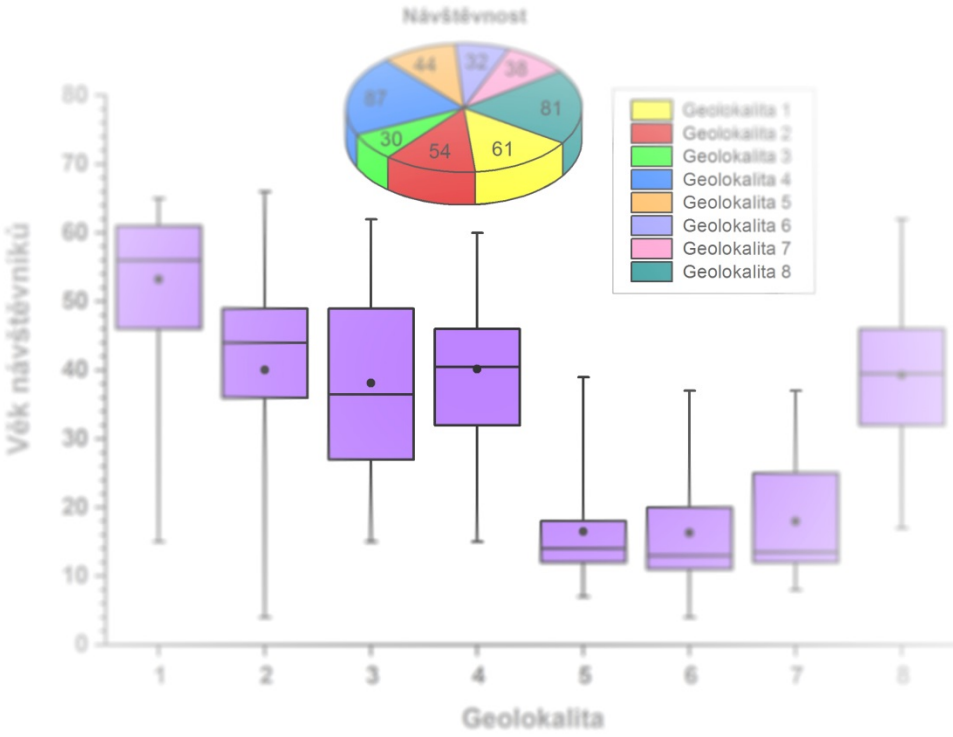
VYSOKÁ ŠKOLA BÁŇSKÁ - TECHNICKÁ UNIVERZITA OSTRAVA
Hornicko-geologická fakulta
Institut geologického inženýrství

Statistika pro Geovědní a montánní turismus

Učební texty předmětu Statistika a informatika - část Statistika



Ing. Jarmila Drozdová, Ph.D.
doc. Dr. Vladimír Homola, Ph.D.

Ostrava 2017

Recenzent: prof. Ing. Ctirad Schejbal, CSc., Dr. h. c.

Tvorba výukových materiálů „Statistika pro Geovědní a montánní turismus“ byla podpořena prostředky z grantového projektu IRP RPP2016/123 „Podpora tvorby multimediálních studijních materiálů pro bakalářské studium u předmětů vyučovaných Institutem geologického inženýrství“.

ISBN 978-80-248-4067-3

Obsah

1 Úvod

Tyto výukové texty jsou určeny studentům 1. ročníku bakalářského studia oboru Geovědní a montánní turismus. Týkají se části Statistika předmětu Statistika a Informatika. Cílem této části předmětu je popsat základní postupy při statistickém vyhodnocení dat pocházejících z oblasti ostatních studovaných předmětů a možnosti interpretace výsledků takových vyhodnocení. Výukové texty jsou pak studijním materiálem, který svým obsahem plně náplň předmětu pokrývá.

Struktura výukových textů i náplně jednotlivých odstavců jsou ovlivněny dvěma skutečnostmi. Jednak se statistická vyhodnocení evidentně opírají o matematické nástroje několika málo matematických disciplín. Dále je však nutno zohlednit fakt, že na obor nastupují studenti nejen matematiku nemilující, ale kteří ji často v předchozím středoškolském studiu neabsolvovali v potřebném rozsahu a kvalitě. Proto jsou v úvodu výukových textů shrnuty alespoň základy těch partií matematiky, které jsou použity v následných statistických metodách (kapitola *Výchozí matematické pojmy*). Studenti, kteří např. absolvovali maturitu z matematiky na gymnáziích, mohou tuto kapitolu přeskočit.

Z předchozího odstavce však vyplývá, že úvodní kapitola musí nutně vybírat z exaktně definované oblasti matematiky takové pojmy a v takových kontextech, které studenti budou schopni vstřebat. Autoři výukových textů zde byli postaveni před nelehkou volbu: zda upřednostnit jednoduchost a názornost (na úkor exaktnosti a preciznosti) nebo striktní ekvivalenci se současným stavem matematických věd (na úkor pochopení podstaty a zejména potřebného rozsahu). Přiklonili se spíše k první možnosti, ovšem v žádném případě nejsou použita zavádějící zjednodušení. K mnoha pojmům (v souvislosti se směřováním k použití ve statistice) lze dojít několika cestami; v tom případě autoři zvolili tu, kterou pokládají za nejsnáze pochopitelnou, aniž by utrpěla obecnost. Např. pojem *Přirozené číslo* by matematik zavedl nejčastěji pomocí algebraických struktur, monoidů nebo alespoň Peanových axiomů. To se pro zamýšlený účel učebních textů zdálo autorům zbytečné, zvláště u přirozených čísel se spoléhají na intuitivní chápání. Obdobná situace nastala u jednoho ze základních pojmů, *množina*.

Učební texty obsahují řadu názorných příkladů včetně interpretace výsledků. Příklady jsou řešeny podle své povahy buď "ručně" (např. kalkulačkou), tabulkovým procesorem (např. programem Excel), nebo specializovaným statistickým software (např. programem Statgraphics). K poslednímu jmenovanému: tento program ve verzi Statgraphic Plus 5.0 je uváděn proto, že VŠB - TU Ostrava zakoupila jeho multilicenci, zaměstnanci ho mají volně ke stažení a je nainstalovaný na mnoha univerzitních učebnách. Týká se to rovněž všech počítačů katedrální učebny J-409, která je studentům běžně přístupná.

2 Výchozí matematické pojmy

V této kapitole je podán souhrn matematických pojmů a značení, který bude použit ve výkladu jednotlivých statistických metod. Většina by měla být známa z předchozího středoškolského studia. Protože se však obecně může používaná symbolika a terminologie v detailech lišit, doporučují autoři kapitolu minimálně zběžně přehlédnout.

2.1 Výroky

Při sdělování skutečností používáme v běžném životě oznamovací věty jednoduché a souvětí podle gramatiky daného jazyka (čeština, angličtina ...). Takové sdělení je tedy tvrzením o nějaké skutečnosti, která však může být pouze domnělá, nemusí mít obecnou platnost. Právě zkoumání platnosti tvrzení a okolností, za kterých se považují za platná, je základním nástrojem vědních oborů.

Důležitá poznámka: *V mnoha publikacích věnovaných výrokům je možno se setkat s termínem "pravdivé tvrzení". V těchto výukových textech se tento termín zásadně nepoužívá (s výjimkou jediného historického místa - "pravdivostní tabulka"). Jednak jde o termín navýsost filozofický; běžný občan vlastně ani nedovede říct, co to pravda vůbec je. Podstatnější však je (a při teoretických úvodech do mnoha partií zvláště matematiky je to zřetelné), že při prvním poslechu některých tvrzení student hned vyhrkne: "to není pravda" popř. "to nemůže být pravda". Příkladem může být tvrzení: dvěma různými body může procházet více různých přímek. Právě z toho důvodu je v těchto textech použito obratu: "tvrzení se považuje za platné" - což je něco zcela jiného, než že to je pravda. Vždyť považuje-li se shora uvedené tvrzení o bodech a přímkách za platné, vybuduje se zcela jiná geometrie než když se považuje za neplatné (neeukleidovské vs. eukleidovské geometrie).*

Příkladem tvrzení, které lze v matematice dovodit za platné, je tato oznamovací věta:

> Druhá mocnina součtu dvou hodnot je rovna součtu druhých mocnin obou hodnot zvětšeného o dvojnásobek součinu obou hodnot.

Obdobný příklad z fyziky: Zjištění dráhy při přímočarém rovnoměrně zrychleném pohybu lze vyjádřit např. touto oznamovací větou:

> Dráha, kterou urazí těleso při rovnoměrně zrychleném pohybu za jistý čas, je rovna polovině součinu hodnoty zrychlení a druhé mocniny doby pohybu.

