10. Základní statistické pojmy.

10.1 Úvodní informace

Statistika je často představována jako pouhý sběr čísel nebo jim podobných údajů. Původní význam toho slova skutečně souvisí se sběrem informací o státu (z latinského status – stát) – počtu obyvatel, sídel, o výběru daní atd. I dnes existují instituce, které se zabývají takovýmto sběrem dat, v ČR je to Český statistický úřad. Sbírá a zveřejňuje některé informace o obcích , průmyslu, ekonomice, o demografickém rozvoji státu. Pod pojmem statistika dnes však míníme mnohem více, statistika se v jistém slova smyslu stala jazykem pro práci s daty , pro jejich zpracování a interpretaci. Ze statistiky se stala rozvinutá vědecká metoda analýzy dat, která nachází široké uplatnění v přírodních i společenských vědách i ve společnosti vůbec.

Při vlastní praxi uplatňujeme dva způsoby přístupu k údajům. Především je to přístup k informacím vnějšího prostředí a posléze naše reflexe na tyto údaje ve formě zobecnění. Například při porovnávání sledovanosti televizních kanálů neoslovujeme všechny domácnosti, ale z pečlivě vybraných domácností a jejich sledovanosti televize činíme závěry platné pro všechny domácnosti.

Proces zobecňování poznatků nazýváme ***induktivním způsobem usuzování*** (indukcí) např. zobecnění sledovanosti ve výběru na všechny domácnosti. Schopnost přijímat nové poznatky a z nich se učit a vyvozovat závěry jsou jedním ze základních rysů lidského uvažování. Druhým způsobem uvažování je princip ***deduktivního přístupu k údajům*** (dedukce). Při deduktivním přístupu činíme závěry z obecných zákonitostí.

Závěry myšlenkových procesů induktivního charakteru jsou ovlivněny postojem subjektu . Induktivní statistika se zabývá způsoby jak přenášet závěry takovýchto procesů, umožňuje z pozorovaných dat vytvářet obecné závěry s určením jejich spolehlivosti. Výpočty takových spolehlivostí jsou založeny na poznatcích teorie pravděpodobnosti a jsou proto objektivní.

10.2 Statistický soubor a výběry

Jedním ze základních pojmů , s kterými se budeme setkávat stále jsou **populace (statistický soubor)** a **výběr**.

Populace je množina všech prvků, které jsou předmětem daného statistického zkoumání. Každý z prvků je statistickou jednotkou. Prvky tvořící statistický soubor jsou buď dány prostě výčtem nebo mají určité společné vlastnosti - tzv. identifikační znaky - umožňující určit, zda prvek do daného statistického souboru patří nebo nepatří. Identifikační znaky tedy statistický soubor mohou vymezovat. Z hlediska velikosti je zřejmé , že většina populací bude mít konečný rozsah, nekonečný rozsah budou mít takové populace, které jsou určeny znakem, který můžeme hypoteticky nekonečněkrát opakovat (např. měříme hmotnost po pokusu, teplotu atd.). Podle počtu sledovaných znaků je potom takováto populace jednorozměrné či vícerozměrná (sledujeme dva a více znaků např. teplotu, tlak; komunikativnost, inteligenci atd.). Pro vlastní popsání populací se používá metoda parametrů charakteristik. Jde o číselné hodnoty, které jsou většinou pevná čísla. Jejich hodnota není známa a je nutno ji zjistit či odhadnou vhodnými statistickými metodami.

Znaky, které sledujeme v populaci mají obecně buď charakter ***kvantitativní*** (lze je vyjádřit číslem např. délka , hmotnost, teplota) a ***kvalitativní*** (jsou většinou vyjádřeny textem). Kvantitativní znaky dělíme dále na spojité – výsledky zkoumání mohou nabývat hodnot některého intervalu (teplota, délka) a diskrétní jestliže existuje jen konečně mnoho možných stavů znaku (např. počet dětí v rodině, počet vykvetlých rostlin atd.).

K vlastnímu měření kvantitativních údajů používáme buď intervalových nebo poměrových stupnic. Jestliže chceme zjistit jen rozdíl mezi kvalitativními hodnotami , používáme **intervalovou** stupnici (v takovýchto stupnicích je počátek volen např. 0°C, stupnice výšky tónu, stupnice bolesti atd.). Při takovémto způsobu měření je většinou nesmyslné označení prvek **a** má hodnotu znaku 2x větší než prvek **b** , neboť počátek je možno volit různě (např. teplota). Pokud chceme měřit údaje ve vztahu k pevným jednotkám (váha, vzdálenost) používáme stupnici **poměrovou** .

Kvalitativní znaky se snažíme také měřit , používáme k tomu nominální (pojem) a ordinální (pořadí) stupnici. ***Nominální*** stupnice je složena z nejméně dvou navzájem se vylučujících tříd. Jestliže jsou třídy právě dvě nazývá se dichotomická. Příklady takovéto stupnice: pohlaví / mužské, ženské /; barva / modrá, zelená, červená, bílá /. Příkladem takovéto klasifikace je také. mezinárodní stupnice nemocí, úrazů a příčin smrti. Čísla , která jsou přiřazena jednotlivým chorobám nic nevypovídají o dané chorobě. ***Ordinální*** stupnice je založena opět na neslučitelných třídách, ale ty jsou ještě navzájem uspořádány . Příklady takovýchto stupnic: nejvyšší úroveň vzdělání / negramotný, základní, střední, vysokoškolské / ; srozumitelnost / žádná, malá, střední, uspokojivá, vynikající/.

V tabulkách 10.1 a 10.2 níže jsou uvedeny způsoby použití jednotlivých stupnic.

Tabulka 10.1

Typ stupnice	Použití pro data	Přípustné změny	Charakteristiky rozdělení
Nominální stupnice	Jsme schopni rozhodnout o rozdílu mezi jednotlivými prvky populace a o jejich zařazení do tříd	Permutace , přejmenování	Absolutní četnost, relativní četnost, modus
Ordinální stupnice	Navíc: Umíme určit, který prvek je menší a který větší a zařadit je do správných tříd	Možno změnit pomocí monotóní transformace (rostoucí)	Dále: Kumulativní četnost, pořadí, kvantily, medián, pořadové hodnoty
Intervalová stupnice	Navíc: Umíme stanovit relativní nulový bod (počátek) a zjistit vztah prvků vůči němu (rozdíly!)	Lineární změna - posunutí a zmenšení nebo zvětšení ($y = a x + b$)	Dále: Aritmetický průměr, směrodatná odchylka, šikmost, špičatost
Poměrová stupnice	Navíc: Umíme stanovit absolutní nulový bod (počátek) a zjistit vztah prvků vůči němu (podíly!)	Změna jen zvětšení nebo zmenšení (kladné) tj. $y = a x (a > 0)$	Dále: Ostatní průměry (harmonický, geometrický), variační koeficient

