1 doplníme intervalovým odhadem parametru σ^2. Podmínky: $n = 224$; $S^2 = 6{,}62 \Rightarrow S = 2{,}57$.

Vzhledem k velkému rozsahu výběru použijeme statistiku 6.25

$$\sigma = 2{,}57 \pm 1{,}96 \,.\, \frac{2{,}57}{\sqrt{2 \cdot 224}}$$

$$\sigma = 2{,}57 \pm 0{,}24 \qquad \text{(cm)}$$

6.3.5 Interval spolehlivosti pro relativní četnost w

Relativní četnost vyjadřuje **podíl příznivých výsledků náhodného experimentu k celkovému počtu pokusů**. Je to tedy veličina s binomickým rozdělením a parametry $\mu = p$, $\sigma^2 = [p(1-p)]/n$. Hodnota *p* představuje relativní četnost (tj. pravděpodobnost výskytu) studovaného jevu v základním souboru, hodnota *n* počet pokusů. Hodnota *w* je relativní četnost ve výběrovém souboru. S použitím vlastností binomického rozdělení můžeme intervaly spolehlivosti odhadnout následovně (vzhledem k tomu, že výpočet binomického rozdělení pro vyšší *n* je obtížný, aproximuje se binomické rozdělení jinými rozděleními, obvykle normálním):

1. Pro dosti velká n (obvykle pro $n > 40$) a $p \in (0{,}3;\ 0{,}7)$ můžeme použít odhad

$$p \doteq w \pm z_{1-\alpha/2} \cdot \sqrt{\frac{w(1-w)}{n}} \qquad (6.26)$$

2. Jestliže $p \notin (0{,}3;\ 0{,}7)$, potom je rozdělení veličiny w výrazně nesouměrné a musíme použít Fisherovu transformaci

$$\varphi_w = 2\arcsin\sqrt{w} \qquad (6.27)$$

Potom interval spolehlivosti pro transformovanou veličinu φ_w je

$$\varphi_w = \varphi_w \pm \frac{z_{1-\alpha/2}}{\sqrt{n}} \qquad \textbf{(6.28)}$$

Interval spolehlivosti pro *p* najdeme zpětnou transformací. Tato transformace je tabelována. Při výpočtu musíme pamatovat, že počítáme v radiánech.

Příklad 6.5:

Byla zkoumána klíčivost semen smrku určité provenience. Ze 100 semen vyklíčilo 68. Stanovte intervalový odhad klíčivosti semen smrku této provenience.

Stanovíme výběrovou relativní četnost w = 68/100 = 0,68. Vzhledem k tomu, že tato hodnota spadá do intervalu 0,3 - 0,7 a rozsah výběru je vyšší než 40 (100 semen), můžeme použít vztah 6.26 (počítáme pro $\alpha = 0,05$):

$$p \cong 0,68 \pm 1,96 \cdot \sqrt{\frac{0,68 \cdot 0,32}{100}} = 0,68 \pm 0,09$$

tedy klíčivost semen smrku zkoumané provenience se pohybuje s pravděpodobností 95 % v rozmezí 0,59 – 0,77 (tedy 59 – 77 %).

Příklad 6.6:

Při průzkumu poškození porostů suchem bylo prověřeno 25 náhodně vybraných porostů zkoumané oblasti. V sedmi z nich bylo zjištěno poškození. Jaké procento poškozených porostů je možné očekávat v dané oblasti?

Relativní četnost je w = 7/25 = 0,28. Vzhledem k této hodnotě a k tomu, že velikost výběru je poměrně malá (n = 25), použijeme postup pomocí Fisherovy transformace.

Použitím vzorce 6.27 zjistíme transformovanou hodnotu (výpočtem nebo pomocí statistických tabulek) $\varphi_w = 2\arcsin(0,28^{-1/2}) = 1,1152$. Dosadíme do vzorce 6.28

$$\varphi_w = 1,1152 \pm \frac{1,96}{\sqrt{25}} = 1,1152 \pm 0,392$$

tedy transformovaná veličina φ_w se nachází v intervalu 0,7232 - 1,5072. Tyto hodnoty retransformujeme na hodnoty p (buď podle tabulek nebo pomocí vztahu $p = \left(\sin \frac{\varphi_w}{2} \right)^2$)

a získáme hodnoty intervalu spolehlivosti pro hodnotu *p* (tedy pro relativní četnost základního souboru, tj. podíl poškozených porostů v celé sledované oblasti) 0,125 – 0,468. Můžeme tedy uzavřít, že s pravděpodobností 95 % se poškození porostů suchem ve sledované oblasti pohybuje v rozmezí 12 – 47 %.

---

7 Testování statistických hypotéz

Testování statistických hypotéz je jedním z nejpoužívanějších a nejdůležitějších postupů matematické statistiky. Umožňuje se stanovenou pravděpodobností testovat platnost různých vlastností základního souboru na základě znalosti hodnot a vlastností výběrového souboru.

7.1 Podstata testování hypotéz

Statistická hypotéza je určitá domněnka (předpoklad) o vlastnostech základního souboru. Jako příklady hypotéz můžeme uvést:

- určitý parametr (např. aritmetický průměr) se rovná určité hodnotě (např. dané normou),
- průměry dvou souborů se sobě rovnají,
- určitá náhodná veličina má normální rozdělení, apod.

Z hlediska matematické statistiky je **statistická hypotéza předpoklad o rozdělení pravděpodobnosti jedné nebo více náhodných veličin**. Tento předpoklad se týká parametrů rozdělení náhodné veličiny v základním souboru nebo se může vztahovat pouze k zákonu rozdělení náhodné veličiny (k hustotě pravděpodobnosti nebo distribuční funkci).

O vlastnostech základního souboru můžeme s určitostí soudit jen na základě znalosti všech jeho prvků. To je však, zvláště u velkých souborů, z mnoha důvodů (technických, ekonomických, časových apod.) nereálné. Zpravidla máme k dispozici výběrový soubor jako výsledek náhodného výběru. Z výběrových charakteristik nejsme schopni jednoznačně („stoprocentně“) rozhodnout, zda posuzovaná vlastnost v základním souboru platí.

Jako příklad si můžeme uvést posuzování tloušťkové vyspělosti několika porostů. Tvrdíme (což je naše statistická hypotéza), že všechny porosty jsou stejně tloušťkově vyspělé. Jak můžeme toto tvrzení potvrdit nebo vyvrátit s jistou mírou objektivity? Vzhledem k tomu, že nemáme zpravidla možnost získat základní soubory (tj. změřit tloušťky všech stromů ve všech posuzovaných porostech), podle pravidel výběrového zjišťování získáme výběrové soubory určité velikosti. Pro tyto výběry vypočítáme výběrové charakteristiky, v našem případě budou nejdůležitější průměrné tloušťky jednotlivých porostů (aritmetické průměry všech měřených tlouštěk v daném porostu) a jejich variabilita. Je pravděpodobné, že průměrné tloušťky ve výběrech z jednotlivých porostů se budou navzájem numericky lišit. Nyní musíme odpovědět na otázku, zda tyto rozdíly jsou způsobeny:

- skutečnými rozdíly mezi základními soubory, tj. že posuzované porosty jsou skutečně tloušťkově odlišné (což by naši hypotézu o stejné průměrné tloušťce zamítalo),
- nebo náhodnými odchylkami, které vznikly pouze jako důsledek konkrétního náhodného výběru, ale v základních souborech (celých porostech) ve skutečnosti neexistují (to by naši hypotézu potvrzovalo).

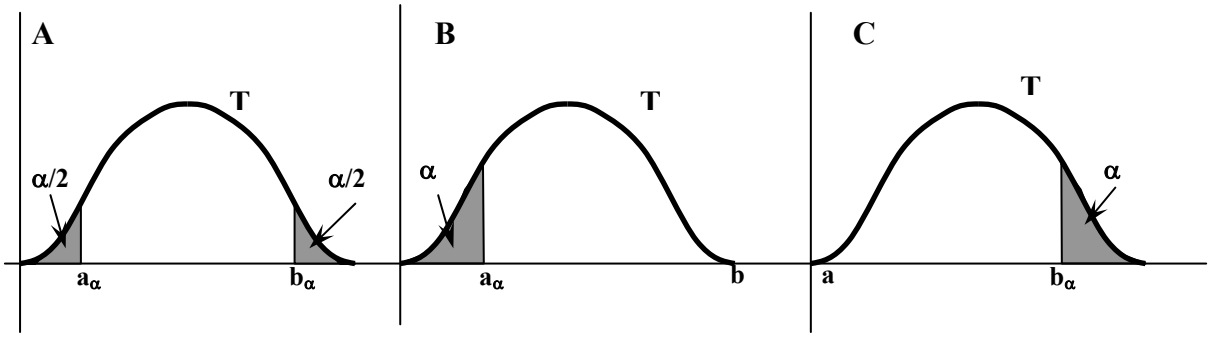
Pokud bychom rozhodovali subjektivně, bylo by toto rozhodnutí zatíženo naší osobní zkušeností, názorem, předjímáním výsledku, naší objektivností apod. Pro objektivizaci rozhodnutí použijeme techniku statistického testování.

Test statistické hypotézy je **pravidlo** (kritérium)**, které na základě výsledků zjištěných z náhodného výběru objektivně předepisuje rozhodnutí, má-li být ověřovaná hypotéza zamítnuta či nikoliv** (MELOUN - MILITKÝ 1994).

Testy statistických hypotéz jsou obecně založeny na tom, že odvodíme (nebo u běžně používaných testů převezmeme z literatury, což je nejčastější případ) pro zkoumanou statistickou vlastnost základního souboru testové kritérium, které je založeno na určité náhodné veličině (obecně ji označme T). Rozdělení této náhodné veličiny je známo pro případ platnosti i neplatnosti nulové hypotézy.

Potom interval $I_p = \langle a_\alpha, b_\alpha \rangle$, kde platí, že $P(a_\alpha \leq T \leq b_\alpha) = 1 - \alpha$ se nazývá interval prakticky možných hodnot náhodné proměnné T při $(100.\alpha)$ % stupni významnosti:

- je-li $P(T < a) = 0$ a $P(T > b_\alpha) = \alpha$, potom b_α nazveme horní kritický bod (viz obrázek 7.1 c),
- je-li $P(T < a_\alpha) = P(T > b_\alpha) = \alpha/2$, potom a_α je dolní a b_α horní kritický bod (viz obrázek 7.1 a),
- je-li $P(T < a_\alpha) = \alpha$ a $P(T > b) = 0$, potom a_α je dolní kritický bod (obrázek 7.1 b).



Obrázek 7.1 – Schématické znázornění oboustranného a jednostranných testů. Bílá plocha pod křivkou je obor přijetí, šedá plocha je obor nepřijetí (kritický)

Mějme statistiku T a určitým způsobem zjištěnou její hodnotu *t*. Buď τ předpokládaná hodnota statistky T a I_p interval prakticky možných hodnot veličiny T při 100α % stupni významnosti. Je tedy $\tau \in I_p$. Potom pokládáme rozdíl $\tau - t$ za:

- náhodně vysvětlitelný, když $t \in I_p$
- statisticky významný, když $t \notin I_p$ při $\alpha = 0{,}05$
- statisticky velmi významný, když $t \notin I_p$ při $\alpha = 0{,}01$.

Uvedené hodnoty meze významnosti a hodnocení významnosti jsou dány úmluvou.

Testů existuje velké množství. Dělí se na dvě základní skupiny:

- **parametrické** - jsou založené na statistických charakteristikách (parametrech) výběrového souboru a jsou vázané na splnění určitých předpokladů o parametrech a určitém typu rozdělení znaku v základním souboru (např. předpoklad normality),
- **neparametrické** - nejsou založené na parametrech a především nejsou závislé na typu rozdělení základního souboru. Používají se především tehdy, jestliže parametrické testy není možné použít pro nesplnění některých podmínek použití. Jejich nevýhodou je především nižší síla testu, tedy schopnost správně zamítnout H_0.

V dalším textu se zaměříme na základní parametrické i neparametrické testy včetně jejich případných modifikací při nesplnění požadovaných podmínek

7.2 Obecný postup při testování statistických hypotéz

Abychom statisticky korektně rozhodli o výsledku statistického testu, musíme zachovat určitý obecně platný postup. Metodologie testování se skládá z následujících kroků:

1. Formulace **nulové hypotézy** ($\mathbf{H_0}$) a **alternativní hypotézy** ($\mathbf{H_1}$). Nulová hypotéza je vlastně posuzované tvrzení o určité vlastnosti základního souboru. Alternativní hypotéza se považuje za platnou v případě, že nulová hypotéza byla zamítnuta.
   Z hlediska provedení testů je důležitý rozdíl mezi jednostrannou a oboustrannou alternativní hypotézou (podle obrázku 7.1). Ukažme si tento rozdíl na příkladu o průměrných tloušťkách porostů, který byl uveden na začátku této kapitoly (pro jednoduchost budeme uvažovat pouze dva porosty): Nulová hypotéza tady zní:
   H_0: *průměrné tloušťky obou porostů se sobě rovnají* nebo - vyjádřeno matematicky - $\mu_1 = \mu_2$ (kde μ jsou střední hodnoty základních souborů, tj. v našem případě průměrné tloušťky vypočítané ze všech stromů obou porostů, μ_1 pro první porost, μ_2 pro druhý porost)
   Alternativní hypotéza může být konstruována trojím způsobem:
   a) H_1: *průměrné tloušťky obou porostů se sobě nerovnají* ($\mu_1 \neq \mu_2$) – toto je příklad tzv. **oboustranného testu**, kdy pro zamítnutí nulové hypotézy stačí, aby se průměrné tloušťky obou porostů nerovnaly a je jedno, která z nich je menší nebo větší.
   b) H_1: *průměrná tloušťka porostu I je větší než průměrná tloušťka porostu II* ($\mu_1 > \mu_2$) – v tomto případě se jedná o **jednostranný test**, kdy k zamítnutí nulové hypotézy musí platit nerovnost $\mu_1 > \mu_2$ (v případě, že se průměrné tloušťky sobě rovnají nebo je tloušťka prvního porostu nižší, nulová hypotéza se nezamítá).
   c) H_1: *průměrná tloušťka porostu I je menší než průměrná tloušťka porostu II* ($\mu_1 < \mu_2$) – také jednostranná hypotéza, ale pro zamítnutí nulové hypotézy musí platit vztah $\mu_1 < \mu_2$.

Konkrétní formulace nulové a alternativní hypotézy závisí na cíli analýzy. Podle toho, jak jsou hypotézy formulovány, se liší rozhodování o výsledku testu. Proto je nutné pečlivé formulaci hypotéz věnovat pozornost.