Je zřejmé, že takové vyjádření sice možné je. Jednak je ale vázáno na konkrétní humánní jazyk a tatáž skutečnost může být vyjádřena několika různě stavěnými větami. Za druhé, samo o sobě není příliš přehledné. To - spolu s rozvojem vědních oborů - vedlo k nutnosti formalizovat zápisy vyjadřující skutečnosti (nejen závěry, ale i předpoklady) daného vědního oboru. Pomocí takové formalizace lze shora uvedený příklad z matematiky zapsat

$$(A + B)^2 = A^2 + 2.A.B + B^2$$

a shora uvedený příklad z fyziky zapsat

$$s = ½ . a . t^2$$

Tyto učební texty se týkají některých oborů matematiky, proto evidentně musí používat symboliku umožňující formalizovaně popsat v těchto oborech používané pojmy (definice) a vztahy mezi nimi (věty). Základním nástrojem je oblast matematické logiky - hraniční disciplíny mezi matematikou a logikou, zkoumající aplikaci formální logiky v matematice. Zde budou podány jen nejnutnější partie zvláště výrokové logiky, které jsou potřebné pro zavedení používané symboliky. Jak bylo zdůrazněno v úvodu, nepůjde o ucelený výklad teorie, ale o účelově vybranou skupinu definic a vět (většinou bez důkazů).

2.1.1 Jednoduchý a složený výrok

Výroková logika: obor, který se zabývá studiem výroků z hlediska jejich platnosti. Je základem formálního odvozovacího systému.

Jednoduchý výrok, podle některých autorů také **Atomický výrok**: takové - z logického hlediska nedělitelné, neobsahující logické spojky ani "pod-výroky" - tvrzení, které lze ohodnotit z hlediska přijaté platnosti (viz dále). Obvykle je vyjádřeno jednoduchou oznamovací větou nebo ekvivalentní posloupností definovaných symbolů.

> Příklad: *Vidím bílý kulatý Měsíc*. Z hlediska uvažovaného kontextu nejde o jednoduchý výrok. V podstatě je tím totiž řečeno: *Vidím bílý Měsíc a současně vidím kulatý Měsíc*.

Abeceda jazyka výrokové logiky: Nechť S označuje jeden nebo více jednoduchých výroků. Abecedou A_S jsou jednak všechny výroky S, a dále logické spojky: symboly ¬ (symbol **negace**) a → (symbol **implikace**).

Výrok V v abecedě A_S je:

1. Každý jednoduchý výrok z S
2. Zápis ¬ V (přijímá se pro něj označení Logická operace negace)
3. Zápis V → V (přijímá se pro něj označení Logická operace implikace)
4. (V)

Z definice výroku je zřejmé, že každý výrok je složen z konečného nenulového počtu jednoduchých výroků. Řečeno jinak, každý výrok je buď jednoduchým výrokem, nebo ho tvoří konečná posloupnost jednoduchých výroků, oddělených logickými spojkami a (nebo) kulatými závorkami.

Příklad: Jsou-li A, B a C jednoduché výroky, je výrokem např. ¬ A → (B → ¬ C).

Atomy výroku V: Ty jednoduché výroky, které výrok V tvoří ve smyslu předchozího odstavce.

Složený výrok: Výrok, který není jednoduchým výrokem.

2.1.2 Ohodnocení platností, platnost výroku

Ohodnocení platností v abecedě A_S: Mějme dva libovolné, ale různé symboly. Pro účely dalšího výkladu použijme např. symboly I a O. Nechť je každému (jednoduchému) výroku z S přiřazen právě jeden z těchto symbolů. Ohodnocení platností je pak pravidlo p, které každému výroku V v abecedě A_s přiřadí právě jeden ze symbolů I a O následovně:

1. p(¬ V) = O, je-li p(V)=I
2. p(¬ V) = I, je-li p(V)=O
3. p(V → W) = I, je=li p(V)=O a p(W)=O
4. p(V → W) = I, je=li p(V)=O a p(W)=I
5. p(V → W) = O, je=li p(V)=I a p(W)=O
6. p(V → W) = I, je=li p(V)=I a p(W)=I
7. p((V)) = p(V)

Platnost výroku: Jestliže pro výrok V je p(V) = I, označuje se výrok V za platný. Jestliže pro výrok V je p(V) = O, označuje se výrok V za neplatný.

Pravdivostní tabulka: Tabulkové vyjádření ohodnocení platností složeného výroku V. Levými sloupci takové tabulky jsou ohodnocení platností atomů výroku V, v pravém sloupci (pravých sloupcích) ohodnocení platností výroku V. Jednoduchým příkladem může být pravdivostní tabulka výrazu, který je logickou operací implikace (viz výše):

X	Y	X → Y
O	O	I
O	I	I
I	O	O
I	I	I

Konjunkce, disjunkce, ekvivalence: Výrok V nazýváme

- konjunkcí výroků X a Y, je-li V tvaru ¬(X → ¬Y), a značíme V = X ∧ Y
- disjunkcí výroků X a Y, je-li V tvaru ¬X → Y, a značíme V = X ∨ Y
- ekvivalencí výroků X a Y, je-li V tvaru (X → Y) ∧ (Y → X), a značíme V = X ≡ Y

Pravdivostní tabulka zavedených pojmů je následující:

X	Y	X ∧ Y	X ∨ Y	X ≡ Y
O	O	O	O	I
O	I	O	I	O
I	O	O	I	O
I	I	I	I	I

Předchozí tabulka (ve spojení s označováním platný - neplatný) dokresluje zavedené verbální vyjádření konjunkce, disjunkce, ekvivalence a implikace:

- ∧ se čte "a současně": výrok X ∧ Y je platný jen tehdy, jsou-li současně platné výroky X i Y.
- ∨ se čte "nebo": výrok X ∨ Y je platný, je-li platný výrok X nebo je platný výrok Y nebo jsou platné oba.
- ≡ se čte "právě tehdy": výrok X ≡ Y je platný, jsou-li oba výroky současně platné nebo současně neplatné (oba výroky jsou si ekvivalentní).
- → se čte "jestliže ... pak": výrok X → Y se tedy čte "jestliže platí výrok X, pak také platí výrok Y". Už z toho je zřejmé, že určitě neplatí situace, kdy při platnosti X je Y neplatný. Ve všech ostatních případech implikace platí.

Tautologie: Složený výrok, jehož platnost je vždy I bez ohledu na platnosti jeho atomů.

Důležitým příkladem tautologie může být výraz V: (X → Y) ≡ (¬Y → ¬X). Jeho pravdivostní tabulka je:

X	Y	¬ X	¬ Y	A = X → Y	B = ¬ Y → ¬ X	A ≡ B
O	O	I	I	I	I	I
O	I	I	O	I	I	I
I	O	O	I	O	O	I
I	I	O	O	I	I	I

Důležitost této tautologie spočívá v možném ekvivalentním vyjádření výroku X → Y (jestliže platí X, pak také platí Y) výrokem "jestliže neplatí Y, pak také neplatí X".

Příklad 1: Výrok "Jestliže je číslo dělitelné šesti, pak je také dělitelné třemi" je ekvivalentní výroku "jestliže číslo není dělitelné třemi, pak není dělitelné šesti".

Příklad 2: Nechť V je výrok "Nebude-li pršet, nezmoknem", což je jen jinak zapsaný výrok "Jestliže nebude pršet, pak nezmokneme". Označíme-li A výrok "Nebude pršet" (jeho negace je výrok "Bude pršet") a B výrok "Nezmokneme" (jeho negace je výrok "Zmokneme"), pak podle předchozího je výrok V ekvivalentní výroku "Jestliže zmokneme, pak bude pršet" - což je celkem logické: jak jinak bychom zmokli, když ne v dešti ☺

2.1.3 Kvantifikátory

Až doposud vypovídaly výroky o vlastnostech několika konkrétních objektů (bez ohledu na to, zda šlo o výroky platné nebo neplatné). Exaktní vědy jsou však budovány na výrocích, které přisuzují přítomnost nebo nepřítomnost nějaké vlastnosti ne jednomu nebo několika málo objektům, ale celé třídě velmi mnoha až nekonečně mnoha objektů. Je zřejmé, že shora uvedenými prostředky pak nelze reálně sestavit výrok, vyjadřující přítomnost nebo nepřítomnost vlastnosti pro každý jednotlivý objekt.

Míra přítomnosti nějaké vlastnosti studovaných objektů popsané výrokem V je pak principielně dvojí:

a. **Všechny** studované objekty mají danou vlastnost (pro všechny studované objekty platí výrok V)
b. **Alespoň jeden** studovaný objekt má danou vlastnost (existuje alespoň jeden ze studovaných objektů, pro který platí výrok V).

***Poznámka**: je třeba si uvědomit, že samotná tvrzení a a b jsou výroky, které mohou být platné nebo neplatné. Aby mohly být použity pro další budování daného vědního oboru, je třeba jejich platnost nebo neplatnost dokázat. Důkaz jako demonstrace platnosti výrazu za určitých předpokladů je založen výhradně na použití dříve dokázaných platných výrazů nebo takových výrazů, které se za platná považují z důvodu nezpochybňovaných, životem mnohokrát ověřených a obecně přijatých pravidel (tzv. axiomů). Otázky soustav axiomů a způsobů důkazů však přesahují rozsah a zaměření této publikace.*

Všeobecný kvantifikátor: symbol ∀ použitý při konstrukci výroku ad a. ve výčtu shora.