Tabulka 10.2

Typ stupnice	Testy	Závislost , nezávislost
Nominální stupnice	χ^2 - testy	Kontingenční koeficienty, čtyřpolíčkový koeficient
Ordinální stupnice	Dále: Pořadové testy, Kolmogor - Smirnův test, U - test	Pořadový korelační koeficient
Intervalová stupnice	Dále: Parametrické testy odvozené z N(0,1)	Korelační koeficient, biseriální koeficienty
Poměrová stupnice	Stejně jako výše	Stejně jako výše

Pro vyšetření populace používáme různý způsob přístupu k datům : Provádíme buď **statistický pokus**, **statistické šetření** nebo **pozorovací studii**. Účelem statistického pokusu je plánovitě měnit faktory (podmínky) a sledovat jejich vliv na změnu vyšetřovaných znaků. Výběr prvků s nimiž experimentujme provádíme zásadně náhodně, aby nedošlo k vychýlení výsledných hodnot. Při tzv. kontrolovaném pokusu rozdělíme vyšetřované skupiny na

pokusné a kontrolní. U pokusné skupiny byla provedeny změna , u kontrolní nikoli. Aby byl pokus dostatečně objektivní , je nutno, aby obě skupiny byly rovnocenné jak na začátku pokusu, tak i v jeho průběhu. Chceme – li zabránit přínosu subjektivní informací volíme často princip tzv. slepého pokusu, kdy ten kdo údaje vyhodnocuje (např. lékař) nevěděl, která skupina je kontrolní a která je pokusná. Jestliže ani vyšetřovaný subjekt neví zda je v pokusné nebo kontrolní skupině nazýváme tento princip dvojité utajení nebo dvojitý slepý pokus.

Je vidět , že princip náhodného výběru a rozdělení na pokusnou a kontrolní skupinu zlepšuje výsledky (odstraňujeme neobjektivitu a závislost).

Někdy ovšem není možné získávat data manipulací s prvky populace . Není možno provádět statistický pokus, můžeme však jednoduše pozorovat jak probíhají změny a registrovat je . Takovému přístupu říkáme statistické **šetření** nebo pozorovací studie. Používáme ho tehdy , kdy nemůžeme využít princip náhody (případy , kdy rozložení znaků v populaci je dáno – např. vzdělání, pohlaví a v pokusu by nebylo respektováno ; někdy není možno realizovat **statistický pokus** z etických důvodů (manipulace s lidmi). Vidíme tedy , že v případě statistického šetření se spokojujeme s pasivním sběrem dat. Problémem takovýchto studií je , že pozorovaný jev je velmi často ovlivněn nežádoucími znaky. Pro pojem úplného šetření tj. šetření provedeného na celé populaci se vžil pojem **census** (sčítání lidu). Pro jeho vysoké ekonomické náklady se provádí v naší republice jednou za deset let.

Každé statistické šetření v podobě censu by bylo především ekonomicky velmi náročné. Ve většině případů ten, kdo chce provést statistické šetření má omezené zdroje (finance, čas). Někdy je k dispozici jen málo údajů (šetření vzácné choroby nebo zvláštního chování pacientů). Při dalších šetřeních bychom museli populaci zničit (například sledování životnosti výrobků). Výběr může nést přesnější výsledky než úplné šetření (při velkém množství chyb vinou neodborných špatně proškolených pozorovatelů vznikne chyba neodstranitelná).

Jakákoli část populace , která dobře odráží její strukturu (především vyšetřované znaky) bude nazvána **reprezentativním** výběrem. Ostatní typy výběru se nazývají **selektivní** výběry, většinou dávají zkreslený obrázek o vyšetřované populaci. Příkladem selektivního výběru je vzorek vysokoškolských profesorů, z něhož budeme usuzovat na vzdělanost celé populace. Je jisté , že struktura vzdělanosti v našem výběru bude značně vychýlena proti celé populaci. Výběry pořizujeme metodami náhodného výběru nebo metodami záměrného výběru.

Metoda **záměrného výběru** se opírá expertní stanoviska k vytvoření representativního výběru (prováděna často v euchologii, sociologii). Jsou často závislé na subjektu experta.

Metoda **náhodného výběru** umožňuje vybírat prvky populace náhodně a nezávisle na subjektech. Podle způsobu provedení rozlišujeme několik druhů náhodného výběru:

Prostý náhodný výběr – prováděn většinou metodou losování (každý prvek populace může být vylosován). Dříve se prováděl i pomocí tabulek náhodných čísel, dnes možno použít i vhodný generátor náhodných čísel různých statistických, ale i nestatistických programů.

Mechanický výběr – jde o jistou formu prostého výběru, nejdříve náhodně očísluji prvky populace a poté zvolím pevné číslo . Všechny prvky, které získám vždy o pevný zadaný krok budou v daném výběru. Pokud neprovedeme na začátku náhodné očíslování, ale číslování je už vytvořeno musí dbát na to , aby krok výběru nesouvisel s číslováním.

Oblastní výběr . Celá populace je rozdělena do částí – oblastí tak , aby se ve sledovaných znacích se od sebe velmi odlišovali, v rámci jedné oblasti jsou sledované znaky málo odlišné. V jednotlivých oblastech potom provedeme prostý výběr. Spojením všech takovýchto dílčích výběrů získáme celý hledaný výběr.

Skupinový výběr. V případě populací, které čítají statisíce nebo miliony prvků je skoro nemožné předchozími metodami vytvořit náhodný výběr. Vyžíváme proto přirozené

rozdělení populace na menší celky nebo vytváříme vlastní umělé dělení. Požadujeme, aby prvky (skupiny) dělení byly pokud možno stejně velké a vyšetřované znaky heterogenní v rámci jedné skupiny. Variabilita mezi jednotlivými skupinami by měla být co nejmenší.

Vícestupňový výběr. Provádí se tehdy, kdy existuje hierarchický popis celé populace (geografický, sociální model).

10.3 Popisná statistika

Popisná statistika (deskriptivní statistika) se zabývá popisem stavu nebo vývoje hromadných jevů. Nejprve se vymezí soubor prvků, na nichž se bude uvažovaný jev zkoumat. Následně se všechny prvky vyšetří z hlediska studovaného jevu. Výsledky šetření - kvalitativní i kvantitativní, vyjádřeny především číselným popisem - tvoří obraz studovaného hromadného jevu vzhledem k vyšetřovanému souboru.

V předchozí části jsme studovali pojem statistického výběru. V této části budeme předpokládat, že jsme provedli výběr z populace a budeme se snažit z těchto dat získat údaje o vlastnostech základního souboru.