2. Volba **hladiny významnosti α.** Hodnota α má zde velmi podobný význam jako u intervalových odhadů (viz kapitola 6), tedy je to riziko, že rozhodnutí učiněné na základě výsledku testu bude chybné (zamítneme nulovou hypotézu, která ve skutečnosti platí). U testování je ovšem situace složitější a hodnota α souvisí s tzv. chybou I. i II. druhu, což bude podrobněji vysvětleno v následující kapitole. Obvyklé hodnoty $\alpha = 0{,}05$ nebo 0,01.
3. Volba **druhu testu** a **testového kritéria**, tj. funkce hodnot náhodného výběru se známým rozdělením pravděpodobností v případě platnosti i neplatnosti nulové hypotézy (na obrázku 7.1 je to náhodná veličina T). Pro často užívané statistické úlohy jsou vyvinuty standardně používané testy včetně testového kritéria, jehož výsledná hodnota je dále porovnávána s kritickou hodnotou.
4. Určení **kritického oboru** (oboru nepřijetí) testového kritéria na základě jejího rozdělení pravděpodobnosti a hladiny významnosti. Kritický obor je určován na především na základě typu testu – zdali se jedná o jednostranný test nebo o oboustranný test. V případě oboustranného testu (alternativní hypotéza je formulována typově podle příkladu v bodě 1a)) se hodnota α rozdělí rovným dílem na obě strany rozdělení, protože nevíme, na které straně by testové kritérium mohlo přesáhnout kritic-

kou hodnotu (viz obrázek 7.1 a - zde jsou kritické obory vyznačeny šedě, obor přijetí nulové hypotézy je bílá plocha pod křivkou). V případě jednostranných testů (alternativní hypotéza je formulována typově podle příkladu 1b) nebo 1c)) se určuje kritická hodnota podle hladiny významnosti α. Konkrétní kritické hodnoty se při praktickém provádění testů vyhledají ve statistických tabulkách nebo jsou určeny speciálními funkcemi ve statistických programech, příp. tabulkových procesorech (Excel, Quattro Pro, apod.). Kritická hodnota rozděluje celý obor testového kritéria na dvě části – obor přijetí a obor nepřijetí.

5. Rozhodnutí o **výsledku testu**, tj. zda
   a) zamítnout nulovou hypotézu (jestliže vypočítaná hodnota testového kritéria padne do oboru nepřijetí),
   b) nezamítnout nulovou hypotézu (jestliže vypočítaná hodnota testového kritéria padne do oboru přijetí).

V této souvislosti je nutné si uvědomit, že nezamítneme-li nulovou hypotézu, neznamená to absolutní potvrzení její platnosti ve skutečnosti. Pouze to znamená, že výsledek testu neukázal tak velkou neshodu mezi zjištěnou skutečností ve výběrovém souboru a testovanou hypotézou o určité vlastnosti základního souboru, která by dala dostatečný důvod k zamítnutí hypotézy. V další práci tedy s určitou pravděpodobností předpokládáme, že platí nulová hypotéza.

Postup testování si ukážeme na příkladu.

Příklad 7.1:

Byla prověřována správnost měření výškoměru. Byla 15 x změřena výška, jejíž hodnota je přesně známa ($\mu_0 = 20$ m). Z výsledků měření se získal průměr měřených výšek $\bar{x} = 19{,}2$ m se směrodatnou odchylkou $S = 1{,}1$ m. Stanovte, zda-li výškoměr měří správně.

Správnost měření určitým přístrojem se posuzuje podle toho, zdali střední hodnota jednotlivých měření se shoduje s normovanou hodnotou (etalonem). V našem případě je normovanou hodnotou předem stanovená výška 20 m. Abychom mohli odpovědět na položenou otázku se „100 %“ určitostí, museli bychom provést teoreticky nekonečně mnoho měření výšky (tím bychom získali základní soubor veličiny „výška“). Tento postup je nemožný, proto se musíme spokojit s konečným počtem měření - s výběrovým souborem. Na základě dat poskytnutým výběrovým souborem je nutné rozhodnout, zda

- průměr naměřených výšek (19,2 m) se odchyluje pouze náhodně od normované hodnoty 20 m (což by bylo způsobeno konkrétním výběrem, tím, že jsem naměřili určitých 15 hodnot, které se „náhodou“ liší od normované hodnoty; pokud bychom naměřili jiné výběry, mohly by se s normovanou hodnotou shodovat nebo se jí alespoň více blížit),
- nebo zda tuto odchylku je nutno považovat za systematickou (to by bylo způsobeno tím, že výškoměr skutečně systematicky podhodnocuje měřené výšky, tj. nižší naměřená výška je vlastností nikoli jednoho náhodného výběru, ale základního souboru naměřených hodnot).

Toto rozhodnutí nám pomůže objektivizovat statistický test.

Prvním krokem je stanovení nulové a alternativní hypotézy. Nulovou hypotézu stanovíme takto:

H_0: Výškoměr měří správně.

nebo přepsáno „matematicky“:

H_0: $\mu = \mu_0$ (*slovně vyjádřeno – výběrový soubor o parametrech* $\bar{x}$ *a* S^2 *pochází ze základního souboru s normálním rozdělením o parametrech* μ *a* σ^2*, přičemž parametr* μ *se rovná normované hodnotě* μ_0*).*

Nyní musíme stanovit alternativní hypotézu, tj. tu vlastnost základního souboru, která bude platit v případě, že zamítneme platnost nulové hypotézy. V našem případě to bude tzv. oboustranná hypotéza, protože výškoměr bude měřit správně jen v tom případě, že průměr naměřených výšek se bude rovnat normované hodnotě. Pokud bude podhodnocovat nebo nadhodnocovat výšky, jedná se o nekvalitní výrobek, a je nám jedno, na kterou stranu vychýlení půjde. Proto stanovíme alternativní hypotézu jako negaci nulové hypotézy:

H_1: *Výškoměr* ***ne****měří správně.*

nebo přepsáno „matematicky“:

H_1: $\mu \neq \mu_0$ *(slovně vyjádřeno – výběrový soubor o parametrech* $\bar{x}$ *a* S^2 *pochází ze základního souboru s normálním rozdělením o parametrech* μ *a* σ^2*, přičemž parametr* μ *se* ***ne****rovná normované hodnotě* μ_0*).*

Nyní stanovíme hladinu významnosti α. Ve shodě se statistickou praxí zvolíme hodnotu $\alpha = 0{,}05$, tedy tento test provedeme se statistickou jistotou (pravděpodobností) 95 %.

Jako další krok stanovíme testovací kritérium, tj. náhodnou veličinu, jejíž rozdělení je známo pro případ platnosti i neplatnosti nulové hypotézy. V případě tohoto testu (říká se mu test průměru pro jeden výběr – dále bude ještě podrobněji rozebrán) je to statistika

$$t = \bar{x} - \mu_0 \cdot \frac{\sqrt{n-1}}{S},$$

která má Studentovo t- rozdělení s (n-1) stupni volnosti (všimněte si, nakolik se tento vzorec podobá vzorci 6.20 pro intervalový odhad průměru – tyto dvě metody – intervalový odhad a testování – spolu skutečně úzce souvisí).

Jestliže dosadíme příslušné hodnoty, dostaneme výslednou hodnotu testového kritéria

$$t = (19{,}2 - 20) \cdot \frac{\sqrt{15-1}}{1{,}1} = -2{,}72$$

Nyní musíme určit kritickou hodnotu (tj. hranici kritického oboru) – viz obrázek 7.1 A. Pokud bude hodnota testového kritéria vyšší (testové kritérium bude dále od normované hodnoty než kritická hodnota, tedy padne do kritického oboru), potom zamítneme nulovou hypotézu, v opačném případě ji nemůžeme zamítnout, údaje z našeho výběru nám k tomu nedávají dostatečné „důkazy“.

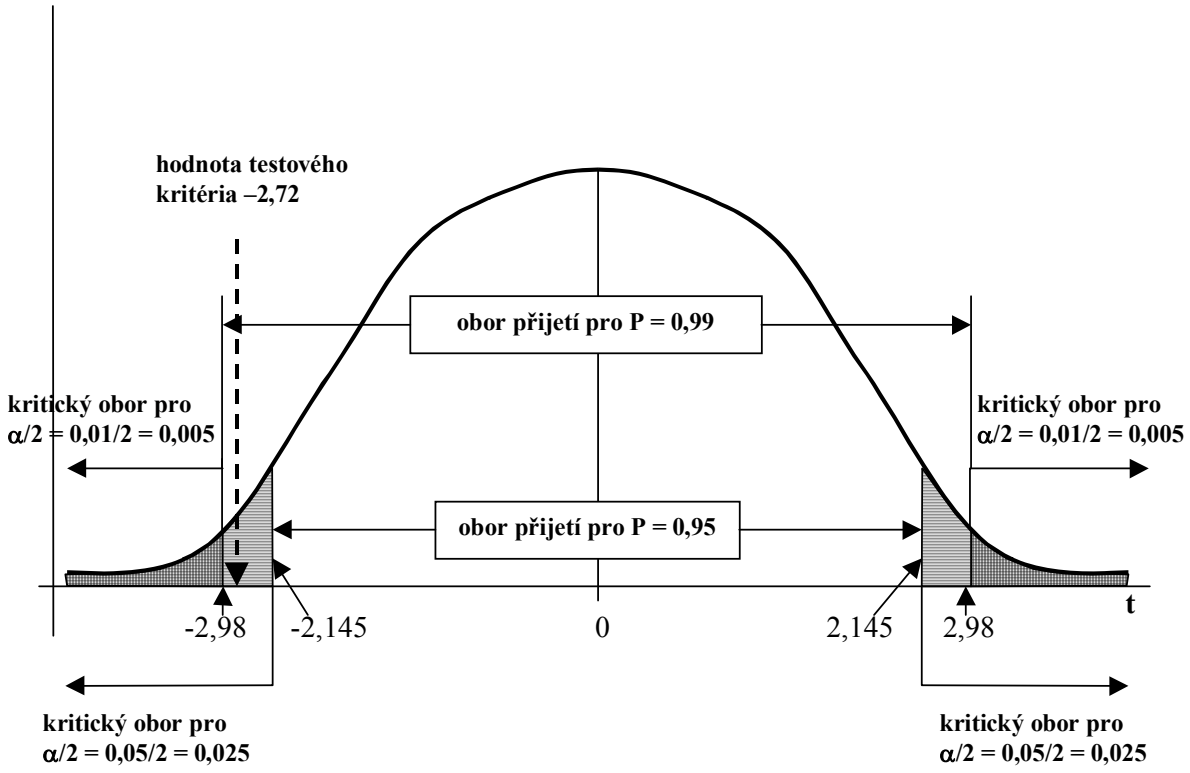
Pracujeme s oboustrannou hypotézou, tedy ve shodě s obrázkem 7.1 A musíme hodnotu α rozdělit na obě strany rozdělení, budeme tedy hledat hodnotu $t_{\alpha/2,n-1} = t_{0{,}025,14}$, což je hodnota 2,145 (buď ve statistických tabulkách nebo funkcí TINV v Excelu). Vzhledem k symetrii t - rozdělení, má kritický obor oboustranného testu dvě hranice (- 2,145 a 2,145 - to jsou hodnoty a_α a b_α na obrázku 7.1 A). Nyní můžeme srovnávat – testové kritérium je t = -2,72, je to hodnota menší než - 2,145, z čehož vyplývá, že padne do dolní poloviny kritického oboru, tedy nulovou hypotézu můžeme s pravděpodobností 95 % zamítnout.

Závěr tedy zní: na základě daného výběru můžeme s pravděpodobností 95 % předpokládat, že výškoměr neměří přesně, systematicky podhodnocuje.

Grafické řešení tohoto příkladu je znázorněno na obrázku 7.2 . Je zde schématicky znázorněna hustota pravděpodobnosti testového kritéria (t-rozdělení pro 14 stupňů volnosti), což je tučná zvonovitá křivka. Na obou koncích jsou vyznačeny kritické obory – čárkovaně pro $\alpha = 0{,}05$ (hranicí jsou kritické hodnoty - 2,145 a 2,145), čtverečkovaným rastrem je vyznačen kritický obor pro $\alpha = 0{,}01$ (hranicí jsou hodnoty $t_{0{,}005,14} =$ 2,98 a –2,98). Vidíme, že testové kritérium –2,72 (vyznačeno čárkovanou šipkou) svou hodnotou spadá mezi obě kritické hod-

noty (mezi –2,145 a-2,98). Znamená to, že na hladině významnosti $\alpha = 0{,}01$ bychom nemohli nulovou hypotézu zamítnout, Můžeme tvrdit s pravděpodobností 95 %, že výškoměr neměří přesně, ale stejné tvrzení bychom si nemohli dovolit s pravděpodobností 99 %. Tento fakt je z obrázku 7.2 také zřejmý, pokud se podíváme na rozsah oborů přijetí – obor přijetí pro P = 99 % je širší než pro 95 %.

Jaká je tedy nejmenší možná hodnota α, kdy ještě můžeme zamítnout nulovou hypotézu? Zjistíme to z distribuční funkce t-rozdělení, kdy pro hodnotu 2,72 získáme hodnotu pravděpodobnosti 0,0166 (buď z podrobných statistických tabulek interpolací nebo např. pomocí funkce Excelu TDIST). Tato hodnota se obvykle ve statistické literatuře a programech nazývá p-hodnota (angl. p-value, p-level, probability level, apod.). Znamená to, že maximální pravděpodobnost, kdy ještě můžeme zamítnout nulovou hypotézu, je v tomto případě $1 - p = 1 - 0{,}0166 = 0{,}9834 = 98{,}3$ %. Toto rozhodnutí můžeme tedy považovat za velmi silné. Pravděpodobnost, že bychom zamítli nulovou hypotézu, která je skutečnosti platí, je pouze 1,66 %.



Obrázek 7.2 – Grafická interpretace příkladu 7.1

7.3 Chyba I. a II. druhu

Vzhledem k tomu, že při testování pracujeme s pravděpodobností, výsledek testování může být zatížen určitou chybou. V podstatě mohou nastat dvě základní situace:

- testovaná nulová hypotéza je **chybně zamítnuta -** tato chyba se nazývá **chyba I. druhu** a její pravděpodobnost je určena předem známou hladinou významnosti α.
- testovaná nulová hypotéza je **chybně nezamítnuta** - tato chyba se nazývá **chyba II. druhu** a její pravděpodobnost se označuje β (není předem známa).

Vztahy mezi platností nulové hypotézy ve skutečnosti a podle rozhodnutí testu uvádí tabulka 7.1 .