Existenční kvantifikátor: symbol ∃ použitý při konstrukci výroku ad b. ve výčtu shora.

Přestože kvantifikované výroky lze vyjádřit běžným hovorovým jazykem, pro účely těchto učebních textů bude podáno pouze použití obecně v matematice. V obou typech kvantifikovaných výroků je především třeba vymezit, kterých objektů se výrok V týká (tedy jednoznačně určit, co znamená shora použitý vágní pojem "studovaný objekt"). Současně se většinou zvolí nějaké symbolické označení pro jeden každý z těchto objektů. Následuje samotný výrok, ve kterém lze zvoleného symbolického označení použít.

Příklad použití všeobecného kvantifikátoru: Je zřejmé, že je-li celé číslo dělitelné šesti, je také dělitelné třemi. To lze formulovat i jinak: pro všechna celá čísla (označme jedno každé např. X) platí: je-li X dělitelné šesti, pak je také dělitelné třemi. Pomocí všeobecného kvantifikátoru se stejné tvrzení zapíše např. takto:

∀ X, X je celé číslo: (X je dělitelné 6) → (X je dělitelné 3)

Analogicky se použije existenční kvantifikátor.

2.2 Základy teorie množin

Cílem kapitoly je precizovat základní pojmy teorie množin, v předchozím studiu často poměrně volně zavedené. Proto v této kapitole jako jediné je ukázána geneze části vědního oboru posloupností definice - věta - důkaz.

2.2.1 Množina

Definice (*Cantorova pseudodefinice množiny*): Množinou rozumíme souhrn libovolných, ale přesně určitelných a rozlišitelných objektů reálného světa nebo našeho nazírání nebo myšlení, shrnutých v jeden logický celek. Tyto objekty nazýváme *prvky množiny*.

Je zřejmé, že pro každou množinu existuje pravidlo, podle kterého lze rozhodnout, zda libovolný objekt reálného nebo imaginárního světa patří nebo nepatří dané množině (*určující pravidlo* množiny). Toto pravidlo mívá mnoho podob: výrok, matematickou formuli, výčet prvků apod.

Označení: Množiny se většinou označují velkými latinskými písmeny, prvky množin malými latinskými písmeny. Skutečnost, že nějaký objekt a je prvkem množiny P, zapisujeme

$$a \in P, \text{ popř. } P \ni a$$

a čteme: *a je elementem (prvkem) P*, popř. *P obsahuje a*. Skutečnost, že nějaký objekt a ***není*** prvkem množiny P, zapisujeme

$$a \notin P$$

Definice: Množina se nazývá *konečná*, má-li konečně mnoho prvků. Každá množina, která není konečná, je *nekonečná*. Přitom konečná množina s nulovým počtem prvků se nazývá *prázdná* množina a označuje se ∅, {∅} nebo jen {}.

Označení: Případ definování (konečné) množiny *výčtem prvků* se symbolicky zapisuje

$$P = \{a_1, a_2, ..., a_n\}$$

Případ definování množiny pomocí *vlastnosti* V se symbolicky zapisuje

$$P = \{ a, a \in D: V(a) \}$$

kde D určuje zdroj prvků; čte se: množina P je množina všech takových a z D, pro které vlastnost V(a) je splněna.

2.2.2 Poznámka k zavedení pojmu Množina

Způsob zavedení pojmu Množina, který je uvedený výše, je jen jeden z možných, a je zřejmě jeden z historicky prvních. Autorem je George Cantor, jehož teorie bývá označována jako *naivní* nebo *intuitivní teorie množin*. V těchto textech je použita proto, že pojmy takto zavedené jsou pro účely vyučovaného předmětu (a nejen jeho, ale i pro většinu ostatních matematických disciplín) zcela postačující, pro naprostou většinu studentů názorné a nerozptylují je při studiu metod statistického zpracování dat. Na druhé straně tvoří seriózní základ pro případné rozšiřující studium této oblasti.

Je zde na místě uvést důvody, proč např. mechanická aplikace takto zavedených pojmů může být zavádějící.

Celkem brzy se totiž ukázalo, že shora uvedená Cantorova pseudodefinice je z hlediska přesnosti nedostatečná, že je třeba daleko přesněji určit co vlastně množina je a co není. Vyniklo to zejména v případech nekonečných množin a množin podmnožin a množin.

Příkladem může být množina definována pomocí vlastnosti: Množina V je množina všech množin, které neobsahují sama sebe. To je ovšem nesmyslné: když se sama neobsahuje tak se má sama obsahovat!

> ***Poznámka**: Protože ve vědě, kterou je matematika, se nemůže předchozí situace označit za "pěknou blbost", používá se termín **paradox**. Právě uvedený příklad je jen do hovorové řeči přepsaný tzv. Russellův paradox: Mějme množinu A všech množin B takových, že $B \notin B$. Pro takto definovanou množinu A nemůže nastat ani případ $A \in A$ ani $A \notin A$. Kdyby totiž bylo $A \notin A$, pak podle definice A do A patří, ovšem kdyby bylo $A \in A$, pak A do A nemůže patřit.*

Právě při shora popsaném způsobu zavedení množin je paradoxů více. Velmi známým je **paradox holiče**:

> Ve kterém městě existuje jediný holič, který holí všechny ty muže, kteří se neholí sami? Uvažme, že pokud se holí sám, tak se neholí sám, ale pokud se neholí sám, tak se holí sám.

Nejasná místa, kterých je v intuitivní teorii množin více, je třeba upřesnit, "dodefinovat". Jak je to např. s případným násobným výskytem prvku: jsou množiny {1, 2, 3} a {1, 1, 1, 2, 2, 3, 3, 3, 3} stejné nebo různé? Jsou množiny {1, 2, 3} a {3, 2, 1} stejné nebo různé? Právě tyto otázky byly řešeny jako jedny z prvních, a to upřesněním pojmu rovnosti dvou množin:

$$A = B \equiv \forall\, x: x \in A \equiv x \in B$$

Mimochodem paradox lze najít už ve výrokové logice popsané shora. Mějme výrok V = (Tento výrok je nepravdivý). Jestliže totiž je výrok V pravdivý, pak je nepravdivý, ovšem je-li výrok V nepravdivý, tak je pravdivý.

2.2.3 Podmnožina, sjednocení, průnik, rozdíl

Definice, označení: Jsou-li P a Q dvě množiny a je-li každý prvek množiny P současně prvkem množiny Q, říkáme, že množina P je *podmnožinou* množiny Q a zapisujeme

$$P \subseteq Q$$

Je-li současně $Q \subseteq P$, pak říkáme, že množiny P a Q jsou totožné a zapisujeme

$$P \equiv Q$$

Označení: Je-li $P \subseteq Q$, ale není-li současně $P \equiv Q$, pak zapisujeme

$$P \subset Q$$

Prázdná množina je podmnožinou každé množiny. Pro libovolnou množinu P tedy platí: $\{\emptyset\} \subseteq P$.

Definice: *Sjednocení* $P \cup Q$ dvou libovolných množin P, Q je množina definovaná takto:

$$P \cup Q = \{ x: x \in P \vee x \in Q \}$$

Definice: *Průnik* $P \cap Q$ dvou libovolných množin P, Q je množina definovaná takto:

$$P \cap Q = \{ x: x \in P \wedge x \in Q \}$$

Definice: *Rozdíl* $P - Q$ dvou libovolných množin P, Q je množina definovaná takto:

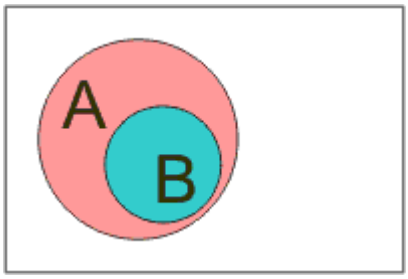
$$P - Q = \{ x: x \in P \wedge x \notin Q \}$$

Definice: Dvě množiny jsou *disjunktní*, je-li jejich průnik prázdný (tedy neobsahují žádný společný prvek).

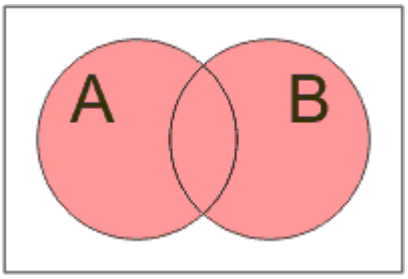
Definice: Nechť **S** je množina množin, tj. $S = \{ A_1, A_2, \ldots, A_n \}$, kde každé A_i je množinou. Takovou množinu nazýváme *systémem množin* A_i. *Sjednocením* $\cup\mathbf{S}$ systému **S** nazýváme množinu $A_1 \cup A_2 \cup \ldots \cup A_n$. *Průnikem* $\cap\mathbf{S}$ systému **S** nazýváme množinu $A_1 \cap A_2 \cap \ldots \cap A_n$.