Grafické znázornění výběrových rozdělení je uvedeno v následující kapitole. V této kapitole budeme využívat data z tabulky 10.3

Tabulka 10.3: Rozdělení měsíčních nákladů studentů na bydlení

Pořadí	Náklady	Pořadí	Náklady	Pořadí	Náklady
1	850	11	1560	21	2900
2	910	12	1560	22	2900
3	920	13	1650	23	3100
4	920	14	1670	24	3150
5	920	15	1780	25	3250
6	1030	16	1790	26	3250
7	1030	17	1850	27	3400
8	1150	18	2200	28	3600
9	1190	19	2600	29	3700
10	1190	20	2800	30	3850

Uveďme dále důležité pojmy, které budeme neustále využívat.

Četnost (absolutní) hodnoty x_i je daná počtem prvků x_i ve výběru.

Relativní četnost hodnoty x_i – je daná podílem absolutní četnosti a celkového počtu prvků ve výběru.

Kumulativní absolutní četnost hodnoty x_i je daná součtem všech absolutních četností prvků, které jsou menší nebo rovny prvku **x_i .**

Kumulativní relativní četnost hodnoty x_i je dána součtem všech relativních četností prvků, které jsou menší nebo rovny prvku **x_i .**

10.3.1 Míry polohy

Jde o číselné hodnoty pomocí , nichž určujeme polohu míst, kolem kterých jsou data nejvíce umístěny.

10.3.1.1 Průměr

Průměr $\bar{x}$ se používá v případě kvantitativních znaků. Je velmi citlivý na odlehlé hodnoty. Průměr n hodnot x_1 , x_2 , … , x_n vypočteme takto

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (10.1).$$

Pro naše data je $\bar{x} = 2422{,}33$.

Někdy jsou data uvedena v tabulce včetně svých absolutních četností (počtu opakování) , potom počítáme průměr jako tzv. **vážený průměr**:

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i . x_i}{n} \qquad (10.2)$$

V tomto případě jsou data rozdělena na k skupin o n_k prvcích.

Pokud jsou data uvedena v tabulce roztříděných dat (původní dat jsou nahrazena příslušností do jednoho z vybraných intervalů) vytvoříme nejprve střed intervalu (bude nahrazovat všechna data uvedená v daném intervalu) a pak z těchto hodnot vytvoříme podle vztahu (10.2) průměr.

Tabulka 10.4 třídní rozdělení četností:

Rozpětí	četnost
0 -500	0
500 - 1000	5
1000 - 1500	5
1500 - 2000	7
2000 - 2500	1
2500 - 3000	4
3000 - 3500	5
3500 - 4000	3
4000 - 4500	0

Hodnota středů intervalů je 250 , 750, …, 4250 . Spočítáme – li průměr podle vzorce (10.2) je hodnota třídního průměru rovna **1733,7**. Je vidět, že hodnota tohoto průměru velmi závisí na správné volbě rozpětí třídy. Pro vytvoření stejně velkých tříd o počtu k z n prvků je možno použít tzv. **Sturgesovo pravidlo**

$$k \approx 1 + 3{,}3 \, . \log_{10} n \qquad (10.3)$$

Například pro náš případ je n = 30 a tedy hodnota k ≈ 5,8745 . Tedy volíme k = 6.

Uveďme dále některé důležité vlastnosti průměru:

a) Jestliže ke každé hodnotě x_i ve výběru přičteme konstantu **k** , zvětší se o konstantu **k** také původní průměr (**k** může být libovolné reálné číslo).

b) Násobíme – li každou hodnotu ve výběru x_i stejnou konstantou **m** , vypočteme nový průměr jako součin starého průměru a konstanty **m**

c) Součet odchylek všech hodnot x_i ve výběru od jejich průměru $\bar{x}$ je roven nule

$$\sum_{i=1}^{n} \left(x_i - \bar{x} \right) = 0 \qquad (10.4)$$

d) Součet čtverců odchylek všech hodnot od jejich průměru je menší než součet čtverců odchylek všech hodnot od libovolné jiné hodnoty.

$$\forall_{a \neq \bar{x}} \quad \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 \leq \sum_{i=1}^{n} \left(x_i - a \right)^2 \qquad (10.5)$$

Těchto vlastností průměru využíváme také k tomu , abychom upravili vstupní hodnoty jejich zmenšením (resp. zvětšením) a posunutím.

Průměr se používá jako číselná charakteristika protože:

a) Je jednoznačný
b) Je lineární
c) Je spolehlivou číselnou hodnotou.

Průměr nepoužijeme , jestliže

a) Rozdělení je vícevrcholové

b) Rozdělení má na krajích otevřené třídy (hodnoty nejsou shora nebo zdola omezené)
c) Údaje nejsou škálované metricky, ale ordinálně
d) Výběr je extrémně malý
e) Rozdělení je asymetrické

10.3.1.2 Modus

Modus $\hat{x}$ je hodnota , která se vyskytuje nejčastěji. Podle tabulky 10.1 ho můžeme zjišťovat i znaků, které jsou kvalitativní, dokonce i nominální. Není ovlivňován všemi prvky ve výběru. Jestliže je četnost všech prvků ve výběru stejná, modus neurčujeme. Jestliže dvě nebo více navzájem sousedících hodnot nabývají stejné největší četnosti, pak aritmetický průměr z těchto hodnot nazveme modusem. Jestliže existují dvě navzájem nesousedící hodnoty s největšími stejnými četnostmi, uvádíme obě jako modus. Rozdělení je pak **dvouvrcholové** (**bimodální**). Již ze samé definice modusu je jasné, že tato charakteristika velmi závisí na výběru a většinou velmi kolísá.

Příklad 10.1

Zjistěte modus šetření výběru barev respondentů – bílá, červená, modrá , červená, zelená, bílá , červená , modrá, bílá, červená.

Odpověď :

Nejčetnější výskyt má a modus je červená.

Příklad 10.2

Zjistěte hodnotu modusu pro data z naší tabulky 10.3.

Odpověď:

Podle tabulky je $\hat{x} = 920$.

Jestliže jsou kvantitativní znaky uspořádány do třídní tabulky , určíme nejdříve modální interval x_D (s nejvyšší četností) a modus stanovíme **interpolací**

$$\hat{x} = x_D + h.\frac{n}{n+m} \qquad (10.6)$$

kde h je délka modálního intervalu, n je četnost , x_D je dolní hranice tohoto intervalu, n je četnost následujícího intervalu a m četnost předchozího intervalu. Aplikujme vzorec (10.6) na data z tabulky 10.4

$$\hat{x} = x_D + h.\frac{n}{n+m} = 1500 + 500.\frac{1}{6} = 1583{,}33\,.$$

Vidíme tedy , že modus zjištěný podle vzorce (10.6) může být výrazně odlišný od modusu skutečného.

10.3.1.3 Kvantily a medián

Přirozenou mírou jsou **kvantily**. Daný výběr se nejdříve seřadí od nejmenší hodnoty po největší a poté určíme pro daný p% kvantil pořadové číslo jednotky n_p , pro které platí

$$n.\frac{p}{100} < n_p < n.\frac{p}{100} + 1, \qquad (10.7)$$

kde n je počet prvků výběru.