Význam chyby I. druhu je zřejmý. Té se dopustíme, jestliže zamítneme nulovou hypotézu, která ve skutečnosti platí. Určení její pravděpodobnosti je jasné – je to hodnota hladiny významnosti α. Jestliže určíme α např. 0,05, potom pravděpodobnost výsledku testu $P = 1 - \alpha = 0{,}95$, tj. 95 %. Pravděpodobnost P si můžeme představit tak, že kdybychom udělali 100 náhodných výběrů ze základního souboru a tyto testovali, tak v pěti případech ze sta chybně zamítneme ve skutečnosti správnou nulovou hypotézu.

		H_0 ve skutečnosti	
		platí	neplatí
H_0 podle rozhodnutí testu	platí	činíme správné rozhodnutí	dopouštíme se chyby II. druhu
	neplatí	dopouštíme se chyby I. druhu	činíme správné rozhodnutí

Tabulka 7.1 - Vztahy mezi platností nulové hypotézy ve skutečnosti a podle rozhodnutí testu

Pravděpodobnost chyby II. druhu se určuje mnohem obtížněji. Hlavní problém spojený s chybou II. druhu spočívá v tom, že její pravděpodobnost je závislá na skutečné hodnotě parametru, tj. čím více se liší hodnoty uvedené v nulové a alternativní hypotéze, tím je pravděpodobnost chyby II. druhu menší. Její podstatu si nejlépe vysvětlíme na jednoduchém příkladu.

Uvažujme, že řešíme test charakterizovaný následujícími hypotézami:

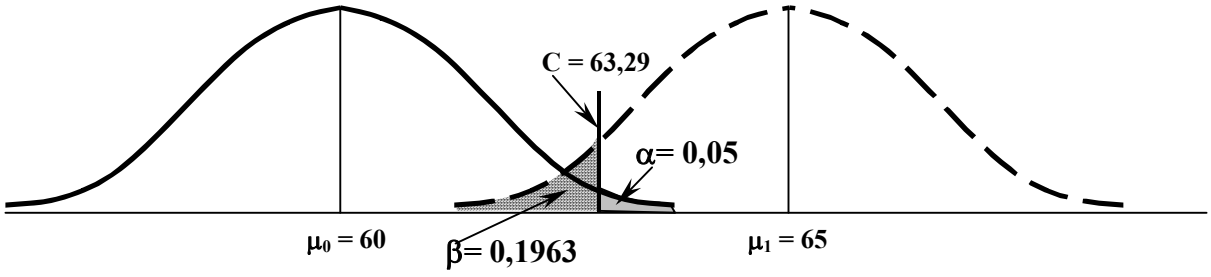
$$H_0: \mu = 60 \qquad H_1: \mu = 65$$

Je nutné zdůraznit, že takto postavený test není obvyklý – skládá se ze dvou jednoduchých hypotéz, obvyklý případ je takový, že v alternativní hypotéze uvažujeme buď hodnotu jakkoli se lišící od hodnoty uvedené v nulové hypotéze (oboustranný test), event. větší nebo menší než hodnota nulové hypotézy (jednostranný test). K těmto případům se dostaneme později na základě vysvětlení provedeného pro tuto jednoduchou hypotézu. Dále budeme pracovat s hladinou významnosti $\alpha = 0{,}05$, s velikostí výběru $n = 100$ a rozptylem základního souboru (předpokládejme, že je znám) $\sigma = 20$.

Prvním krokem je určení kritické hodnoty v měřítku testované veličiny. Pracujeme s jednostranným (pravostranným) testem a vzhledem k velikosti výběru můžeme použít normované normální rozdělení, kde pro hodnotu $\alpha = 0{,}05$ určíme kritickou hodnotu $z = 1{,}645$. Provedeme transformaci do měřítka veličiny, se kterou pracujeme, a zjistíme reálnou hodnotu hranice kritického oboru C:

$$C = \mu_0 + 1{,}645 \cdot \frac{\sigma}{\sqrt{n}} = 60 + 1{,}645 \cdot \frac{20}{10} = 63{,}29$$

Znamená to, že nulová hypotéza nebude přijata, jestliže testové kritérium bude vyšší než



63,29.

Obrázek 7.3– Grafické znázornění chyby I. a II. druhu pro jednostranný test (bližší vysvětlení viz v textu)

Tuto situaci znázorňuje obrázek 7.3. Zde je tučnou plnou čarou znázorněna hustota pravděpodobnosti testového kritéria pro případ platnosti nulové hypotézy, tj. pro $\mu = 60$, a tučnou čár-

kovanou čárou pro případ platnosti alternativní hypotézy, tj. $\mu = 65$. Je zde také vyznačena hranice kritického oboru C, která na obou křivkách hustoty pravděpodobnosti vymezuje dvě důležité oblasti:

- chybu I. druhu – je dána a priori hodnotou α a vyznačena šedou plochou. To je pravděpodobnost že výběrový průměr $\bar{x}$ padne napravo od kritické hranice C (a tedy test odmítne nulovou hypotézu) za předpokladu, že ve skutečnosti nulová hypotéza platí, tj. střední hodnota základního souboru je skutečně 60;
- chybu II. druhu – její hodnota je 0,1963 (výpočet bude proveden níže) a je vyznačena „cihličkovým" rastrem. To je pravděpodobnost že výběrový průměr $\bar{x}$ padne nalevo od kritické hranice C (a tedy test přijme nulovou hypotézu) za předpokladu, že ve skutečnosti nulová hypotéza neplatí, tj. střední hodnota základního souboru je ve skutečnosti 65.

Jak stanovíme velikost chyby II. druhu? Pravděpodobnost výskytu obou druhů chyb můžeme vyjádřit pomocí následujících podmíněných pravděpodobností:

$$\alpha = P(\bar{x} > C \mid \mu = \mu_0) \tag{7.1}$$

$$\beta = P(\bar{x} < C \mid \mu = \mu_1) \tag{7.2}$$

Tento zápis pro rovnici 7.2 (tedy pro chybu II. druhu) čteme: beta je pravděpodobnost, že hodnota $\bar{x}$ bude menší než hranice kritického oboru C za předpokladu, že skutečná střední hodnota základního souboru μ se rovná hodnotě μ_1 (tedy 65). Pokud tuto rovnici vyřešíme, získáme hodnotu β pro alternativní hypotézu $\mu = 65$:

$$\beta = P(\bar{x} < C \mid \mu = \mu_1) = P\left(\frac{\bar{x} - \mu_1}{\sigma \cdot \sqrt{n}} < \frac{C - \mu_1}{\sigma \cdot \sqrt{n}}\right) = P\left(Z < \frac{63{,}29 - 65}{20/10}\right) = P(Z < -0{,}855) = 0{,}1963$$

Při řešení této rovnice jsme vycházeli z úvahy, že za předpokladu platnosti alternativní hypotézy je $\bar{x}$ normálně rozdělenou náhodnou veličinou (na základě centrální limitní věty) s průměrem 65 a standardní chybou (směrodatnou odchylkou výběrového průměru) rovnou $\sigma/\sqrt{n}$. Abychom jednoduše vyjádřili pravděpodobnost β, transformovali jsme $\bar{x}$ na normovanou veličinu Z s využitím hodnoty μ_1 jako průměru veličiny $\bar{x}$. Výsledkem je hodnota 0,1963, což je pravděpodobnost, že nulová hypotéza bude přijata, i když ve skutečnosti platí alternativní hypotéza.

Podobným způsobem bychom mohli vypočítat velikost chyby II.druhu pro jakoukoli jinou hodnotu alternativní hypotézy.

Z hodnoty β se stanovuje hodnota $S = 1 - \beta$, která se nazývá **síla testu**. Je to **pravděpodobnost, že test správně odmítne nulovou hypotézu, jestliže ve skutečnosti platí alternativní hypotéza**. Test je tím silnější („přísnější"), čím je vyšší jeho schopnost odmítnout nesprávnou hypotézu. V našem ukázkovém příkladu je síla testu $S = 1 - \beta = 1 - 0{,}1963 = 0{,}8037$. Jinak řečeno, za předpokladu, že ve skutečnosti platí alternativní hypotéza ($\mu = 65$), máme šanci (pravděpodobnost) asi 80 %, že zamítneme nulovou hypotézu.

Ovšem při praktickém provádění testů jsou zpravidla alternativní hypotézy konstruovány jinak – není zde uváděna konstantní hodnota, ale celý interval, např. obvyklá formulace pro náš příklad by (v případě jednostranného testu) zněla takto:

H_0: $\mu = 60$ $\qquad$ H_1: $\mu > 60$

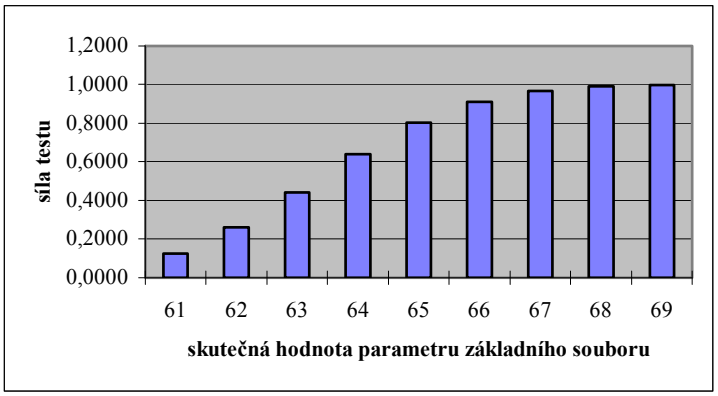
V tomto případě je nulová hypotéza neplatná v případě, že hodnota μ je jakékoli číslo větší než 60. Pro každou hodnotu lze takto stanovit chybu II. druhu a sílu testu. Pro hodnoty 61 - 69 to ukazuje obrázek 7.4 . Z hodnot síly testu a grafu je zřejmé, že pro hodnoty blízké nulové hypotéze je velmi vysoká pravděpodobnost přijetí nulové hypotézy, i když ve skutečnosti bude platit alternativní hypotéza (např. jestliže ve skutečnosti je střední hodnota základního souboru 62, existuje zde asi 74 % pravděpodobnost přijetí nulové hypotézy, tj. předpokladu, že střední hodnota základního souboru je rovna hodnotě 60). Naopak pro hodnoty μ hodně vzdálené od nulové hypotézy (např. 68 a vyšší) je síla testu téměř „absolutní“, tj. blíží se 100 %. Pro tyto hodnoty se prakticky nemůže stát, že by v případě jejich platnosti byla přijata nulová hypotéza.

Abychom nemuseli počítat hodnoty β pro všechny možné hodnoty parametrů, je závislost síly testu na parametru zobecněna jako tzv. silofunkce. Obecný tvar silofunkce pro jednostranný a oboustranný test ukazuje obrázek 7.5 . Na ose x je vždy vyznačena skutečná hodnota parametru, na ose y odpovídající síla testu. Šipkou je vyznačena poloha parametru v nulové hypotéze. Minimální síla testu je právě v tomto bodě (je to vlastně hodnota α, v našem případě 0,05), nahoře je silofunkce ohraničena hodnotou 1. Vlevo silofunkce pro jednostranný test, vpravo pro oboustranný, kde musí být vyznačeny síly testu na obě strany od hodnoty parametru nulové hypotézy.

Síla testu závisí na následujících faktorech:

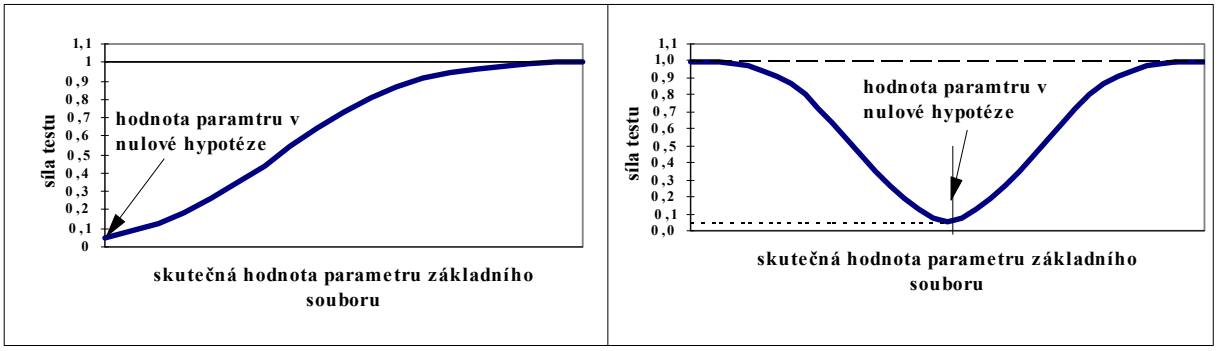
- odchylka mezi hodnotou testovaného parametru v nulové hypotéze a skutečnou hodnotou parametru – čím je tato odchylka větší tím je síla testu vyšší
- variabilitě (směrodatné odchylce nebo rozptylu) základního souboru – čím je variabilita menší, tím je vyšší síla testu
- na velikosti výběrového souboru – čím je výběrový soubor větší, tím je vyšší síla testu (máme více informací o základním souboru)
- na hladině významnosti α - čím je vyšší hladina významnosti (a tedy vzrůstá pravděpodobnost chyby I. druhu), tím je vyšší síla testu (protože klesá β). Tato situace je zřejmá z obrázku 7.3 – pokud bychom zvolili hladinu významnosti např. 0,1, potom se změní a posune doleva i kritická hranice C, čímž se zvětší i šedá plocha představující hodnotu α, a při nezměněném postavení obou hustot pravděpodobnosti bude proto plocha zaujatá chybou II. druhu (cihličkovaná) menší, čímž vzroste síla testu.

Hodnota parametru	Chyba II. druhu	Síla testu
61	0,8739	0,1261
62	0,7405	0,2595
63	0,5577	0,4423
64	0,3613	0,6387
65	0,1963	0,8037
66	0,0877	0,9123
67	0,0318	0,9682
68	0,0092	0,9908
69	0,0021	0,9979



Obrázek 7.4 – Síla testu pro různé hodnoty μ

Z tohoto přehledu plyne, že sílu testu můžeme ovlivnit velikostí výběru a stanovením hodnoty hladiny významnosti. (první dvě podmínky se ovlivnit nedají, jsou určeny základním souborem, „skutečností“). Např. síla testu pro hodnotu 62 v našem příkladu pro velikost výběru n = 100 je asi 0,26, tedy máme jen asi 26 % naději, že odmítneme nulovou hypotézu v případě, že skutečná hodnota parametru je 62. Jestliže zvýšíme velikost výběru na n = 400, potom (s přepočítáním nové hranice kritického oboru na 61,645) na základě vyřešení vztahu 7.2 získáme hodnotu $\beta = 0{,}3613$ a tedy sílu testu $1 - \beta = 0{,}6387$, tj. v po čtyřnásobném zvýšení velikosti výběru vzrostla naše naděje na zamítnutí nulové hypotézy, za předpokladu, že skutečná hodnota parametru je 62, z 26 % na asi 64 %.