Definice: Systém **S** je *disjunktním systémem*, platí-li pro libovolné dvě množiny A_i, A_j z S, že jejich průnik je prázdný: $A_i \cap A_j = \{\emptyset\}$ - tedy že žádné dvě nemají společný prvek.

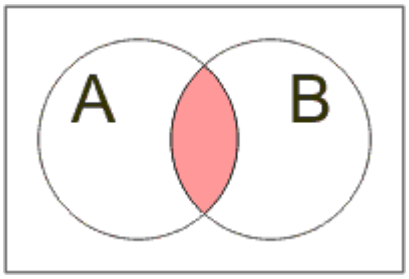
Vzájemné vztahy mezi množinami názorně graficky vyjadřují Vennovy diagramy (poprvé je kolem roku 1880 použil John Venn):



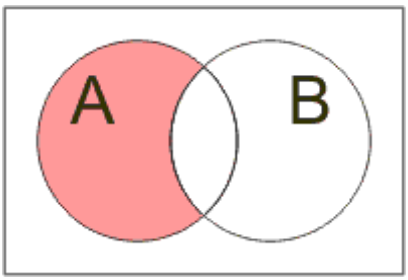
Podmnožina: $B \subset A$



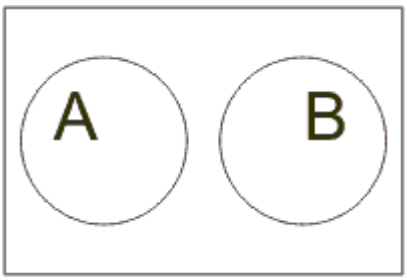
Sjednocení: $A \cup B$



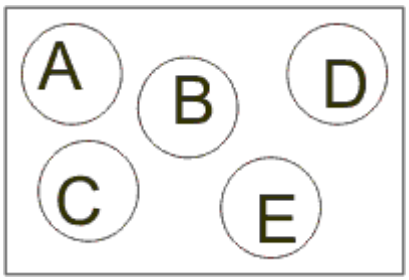
Průnik: $A \cap B$



Rozdíl: A − B



Disjunktní A a B



Disjunktní systém množin

2.2.4 Kartézský součin

Definice: *Kartézský součin* množin A, B (značení: A × B) je množina všech uspořádaných dvojic takových, že první prvek dvojice je prvkem A a druhý prvek dvojice prvkem B:

$$A \times B = \{ [a,b]: a \in A \wedge b \in B \}$$

Obdobně kartézský součin množin A, B, C (značení: A × B × C) je množina všech uspořádaných trojic:

$$A \times B \times C = \{ [a,b,c]: a \in A \wedge b \in B \wedge c \in C \}$$

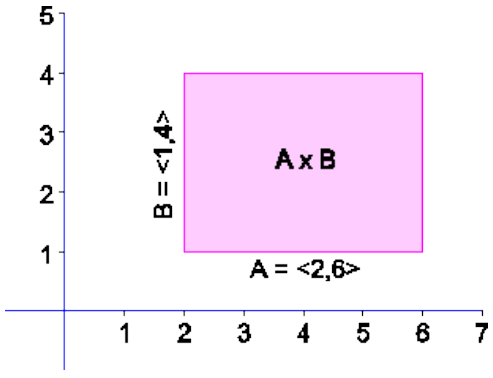
atd. Je-li jedna z množin prázdná, je i kartézský součin prázdná množina.

Označení: Součin A × A označujeme také $\mathbf{A^2}$, M × M × M × M × M označujeme také $\mathbf{M^5}$ atd.

Jsou-li množiny A a B konečné s počty prvků n_A a n_B, je i jejich kartézský součin A × B konečná množina; počet jejich prvků (=počet uspořádaných dvojic) je roven $n_A \times n_B$ - dvojice obsahují kombinace "každý z A s každým z B". Pro grafické zobrazení kartézského součinu se pro množiny s malým počtem prvků může použít tabulka; např. pro {červená, zelená, žlutá} × {jablko, hruška}:

	jablko	**hruška**
červená	červené jablko	červená hruška
zelená	zelené jablko	zelená hruška
žlutá	žluté jablko	žlutá hruška

I v některých případech nekonečných množin lze s výhodou kartézský součin znázornit graficky. Je-li např. A=<2,6> a B=<1,4> (uzavřené intervaly reálných čísel), pak znázornění A × B může být např. následující:



2.2.5 Relace

Definice: *Binární relace* R z množiny A do množiny B je libovolná podmnožina kartézského součinu A × B:

$$R \subseteq A \times B$$

Poznámka: *Kartézský součin dvou množin jsou tedy všechny kombinace typu [a, b], kdežto relace jen některé.*

Označení: Je-li [a,b] ∈ R, píšeme: aRb a čteme: prvku a je v relaci R přiřazen prvek b, nebo: prvek a je v relaci R s prvkem b. Je-li naopak [a,b] ∉ R, píšeme a$\bar{R}$b a čteme: prvek a není v relaci R s prvkem b.

Označení: Je-li R ⊆ A × A, pak R nazýváme *relací v množině* A.

> Příklad: V množině A = {3, 5, 7, 9} je dána relace R = { [3,5], [3,7], [3,9], [5,7], [5,9], [7,9] }. Je tedy např. 3R7, 7R9, ale 9$\bar{R}$3.

Jsou-li množiny A a B konečné, lze pro znázornění relací použít několika způsobů. Nejčastěji používané jsou dva: maticový a tabulkový. Maticovým zápisem relace R z předchozího příkladu je následující matice 4x4 (nad matici resp. před matici byly pro přehlednost přidány nadpisy sloupců resp. řádků); v ní hodnota **0** značí, že prvky v relaci nejsou, hodnota **1** značí, že prvky v relaci jsou:

(a↓) R (b→)	**3**	**5**	**7**	**9**
3	0	1	1	1
5	0	0	1	1
7	0	0	0	1
9	0	0	0	0

Tabulkovým zápisem relace R z předchozího příkladu je následující tabulka:

a ∈ A	**b ∈ B**
3	5
3	7
3	9
5	7
5	9
7	9

Definice: *n-ární relace* je libovolná podmnožina R kartézského součinu $A_1 \times A_2 \times ... \times A_n$.

Maticové zobrazení n-árních relací pro větší n je velmi nepraktické a nepřehledné. Proto se relace s konečným (často i značným) počtem prvků zobrazují výhradně jako n-sloupcové tabulky - pokud se samozřejmě nedají vyjádřit jinak, např. symboly výrokového počtu apod.

2.2.6 Vlastnosti binárních relací

Tento odstavec se týká pouze relací tvaru R ⊆ A × A, tj. relací v množině A. Tyto relace mohou mít některé vlastnosti (např. že pro žádný prvek a ∈ A neobsahují dvojici [a, a]). V následující tabulce jsou definovány některé základní typy relací podle svých vlastností:

Relace R je ...	**... právě když platí:**
reflexivní	∀ x ∈ A : xRx
symetrická	∀ x, y ∈ A : xRy → yRx
tranzitivní	∀ x, y, z ∈ A : xRy ∧ yRz → xRz
areflexivní	∀ x, y ∈ A : xRy → x≠y
antisymetrická	∀ x, y ∈ A : xRy ∧ yRx → x=y
ekvivalence	R je reflexivní, symetrická, tranzitivní
(neostré) uspořádání	R je reflexivní, antisymetrická, tranzitivní
ostré uspořádání	R je areflexivní, tranzitivní

> Příklad: Nechť je dána relace R v {3, 5, 7, 9} - viz příklad shora. Tato relace je areflexivní (pro všechna [x,y] ∈ R je x ≠ y) a tranzitivní (3R5 ∧ 5R7 → 3R7; 3R7 ∧ 7R9 → 3R9; 5R7 ∧ 7R9 → 5R9). Relace je tedy **ostré uspořádání**; tato relace se často místo obecného R značí "<". Je tedy 3<5, 3<7, 3<9, 5<7, 5<9, 7<9.

2.2.7 Zobrazení

Definice: F je *zobrazení* z A do B, právě když je F (binární) relace z A x B a současně platí:

$$\forall a \in A, b \in B, c \in B : aFb \wedge aFc \rightarrow b=c$$

jinak řečeno: jednomu a z A "odpovídá" v zobrazení F nanejvýš jedno b z B.

Příklad: Relace **<** z předchozího příkladu není *zobrazení* z A do A, protože je jednak 3<5 a jednak např. 3<7. Prvek 3 není v relaci s nejvýš jedním prvkem.

Označení: Shora definované zobrazení se také značí F: A ⇒ B. Skutečnost, že xFy, se také zapisuje y=F(x).

Definice: Je-li F zobrazení, aFb, pak se prvek a nazývá **vzor** prvku b a prvek b **obrazem** prvku a.