Pro hodnotu p = 50% se daný kvantil označuje **medián** $\tilde{x}$. Jestliže je počet n sudé číslo , vypočteme medián jako průměrnou hodnotu z hodnot stojících vlevo a vpravo od teoretického mediánu určeného vzorcem (10.7). Medián popisuje hodnotu, která dělí daný výběr na dvě stejně velké části. V našem příkladě je $\tilde{x} = \frac{1780 + 1790}{2} = 1785$.

Další významné kvantity jsou :

Dolní kvartil $x_{0,25}$ je určen jako 25% kvantil.

Horní kvartil $x_{0,75}$ je určen jako 75% kvantil.

V našem případě je $x_{0,25} = 1080$ a $x_{0,75} = 3000$.

Pro hodnoty kvartilů vytváříme ještě jednu míru (jde o míru variability) a to **kvartilové rozpětí** $\mathbf{R_q} = x_{0,75} - x_{0,25}$

V našem případě je $\mathbf{R_q} = 3000 - 1080 = 1920$.

Pro hodnoty p=10,20,…,90 nazýváme takto spočtené kvantily názvy **decily**. Pro hodnoty p = 1,2,3,…,99 nazýváme podobně kvantily jako **percentily**.

Pomocí kvartilů je také možno velmi přehledně znázornit data v grafu s názvem Box Plot nebo jinak Krabicový graf nebo Krabicový diagram nebo Vousatá krabička. Pomocí něho můžeme rozdělit data z výběru na **vnitřní**, **vnější** a **odlehlá**. Vytváříme ho následujícím způsobem:

Základním prvkem grafu je obdélník, jehož hrany tvoří hodnoty dolního a horního kvartilu – uvnitř tohoto obdélníku je 50% hodnot výběru. Uvnitř je svislou čarou vyznačen medián, popř. tečkou průměr (křížkem modus) . Z obdélníku vedou dvě úsečky kolmé k hranám, jejichž délka je dána vzdáleností **vnitřních hradeb** od hrany obdélníku. Vnitřní hradby se vypočtou tímto předpisem

$$h_D = x_{0,25} - 1,5 \,.\, (x_{0,75} - x_{0,25}) \qquad (10.8)$$

$$h_H = x_{0,75} + 1,5 \,.\, (x_{0,75} - x_{0,25}) \qquad (10.9)$$

V našem případě jsou $h_D = 1080 - 1,5 \,.\, 1920 = -1800$ a $h_H = 3000+1,5.1920 = 5865$.

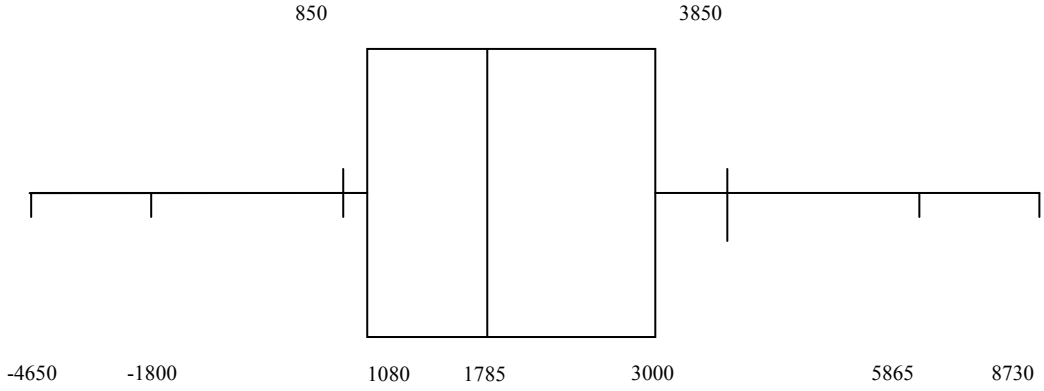
Dále se počítají vnější hradby

$$H_D = x_{0,25} - 2.(1,5 \,.\, (x_{0,75} - x_{0,25})) \qquad (10.10)$$

$$H_H = x_{0,75} + 2.(1,5 \,.\, (x_{0,75} - x_{0,25})) \qquad (10.11)$$

V našem případě je $H_D = 1080-3.1920 = -4680$ a $H_H = 3000+3.1920 = 8730$.

Hradby slouží pro identifikaci dat ve výběru. Hodnoty uvnitř vnitřních hradeb jsou **hodnoty přilehlé**; hodnoty mezi vnitřními a vnějšími hradbami jsou **hodnoty vnější** a hodnoty vně vnějších hradeb jsou **hodnoty vzdálené** nebo jinak odlehlé. Do grafu se zakresluje i minimální a maximální hodnoty jako body.



Jestliže máme data uvedena v třídní tabulce musíme p% kvantil počítat pomocí **lineární interpolace**

$$\frac{x_p - x_D}{x_H - x_D} = \frac{p - n_D}{n_H - n_D}, \qquad (10.12)$$

kde x_D je dolní a x_H je horní mez intervalu v němž leží daný kvantil; n_D je kumulativní relativní četnost odpovídající x_D a n_H je kumulativní relativní četnost odpovídající x_H .Zjistěme hodnotu kvantilu pro náš případ tabulky 10.4:

$$\frac{\tilde{x} - 1500}{2000 - 1500} = \frac{0,5 - 0,33}{0,57 - 0,33} \Rightarrow \tilde{x} = 1854,167 \,.$$

Použití mediánu je vhodné při rozděleních s otevřenými třídami, pro ordinální hodnoty, pro velmi symetrická rozdělení.

10.3.1.4 Geometrický průměr

Provádí se jen pro hodnoty ve výběru, které jsou kladné . Jeho označení je **G** a spočítá se jako n – tá odmocnina ze součinu hodnot x_i. Používáme ho , jak je zřejmé z definice , na kvantifikovatelné znaky měřené na poměrové stupnici. Používá se k určení průměrné změny velikosti, jestliže předpokládáme , že tato změna je konstantní (multiplikativně).

$$G = \sqrt[n]{x_1 . x_2 . \cdots . x_n} \qquad (10.13)$$

Klasickým příkladem k užití je případ výpočtu inflace za několik let známe – li hodnoty inflačních kroků mezi jednotlivými následnými roky.