Obrázek 7.5 – Znázornění silofunkce pro jednostranný (vlevo) a oboustranný test (vpravo)

Tuto kapitolu tedy můžeme uzavřít konstatováním, že u testování statistických hypotéz první správné rozhodnutí (přijetí správné nulové hypotézy) činíme s pravděpodobností $P=1-\alpha$, druhé správné rozhodnutí (správné odmítnutí neplatné nulové hypotézy) činíme s pravděpodobností $S = 1 - \beta$. Dá se říci, že „ideální“ test by měl mít co nejmenší α (a z toho vyplývající co největší P) a co nejmenší β (a tedy co největší S). Protože obě chyby spolu souvisí (pokud snižujeme α, zároveň zvyšujeme β), je nutné nalézt přijatelný kompromis – ten je dán pro většinu aplikací právě hodnotou $\alpha = 0{,}05$.

Tato kapitola měla za cíl pouze pochopení podstaty a významu obou chyb. Konkrétní postupy výpočtu chyby II. druhu pro jednotlivé typy testů vyžadují speciální výpočetní vztahy s využitím statistických tabulek, grafikonů a specializovaných počítačových programů.

7.4 Parametrické testy

Parametrické testy jsou založeny na určitých předpokladech. Zpravidla se předpokládá, že se jedná o nezávislý náhodný výběr z určitého rozdělení, které je specifikováno buď úplně (známe jeho typ i parametry) nebo neúplně (známe typ rozdělení, ne však parametry). U běžných typů testů pro střední hodnoty a charakteristiky variability se předpokládá zpravidla normální rozdělení. Ve většině případů jsou tyto předpoklady splněny. Pokud nejsou, doporučuje se používat neparametrických testů (viz kapitola 7.5) nebo modifikovaných parametrických testů.

7.4.1 Testy hypotéz o parametrech jednoho výběru

7.4.1.1 Hypotézy o rozptylech

H_0: $\sigma^2 = \sigma_0^2$

Výběrový soubor s rozptylem S^2 pochází ze základního souboru s rozdělením $N(\mu,\sigma^2)$, přičemž platí, že rozptyl základního souboru σ^2 se rovná známé konstantě σ_0^2.

Tyto testy se také nazývají testy přesnosti. Vyplývá to z toho, že se používají pro zjištění, zdali variabilita nějaké veličiny nekolísá více, než stanovuje příslušná norma nebo zda je měření dostatečně přesné (měření je tím přesnější, čím menší má variabilitu). Normované hodnoty (dané předpisem, normou, dohodou, apod.) jsou hodnoty σ_0^2, skutečně zjištěné hodnoty variability – rozptylu – se označují S^2.

a) pro $n \leq 30$ se používá statistika

$$\chi^2 = \frac{(n-1).S^2}{\sigma_0^2}. \tag{7.3}$$

Kritické obory uvádí tabulka 7.2 . Pokud testové kritérium padne do kritického oboru, zamítá se nulová hypotéza.

Nulová hypotéza	Alternativní hypotéza	Kritický obor
$\sigma^2 = \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$\chi^2 \geq \chi^2_{\alpha}$ (n - 1)
	$\sigma^2 < \sigma_0^2$	$\chi^2 < \chi^2_{(1-\alpha)}$ (n - 1)
	$\sigma^2 \neq \sigma_0^2$	$\chi^2 > \chi^2_{\alpha/2}$ (n - 1) nebo $\chi^2 < \chi^2_{(1-\alpha/2)}$ (n - 1)

Tabulka 7.2 – Kritické obory pro hypotézy o rozptylech pro jeden výběr do 30 prvků

b) pro $n > 30$ se používá statistika

$$Z = \sqrt{2\chi^2} - \sqrt{2f-1}, f = n-1 \tag{7.4}$$

Nulová hypotéza se přijímá, jestliže testové kritérium Z padne do intervalu $< -z_{\alpha}, z_{\alpha} >$ pro jednostranný test nebo do intervalu $< -z_{\alpha/2}, z_{\alpha/2} >$ pro oboustranný test.

Příklad 7.2 (podle GROFÍKA 1987):

Variabilita teploty vzduchu v laboratoři je stanovena směrodatnou odchylkou $\sigma_0 = 3°C$. Při kontrole laboratorních podmínek bylo provedeno 20 měření teploty a ze zjištěných údajů byla vypočtena výběrová směrodatná odchylka $S = 3,27$ °C. Je třeba posoudit, zda tato hodnota nesignalizuje zvýšení variability teploty vzduchu.

Vzhledem k formulaci úlohy testujeme nulovou hypotézu H_0: $\sigma^2 = \sigma_0^2$ oproti pravostranné alternativní hypotéze H_1: $\sigma^2 > \sigma_0^2$ (jde nám pouze o problém zvýšení variability, po-

kud bude variabilita teplot nižší, je to v pořádku). Použijeme testové kritérium 7.3 (protože n < 30)

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} = \frac{(20-1)\cdot 3{,}27^2}{3^2} = 22{,}57$$

Tuto hodnotu srovnáme s kritickou hodnotou $\chi^2_{0,05;19} = 30{,}14$. Protože testové kritérium je nižší než kritická hodnota, nezamítáme nulovou hypotézu a přijímáme závěr, že s pravděpodobností 95 % není prokázána zvýšená variabilita teploty vzduchu v laboratoři.

7.4.1.2 Hypotézy o průměrech

H_0: $\mu = \mu_0$

Výběrový soubor s rozsahem n, aritmetickým průměrem $\bar{x}$ *a rozptylem* S^2 *pochází ze základního souboru s rozdělením* $N(\mu_0, \sigma^2)$.

Tento typ testu se nazývá test správnosti. Použije se statistika

$$t = \frac{|\bar{x} - \mu_0|}{\frac{S}{\sqrt{n-1}}} = |\bar{x} - \mu_0| \cdot \frac{\sqrt{n-1}}{S} \qquad (7.5)$$

Kritické obory jsou uvedeny v tabulce 7.3 . Příklad na tento typ testu s podrobným vysvětlením je příklad 7.1.

Nulová hypotéza	Alternativní hypotéza	Kritický obor
$\mu = \mu_0$	$\mu > \mu_0$	$t \geq t_\alpha (n-1)$
	$\mu < \mu_0$	$t < -t_\alpha (n-1)$
	$\mu \neq \mu_0$	$\lvert t \rvert \geq t_{(\alpha/2)} (n-1)$

Tabulka 7.3 - Přehled kritických oborů pro hypotézy testů průměrů pro jeden výběr

7.4.1.3 Hypotézy o relativních četnostech

Relativní četnosti můžeme považovat za pravděpodobnosti, proto jde v podstatě o testování hypotéz o pravděpodobnostech.

H_0 : $p = p_0$

Pravděpodobnost nastoupení jevu v základním souboru p je rovna předpokládané (známé) pravděpodobnosti (p_0).

Je-li $n > 40$ a $p \in (0{,}3 ; 0{,}7)$, přičemž výběrový podíl označíme w, potom použijeme statistiku

$$Z = (w - p_0)\sqrt{\frac{n}{w.(1-w)}} \qquad (7.6)$$

která má (přibližně) N(0,1) rozdělení. Hypotézu přijímáme, když $Z \in < -z_{\alpha/2} ; z_{\alpha/2}>$ pro oboustrannou hypotézu nebo $Z \in < -z_{\alpha} ; z_{\alpha}>$ pro jednostrannou hypotézu.

Je-li $p \notin (0,3;\ 0,7)$ použijeme statistiku

$$Z = \left(2.\arcsin\sqrt{w} - 2.\arcsin\sqrt{p_0}\right).\sqrt{n} \qquad (7.7)$$

Kritické hodnoty jsou stejné jako v předchozím případě.

Příklad 7.3:

Při honu na zajíce bylo uloveno celkem 565 zajíců. Z toho počtu bylo 302 samců a 263 zaječek. Je správný závěr, že v honitbě není poměr pohlaví 1 : 1?

Poměru pohlaví 1: 1 odpovídá relativní četnost (pravděpodobnost) zajíce $p_0 = 0,50$. Z výsledku honu je $p = w_{(zajíc)} = 302 : 565 = 0,535$. Vyslovme hypotézu $H_0 : p = p_0$, požijeme statistiku 7.6 a hladinu významnosti $\alpha = 0,05$.

Vypočteme

$$Z = (0,535 - 0,500).\sqrt{565:(0,535.0,465)} = 1,668.$$

Tabulková hodnota $z_{0,05} = 1,96$. Protože $z = 1,668 < z_{0,05} = 1,96$, hypotézu o nevýznamné odchylce pravděpodobností přijímáme. Z výsledku honu neplyne, že v honitbě není poměr pohlaví zajíců 1 : 1.

7.4.1.4 Grubbsův test extrémních odchylek

V souboru pozorovaných nebo měřených hodnot se někdy objeví hodnota, která je nápadná svou extrémně odlišnou velikostí. V zájmu správného statistického zpracování je nutné posoudit, zda tato odchylka je náhodná nebo zda je zatížena hrubou chybou a do souboru proto nepatří. Pro použití Grubbsova testu se předpokládá normální rozdělení souboru hodnot.

Testujeme nulovou hypotézu

H_0: *Odchylka extrémní hodnoty je náhodná*

Je-li extrémní největší hodnota, volíme testovací kritérium tvaru

$$T_n = \frac{x_n - \overline{x}}{S} \qquad (7.8)$$

při extrémní nejmenší hodnotě volíme test ve tvaru

$$T_1 = \frac{\overline{x} - x_1}{S} \qquad (7.9)$$

V tabulce kritických hodnot $T_{n,\alpha} = T_{1,\alpha}$ pro Grubbsův test vyhledáme kritickou hodnotu. Nulová hypotéza je přijata, když $T_1 < T_{1,\alpha}$, resp. $T_n < T_{n.\alpha}$. Nepřijetí nulové hypotézy je podkladem pro možné vyloučení hodnoty ze souboru (ale přesto musíme důkladně zvážit, jakým způsobem tato hodnota vznikla – podrobněji v kapitole 4 při rozboru příkladu s extrémními hodnotami).

Příklad 7.4:

V souboru pěti měření 1,73; 1,86; 1,78; 2,14; 1,85 je podezřele odlišná hodnota 2,14. Zda je zatížena hrubou chybou, ověříme Grubbsovým testem.

$\alpha = 0{,}05$; $T_{5;\,0{,}05} = 1{,}869$; $\bar{x} = 1{,}872$; $S = 0{,}142$

$$T_5 = \frac{2{,}14 - 1{,}872}{0{,}142} = 1{,}887$$

Protože platí 1,887 > 1,869 nulovou hypotézu o náhodné odchylce zamítáme a hodnotu 2,14 považujeme za extrémní.

7.4.1.5 Testy normality

Jak již bylo několikrát zdůrazněno, je předpoklad, že výběr pochází ze základního souboru s normálním rozdělením, jedním z nejdůležitějších. Pokud není předpoklad normality splněn, značně to komplikuje následnou analýzu (nutnost transformace, použití kvantilových charakteristik, apod.). Proto je nutné vždy, kdy vznikne pochybnost o normalitě základního souboru, ze kterého byl proveden analyzovaný výběr, provést test normality.

Testů normality existuje celá řada a jsou založeny na různých principech a předpokladech (např. Shapiro-Wilkův test, Anderson-Darlingův test, D'Agostinův omnibus test apod.). Zde uvedeme jeden z nejjednodušších testů, který je založen na odděleném posuzovaní šikmosti a špičatosti. Tento typ testu má tu výhodu, že můžeme zjistit, která charakteristika tvaru (zda šikmost nebo špičatost) způsobuje odchylku od normality.

Testujeme nulovou hypotézu

H_0: *Výběr pochází ze základního souboru s normálním rozdělením.*

Testové kritérium má dvě části, část A_1 testuje šikmost, část E_1 testuje nesouměrnost:

$$A_1 = \frac{|A|}{\sqrt{\frac{6\cdot(n-2)}{(n+1)\cdot(n+3)}}} \tag{7.10}$$

$$E_1 = \frac{|E| - \frac{6}{n+1}}{\sqrt{\frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}}} \tag{7.11}$$

Nulovou hypotézu přijímáme, jestliže platí: A_1 a současně $E_1 \le z_{\alpha/2}$, kde $z_{\alpha/2}$ je kvantil normovaného normálního rozdělení N(0,1). Pokud alespoň jedno testové kritérium nevyhoví této nerovnosti, nulová hypotéza se zamítá.

Příklad 7.5:

Provedeme test normality pro zadání z kapitoly 4.5. Údaje, potřebné pro výpočet, jsou v následující tabulce:

Testovaný soubor	Výběrová šikmost	Výběrová špičatost	Počet prvků	Testové kritérium A_1	Testové kritérium E_1	Kritická hodnota $z_{0,025}$	Výsledek testu
Porost I	-0,05	-0,23	60	0,17	0,24	1,96	H_0 přijata
Porost II	2,42	4,65	42	6,87	7,12	1,96	H_0 zamítnuta

Test potvrdil závěry, ke kterým jsme dospěli již při hodnocení těchto souborů v kapitole 4.5, že soubor tlouštěk v Porostu I je normální, zatímco soubor v Porostu II nevykazuje normalitu vzhledem k extrémním hodnotám, které obsahuje.

7.4.1.6 Testy náhodnosti

Jedním ze zásadních požadavků na kvalitní výběr je vzájemná nezávislost jeho prvků. Pokud není tato podmínka splněna, je nutno prověřit celý proces plánování experimentu a sběru dat, neboť vzájemná závislost může být způsobena různými příčinami (např. změnou stavu měřícího zařízení, nekonstantností podmínek měření, nesprávně provedeným výběrem apod.). V každém případě, pokud je to technicky možné, je lepší získat nová data podle postupu upraveného na základě pečlivé analýzy možností vzniku závislosti v předchozím měření.