Definice: Množina všech vzorů zobrazení F se nazývá *definiční obor* Δ(F) zobrazení F. Množina všech obrazů zobrazení F se nazývá *obor hodnot* Φ(F) zobrazení F:

$$\Delta (F) = \{ a : a \in A \wedge \exists b \in B : [a,b] \in F \}$$

$$\Phi (F) = \{ b : b \in B \wedge \exists a \in A : [a,b] \in F \}$$

Definice: Je-li Δ(F) = A, pak se F nazývá zobrazení (celé) A. Je-li Φ(F) = B, pak se F nazývá zobrazení na B. Existují tedy celkem čtyři zobrazení, terminologicky zachycená takto:

F : A⇒B	**Δ(F) ⊂ A**	**Δ(F) = A**
Φ(F) ⊂ B	zobrazení z A do B	zobrazení A do B
Φ(F) = B	zobrazení z A na B	zobrazení A na B

Definice: Mějme F: A ⇒ B. Zobrazení F se nazývá *prosté*, platí-li:

$$\forall a \in A, b \in A, c \in B : aFc \wedge bFc \rightarrow a=b$$

jinak řečeno: dva různé vzory z A nemohou mít v prostém zobrazení F stejný obraz v B.

Příklad: Na množině I přirozených čísel (I = {1, 2, 3, ...}) je dána relace (označme ji např. symbolem trojlístku ♣) takto: x ♣ y ≡ y=x+1. Tato relace je tedy množinou dvojic typu [x, x+1], kde x ∈ I. Je-li x♣y (tedy y=x+1) a současně x♣z (tedy z=x+1), je evidentně y=z. Relace ♣ je tedy zobrazení. Definičním oborem je celá množina I: každému vzoru x∈I odpovídá - dokonce právě jeden - obraz y=x+1∈I. Oborem hodnot však není celá množina I (prvek 1∈I není obrazem žádného vzoru x∈I; mělo by být 0♣1, ale 0∉I). Zobrazení ♣ je tedy zobrazení (celé) I do I (nikoliv na I).

2.2.8 Operace

Definice: Nechť $\mathbf{M}_1$, $\mathbf{M}_2$, ... , $\mathbf{M}_n$, **V** jsou libovolné množiny. **F** *je n-ární operace* z $M_1 \times M_2 \times ... \times M_n$ do V, je-li F (n+1)-ární relace, $F \subseteq M_1 \times M_2 \times ... \times M_n \times V$, a platí-li:

$$\forall a_i \in M_i, \forall z,u \in V : [a_1, a_2, ... , a_n, z] \in F \wedge [a_1, a_2, ... , a_n, u] \in F \rightarrow z=u$$

Definice: n-tice $\mathbf{a}=[a_1, a_2, ... , a_n]$ se nazývá *operandy* operace F, prvek z∈V se nazývá *hodnota operace* F na operandech **a**. Velmi často se zapisuje z = F (**a**) nebo $z = F (a_1, a_2, ... , a_n)$.

Poznámka: Zobrazení definované v předchozím odstavci je unární operací ve smyslu definice n-ární operace. Obecně n-ární operace má n operandů. Specielně nulární operace nemá žádný operand; je pak z=F(), a protože z∈V, je nulární operace množina s nejvýš jedním prvkem. Nulární operace slouží k výběru tohoto prvku.

Příklad: Zobrazení ♣ z příkladu předchozího odstavce je tedy unární operací z I do I. Je-li y=♣ x, je y=x+1. Je proto ♣3=4, ♣28=29, ♣(x+5)=x+6 pro x∈I atd.

2.2.9 Rozklad na třídy

Definice: Nechť je dána neprázdná množina P a neprázdný systém **T** = { A_1, A_2, ... , A_n }, kde pro každé i je $A_i \subseteq P$ (**T** je tedy systémem podmnožin množiny P). Tento systém **T** nazýváme *rozkladem množiny* P *na třídy*, jestliže **T** je disjunktním systémem a $\cup$**T** = P. Každá $A_i \in$ **T** se nazývá třída v **T**.

T je tedy rozkladem P na třídy, jestliže průnik libovolných dvou tříd je prázdný (žádné dvě nemají společný prvek) a jejich sjednocením je celá P (tedy celá původní rozkládaná množina).

Jak bylo shora uvedeno, každé množině přísluší určující pravidlo, podle kterého byla množina vytvořena. Rozklad na třídy **T** je systémem podmnožin, je tedy sám množinou, existuje tedy i pro něj určující pravidlo. Velmi často mívá toto určující pravidlo matematickou podobu.

> Příklad: Barvy {červená, modrá, zelená, žlutá, jiná} rozkládají množinu všech jednobarevných aut na třídy. Tento rozklad **B** je tvořen pěti třídami (podmnožina červených aut, podmnožina modrých aut, podmnožina zelených aut, podmnožina žlutých aut, podmnožina aut jakékoliv jiné barvy). Sjednocení všech těchto tříd je celá množina všech jednobarevných aut, a každé dvě třídy jsou disjunktní (protože jde o jednobarevná auta, není žádné auto např. červené a zelené současně).

Věta: Každý rozklad množiny P na třídy definuje na P relaci ekvivalence. Každá ekvivalence na P definuje rozklad P na třídy.

Důkaz:

> **A**. Mějme rozklad **T** množiny P na třídy. Definujme relaci ≈ na P takto: $a \approx b \equiv \exists$ i: $A_i \in$**T** $\wedge$ a $\in A_i \wedge$ b $\in A_i$ (a≈b, právě když "patří do stejné třídy"). Relace ≈ je ekvivalence na P: je reflexivní (každé a∈P patří právě do jedné třídy a je tedy a≈a), symetrická (je-li a≈b, pak a i b patří do stejné třídy, je tedy současně i b≈a) a tranzitivní (je-li a≈b - a i b patří do stejné třídy - a současně b≈c - b i c patří do stejné třídy - pak tedy evidentně a i c patří do téže třídy a je tedy a≈c). Relace ≈ je tedy ekvivalence.
>
> **B.** Nechť naopak je na P definována relace ekvivalence ↔.
>
> Definujme systém **T** podmnožin $A_i \subseteq P$ takto: je-li a∈P, b∈P, a↔b, pak existuje i tak, že $a \in A_i \subseteq P$, $b \in A_i \subseteq P$ (každá A_i je množina všech prvků z P, které jsou spolu ekvivalentní).
>
> Tento systém je především disjunktní. Mějme totiž dvě různé podmnožiny A_i a A_k a předpokládejme, že nejsou disjunktní (existuje tedy alespoň jeden prvek z∈P takový, že $z \in A_i \wedge z \in A_k$). Protože A_i a A_k jsou podle předpokladu různé, existuje v A_i jeden prvek $a \in A_i$ takový, že $a \notin A_k$; dále existuje v A_k jeden prvek $b \in A_k$ takový, že $b \notin A_i$.
>
> Protože $a \in A_i$ a současně $z \in A_i$, je a↔z. Protože $b \in A_k$ a současně $z \in A_k$, je b↔z. Protože ↔ je ekvivalence a je tedy symetrická, je i z↔b; protože dále je ↔ tranzitivní a je a↔z a z↔b, je tedy a↔b. To je však podle definice ve sporu s předpokladem, že podmnožiny A_i a A_k jsou různé - podmnožiny jsou tedy disjunktní.
>
> Tvrzení, že $\cup$**T**=P, je zřejmé. Kdyby neplatilo, existoval by prvek a∈P takový, že by nebyl prvkem žádné třídy A_i. Protože však ↔ je ekvivalence a tedy je reflexivní, platí pro každý prvek (tedy i pro ono a) z P, že a↔a. Musí tedy existovat A_k tak, že $a \in A_k$ (byť by A_k měla být jednoprvková množina). To je ovšem ve sporu s tím, že prvek a v žádné A_k není.
>
> Což bylo dokázat.

2.3 Základy počtu pravděpodobnosti

2.3.1 Náhodný pokus, náhodný jev

Pokus je v teorii pravděpodobnosti chápán jako jakákoliv činnost uskutečněná za přesně definovaných podmínek, která může být při těchto podmínkách libovolně opakována.

Jako **elementární jev** se označuje každý výsledek pokusu, přičemž lze konstatovat, zda nastal nebo nenastal.

*Upozornění: Korektně by se mněl definovat **elementární jev daného pokusu P za systému podmínek C**. To se týká i všech dále zaváděných pojmů. Při rozboru, vyhodnocení a práci s elementárními jevy je však většinou z kontextu zřejmé, že se týkají stejného pokusu za stejných podmínek, proto se v označení pojmu část "daného pokusu P za podmínek C" vynechává.*

Deterministický pokus je takový pokus, který má vždy jediný (stejný) výsledek. Ten je tedy přesně určen podmínkami, za kterých se pokus provádí.

Náhodný pokus je takový pokus, který má více možných výsledků závisejících na náhodě. Protože při opakování pokusu vždy za stejných podmínek jsou získány různé výsledky, nelze z daných podmínek konkrétní výsledek předem určit. U náhodného pokusu je ovšem známa množina výsledků, které po provedení pokusu mohou nastat.

Příklad: Snad nejjednodušším a klasickým příkladem na náhodný pokus je házení kostkou.