10.3.1.5 Harmonický průměr

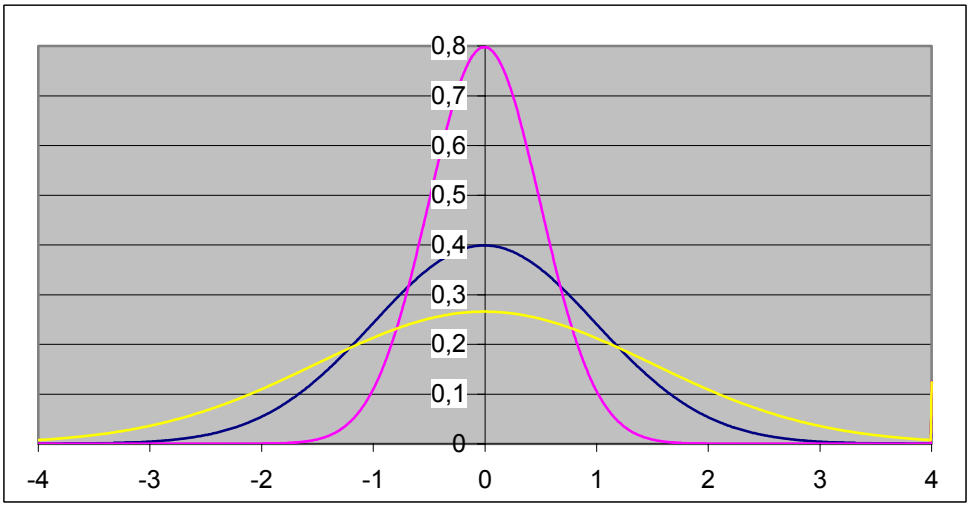
Harmonický průměr **H** zjistíme jako podíl počtu hodnot n a součtu převrácených hodnot výběru.

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \qquad (10.14)$$

Pomocí harmonického průměru lze například počítat úlohy na průměrné rychlosti , jestliže jsou známy tyto hodnoty na jednotlivých úsecích trati a chceme získat průměrnou rychlost celkovou.

10.3.2 Míry variability

Pomocí jen měr polohy nelze přesně popsat výběr, protože mnoho dat má stejné nebo přibližně stejné hodnoty jednotlivých parametrů měr polohy, přesto jsou na první pohled odlišné . Na obrázku níže je uveden případ tří skupin dat, která mají stejný průměr, modus, medián a přesto jsou odlišná. Odlišnost vidíme v soustředění hodnot kolem průměru. Toto soustředění budeme studovat pomocí různých měr variability.



10.3.2.1 Variační rozpětí

Variační rozpětí **R** se vypočte jako rozdíl mezi největší a nejmenší hodnotou výběru.

$$\mathbf{R} = x_{max} - x_{min} \qquad (10.15)$$

Pokračujme dále v našem příkladě , hodnota **R** = 3 850 – 850 = 3 000

Výhodou této míry je jednoduchost určení a porozumění. Je však málo stabilní vzhledem k počtu členů výběru. Používá se proto jen u malých výběrů (n ≤ 12). Výrazně závisí na velikosti výběru. Proto nemůžeme mezi sebou porovnávat jednotlivé hodnoty variačního rozpětí z různě velkých výběrů. Nedává spolehlivé odhady rozptylu základního souboru.

10.3.2.2 Průměrná odchylka

Průměrnou odchylku **e** výběru definujeme jako aritmetický průměr z absolutních hodnot odchylek všech hodnot výběru od průměru

$$e = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} \qquad (10.16)$$

Uvádíme ji jen pro úplnost. Je málo stabilní vzhledem k velikosti výběru a dává nespolehlivé odhady pro rozptyl.

10.3.2.3 Rozptyl a směrodatná odchylka

Nejužívanější mírou variability je rozptyl (resp. směrodatná odchylka). Pomocí něho měříme velikost čtverců odchylek jednotlivých hodnot výběru od průměru. Označujeme ho většinou symbolem $\mathbf{s^2}$ a nazýváme ho výběrovým rozptylem

$$s^2 = \frac{1}{n-1} . \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 , \qquad (10.17)$$

Všimněme si , že při výpočtu nedělíme součet odchylek čtverců hodnotou n (jako při definici klasického rozptylu) , ale hodnotou n – 1 (nazývanou také počtem stupňů volnosti). Je to provedeno proto, že získáme lepší odhad skutečného rozptylu σ^2 populace.

Výběrová směrodatná odchylka se označuje symbolem **s** a je rovna odmocnině z výběrového rozptylu

$$s = \sqrt{\frac{1}{n-1} . \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2} , \qquad (10.18)$$

Pro vlastní výpočet se hodí i jiná forma vzorce (10.17)

$$s^2 = \overline{x^2} - \bar{x}^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n} \right)^2 , \mathrm{i} = 1,2,\cdots,\mathrm{n} \qquad (10.19)$$

Použijeme – li vzorce na určení rozptylu pro data z tabulky 10.3 získáme $\mathbf{s^2}$ = 1019733,448 a hodnota **s** = 1009,82 .

Jsou – li hodnoty x_i výběru uvedené včetně četností n_i potom přejde vzorec (10.16) na

$$s^2 = \frac{1}{n-1} . \sum_{i=1}^{k} n_i . \left(x_i - \bar{x} \right)^2 = \frac{1}{n-1} . \left(\sum_{i=1}^{k} n_i . x_i^2 - n . \bar{x}^2 \right) , \qquad (10.20)$$

kde k je počet všech různých hodnot ve výběru a n je celkový počet prvků výběru. Jestliže jsou data uvedena pomocí třídění do intervalů např. data z tabulky 10.4 , potom většinou hodnoty x_i znamenají středy třídních intervalů a n_i počet dat v tomto intervalu. Pokud jsou třídní intervaly **ekvidistantní** (mají pevnou délku) s rozměrem h bude výpočet podle vzorce (10.20) zatížen chybou . Tuto chybu opravujeme pomocí tzv. **Sheppardovy korekce**

$$s^2_{kor} = s^2 - \frac{h^2}{12} \qquad (10.21)$$

Použijeme – li opět naše data z tabulky 10.4 získáme :
Nekorigované hodnoty s^2 = 1002500 a s = 1001,249;
Korigované hodnoty s^2_{kor} = 981666,7 a s_{kor} = 990,7909.

Velmi často nastává případ , že celý výběr je z určitých důvodů rozdělen do **k** dílčích částí . V i – té části je počet prvků roven n_i , průměr je roven $\overline{x_i}$ a výběrový rozptyl s_i^2 . Potom můžeme počítat celkový výběrový rozptyl $\mathbf{s^2}$ jako

$$s^2 = \frac{1}{n-1}.\left(\sum_{i=1}^{k} (n_i - 1).s_i^{\,2} + \sum_{i=1}^{k} n_i.(\overline{x}_i - \overline{x})^2 \right) \qquad (10.22)$$

Z předchozího vzorce vyplývá, že celkový výběrový rozptyl $\mathbf{s^2}$můžeme rozložit na dvě části – na **vnitroskupinový** a **meziskupinový**. Vnitroskupinovým výběrovým rozptylem sledujeme variabilitu uvnitř jednotlivých skupin a meziskupinovým výběrovým rozptylem variabilitu mezi těmito skupinami. Takovéto metody rozdělení celkové variability na nezávislé části budeme dále využívat v části Analýza rozptylu (**ANOVA**).