Pokud se uvedené faktory mění s časem, vzniká závislost mezi prvky výběru uspořádanými v časovém sledu. K identifikaci této závislosti se testuje významnost autokorelačního koeficientu. Autokorelace je obecně závislost mezi následujícími prvky a označuje se řády (např. autokorelace I. řádu je závislost předchozího a následujícího prvku). K identifikaci časové závislosti prvků výběru nebo závislostí související s pořadím jednotlivých měření se autokorelační koeficient I. řádu testuje pomocí testovacího kritéria (MELOUN - MILITKÝ 1994)

$$t_n = \frac{T_1\sqrt{n+1}}{\sqrt{1-T_1}} \tag{7.12}$$

kde

$$T_1 = \left(1-\frac{T}{2}\right)\sqrt{\frac{n^2-1}{n^2-4}} \tag{7.13}$$

kde T je von Neumannův poměr

$$T = \frac{\sum_{i=1}^{n-1}(x_{i+1}-x_i)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \tag{7.14}$$

Nulová hypotéza říká, že autokorelační koeficient I. řádu je roven nule, tj. že za sebou jdoucí prvky výběru jsou nezávislé. Pokud platí, že

$$|t_n| > t_{\alpha/2,n+1}$$

potom je nutno hypotézu o nezávislosti prvků na hladině významnosti α zamítnout. $t_{\alpha/2,n+1}$ je kritická hodnota Studentova rozdělení pro (n + 1) stupňů volnosti.

Existují ještě další parametrické i neparametrické testy nezávislosti prvků (v části věnované neparametrickým testů bude ještě uveden test bodů zvratu). Zde uvedený test patří mezi nejpoužívanější, protože autokorelace I. řádu bývá zpravidla nejčastější a nejsilnější.

Příklad 7.6
Provedeme test nezávislosti prvků z kapitoly 4.5 (Porost I).

Test provedeme pomocí von Neumanova kritéria 7.12, které bylo vyčísleno na hodnotu t_n =0,854. Kritická hodnota $t_{0,05;61}$ = 2, nulovou hypotézu přijímáme, tedy předpokládáme, že prvky studovaného výběru jsou nezávislé.

7.4.2 Testy hypotéz o parametrech dvou výběrů

Tyto testy patří ve statistické praxi k nejpoužívanějším. Vzhledem k jejich důležitosti zde uvedeme také různé modifikace, které se používají při nesplnění podmínek daných nulovou hypotézou (týká se to především podmínky normality). Vzhledem k tomu, že parametrické testy (zvláště F-test) jsou citlivé na nesplnění podmínek, je vhodné před jejich provedením vyšetřit výběry pomocí postupů exploratorní analýzy dat.

7.4.2.1 Hypotézy o rozptylech

Tyto testy se také nazývají testy homogenity rozptylů. Používají se jednak samostatně, jednak jako první krok v testech středních hodnot. Testuje se hypotéza

H_0: $\sigma_1^2 = \sigma_2^2$

Dva výběrové soubory s rozsahy n_1 a n_2 a s výběrovými rozptyly S_1^2 a S_2^2 pocházejí ze dvou základních souborů s normálním rozdělením s rozptyly σ_1^2 a σ_2^2, přičemž platí, že $\sigma_1^2 = \sigma_2^2$.

Alternativní hypotéza H_1 je buď $\sigma_1^2 > \sigma_2^2$ nebo $\sigma_1^2 \neq \sigma_2^2$. Běžně se používá **F - test** využívající testového kritéria

$$F = \left(\frac{\hat{s}_x^2}{\hat{s}_y^2} \right), \hat{s}_x^2 \geq \hat{s}_y^2 \qquad (7.15)$$

kde $\hat{s}_x^2, \hat{s}_y^2$ jsou bodové odhady rozptylů základního souboru (vypočítané podle vztahu 6.7).

Kritické obory pro obě varianty H_1 uvádí tabulka

F-test je značně citlivý na předpoklad normality (MELOUN - MILITKÝ 1994). V případě, že tento předpoklad není splněn (nutno ověřit testy normality nebo pomocí exploratorních grafů), je nutné F-test modifikovat nebo použít test jiný.

Nulová hypotéza	Alternativní hypotéza	Kritický obor
$\sigma_1^2 = \sigma_2^2$.	$\sigma_1^2 \neq \sigma_2^2$	$F > F_{\alpha/2}(n_1 - 1, n_2 - 1)$
	$\sigma_1^2 > \sigma_2^2$	$F > F_{\alpha}(n_1 - 1, n_2 - 1)$

Tabulka 7.4 – Kritické obory pro F-test

Jestliže mají zkoumané výběry jinou špičatost než odpovídá normálnímu rozdělení, je nutno užít kvantil $F_{\alpha/2}(\nu_1, \nu_2)$, kde se stupně volnosti ν_1, ν_2 vypočítají podle vztahů

$$\nu_1 = \frac{n_1 - 1}{1 + \frac{g_{2c}}{2}} \text{ a } \nu_2 = \frac{n_2 - 1}{1 + \frac{g_{2c}}{2}}, \text{ kde } g_{2c} = \frac{2(n_1 + n_2)\left[\sum_{i=1}^{n_1}(x_i - \bar{x})^4 + \sum_{i=1}^{n_2}(y_i - \bar{y})^4\right]}{\left[\sum_{i=1}^{n_1}(x_i - \bar{x})^2 + \sum_{i=1}^{n_2}(y_i - \bar{y})^2\right]} - 3$$

Při testování nenormálních výběrů je možné použít Jackknife test. Podrobnosti k jeho výpočtu lze nalézt např. v MELOUN - MILITKÝ 1994 Vzhledem k velké výpočetní náročnosti se tento test běžně nepoužívá, k jeho výpočtu se doporučuje použít vhodný program.

7.4.2.2 Hypotézy o shodě středních hodnot nezávislých výběrů

Tyto testy (zvané také **Studentovy t-testy**) jsou z hlediska nulové hypotézy obdobou F-testů, ale místo rozptylu se zde testuje shoda středních hodnot, zpravidla aritmetických průměrů. Také zde se předpokládá, že základní soubory mají normální rozdělení. Obecně platí, že t-testy jsou robustnější než F-testy, tj. jejich výsledek je méně ovlivněn mírným nesplněním základních podmínek, především normality. Přesto byly pro výběry silněji odchylné od normality vyvinuty speciální modifikace.

Před použitím vlastního t-testu je nutné provést testování homogenity rozptylů pomocí vhodného testu (viz kap.7.4.2.1).

Nezamítneme-li hypotézu o homogenitě rozptylů ($\sigma_1^2 = \sigma_2^2$), potom použijeme testové kritérium

$$T_1 = \frac{|\bar{x} - \bar{y}|}{\sqrt{(n_1 - 1)s_x^2 + (n_2 - 1)s_y^2}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}} \tag{7.16}$$

Tabulka 7.5 uvádí kritické obory pro jednotlivé alternativní hypotézy. Používají se kvantily Studentova t-rozdělení pro ($n_1 + n_2 - 2$) stupňů volnosti.

Nulová hypotéza	Alternativní hypotéza	Kritický obor
	$\mu > \mu_0$	$t \geq t_\alpha (n_1 + n_2 - 2)$
$\mu=\mu_0$	$\mu < \mu_0$	$t < t_\alpha (n_1 + n_2 - 2)$
	$\mu \neq \mu_0$	$\lvert t \rvert \geq t_{\alpha/2} (n_1 + n_2 - 2)$

Tabulka 7.5 - Kritické obory pro t-test

Zamítneme-li hypotézu o homogenitě rozptylů ($\sigma_1^2 \neq \sigma_2^2$), potom použijeme testové kritérium

$$T_2 = \frac{(\bar{x} - \bar{y})}{\sqrt{\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}}} \tag{7.17}$$

Počet **ekvivalentních stupňů volnosti** ν se vypočítá podle vzorce

$$\nu = \frac{\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}}{\frac{s_x^4}{n_1^2(n_1-1)} + \frac{s_y^4}{n_2^2(n_2-1)}} \qquad (7.18)$$

Kritické obory jsou stejné (podle tabulky 7.5) pro ν stupňů volnosti.

Pro správné provedení t-testu tedy potřebujeme znát kromě středních hodnot i rozptyly obou výběrů. Bylo však zjištěno, že při stejných rozsazích obou výběrů je možné použít testové kritérium T_1 i při nesplnění podmínky homogenity rozptylů. Pokud se rozsahy výběrů výrazně liší, potom lze tento postup použít jen omezeně.

Při nesplnění podmínky normality je nutné rozlišovat mezi nenormalitou způsobenou šikmostí a špičatostí. Vůči odchylkám od normality ve špičatosti jsou testová kritéria T_1 i T_2 dostatečně robustní.

Pokud zjistíme odchylku od normality vlivem šikmosti výběrového rozdělení, potom se doporučuje použít upraveného testového kritéria. Výpočet je poměrně složitý.

Pokud výběry obsahují vybočující měření, která nelze vyloučit, existují speciální modifikace t-testu založené na tzv. uřezaných průměrech a winsorizovaném rozptylu.

Podrobnější informace o těchto speciálních úpravách testového kritéria t – testu lze nalézt např. v MELOUN - MILITKÝ1994, DRÁPELA-ZACH 1996.

Příklad 7.7:

Na dvou plochách byly zjišťovány výčetní tloušťky modřínu. Na první ploše bylo změřeno 90 stromů a zjištěna průměrná tloušťka $\bar{d}_1 = 40{,}8$ *cm a rozptyl* $S_1^2 = 29{,}5$ *cm*^{2}. *Na druhé ploše bylo změřeno 50 stromů a zjištěna průměrná tloušťka* $\bar{d}_2 = 36{,}0$ *cm a rozptyl* $S_2^2 = 22{,}8$ *cm*^{2}. *Je třeba určit, zda lze modřínové porosty na obou plochách pokládat za stejně tloušťkově vyspělé.*

Odpověď na otázku bude kladná, když rozdíly mezi $\bar{d}_1$ a $\bar{d}_2$ budou statisticky nevýznamné. Volíme tedy $H_0 : \mu_1 = \mu_2$, a $\alpha = 0{,}05$. Abychom zvolili správný test, musíme nejdříve ověřit hypotézu $H_0 : \sigma_1^2 = \sigma_2^2$. K ověření použijeme statistiku 7.15

$$F = \frac{\hat{S}_1^2}{\hat{S}_2^2} = \left(S_1^2 \cdot \frac{n_1}{n_1 - 1} \right) : \left(S_2^2 \cdot \frac{n_2}{n_1 - 1} \right) = \left(29{,}5 \cdot \frac{90}{89} \right) : \left(22{,}8 \cdot \frac{50}{49} \right) = 1{,}28.$$

Protože $F_{0,05;89,49} = 1{,}55$, hypotézu o homogenitě rozptylů přijímáme.

Nulovou hypotézu o průměrech tedy ověříme pomocí testového kritéria 7.16

$$t = \frac{|40{,}8 - 36{,}0|}{\sqrt{90.29{,}5 + 50.22{,}8}} \cdot \sqrt{\frac{90.50.(90+50-2)}{90+50}} = 5{,}19$$

$t_{0,05;138} = 1{,}98.$

Protože 5,19 > 1,98, hypotézu o rovnosti aritmetických průměrů zamítáme. Porosty nejsou stejně tloušťkově vyspělé. Porost na ploše 1 je „tlustší“.

7.4.2.3 Hypotézy o shodě středních hodnot závislých výběrů

Dosud jsme se zabývali případy, kdy všechny výběry byly mezi sebou nezávislé, tj. hodnoty jednoho výběru nebyly v žádném vztahu k hodnotám jiných výběrů. Zvláštním případem t - testů je **testování závislých výběrů.** V tomto případě se používá testové kritérium, které zohledňuje těsnost závislosti obou výběrů (ŠMELKO 1991):

$$T_3 = \frac{|\bar{x} - \bar{y}|}{\sqrt{\left(\frac{s_x^2}{n_1 - 1}\right) + \left(\frac{s_y^2}{n_2 - 1}\right) - 2 \cdot r_{xy} \cdot \left(\frac{s_x}{\sqrt{n_1 - 1}}\right) + \left(\frac{s_y}{\sqrt{n_2 - 1}}\right)}} \qquad (7.19)$$

kde

r_{xy} je korelační koeficient - míra závislosti mezi výběry X a Y

Kritický obor se v případě homogenity rozptylů určí podle tabulky 7.5 , pro případ nehomogenity rozptylů se počet stupňů volnosti ν určí podle vzorce

$$\nu = \frac{t_{\alpha,n_1-1} \cdot \frac{s_x^2}{n_1 - 1} + t_{\alpha,n_2-1} \cdot \frac{s_y^2}{n_2 - 1}}{\frac{s_x^2}{n_1 - 1} \cdot \frac{s_y^2}{n_2 - 1}} \qquad (7.20)$$

Příklad 7.8 *(ŠMELKO 1991):*

Při stanovení nejvhodnější metodiky pro odběr vývrtů pro určení tloušťkového přírůstu byla také zkoumána otázka tvorby tloušťkového přírůstu na stromech v závislosti na svahu terénu. Na 301 buku byl změřen 5-letý tloušťkový přírůst po svahu (X_1) a proti svahu(X_2). Byly získány následující výběrové charakteristiky:

	Počet stromů	*Aritmetický průměr (cm)*	*Směrodatná odchylka (cm^2)*
tloušťkový přírůst po svahu	*301*	*1.898*	*1.100*
tloušťkový přírůst proti svahu	*301*	*1.807*	*1.030*

Úkolem je rozhodnout, zda je mezi X_1 a X_2 statisticky významný rozdíl, tj. zda je možné předpokládat, že ukládání tloušťkového přírůstu je ovlivněno svahem terénu.

Byla stanovena H_0: Rozdíl mezi tloušťkovým přírůstem po svahu a proti svahu je náhodný, tj. $\mu_1 = \mu_2$, proti H_1: Rozdíl mezi tloušťkovým přírůstem po svahu a proti svahu je významný, tj. $\mu_1 \neq \mu_2$.

Vzhledem k charakteru problému byla při řešení předpokládána závislost mezi X_1 a X_2, což vypočítaný korelační koeficient $R_{x1,x2} = 0.82$ potvrdil (testování prokázalo jeho významnost). Hladina významnosti $\alpha = 0.05$. Proto bylo použito testové kritérium T_3 podle vzorce 7.19:

$$T_3 = \frac{1.898 - 1.807}{\sqrt{\frac{1.1^2}{301-1} + \frac{1.03^2}{301-1} - 2 \cdot 0.84 \cdot \frac{1.1}{\sqrt{301-1}} \cdot \frac{1.03}{\sqrt{301-1}}}} = 2.64$$

Vzhledem k tomu, že F-test prokázal homogenitu rozptylu, výsledek testu určíme porovnáním hodnoty T_3 s kritickou hodnotou $T_{0.025,600} = 1.96$. Protože platí, že 2.64 > 1.96, nulovou hypotézu zamítáme a můžeme konstatovat, že u buků rostoucích na svahu se tloušťkový přírůst po svahu a proti stahu významně liší.

Pokud bychom použili t-test pro nezávislé výběry podle vztahu 7.16, získali bychom hodnotu $T_1 = 1.05$ a tedy dospěli k opačnému výsledku testu, kdy bychom H_0 přijali.