Elementární náhodný jev je každý možný výsledek náhodného pokusu.

Příklad: Elementárním náhodným jevem je: padla 6-ka.

Základní prostor elementárních jevů (a pokud je z kontextu zřejmé, zkráceně jen **Základní prostor**) náhodného pokusu je množina všech elementárních náhodných jevů náhodného pokusu. Značí se obvykle **Ω**.

Příklad: Je zřejmé, že základním prostorem elementárních jevů při házení kostkou je množina Ω={padla 1-ka, padla 2-ka, padla 3-ka, padla 4-ka, padla 5-ka, padla 6-ka}.

Náhodný jev A je jakákoliv podmnožina základního prostoru Ω: $A \subset \Omega$. Elementární jev je tedy jeden z možných výsledků, náhodný jev je jeden nebo více různých možných výsledků.

Příklad: Náhodným jevem při házení kostkou je např. {padla 1-ka, padla 3-ka, padla 5-ka} $\subset \Omega$. V občanské mluvě by se řeklo: padlo liché číslo.

Jistý jev A je celý základní prostor Ω. V tomto kontextu se označuje většinou I.

Nemožný jev A je prázdná podmnožina základního prostoru Ω. Označuje se většinou ∅.

Opačný jev ¬A k jevu A: $\neg A = \Omega - A$.

Součet (sjednocení) jevů A a B: takový jev C, který nastane tehdy, když nastane jev A nebo jev B. Značí se $C = A + B$ nebo $C = A \cup B$.

Součin (průnik) jevů A a B: takový jev C, který nastane tehdy, když nastane jev A a současně jev B. Značí se $C = A . B$ nebo $C = A \cap B$.

Neslučitelné (disjunktní) jevy A a B: $A \cap B = \emptyset$ - tedy nenastanou současně.

Příklad: Mějme při házení kostkou (základní prostor Ω={padla 1-ka, padla 2-ka, padla 3-ka, padla 4-ka, padla 5-ka, padla 6-ka}) dva jevy: S={padla 2-ka, padla 4-ka, padla 6-ka}, L={padla 1-ka, padla 3-ka, padla 5-ka}. Tyto jevy jsou neslučitelné (jejich průnik je nemožný jev), jejich sjednocení je jistý jev. Je to zřejmé, pokud si uvědomíme, že S je "padlo sudé číslo" a L je "padlo liché číslo". Je nemožné, aby padlo sudé a liché současně, a je jisté, že vždy padne sudé číslo nebo liché číslo.

Systém neslučitelných (disjunktních) jevů: množina jevů $\{A_i\}$ po dvou disjunktních - $A_k \cap A_j = \emptyset$ pro všechny $k \neq j$.

Úplný systém neslučitelných (disjunktních) jevů: Systém $\{A_i\}$ n disjunktních jevů, pro který platí $A_1 \cup A_2 \cup ... \cup A_n = \Omega$.

2.3.2 Pravděpodobnost náhodného jevu

Konáme-li náhodný pokus a posuzujeme-li jeho jednotlivé možné výsledky z nějakého - byť subjektivního - úhlu pohledu, vnímáme některé výsledky jako příznivější, některé jako méně příznivé. V reálném životě se často řídíme právě mírou očekávání konkrétních výsledků. Je-li tato míra očekávání velmi malá, pak třeba některé navazující činnosti zcela vynecháme. Je tedy třeba přesně definovat, co se rozumí pod pojmem "míra očekávání" a jak tuto míru kvantifikovat.

K ohodnocení nějakého výsledku se i v běžné řeči použije obrat typu "výsledek je málo pravděpodobný". A tento pojem - pravděpodobnost - je základem jednoho oboru matematiky, počtu pravděpodobnosti. V něm se pravděpodobnost jako míra očekávání hodnotí reálným číslem z intervalu <0, 1>. Jev s pravděpodobností 0 nikdy nenastane, jev s pravděpodobností 1 nastane vždy (viz shora jev nemožný, jev jistý).

Existuje řada definic pravděpodobnosti (Laplaceova klasická, statistická, geometrická, Kolmogorovova axiomatická). Pro účely těchto učebních textů uveďme nejprve *statistickou* definici pravděpodobnosti:

> Nechť je náhodný pokus proveden M-krát. Nechť se nějaký jev J vyskytl N-krát. Číslo
>
> $R(J) = N / M$
>
> se nazývá relativní četnost jevu J. Nechť se počet provedení pokusu M zvětšuje nade všechny meze. Blíží-li se při tomto zvětšování relativní četnost R nějakému číslu P, pak se toto číslo nazývá pravděpodobnost jevu J.

Uveďme ještě *klasickou* definici:

> Mějme základní prostor M elementárních jevů $\Omega = \{A_i\}$. Mějme jev $J = A_{i1} \cup A_{i2} \cup ... \cup A_{iN}$, $N \leq M$. Pravděpodobností jevu J se pak nazývá číslo
>
> $P(J) = N / M$
>
> V praxi se často elementární jevy, jichž je J součtem, nazývají *příznivé* elementární jevy. Uvedená rovnice definující pravděpodobnost jevu pak přechází na známé slovní vyjádření: počet příznivých ku počtu všech.

Z teorie (která však překračuje pojetí těchto učebních textů) plyne, že za jistých předpokladů jsou hodnoty pravděpodobnosti podle statistické i klasické definice shodné.

> Příklad: Při házení kostkou je pravděpodobnost, že padne 3 (= příznivý výsledek) rovna podle klasické definice 1/6. Protože při "spravedlivé" kostce jsou všechna čísla rovnocenná, pak při zvětšujícím se počtu hodů padne každé číslo víceméně stejně krát. Protože čísel je šest, pak se podle statistické definice pravděpodobnost blíží rovněž 1/6.

Z definice pravděpodobnosti lze dovodit (a zavést) následující; viz také označení a definice shora:

a. Pravděpodobnost $P(J) \in <0, 1>$
b. Velmi často se pravděpodobnost vyjadřuje v procentech: $P_{\%}(J) = 100 * P(j) = 100 * N / M$.
c. Má-li jev J pravděpodobnost P(J)=0 (tedy 0%), nazývá se **nemožný**.
d. Má-li jev J pravděpodobnost P(J)=1 (tedy 100%), nazývá se **jistý**.
e. Má-li jev J pravděpodobnost P(J) (tj. že nastane), má pravděpodobnost 1-P(J) že nenastane.
f. **Šance** je podíl pravděpodobnosti, že jev nastane, ku pravděpodobnosti, že nenastane. Nejčastěji se šance vyjadřuje a : b, např. 1 : 3. Jev se šancí a : b má pravděpodobnost a/(a+b).

2.3.3 Náhodná veličina

Elementární náhodný jev, tak jak byl shora zaveden, je každý možný výsledek náhodného pokusu. Základní prostor je množina všech elementárních jevů.

Pro hodnocení výsledků pokusu lze aplikovat numerické postupy, pokud jsou elementární náhodné jevy kvantitativního charakteru (jsou vyjádřeny číslem). Mnohé výsledky však číselný charakter nemají (např. účastník zájezdu zvolil Kypr, při hodu mincí padl rub apod.). Je proto žádoucí transformovat i takové výsledky na čísla.

Náhodná veličina X je každé zobrazení X základního prostoru Ω elementárních jevů určitého pokusu do množiny R reálných čísel: $X : \Omega \Rightarrow R$. Je-li tedy $\omega \in \Omega$ elementární náhodný jev, pak $x = X(\omega) \in R$.

> *Poznámka: Jsou-li elementární náhodné jevy kvantitativního charakteru (t.j. výsledkem náhodného pokusu je - nejobecněji reálné - číslo ω), je přirozeně minimálně funkce X(ω)=ω náhodnou veličinou přiřazenou prováděnému pokusu. V praxi se proto setkáváme s případy, kdy se pojem Veličina X a Hodnota x veličiny X ztotožňují.*
>
> Příklad 1: Nechť je Ψ={padla 1-ka, padla 2-ka, padla 3-ka, padla 4-ka, padla 5-ka, padla 6-ka} základním prostorem. Pak náhodnou veličinou $K : \Psi \Rightarrow R$ může být zobrazení dané tabulkově: K(padla 1-ka)=1, K(padla 2-ka)=2, K(padla 3-ka)=3, K(padla 4-ka)=4, K(padla 5-ka)=5, K(padla 6-ka)=6.

Příklad 2: Nechť je Θ={Spojitě naměřené venkovní teploty Z dne 13/6/2016} základním prostorem. Pak náhodnou veličinou T : Θ ⇒ R může být zobrazení T(Naměřena venkovní teplota Z) = Z + 273,16 (tedy reálné číslo vyjadřující naměřenou teplotu ve °K).

Definičním oborem Δ(X) náhodné veličiny X je tedy celý základní prostor elementárních jevů (každému elementárnímu náhodnému jevu je přiřazeno právě jedno reálné číslo).

Oborem hodnot Φ(X) je množina těch reálných čísel, které jsou obrazy jednotlivých elementárních náhodných jevů. Přitom každá z těchto hodnot má určitou pravděpodobnost.