Výběrový rozptyl nezávisí na zvětšení či zmenšení všech hodnot výběru o konstantu. Jestliže všechny hodnoty výběru zvětšíte **m** - krát , zvětší se výběrový rozptyl $\mathbf{m^2}$ – krát. Těchto vlastností velmi často využíváme pro úpravu původní tabulky dat tím, že všechny hodnoty posuneme - volba nového počátku a výrazně zmenšíme (zvětšíme) – volba nové jednotky.

10.3.2.4 Variační koeficient

Nechť má výběr n členů s průměrem $\overline{x}$ a směrodatnou odchylkou **s**. Potom variační koeficient výběru **v** je daný vztahem

$$v = \frac{s}{\overline{x}}.100\% \qquad (10.23)$$

Používáme ho , když chceme porovnat variabilitu různých znaků ve výběru nebo mezi různými výběry.

10.3.3 Charakteristiky tvaru rozdělení

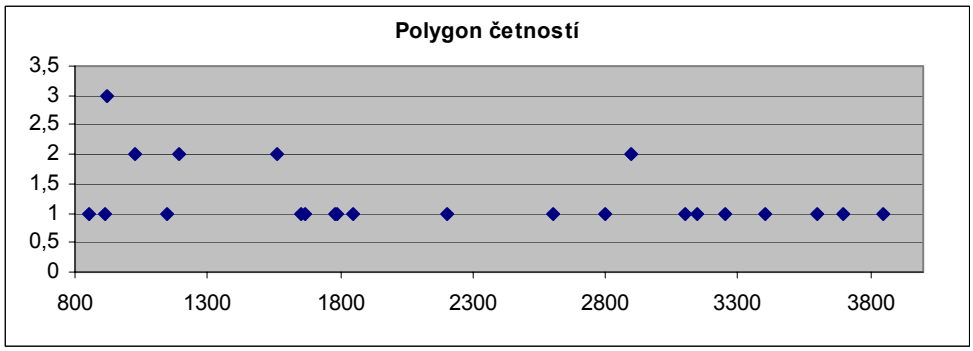
10.3.3.1 Výběrová míra šikmosti

Jde o číselný údaj, který vypovídá o o souměrnosti či nesouměrnosti tvaru rozdělení. Označuje se symbolem **a** .

$$a = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^3}{n.s^3}, \qquad (10.24)$$

kde n je počet členů výběru, s je hodnota výběrové směrodatné odchylky, $\overline{x}$ je průměr a x_i je konkrétní hodnota výběru. Je – li rozdělení souměrné, je hodnota **a** = 0. Rozdělení je tím nesousměrnější , čím se hodnota **a** více liší od nuly. Je – li jeho hodnota kladná, potom je rozdělení zešikmeno kladně (ve výběru je větší koncentrace menších hodnot). Je – li jeho hodnota záporná, potom je zešikmeno záporně (ve výběru je větší koncentrace větších hodnot).

Pokračujme s naším příkladem , s daty z tabulky 10.3. Níže vidíme data v grafu.



Hodnota míry šikmosti pro naše hodnoty **a** = 1. Je tedy kladná a data jsou zešikmena kladně.

10.3.3.2 Výběrová míra špičatosti.

Tato míra popisuje stupeň koncentrace hodnot znaku kolem charakteristiky úrovně (kolem průměru). Stejné nahuštění prostředních i krajních hodnot vede k plochosti (hodnota míry je potom záporná), větší nahuštění prostředních hodnot se projevuje špičatostí rozdělení(hodnota míry je kladná. Tato míra porovnává dané rozdělení s normovaným normálním rozdělením N(0,1) (má hodnotu špičatosti rovnu nule). Vypočte se podle vztahu

$$\mathbf{b} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^4}{n.s^4} - 3, \qquad (10.25)$$

označuje se symbolem **b**.

Hodnota špičatosti pro naše data z tabulky 10.3 je rovna – 0,93 . Rozdělení je ploché, což je vidět i z polygonu četností.

10.4 Grafické zobrazení dat

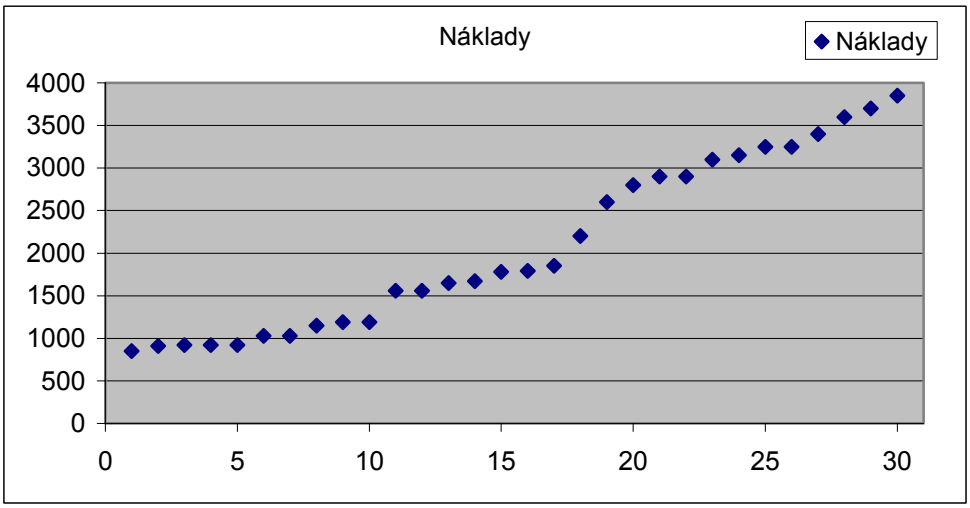
Pro presentaci statistických údajů je velmi působivé používat různé grafické způsoby. Každý typ grafického zobrazení hodnot má svoje omezení, ale zároveň i svoje výhody. Kromě klasických typů se k zobrazování statistických dat hodí speciální grafy, jeden typ jsme už měli možnost vidět v části 10.3.1.3 Kvantily a medián šlo o tzv. Box Plot neboli Krabicový graf. V dalším si ukážeme možné grafy pro presentaci údajů.

Běžné grafy

10.4.1 Bodový graf

Znázorňuje hodnoty pomocí bodů,většinou v pravoúhlé soustavě. Používá se většinou k zachycení závislostí právě dvou statistických znaků. Při více než dvou znacích jeho jednoduchost mizí a stává se méně přehledným. Nelze pomocí něho vystihnout data s větší četností.