7.4.2.4 Párový t - test

Zvláštním druhem závislých výběrů je ten případ, kdy hodnoty sledovaného znaku tvoří páry, tj. **byly získány dvěma způsoby na stejném jedinci**. Symbolicky:

A značí jeden typ realizace (např. měření jedním typem přístroje, první metodou apod.,)

B značí druhý typ realizace. (např. měření druhým typem přístroje, jinou metodou apod.)

Vytvoříme soubor dvojic (x_{A1}, x_{B1}), (x_{A2}, x_{B2}),..... Zkonstruujeme náhodnou proměnnou $D = X_A - X_B$, tedy rozdíly mezi měřenými hodnotami na každém jedinci. Vzhledem k tomu, že v případě obou realizací (tj. způsobů měření nebo různých měřících přístrojů) jde o stejnou veličinu, je možné předpokládat, že obě realizace by měly dospět ke stejnému výsledku, tedy že hodnota D = 0. U proměnné D předpokládáme normální rozdělení $N(0,\sigma^2)$. Testujeme hypotézu $H_0 : \mu_D = 0$ proti $H_1 : \mu_D \neq 0$.

Použijeme statistiku

$$T = |\overline{D} - 0| \cdot \frac{\sqrt{n-1}}{S_D} \qquad (7.21)$$

Hypotézu přijímáme, když $t < t_{\alpha/2;\, n-1}$

Příklad 7.9:

Byla změřena výška 10 stromů vždy dvěma výškoměry. Máme ověřit předpoklad, že oba přístroje měří stejně, tj. rozdíly v naměřených hodnotách jsou náhodné.

h_A	21,8	18,3	25,6	20,1	19,8	16,1	12,7	18,2	12,3	18,3
h_B	22,4	18,7	25,3	20,3	19,6	16,2	12,5	18,4	12,5	19,0
D	-0,6	-0,4	0,3	-0,2	0,2	-0,1	0,2	-0,2	-0,2	-0,7

Je n = 10 , $\overline{D}$ = -0,17, S_d = 0,3195. Podle statistiky 7.21 vypočítáme

T = 0,17 . (3 : 0,3195) = 1,596.

Tabulková hodnota $t_{0,025;\, 9} = 2,262$. Protože 1,596 < 2,262, nulovou hypotézu přijímáme. Oba přístroje měří stejně.

7.4.2.5 Hypotézy o relativních četnostech

H_0: $p_1 = p_2$

Pravděpodobnosti dané relativními četnostmi w_1, w_2 dvou nezávislých výběrů o rozsazích n_1, n_2 jsou si rovny.

Je-li $p \in (0,3;\ 0,7)$, volíme statistiku

$$Z = (w_1 - w_2) : \sqrt{\frac{w_1(1-w_1)}{n_1} + \frac{w_2(1-w_2)}{n_2}} \qquad (7.22)$$

která má N(0,1) rozdělení. Hypotézu přijímáme, když $Z \in < -z_{\alpha/2} ; z_{\alpha/2} >$ pro oboustrannou hypotézu nebo $Z \in < -z_{\alpha} ; z_{\alpha} >$ pro jednostrannou hypotézu.

Je-li $p \notin (0,3;\ 0,7)$, volíme statistiku

$$Z = \left(2.\arcsin\sqrt{w_1} - 2.\arcsin\sqrt{w_2}\right) : \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \qquad (7.23)$$

se stejnými kritickými hodnotami jako v předchozím případě.

Příklad 7.10:

Ve dvou porostních skupinách smrku bylo sledováno napadení stromů václavkou. Ve skupině A bylo z 256 prověřovaných stromů napadeno 151, ve skupině B ze 195 stromů napadeno 111 stromů. Liší se porosty v pravděpodobnosti napadení?

Prověříme hypotézu $H_0 : p_A = p_B$. Protože $n > 40$, $w_A = 151/256 = 0,590$, $w_B = 111/195 = 0,569$, tj. $w_A, w_B \in < 0,3;\ 0,7 >$, tedy použijeme statistiku 7.22. Volíme $\alpha = 0,05$.

Vypočteme

$$z = (0,590 - 0,569) : \sqrt{\frac{0,590.0,410}{256} + \frac{0,569.0,431}{195}} = 0,447$$

Protože $z_{0,025} = 1,96$ je $z < z_{0,025}$, hypotézu přijímáme.

V obou porostech je pravděpodobnost napadení václavkou stejná.

Příklad 7.11:

Při kontrole dvou zásilek výrobků bylo shledáno, že v zásilce ze závodu A bylo z 250 výrobků 28 mimo povolenou toleranci, v zásilce ze závodu B z 400 výrobků 35 mimo toleranci. Liší se výroba závodů A,B v ukazateli procenta nevyhovujících výrobků?

Vystavme hypotézu $H_0 : w_A = w_B$ proti $H_1 : w_A \neq w_B$

$$w_A = \frac{28}{250} = 0,112 \notin (0,3;0,7),$$

Použijeme statistiku 7.23

$$Z = \left(2.\arcsin\sqrt{0,112} - 2.\arcsin\sqrt{0,0875}\right) : \sqrt{\frac{1}{250} + \frac{1}{400}} = 0,0063$$

přičemž kritická hodnota je $z_{0,025} = 1,96$. Znamená to, že přijímáme závěr, že výroba závodů A a B se neliší z hlediska procenta nevyhovujících výrobků.

7.4.3 Testy hypotéz o parametrech více výběrů

V testování hypotéz se pojmem více výběrů myslí „více než dva“, tj. následující testy budou platné pro porovnávání 3 a více výběrů. Je nutné zdůraznit, že pro testování tří a více výběrů je nepřípustné používat opakované testy pro dva výběry. Jestliže chceme např. porovnávat tři výběrové průměry a testovat jejich shodnost, není možné použít postupně t-testu pro porovnání prvního s druhým, druhého se třetím a prvního se třetím. Důvodem k zamítnutí nulové hypotézy o shodě je její nepřijetí v alespoň jedné kombinaci výběrů. Z toho plyne hlavní důvod nepřípustnosti tohoto postupu - zvyšování pravděpodobnosti chyby I. druhu. T-test je konstruován tak, že pravděpodobnost chyby I. druhu (hladina významnosti α) je nastavena např. na 5 % pro dva výběry. Je dokázáno (viz např. ZAR 1984), že pokud bychom použili výše uvedenou kombinaci dvouvýběrových t-testů pro testování shody tří průměrů, vzroste pravděpodobnost chyby I. druhu na 13%, pokud bychom tento postup aplikovali na 10 výběrů, byla by pravděpodobnost chyby I. druhu již 63 % atd. Znamená to, že pro větší počet výběrů bychom téměř jistě zamítli nulovou hypotézu o rovnosti průměrů jednotlivých výběrů, i když by ve skutečnosti pocházely ze stejného základního souboru. Je proto nutné zdůraznit, že pro testování více výběrů je nutné použít speciální testy. V této kapitole budou probrány podrobně testy pro rozptyly, odpovídající testy pro střední hodnoty (tzv. analýza rozptylu) budou pro svou náročnost a rozsah probrány ve zvláštní kapitole ve II. díle skript.

7.4.3.1 Testy hypotéz o rozptylech

7.4.3.1.1 Pro více výběrů stejného rozsahu

H_0: $\sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$

Více než dva (obecně k) výběrových souborů s rozsahem n a rozptyly $S_1^2, S_2^2, ...S_k^2$ pochází ze základních souborů s rozptyly $\sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$ proti alternativě, že alespoň mezi dvěma výběry existuje statisticky průkazný rozdíl rozptylů.“

Pro testování této hypotézy se používá Cochranův test, který má testové kritérium:

$$Q = \frac{\max\left(S_1^2, S_2^2, ...S_k^2\right)}{\sum_{i=1}^{k} S_i^2} \qquad (7.24)$$

kde max(.) znamená maximální hodnotu výrazu v závorce, v našem případě je to tesy nejvyšší hodnota výběrového rozptylu ze všech porovnávaných.

Kritické hodnoty rozdělení Q pro α, k, n-1 jsou tabelovány. Je-li vypočítaná hodnota Q menší než tabelovaná hodnota $q_{\alpha;\, k,\, n-1}$, nulovou hypotézu nezamítáme.

7.4.3.1.2 Pro více výběrů různého rozsahu

H_0: $\sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$

Více než dva (obecně k) výběrových souborů s rozsahy n_1, n_2, ..., n_k a rozptyly $S_1^2, S_2^2, ...S_k^2$ pochází ze základních souborů s rozptyly $\sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$ proti alternativě, že alespoň mezi dvěma výběry existuje statisticky průkazný rozdíl rozptylů."

Jestliže mají výběry různý rozsah, použije se Bartlettův test s testovým kritériem

$$B = \frac{\ln 10}{C}\left[(n-k)\log S^2 - \sum_{i=1}^{k}(n_i - 1)\log S_i^2\right] \qquad (7.25)$$

kde $n = \sum_{i=1}^{k} n_i$, $S^2 = \frac{1}{n-k}\sum_{i=1}^{k}(n_i - 1)S_i^2$, $C = 1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{n_i - 1} - \frac{1}{n-k}\right)$.

Statistika B má za předpokladu platnosti nulové hypotézy přibližně χ^2 rozdělení, když $n_i \geq 6$ pro i = 1, ...,k. Podmínka pro n_i musí být splněna, abychom test mohli použít. Při výpočtu můžeme nejdříve vypočítat hodnotu B pro C = 1. Je-li vypočítaná hodnota B menší než kritická hodnota, hypotézu přijímáme. Je-li B větší než kritická hodnota, počítáme ji znovu pro vypočítanou hodnotu C. Kritická hodnota pro statistiku B je $\chi^2_{\alpha,k-1}$.

Příklad 7.12:

Při výběru výškoměru k měření výšek stromů na výzkumné ploše byla pěti přístroji měřena výška stejného stromu vždy pětkrát. Z naměřených hodnot byly vypočteny výběrové rozptyly $S_1^2 = 0{,}11m^2$, $S_2^2 = 0{,}18m^2$, $S_3^2 = 0{,}39m^2$, $S_4^2 = 0{,}25m^2$, $S_5^2 = 0{,}32m^2$. Je možné pokládat všechny přístroje za stejně přesné?

Rozptyly jsou mírou přesnosti měření. Otázce odpovídá hypotéza o statistické nevýznamnosti rozdílů rozptylů. O souborech měření můžeme předpokládat jejich normální rozdělení. K řešení můžeme použít Cochranův test, protože rozsahy výběrů jsou stejné.

$Q = \frac{0{,}39}{1{,}25} = 0{,}312$, z tabulek $q_{0{,}05;\,5{,}4} = 0{,}544$.

Protože platí, že 0,312 < 0,544, nulovou hypotézu nezamítáme. Rozdíly mezi rozptyly pokládáme za náhodné a přístroje za stejně přesné.

Příklad 7.13:

	Záhon 1	Záhon 2	Záhon 3	Záhon 4	Záhon 5
Výška semenáčků v cm	4,2	6,7	7,9	9,0	9,8
	6,9	6,7	6,4	7,0	9,6
	8,0	5,5	8,1	7,9	9,1
	4,5	4,2	7,0	8,8	6,6
	7,5	4,7	8,0	9,6	7,7
	6,5	5,3	9,7	10,2	8,1
	6,5	4,4	5,2	8,2	8,4
		8,5		7,5	5,8
				6,2	
počet	7	8	7	9	8
S_i^2	1,774	1,870	1,765	1,438	1,740

Při výzkumu nejvhodnějšího způsobu zastínění na výškový růst semenáčků byly na 5 srovnávaných záhonech (na každém byl použit jiný způsob zastínění) měřeny výšky semenáčků (údaje viz v tabulce vlevo). Tato úloha se řeší analýzou rozptylu. Pro její použití musíme nejprve určit, zda můžeme předpokládat homogenitu rozptylu v celém souboru. Zjistěte Bartlettovým testem, zda tento předpoklad platí.

Nejprve vypočítáme hodnoty n = 39, S^2 = 2,637 a dosadíme do vztahu 7.25 (zatím nepočítáme hodnotu C, uvažujeme C = 1). Výsledkem je hodnota testového kritéria B = 3,32. Tato hodnota je menší než kritická hodnota $\chi^2_{0,05;4}$ = 9,49, takže nemusíme již počítat C a můžeme konstatovat, že všechny výběry pochází ze základních souborů se stejnými rozptyly.

7.4.3.2 Testy hypotéz o střední hodnotě

Tuto problematiku řeší **analýza rozptylu**. Problematika této metodologie je velmi rozsáhlá a složitá, proto bude probrána podrobně v II. díle tohoto učebního textu v samostatné kapitole.

7.5 Neparametrické testy

Neparametrické testy se používají zpravidla v těchto případech:

- pro velmi malé výběry,
- pro výběry pocházející z výrazně nenormálních základních souborů,
- pro výběry pocházející ze základních souborů, jejichž rozdělení je neznámé.

Neparametrické testy jsou nezávislé na tvaru rozdělení základního souboru. Vycházejí z velmi obecných předpokladů, zpravidla se pouze požaduje, aby rozdělení náhodné veličiny bylo spojité.

Mezi hlavní výhody neparametrických testů patří:

- nezávislost na tvaru rozdělení
- v mnoha případech jsou použitelné jak pro kvantitativní, tak i pro kvalitativní znaky
- zpravidla mají jednoduchý výpočet

Hlavní nevýhodou neparametrických testů patří menší síla testů. K dosažení stejné síly testu oproti parametrickému testu je zapotřebí větší rozsah výběru.

V následujícím textu si uvedeme nejběžnější neparametrické testy. Jsou to většinou neparametrické obdoby parametrických testů střední hodnoty, které se nejčastěji používají..

7.5.1 Znaménkový test pro jeden výběr

Testujeme hypotézu

H_0: *Výběrový soubor o rozsahu n pochází ze základního souboru se spojitým rozdělením, jehož medián $x_{0.50}$ se rovná známé hodnotě C.*

Medián zde používáme jako robustní charakteristiku polohy, protože neznáme typ rozdělení základního souboru. Je to vlastně neparametrická obdoba testu střední hodnoty pro jeden výběr.

Vycházíme z počtu kladných odchylek prvků výběrového souboru od výběrového mediánu $Z = x_i - C$ (tedy „+“ rozdíly - proto znaménkový test). Za předpokladu platnosti H_0 má veličina Z binomické rozdělení se střední hodnotou $E(Z) = n/2$ a rozptylem $D^2(Z) = n/4$.Testové kritérium

$$U = \frac{2 \cdot Z - n}{\sqrt{n}} \tag{7.26}$$

má v případě platnosti normální aproximace normované normální rozdělení N (0,1).
Hypotézu H_0 přijímáme, jestliže platí $z_{\alpha/2} < U < z_{1-\alpha/2}$.