Příklad: V předchozím příkladu 1 je oborem hodnot náhodné veličiny K množina {1, 2, 3, 4, 5, 6} s pravděpodobnostmi {1/6, 1/6, 1/6, 1/6, 1/6, 1/6}.

Diskrétní náhodná veličina je taková náhodná veličina, pro kterou existuje zobrazení, jehož oborem hodnot je spočetná množina reálných čísel.

Příklad: V předchozím příkladu 1 je oborem hodnot náhodné veličiny K množina {1, 2, 3, 4, 5, 6}, veličina K je tedy diskrétní.

Spojitá náhodná veličina je taková náhodná veličina, pro kterou existuje zobrazení, jehož oborem hodnot je interval reálných čísel.

Příklad: V předchozím příkladu 2 je oborem hodnot náhodné veličiny T interval $<Z_{min}+273,16; Z_{max}+273,16>$, veličina T je tedy spojitá.

2.3.4 Rozdělení pravděpodobnosti náhodné veličiny

Přestože náhodná veličina zobrazuje základní pravděpodobnostní prostor náhodných elementárních jevů majících každý přiřazenou jistou pravděpodobnost, není zobrazením X elementárního jevu jeho pravděpodobnost. Pravděpodobnost různých jevů je dána mírou pravděpodobnosti P na základním prostoru Ω. Zobrazení X popisuje nějakou číselnou vlastnost, kterou jevy v Ω mohou mít (např. váha náhodně vybraného účastníka zájezdu, počet bělochů v náhodném okamžiku v restauraci Red Elephant v Nigérii apod.).

Pravděpodobnost náhodné veličiny X (jako obrazu) je odvozena od pravděpodobnosti jevu (jako vzoru). Pravděpodobnost P, že náhodná veličina X nabude např. hodnot mezi 5 a 7, je rovna

$$P (\{ \omega \in \Omega: X(5 \leq \omega \leq 7) \})$$

Rozdělením pravděpodobnosti náhodné veličiny se rozumí matematický popis (především funkčním vztahem) pravděpodobnosti výskytu jednotlivých možných výsledků pozorovaného náhodného jevu popisovaného náhodnou veličinou. Je tedy definováno na základním prostoru, což je množina všech možných výsledků zkoumaného jevu. Rozdělení pravděpodobnosti může být popsáno různými způsoby: pravděpodobnostní funkcí resp. hustotou pravděpodobnosti, distribuční funkcí, charakteristickou funkcí a dalšími.

Některé funkční tvary se při popisu rozdělení pravděpodobnosti vyskytují v praxi tak často, že byly pojmenovány podle charakteru zkoumané veličiny resp. autora (rozdělení alternativní, Poissonovo, binomické, rovnoměrné, Studentovo atd). V dalších odstavcích bude popsáno jedno z nejdůležitějších rozdělení, a to *normální* (známé také jako Gaussovo).

Rozdělení pravděpodobnosti diskrétní náhodné veličiny je často - zejména pro obor hodnot malého rozsahu - popsáno tabulkou.

2.3.5 Distribuční funkce

Distribuční funkce (Distribution function) F náhodné veličiny Z pro její hodnotu x je pravděpodobnost, že hodnota náhodné veličiny Z bude nanejvýš rovna x: $F_Z(x) = P(Z \leq x)$. Je-li zřejmé, o jakou náhodnou veličinu Z se jedná, zapisuje se jen $F(x) = P(Z \leq x)$.

Mezi některé obecné vlastnosti distribuční funkce patří:

1. $F(x) \in <0, 1>$

2. $F(-\infty) = 0$, $F(+\infty) = 1$
3. Je-li $x_1 < x_2$, je $F(x_1) \leq F(x_2)$ (F je neklesající)
4. Je-li $x_1 < x_2$, je $P(x_1 < X \leq x_2) = F(x_2) - F(x_1)$

Distribuční funkce tedy popisuje, jak je pravděpodobnost hodnot náhodné veličiny kumulativně rozdělena ve svém definičním oboru.

2.3.6 Pravděpodobnostní funkce, hustota pravděpodobnosti

Distribuční funkce udává pravděpodobnost, s jakou náhodná veličina nabude nanejvýš konkrétní hodnoty. S použitím posledního vztahu předchozího odstavce umožňuje rovněž zjistit, s jakou pravděpodobností nabude náhodná veličina hodnoty z konkrétního intervalu. V praxi je však často kladen požadavek na zjištění pravděpodobnosti, s jakou nabude náhodná veličina jedné konkrétní hodnoty. V takovém případě je však nutno rozlišovat, zda jde o veličinu diskrétní nebo spojitou.

Diskrétní náhodná veličina

Nechť pro jednotlivé možné hodnoty diskrétní náhodné veličiny X

$$x_1 < x_2 < \ldots < x_n$$

má její distribuční F funkce hodnoty

$$F(x_1) \leq F(x_2) \leq \ldots \leq F(x_n)$$

Označme f(x) pravděpodobnost P(X=x). Je pak evidentně pro k=2, 3, ..., n

$$f(x_1) = F(x_1),\ f(x_k) = F(x_k) - F(x_{k-1})$$

Tím je však definována (pro diskrétní veličinu např. tabulkově) nová funkce f(x), která každé možné hodnotě x náhodné veličiny X přiřazuje pravděpodobnost, že tato veličina nabude hodnoty právě x.

Pravděpodobnostní funkcí (Probability mass function) diskrétní náhodné veličiny se rozumí právě popsaná funkce f(x). Mezi některé její vlastnosti patří:

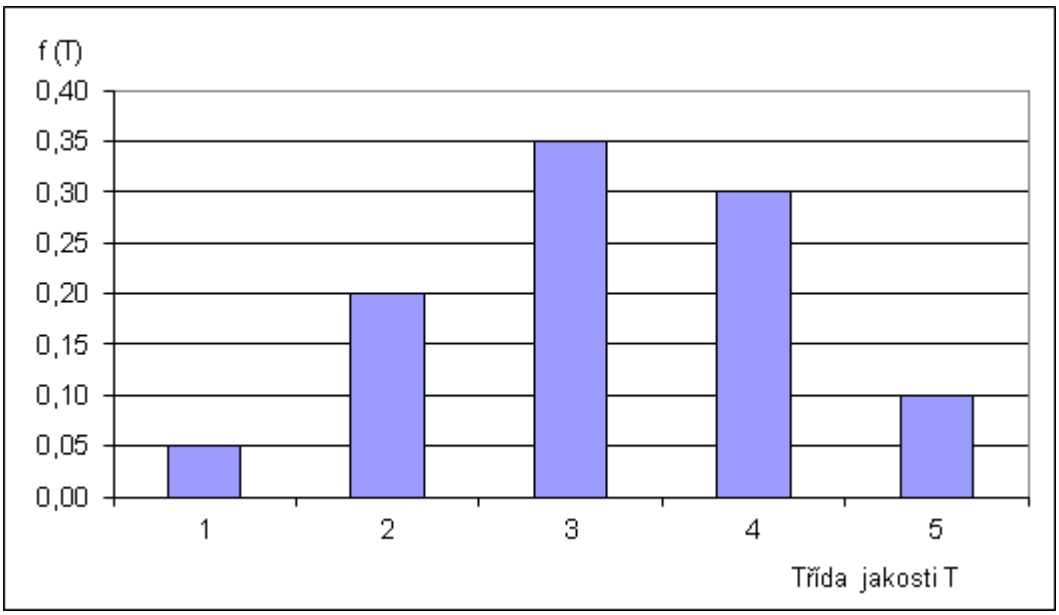
1. $F(x_k) = \sum f(x_i)$ pro i = 1, 2, ..., k; k = 1, 2, ..., n
2. $f(x_i) \geq 0$
3. $\sum f(x_i) = 1$ pro i = 1, 2, ..., n

Vlastnost ad 1. plyne přímo ze způsobu zavedení funkce f(x). Vlastnost ad 2. plyne z toho, že hodnotou f(x) je pravděpodobnost. Vlastnost ad 3. plyne z toho, že jevy J_i, jimž veličina X přiřazuje hodnotu x_i, tvoří úplný systém navzájem disjunktních jevů a při každém pokusu s jistotou nastane vždy jeden z nich.