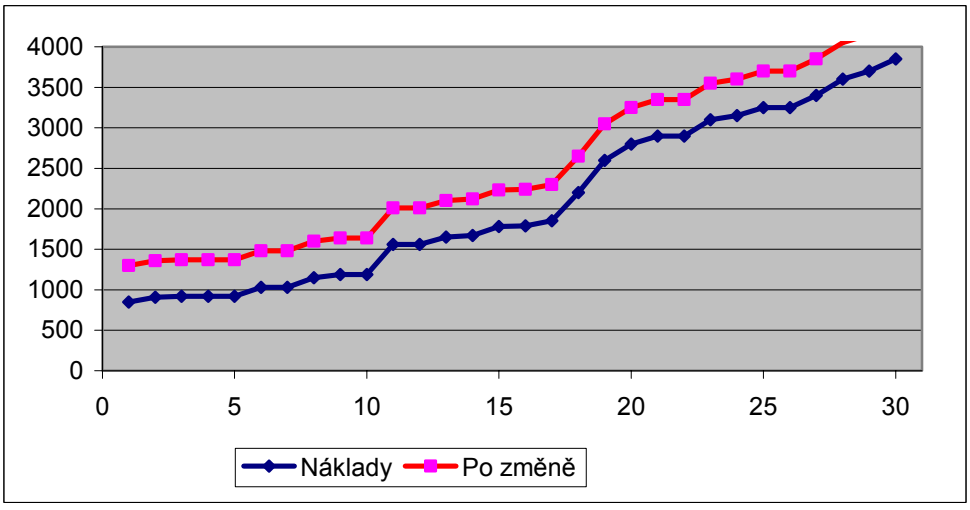
Graf 10.1 – velikost nákladů v závislosti na pořadí



10.4.2 Spojnicový graf

Jestliže chceme znázornit velké množství hodnot, chceme – li vystihnout průběh časové řady hodí se k tomu více spojnicový graf. Používá se také k vyjádření předpokladu o spojitosti vyšetřovaného znaku. Jestliže se pomocí něho vyjadřuje rozložení absolutních nebo relativních četností ve výběru , nazýváme se polygon četností.

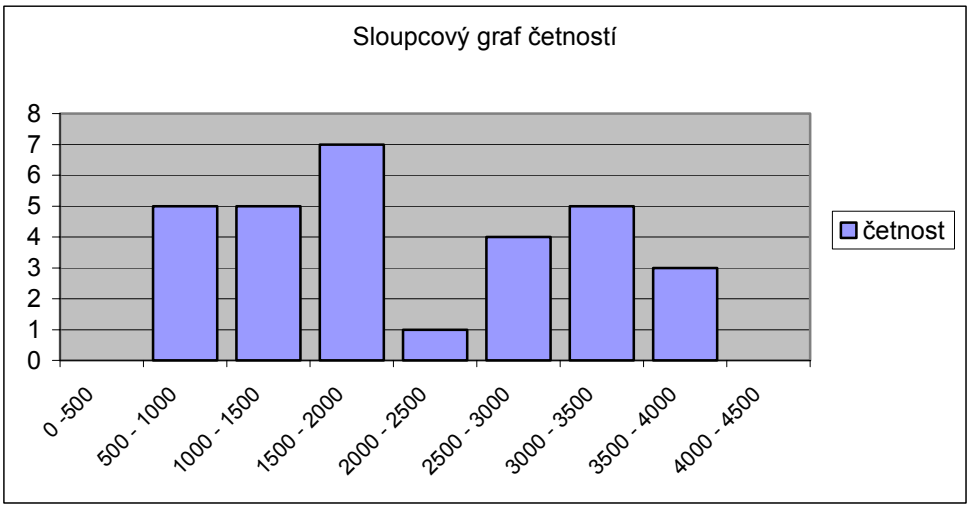
Graf 10.2 – spojnicový graf, vyjadřuje změnu nákladů



10.4.3 Sloupcový graf

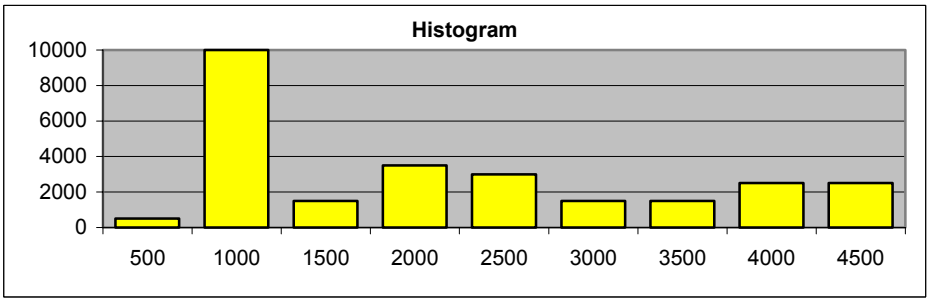
Sloupcový graf vyjadřuje jednoduché závislosti mezi dvěma hodnotami, velmi často jsou jednotlivé prvky výběru seskupovány do tříd. Existuje několik typů těchto grafů – klasické sloupcové, sloupcové s procentním rozložením, trojrozměrné sloupcové grafy. Klasická ukázka je uvedena v grafu 10.3

Graf 10.3- rozdělení nákladů do tříd



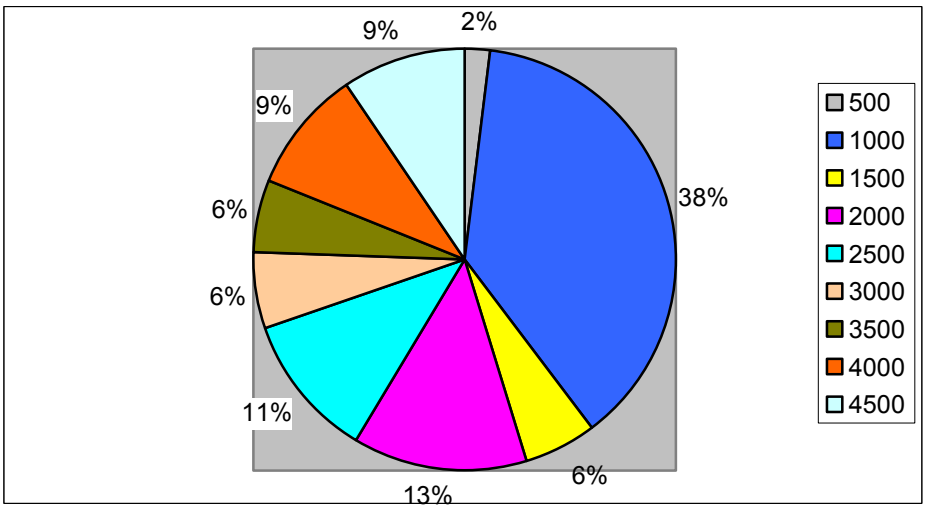
10.4.4 Histogram

Svou definicí je to sloupcový graf , který se používá k znázornění absolutních nebo relativních četností (většinou)spojitého znaku. Sloupce v grafu jsou zásadně vertikální,šířka sloupce odpovídá velikosti třídy a celková plocha sloupce odpovídá četnosti prvků třídy ve výběru.