Příklad 7.14:

Při zkouškách přesnosti nového typu výškoměru byla několikrát změřena předem známá výška (20 m). Při celkem sedmi měřeních byly dosaženy tyto výsledky:

Číslo měření	1	2	3	4	5	6	7
Naměřená výška (m)	20.5	19.4	18.9	21.1	20.6	20.8	20.2

Určete, zda posuzovaný výškoměr měří přesně.

Odpověď na danou otázku bude kladná, jestliže rozdíl stanovené výšky (20) a střední hodnoty bude náhodný (což stanovíme jako H_0). Vzhledem k malému počtu prvků výběru použijeme neparametrický znaménkový test. Z = 5, C = 20, $\alpha = 0.05$, potom testová statistika je

$$U = \frac{2 \cdot 5 - 7}{\sqrt{7}} = 1.13$$

Kritická hodnota $U_{0.975} = 1.96$, tedy H_0 nezamítáme. Výškoměr můžeme považovat za přesný.

7.5.2 Wilcoxonův test po dva výběry

Je to neparametrická obdoba t-testu. Vycházíme z předpokladu, že dva výběry o rozsahu n_1, resp. n_2 (kdy $n_1 \le n_2$) se spojitým rozdělením mají distribuční funkce $F_1(x)$ a $F_2(x)$. Testujeme hypotézu H_0 o rovnosti těchto distribučních funkcí $F_1(x) = F_2(x)$ oproti H_1 posunutí v poloze $F_1(x) = F_2(x - \Delta)$ pro všechna reálná x a $\Delta > 0$.

Vytvoříme sdružený výběr, tj. oba výběry spojíme do jednoho o rozsahu

$$N = n_1 + n_2.$$

Poté jednotlivým hodnotám přiřadíme pořadová čísla (tj. vzestupné uspořádání 1,2, ..., N). Pokud je několik hodnot stejných, dostávají průměrnou hodnotu svých pořadí (např. hodnoty 6, 3, 3, 2, 8 seřadíme podle velikosti 2, 3, 3, 6, 8 s pořadovými čísly 1,2.5, 2.5, 4, 5. Obě stejné hodnoty - 3 - měly obdržet pořadová čísla 2 a 3, dostaly tedy průměrnou hodnotu - 2.5). Pořadová čísla prvního výběru dostanou označení $R_{x1}, R_{x2}, ..., R_{x,n_1}$, druhého výběru $R_{y1}, R_{y2}, ..., R_{y,n_2}$. Vypočítáme součty

$$T_x = R_{x1} + R_{x2} + ... + R_{x,n_1}$$

$$T_y = R_{y1} + R_{y2} + ... + R_{y,n_2}.$$

Poté vypočítáme

$$U_x = n_1 n_2 + \frac{n_1(n_1+1)}{2} - T_x \qquad (7.27)$$

$$U_y = n_1 n_2 + \frac{n_2(n_2+1)}{2} - T_y \qquad (7.28)$$

Použijeme testové kritérium

$$U = \min (U_x, U_y)$$

Je-li hodnota U menší nebo rovna kritické hodnotě U_α (nalezneme ve speciálních tabulkách na základě α a rozsahu obou výběrů), zamítáme H_0.

Při velkých rozsazích výběrů, které nejsou tabelovány, se vypočítá testové kritérium

$$U_0 = \frac{U_x - \frac{1}{2} n_1 n_2}{\sqrt{\frac{n_1 n_2}{12}(n_1 + n_2 + 1)}} \tag{7.29}$$

které má asymptoticky normální rozdělení N(0,1). Jestliže platí, že $|U_0| > u_\alpha$, zamítáme H_0.

Příklad 7.15 (upraveno podle GROFÍKA 1987)*:*

Při výzkumu dvou proveniencí smrku byl mimo jiné zkoumám výškový růst. Následující tabulka uvádí výšky v cm pro obě provenience, které byly vysazeny na 5, resp. 6 plochách ve srovnatelných podmínkách.

Číslo plochy	1	2	3	4	5	6
Provenience I	340	440	310	358	401	
Provenience II	350	315	405	339	374	380

Posuďte, zda jsou mezi oběma proveniencemi významné rozdíly ve výškovém růstu.

Jde o posouzení stejného problému jako u t-testu, tedy zkoumáme, zda rozdíl mezi středními hodnotami obou výběrů je statisticky významný nebo není.
Jako H_0 zvolíme tvrzení o jejich rovnosti. Vzhledem k malému rozsahu obou výběrů a neznalosti jejich rozdělení provedeme neparametrický Wilcoxonův test.
Zvolíme hladinu významnosti $\alpha = 0.05$. Jako první krok vytvoříme z obou výběrových souborů sdružený výběr, který uspořádáme podle velikosti a jednotlivým hodnotám přiřadíme pořadí. Výsledek je v následující tabulce:

Provenience I (x_i)	310			340		358			401		440
Provenience II (y_i)		315	339		350		374	380		405	
Pořadí R_x	1			4		6			9		11
Pořadí R_y		2	3		5		7	8		10	

Dále vypočítáme pomocné výpočty a testové kritérium U:

$T_x = 1 + 4 + 6 + 9 + 11 = 31 \quad T_y = 2 + 3 + 5 + 7 + 8 + 10 = 35$

$U_x = 5.6 + \frac{5 \cdot 6}{2} - 31 = 14 \qquad U_y = 5.6 + \frac{6 \cdot 7}{2} - 35 = 16$

Testové kritérium stanovíme jako U = min(14,16) = 14. V tabulce Wilcoxonovy statistiky najdeme $U_{0.05;5,6} = 3$. Protože 14 > 3, nezamítáme H_0 a přijímáme závěr, že obě provenience jsou stejně výškově vyspělé.

7.5.3 Wilcoxonův test pro párové hodnoty

Je to neparametrická obdoba párového t-testu. Testujeme hypotézu
H_0: *Oba výběry pocházejí z jednoho základního souboru*.

Podobně jako při t-testu pro párové hodnoty vypočítáme rozdíly (d_i) mezi naměřenými hodnotami, které tvoří pár. Hodnoty d_i mohou být záporné, rovny nule nebo kladné. Hodnoty nulové a jim odpovídající páry hodnot při dalším postupu neužijeme. Nenulové hodnoty seřadíme podle velikosti (bez ohledu na znaménko) a přiřadíme jim pořadové číslo stejným způsobem jako u Wilcoxonova testu pro dva výběry. S přiřazenými pořadovými čísly pracujeme jako hodnotami souboru.

K přiřazeným pořadovým číslům potom připíšeme znaménko příslušné hodnoty. Kladné hodnoty pořadových čísel a rovněž záporné hodnoty pořadových čísel sečteme.

Platí-li nulová hypotéza, je rozdíl mezi součtem kladných a součtem záporných pořadových čísel minimální. Čím více se oba výběry od sebe liší, tím větší je rozdíl obou součtů.

Označíme-li menší hodnotu obou součtů písmenem W a platí $W \le W_{\alpha, n}$, kde W_α je kritická hodnoty stability W při platnosti H_0 pro mez významnosti α a počet nenulových diferencí n, nulovou hypotézu nepřijímáme.

Tabulky $W_{\alpha, n}$ jsou v učebnicích uváděny do různého počtu n (25 až po 65). Při větším počtu nenulových diferencí, než je uvedeno v tabulkách, postupuje se následovně:

Vypočítá se hodnota statistiky

$$U_w = \frac{W^+ - \frac{1}{4}n(n+1)}{\sqrt{\frac{1}{24}n(n+1)(2n+1)}} \qquad (7.30)$$

Statistika U_w má asymptoticky normované normální rozdělení N(0,1). Je-li tedy

$$|U_w| \le z_\alpha \qquad (7.31)$$

nulovou hypotézu H_0 přijímáme na hladině významnosti, která se asymptoticky blíží α.

Příklad 7.16:

Pro posuzování stupně kvality určitého výrobku byli určeni dva odborníci. Bylo ověřováno, zda se jejich subjektivní hlediska liší významně nebo jen náhodně. Na zkušebním vzorku 15 výrobků hodnotili oba kvalitu počtem bodů od 0 do 10. Výsledky hodnocení jsou v tabulce 7.6 *spolu s následným zpracováním*.

$n = 10$; $\alpha = 0{,}05$; $W = 22{,}5 \quad \ge \quad W_{0{,}05,\ 10} = 8$

Nulovou hypotézu nezamítáme. Oba odborníci se při hodnocení významně neliší.

7.5.4 Testy shody

Označují se tak testy hypotéz o shodě rozdělení. Testuje se, zda výběr pochází z určitého základního souboru se známým teoretickým rozdělením (hustotou pravděpodobnosti).

Číslo vzorku	Hodnocení		d_i	Pořadí d_i		
	A	B		±	+	-
1	8	7	+1	1	2,5	
2	7	9	-2	5		7,5
3	10	9	+1	2	2,5	
4	9	8	+1	3	2,5	
5	7	7	0			
6	6	6	0			
7	8	7	+1	4	2,5	
8	9	7	+2	6	7,5	
9	6	8	-2	7		7,5
10	8	8	0			
11	7	9	-2	8		7,5
12	10	8	+2	9	7,5	
13	9	9	0			
14	7	5	+2	10	7,5	
15	5	5	0			
Součet pořadí					**32,5**	**22,5**

Tabulka 7.6 – Zadání příkladu 7.16 a pomocné výpočty (stanovení pořadí)

7.5.4.1 χ^2 - test pro jeden výběr

Mějme určité experimentální rozdělení dané četnostmi n_{ei} (dané např. tříděním souboru do tříd, potom n_{ei} jsou absolutní třídní četnosti). Uvažujme určité teoretické rozdělení, které považujeme za model pro experimentální výběr, s teoretickými četnostmi n_{oi} (např. normální rozdělení, ale může být i jakékoli jiné). Vyslovme hypotézu

H_0: *Četnosti n_{ei} a n_{oi} se liší pouze náhodně.*

Pro ověření této nulové hypotézy používáme kritérium

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_{ei} - n_{oi})^2}{n_{oi}} \tag{7.32}$$

Testovací kritérium má rozdělení χ^2 s *f* stupni volnosti. Hodnota f záleží na následujícím pravidlu:

- f = k – 1, kde k je počet tříd, za předpokladu, že známe parametry teoretického rozdělení v základním souboru (např. pro normální rozdělení střední hodnotu a rozptyl základního souboru)
- f = k - 1 - c, kde c je počet parametrů teoretického rozdělení (např. 2 – střední hodnota a rozptyl - pro normální rozdělení, 1 - λ - pro Poissonovo apod.) a uvažuje se tehdy, jestliže parametry teoretického rozdělení neznáme a pouze je odhadujeme na základě výběrových charakteristik (např. neznáme parametry μ a σ pro normální rozdělení a aproximujeme je pomocí výběrových charakteristik $\bar{x}$ a s^2).

Při aplikaci χ^2 testu je nutno respektovat podmínky, za kterých jej lze použít:

a) při k = 2 nesmí být $n_{oi} < 5$,

b) při k > 2 nemá být více než 20% $n_{oi} < 5$ a žádná menší než 1.

Nevyhovují-li některé četnosti této podmínce, lze sloučit sousední třídy tak, aby podmínce bylo vyhověno. Číslo k je potom rovno novému počtu tříd po sloučení.

Hypotézu H_0 přijímáme, když $\chi^2_{vyp} < \chi^2_{\alpha,f}$

Příklad 7.17:

V porostu borovice bylo změřeno 98 výšek a roztříděno do 2 m tříd. Řešila se otázka, zda pro rozdělení výšek v porostu je vhodným teoretickým modelem normální rozdělení.

Vypočítaly se teoretické četnosti normálního rozdělení pro $\mu = 28{,}1$ m a $\sigma^2 = 7{,}24$ m^2, stanovila se hypotéza o shodě rozdělení a použilo se kritérium7.32. Hladina významnosti $\alpha = 0{,}05$.

i	1	2	3	4	5	6	7	8
h_i	22	24	26	28	30	32	34	36
n_{ei}	2	9	22	31	22	9	2	1
n_{oi}	2	9	21	29	23	10	3	1

Vzhledem k tomu, že zadání příkladu nevyhovuje podmínce, že nejvýše 20 % tříd může mít četnost menší než 5 (podle tohoto pravidla mohou mít v našem případě pouze 2 třídy četnost menší než 5), sloučíme dohromady třídy 1 + 2 a třídy 6 + 7 + 8, čímž získáme novou tabulku (je nutno znovu vypočítat i teoretické četnosti normálního rozdělení):

n_{ei}	11	22	31	22	12
n_{oi}	11	21	29	23	14
$\frac{(n_e - n_o)^2}{n_o}$	0,00	0,05	0,14	0,04	0,29

Vzhledem k tomu, že parametry normálního rozdělení byly odhadnuty pomocí výběrových charakteristik, je počet stupňů volnosti $f = k - 1 - c = 5 - 1 - 2 = 2$.

$$\chi^2 = 0{,}52; \qquad \chi^2_{0{,}05;\,2} = 6{,}0$$

Hypotézu o náhodných rozdílech četností přijímáme. Normální rozdělení o parametrech $\mu = 28{,}1$ m a $\sigma^2 = 7{,}24$ m^2 je vhodným modelem rozdělení výšek ve 2 m třídách.

7.5.4.2 Kolmogorovův-Smirnovovův test pro jeden výběr

Tento test se používá ke stejnému účelu jako předchozí, ale má výhodu v tom, že není omezen různými podmínkami použití jako χ^2 test. Proto se v případě nesplnění podmínek χ^2 testu doporučuje použít Kolmogorovův-Smirnovovův test pro jeden výběr. Testuje se opět hypotéza

H_0: *Četnosti n_{ei}, n_{oi} se liší pouze náhodně.*

Testovací kritérium má tvar

$$D_1 = \frac{1}{n}.\max\left|N_{ei} - N_{oi}\right|, \tag{7.33}$$

kde N_{ei}, N_{oi} značí hodnoty třídních součtových četností, n je rozsah celého souboru. Výraz max $|\,.\,|$ značí nejvyšší (maximální) absolutní hodnotu rozdílu součtových četností experimentálního a modelového rozdělení.

Veličina D_1 má své speciální rozdělení, závislé na rozsahu výběru n. Kritické hodnoty D_1 pro $\alpha = 0{,}05$ a $0{,}01$ jsou pro $n \leq 40$ tabelovány. Pro $n > 40$ se počítají podle přibližného vzorce

$$D_{1,\alpha} \doteq \frac{1}{\sqrt{n}} \cdot \sqrt{-\frac{1}{2} \ln \frac{\alpha}{2}} \tag{7.34}$$

Hypotézu o shodnosti rozdělení přijímáme, když $D_1 < D_{1,\alpha}$.