10.4.5 Kruhový graf

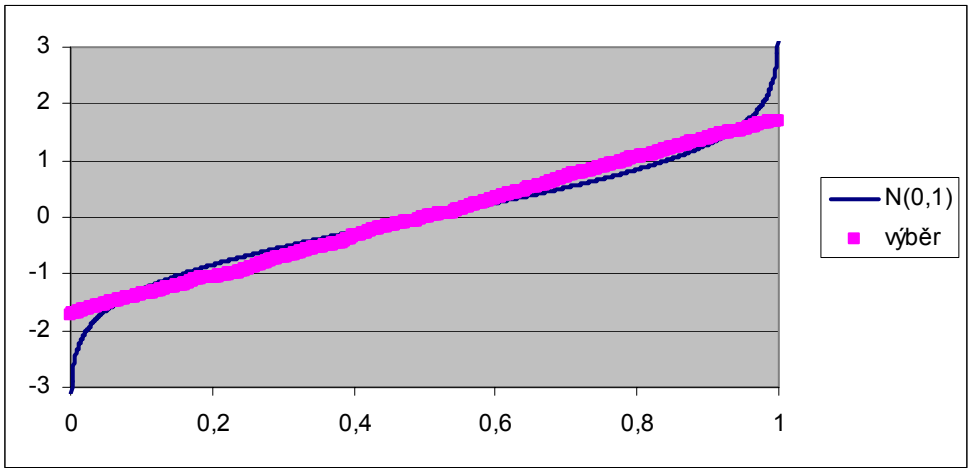
Zobrazuje hodnoty jako výseče v kruhu a tím se zachytí struktura výběru. Předchozí data jsou zobrazena v kruhovém grafu (koláč, výsečový graf) takto



Jedním z užívaných grafických způsobů je dříve uvedený **histogram**. V současné době existuje mnoho profesionálních způsobů presentace statistických dat. V části 10.3.1.3 Kvantily a medián jsme zavedli velmi užitečný typ Box Plot – český ekvivalent názvu je Krabicový graf. Statistických grafů existuje velké množství, zaměříme se na některé speciální.

10.4.6 Kvantilový graf

Jde typ grafu , kterým můžeme přehledně znázornit data, porovnat je se známými rozděleními, najít vybočující hodnoty atd. Na osu x nanášíme pořadovou pravděpodobnost teoretického rozdělení, na osu y skutečné kvantily daných dat. Na grafu níže je uvedeno porovnání výběru s N(0,1). Data se s hodnotami teoretického rozdělení neshodují, zjevně



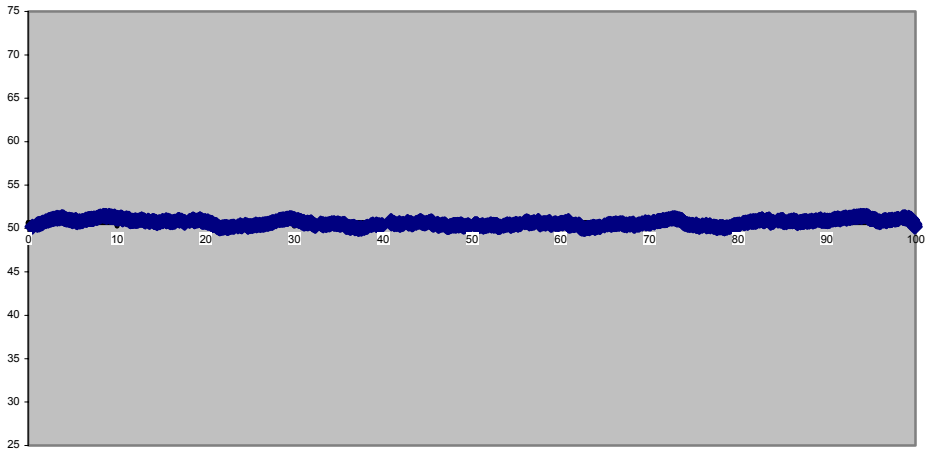
vybočují na krajích.

Tento typ grafu se velmi často užívá pro první porovnání údajů především s normálním normovaným rozdělením. Dříve se k takovému porovnání používal tzv. pravděpodobnostní papír, dnes ho provádíme s pomocí počítače.

Mezi základní statistická vyšetřování patří rozhodnutí, zda daný výběr patří nebo nepatří k rozdělím symetrickým. K takovému rozhodnutí nám pomáhá následující typ grafu:

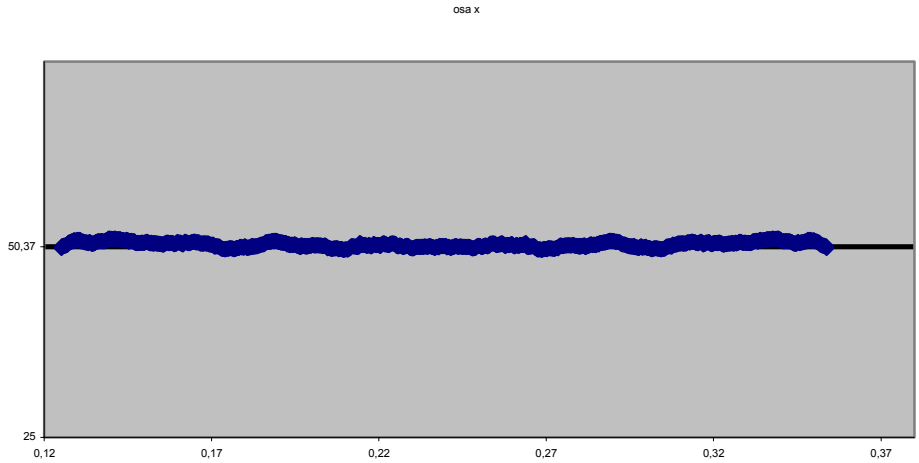
10.4.7 Graf polosum

Jeho konstrukce je založena na myšlence, že u symetrického rozdělení je aritmetický průměr kvantilu p% a kvantilu (1-p)% stejný a je roven mediánu. Níže je uveden daný graf pro data vyšetřovaná v předchozí části. Symetrická rozdělení jsou tedy charakterizována přímkou y= $\tilde{x}$. Celkově je zřejmé,že data pochází ze symetrického rozdělení.



10.4.8 Graf symetrie

Pomocí tohoto grafu je možno sledovat znak symetrie výběru. Na osu x nanášíme hodnoty $\frac{u^2_{P_i}}{2}$ *pro* $P_i = \frac{i}{n+1}$ a na osu y stejné hodnoty jako u předchozího grafu tedy hodnoty $\frac{(x_{(n+1-i)} x_{(i)})}{2}$



Opět je zřejmé, že hodnoty výběru jsou symetrické , s výjimkou krajních hodnot.
Pomocí dalšího grafu je možno srovnávat parametr špičatosti s rozdělením N(0,1).

10.4.9 Graf špičatosti

Za předpokladu symetrie je pro normální rozdělení grafem přímka. Pokud leží body na přímce s nenulovou směrnicí, je hodnota této směrnice odhadem výběrového parametru špičatosti. Opět je zřejmé, že data odpovídají symetrii, navíc můžeme z grafu odhadnout výběrovou špičatost.