Příklad 7.18:

Ověříme výsledek příkladu 7.17 pomocí Kolmogorov – Smirnovova testu:

i	1	2	3	4	5	6	7	8
N_{ei}	2	11	33	64	86	95	97	98
N_{oi}	2	11	32	61	84	94	97	98
$\lvert N_{ei} - N_{oi} \rvert$	0	0	1	3	2	1	0	0

max $|\ N_{ei} - N_{oi}|\ = 3$; $D_1 = 3 : 98 = 0{,}031$.

$D_{1;\,0,05} = 1{,}36 . (1/\sqrt{98}) = 0{,}137$. $D_1 < D_{1;\,0,05}$.

Nulovou hypotézu přijímáme, oba testy shody podaly stejný výsledek.

7.5.4.3 Kolmogorovův - Smirnovovův test pro dva výběry

Dvouvýběrové testy se používají tehdy, jestliže porovnáváme dvě výběrová (experimentální) rozdělení (nikoli výběrové a teoretické rozdělení jako v případě jednovýběrových testů shody). Testuje se hypotéza

H_0: *Četnosti dvou výběrových rozdělení se statisticky významně neliší.*

Používá se testové kritérium

$$D_2 = \max |\ W_{1,i} - W_{2,i}\ |\,, \tag{7.35}$$

kde $W_{1,i}$, $W_{2,i}$ jsou relativní kumulativní třídní četnosti 1. a 2. souboru. Hypotézu přijímáme, když $D_2 < D_{2;\alpha}$.

Kolmogorov - Smirnovův test pro dva výběry můžeme použít v těchto případech.

a) $n_1 = n_2 \leq 40$; při malých výběrech musí být $n_1 = n_2$.

b) $n_1 > 40$, $n_2 > 40$, může být $n_1 \neq n_2$.

Kritické hodnoty $D_{2,\,\alpha}$ pro případ a) jsou tabelovány, pro případ b) se počítají ze vzorce

$$D_{2;\alpha} \cong \sqrt{-\frac{1}{2} \ln \frac{\alpha}{2}} \cdot \sqrt{\frac{n_1 + n_2}{n_1 \cdot n_2}} \tag{7.36}$$

Příklad 7.19:

Ve dvou porostech byly měřeny výčetní tloušťky ve 4-cm tloušťkových třídách. Zjištěné počty v jednotlivých tloušťkových třídách jsou uvedeny v tabulce 7.7 . Posuďte, zda lze rozdělení tlouštěk v obou porostech považovat za shodné.

Číslo tloušťkové třídy	Tloušťková třída	Četnosti v tloušťkové tříd		Relativní součtové četnosti		Rozdíly relativních součtových četností
		Porost I	Porost II	Porost I	Porost II	
1	18	8	5	0,048	0,029	0,019
2	22	16	7	0,145	0,070	0,074
3	26	25	18	0,295	0,175	0,120
4	30	31	25	0,482	0,322	0,160
5	34	38	48	0,711	0,602	0,109
6	38	24	42	0,855	0,848	0,007
7	42	12	15	0,928	0,936	- 0,008
8	46	6	5	0,964	0,965	- 0,001
9	50	4	5	0,988	0,994	- 0,006
10	54	2	1	1,000	1,000	0,000

Tabulka 7.7 – Zadání příkladu 7.19 a pomocné výpočty

K řešení úlohy použijeme testové kritérium 7.35, které vlastně spočívá ve vybrání největší absolutní hodnoty rozdílu mezi relativními součtovými četnostmi porovnávaných rozdělení. Je to hodnota $D_2 = 0{,}160$ (v tabulce 7.7 je vyznačena šedě). Tuto hodnotu porovnáme s kritickou hodnotou podle vztahu 7.36, kde za α dosadíme 0,05 a za $n_1 = 166$, $n_2 = 171$.

Výsledná hodnota $D_{2,\alpha} = 0{,}15$ je menší než hodnota 0,16, tedy nulovou hypotézu zamítáme. Oba porosty mají jiné rozdělení četností tlouštěk.

7.5.5 Dixonův test extrémních odchylek

Je to neparametrická obdoba Grubbsova testu. Základem je uspořádání souboru podle velikosti.

Je-li extrémní největší hodnota, je testovací kritérium tvaru

$$Q_n = \frac{x_n - x_{n-1}}{x_n - x_1} \tag{7.37}$$

Je-li extrémní nejmenší hodnota, je kritérium

$$Q_1 = \frac{x_2 - x_1}{x_n - x_1} \tag{7.38}$$

Kritické hodnoty $Q_{n,\alpha} = Q_{1,\alpha}$ pro Dixonův test vyhledáme ves speciální tabulce.

5. Nulová hypotéza se zamítá když $Q_1 > Q_{1,\alpha}$ resp. $Q_n > Q_{n,\alpha}$.

Příklad 7.20:
Úkol příkladu 7.4 vyřešíme Dixonovým testem.

$\alpha = 0{,}05;$ $\quad Q_{5;\,0{,}05} = 0{,}642$

$$Q_5 = \frac{2{,}14 - 1{,}86}{2{,}14 - 1{,}73} = 0{,}683$$

$0{,}683 > 0{,}642$ se stejným závěrem jako u příkladu 7.4 - hodnota 2,14 je extrémní.

7.5.6 Testy náhodnosti

Testujeme hypotézu

H_0: P*rvky výběru jsou nezávislé.*

Neparametrická verze testu náhodnosti – tj. testu nezávislosti jednotlivých prvků výběru – je založena na tzv. bodech zvratu. Řekneme, že číslo x_i ($2 \le i \le n - 1$) se nazývá bodem zvratu v posloupnosti různých čísel $x_1, x_2, \ldots, x_n$, jestliže platí buď $x_{i-1} < x_i > x_{i+1}$ nebo $x_{i-1} > x_i < x_{i+1}$, tedy body zvratu jsou takové hodnoty, kde se mění trend hodnot, tj. po stoupajících hodnotách následují klesající nebo naopak. Body zvratu jsou vyznačeny na obrázku 7.6 rámečky.

Používá se testové kritérium

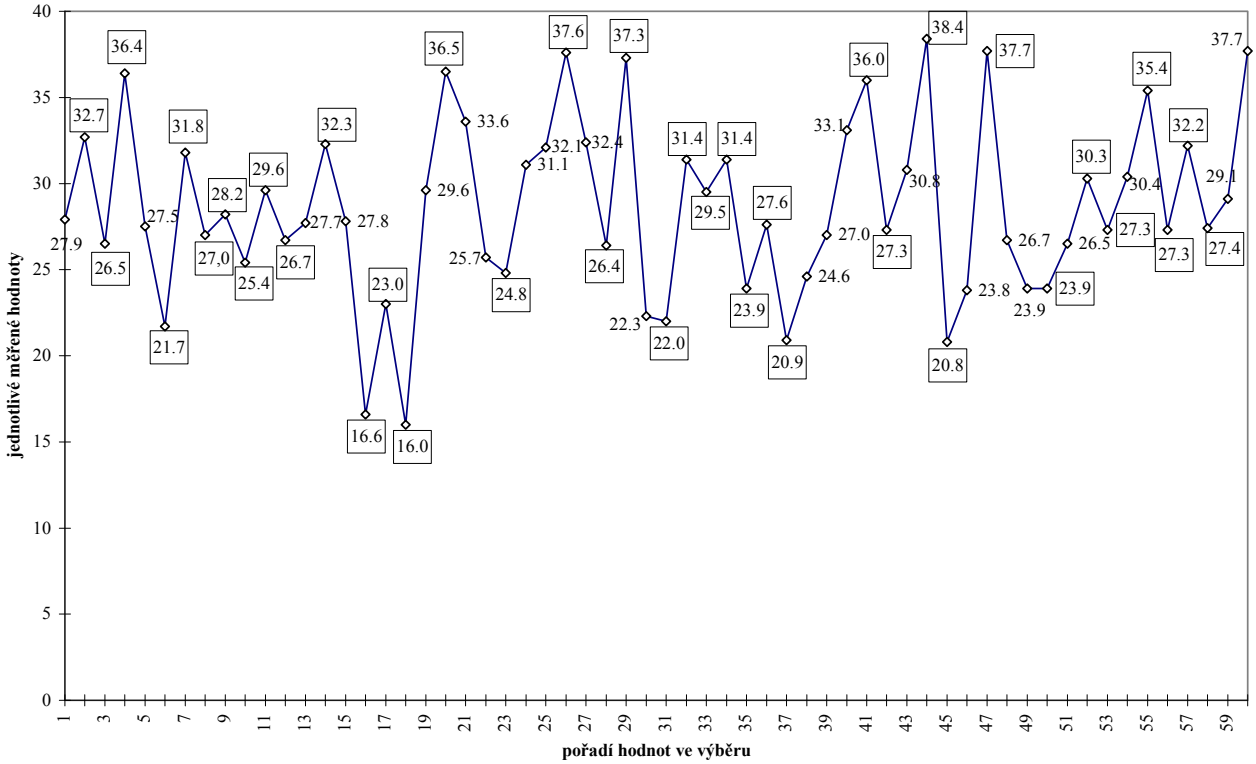
$$U = \frac{Z - \dfrac{2n - 4}{3}}{\sqrt{\dfrac{16n - 29}{90}}} \qquad (7.39)$$

které má pro velké výběry (teoreticky pro $n \to \infty$) asymptoticky normované normální rozdělení N(0,1). Jestliže $|U| < z_{\alpha/2}$, potom přijímáme nulovou hypotézu, tedy závěr, že prvky výběru jsou nezávislé.

Příklad 7.21:
Ověříme nezávislost prvků výběru z kapitoly 4.5 (Porost I) podle testu bodů zvratu.

Jedná se o výběr 60 prvků (měřených tlouštěk stromů), jehož body zvratu jsou na obrázku 7.6 (body, u nichž jsou čísla v rámečku, tedy „hroty“ grafu).

Z tohoto obrázku vyplývá, že bohů zvratu je celkem 38, tedy $Z = 38$ a $n = 60$. Dosadíme do vztahu 7.39 a získáme výslednou hodnotu $U = -0{,}207$. Vzhledem k tomu, že absolutní hodnota U je menší než kritická hodnota $z_{0,05} = 1{,}96$, přijímáme nulovou hypotézu a prohlásíme prvky tohoto výběru za náhodné.



Obrázek 7.6 – Body zvratu (vyznačeny čísly v rámečku) pro data příkladu 7.21

8 Použitá a doporučená literatura

ANDĚL, J., 1978: Matematická statistika. Praha, SNTL -Alfa .

BENEDÍK, J., 1989: Biostatistika. Brno, UJEP, 233 s.

CIPRA, T., 1986: Analýza časových řad s aplikacemi v ekonomii. Praha, SNTL-Alfa

CYHELSKÝ, L., NOVÁK,I., 1967: Statistika. Praha, SNTL, 288 s.

ČERMÁK, V., 1968: Statistika. Praha, SNTL, 208 s.

DRÁPELA, K., ZACH, J., 1995: Dendrometrie (dendrochronologie). Skriptum MZLU Brno, 152 s.

DRÁPELA, K. ZACH, J., 1996: Biometrie (biostatistika) – vybrané části, Skriptum MZLU Brno, 153 s.

GROFÍK, R. a kol., 1987: Štatistika. Bratislava, Príroda, 520 s.

HALD, A., 1956: Matematičeskaja statistika s techničeskimi priloženijami. Moskva, Izdavatělstvo inostrannoj litertury, 664 s.

HÁTLE, J., LIKEŠ, J., 1972: Základy počtu pravděpodobnosti a matematické statistiky, Praha, SNTL, 464 s.

HEBÁK, P., KAHOUNOVÁ, J., 1988: Počet pravděpodobnosti v příkladech. SNTL, Praha, 312 s.

CHAMBERS. J.M. a kol., 1983: Graphical Methods for Data Analysis. Belmont, Duxburry Press.

CHATFIELD, C., 1984: The Analysis of Time Series. An Introduction. London, Chapman and Hall, 286 s.

KUBÁČEK, L., PÁZMAN, A., 1979: Štatistické metódy v meraní. Bratislava, Veda, 148 s.

LAAR, A., 1979: Biometrische Methoden in der Forstwissenschaft. München, 633 s.

LEPORSKÝ, A., 1953: Statistické metody. Učební texty vysokých škol. Lesnická fakulta VŠZ Brno, SPN, Praha

MELOUN, M., MILITKÝ, J., 1994: Statistické zpracování experimentálních dat. Praha, Plus, 839 s.

MICHÁLEK a kol., 1982: Biometrika. Praha, SPN, 404 s.

MINAŘÍK, B., 1995: Statistika I pro ekonomy a manažery. Skriptum MZLU Brno, 160 s.

MINAŘÍK, B.,1996: Statistika II pro ekonomy a manažery. Skriptum MZLU Brno, 144 s.

MINAŘÍK, B., 1996: Statistika III. Skriptum MZLU Brno, 156 s.

MYSLIVEC, V., 1957: Statistické metody zemědělského a lesnického výzkumnictví. Praha, SZN

REISENAUER, R., 1970: Metody matematické statistiky a jejich aplikace v technice. Praha, SNTL, 240 s.

SACHS, L., 1972: Statistische Auswertugsmethoden. Berlin, Heidelberg, New York, Springer - Verlag , 506 s.

ŠMELKO, Š. 1991: Štatistické metódy v lesníctve. Skriptum VŠLD Zvolen, 276 s.

ŠMELKO, Š., WOLF, J., 1977: Štatistické metódy v lesníctve. Bratislava, Príroda, 330 s.

ŠTULAJTER, F.,1989: Odhady v náhodných procesoch. Bratislava, ALFA, 288 s.

TUKEY, J. W., 1977: Exploratory Data Analysis. Adison-Wesley, 670 s.

ÜBERLA, K., 1974: Faktorová analýza. Bratislava, ALFA.

ZACH, J., 1990 A: Statistické metody - cvičení. Skriptum VŠZ Brno, 74 s.

ZACH, J., 1990 B: Statistické metody - vybrané části. Skriptum VŠZ Brno, 74 s.

ZACH, J., 1993: Statistické metody. Skriptum VŠZ Brno, 165 s.

ZACH, J., DRÁPELA, K., SIMON, J., 1994: Dendrometrie (cvičení). Skriptum VŠZ Brno, 167 s.

ZAR, J.H., 1984: Biostatistical Analysis, Prentice-Hall Int., New Jersey, 718 s.

Obsah