Příklad 1: Mějme dány pravděpodobnosti, že třída jakosti nějaké imaginární řeky v nějakém profilu nabude hodnoty T, následující tabulkou vycházející z dlouhodobého pozorování a měření:

Třída jakosti T	1	2	3	4	5
Pravděpodobnostní funkce f(T)	0,05	0,20	0,35	0,30	0,10

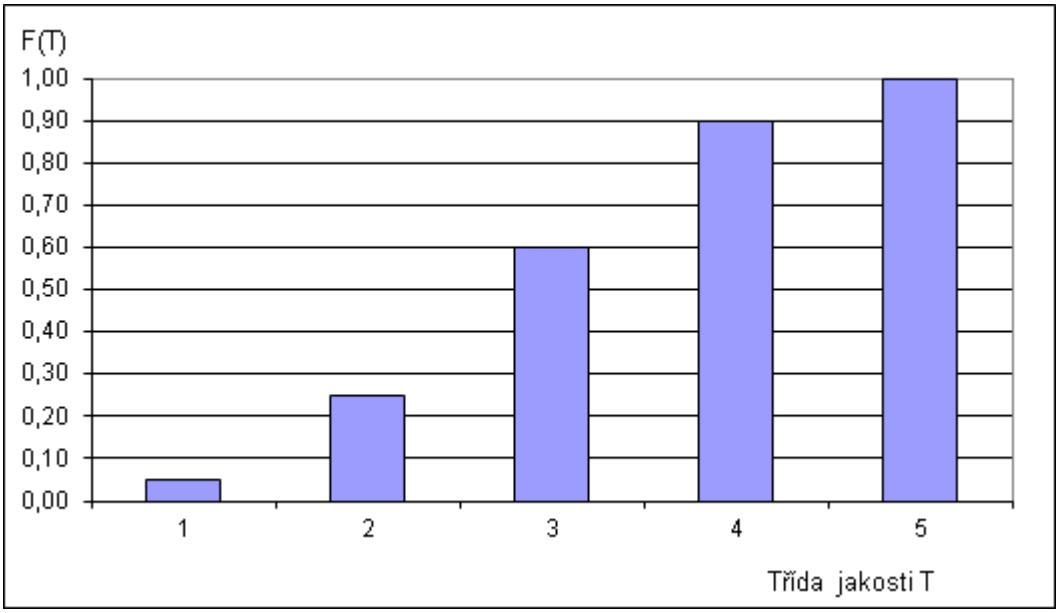
Pravděpodobnostní funkce je tedy definována takto (jde o diskrétní funkci): f(1)=0,05; f(2)=0,20; f(3)=0,35; f(4)=0,30; f(5)=0,10. Pravděpodobnostní funkce má i své grafické vyjádření, např. následujícím grafem:



Obdobně distribuční funkce na týchž datech dá následující tabulku:

Třída jakosti T	1	2	3	4	5
Distribuční funkce F(T)	0,05	0,25	0,60	0,90	1,00

a příslušný graf:



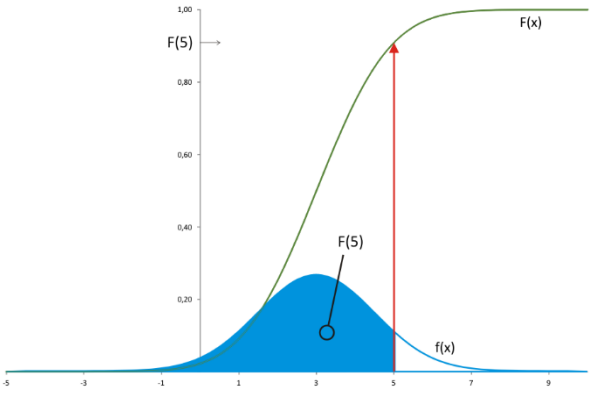
Spojité náhodné veličiny

Spojitá náhodná veličina je zobrazením do množiny reálných čísel. Její distribuční funkce pak umožní zjistit pravděpodobnost, že náhodná veličina nabude hodnoty z konkrétního intervalu (a, b> reálných čísel - viz shora vlastnost ad 4. distribuční funkce: $P(a<X\leq b) = F(b) - F(a)$.

Zjistit pravděpodobnost pro jedinou konkrétní hodnotu spojité náhodné veličiny (např. a) by zdánlivě šlo zjištěním pravděpodobnosti na intervalu (a, a+h> a následným zmenšováním tohoto intervalu (tj. přibližováním $h\rightarrow 0$). V limitě ovšem bude pravděpodobnost rovna F(a)-F(a)=0. Není tedy smysluplné definovat pro spojitou náhodnou veličinu pravděpodobnostní funkci analogicky diskrétní náhodné veličině.

Hodnota distribuční funkce diskrétní veličiny F(a) je rovna součtu pravděpodobností všech takových hodnot x zkoumané veličiny, které nepřesáhnou hodnotu a - hodnoty x jsou však konkrétní reálná čísla. U spojité veličiny však jde o interval (-∞; a> "nekonečně mnoha nekonečně blízkých" reálných čísel a nejde tedy použít prostý součet jejich pravděpodobností.

Z integrálního počtu je však známo, že takový "nekonečný součet" poskytne - např. formou velikosti plochy - určitý integrál. Pro účely spojité náhodné veličiny opačně: Je-li známa distribuční funkce F jakožto kumulace pravděpodobností až do hodnoty a včetně, určitě existuje funkce f taková, že integrována až do hodnoty a dá hodnotu F(a) - viz následující obrázek:



Tato funkce se nazývá **hustota pravděpodobnosti (Probability density function)**. Mezi distribuční funkcí F a hustotou pravděpodobnosti f spojité náhodné veličiny tedy platí vztah:

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

Naopak, mezi hustotou pravděpodobnosti f(x) a distribuční funkcí F(x) platí vztah

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

2.3.7 Střední hodnota, rozptyl, směrodatná odchylka

Při zkoumání veličin je jedním z cílů určit jakousi jejich reprezentativní (vztažnou, očekávanou, charakteristickou, teoretickou) hodnotu, ve statistice nazývané **mírou polohy**. Míru polohy se snažíme odhadnout např. opakovaným měřením dané veličiny. Považujeme tedy veličinu za náhodnou veličinu a zjišťujeme pravděpodobnosti výskytu jednotlivých hodnot, tedy pravděpodobnostní resp. distribuční funkci. Ukazuje se, že míra polohy závisí na pravděpodobnostní funkci zkoumané veličiny.

Mezi nejznámější míry polohy diskrétních veličin patří aritmetický průměr, modus, medián a další. Ovšem ne každou míru polohy lze použít pro jakoukoliv veličinu - viz dále.

Míra polohy je tedy jakousi "střední" hodnotou. Je ovšem podstatný rozdíl mezi veličinou, jejíž hodnoty jsou téměř všechny skoro rovny stanovené míře polohy, a mezi veličinou - byť se stejnou mírou polohy, jejíž hodnoty jsou rozvrstveny v širokém pásmu kolem ní. Ve statistice je tato větší či menší proměnlivost dat charakterizována **mírou variability**. Mezi často používané (ale stejně tak často chybně interpretované) míry variability patří rozptyl a směrodatná odchylka. Stejně jako míra polohy závisí na pravděpodobnostní resp. distribuční funkci.

Hodnota vyjadřující střed (centrální tendenci, těžiště, průměrnou velikost apod.) je definována jako střední hodnota:

Definice: *Střední hodnota* E(x) diskrétní náhodné veličiny x je definována takto:

$$E(x) = \sum x_i \,.\, p(x_i)$$

kde se sčítá přes všechna i.

Příklad: Třídy jakosti ve shora uvedeném profilu hypotetické řeky mají střední hodnotu 1.0,05+2.0,20+3.0,35+4.0,30+5.0,10=**3,20.**

Definice: *Střední hodnota* E(x) spojité náhodné veličiny x je definována takto:

$$E(x) = \int_{-\infty}^{+\infty} x \cdot f(x) \cdot dx$$

Poznámka: Je-li z kontextu zcela jasné, o jakou náhodnou veličinu x se jedná, pak se pro přehlednost dalších vzorců často označuje E(x) symbolem μ.

Hodnoty, jistým způsobem vyjadřující rozptýlení, souhrnnou odchylku od střední hodnoty, jsou definovány jako rozptyl (angl. Variance) resp. směrodatná odchylka (angl. Standard deviation).

Definice: *Rozptyl* D(x) diskrétní náhodné veličiny x je definován takto:

$$D(x) = \sum((x_i - \mu)^2 \,.\, p_i) = (\sum x_i^2 \,.\, p_i) - \mu^2$$

kde symbolem μ je označena střední hodnota E(x) a kde se sčítá přes všechna i.

Příklad: Třídy jakosti ve shora uvedeném profilu hypotetické řeky mají rozptyl 1^2.0,05+2^2.0,20+3^2.0,35+4^2.0,30+5^2.0,10-3,20^2=**1,06.**

Definice: *Rozptyl* D(x) spojité náhodné veličiny x je definován takto:

$$D(x) = \int_{-\infty}^{+\infty} (x-\mu)^2 \cdot f(x) \cdot dx = \left(\int_{-\infty}^{+\infty} x^2 \cdot f(x) \cdot dx \right) - \mu^2$$

kde symbolem μ je označena stejně jako v předchozím vztahu střední hodnota E(x).

Poznámka: Je-li z kontextu zcela jasné, o jakou náhodnou veličinu x se jedná, pak se pro přehlednost dalších vzorců často označuje D(x) symbolem σ^2.

Definice: *Směrodatná odchylka* je definována jako druhá odmocnina z rozptylu - √D(x) - a je tedy rovna hodnotě σ.

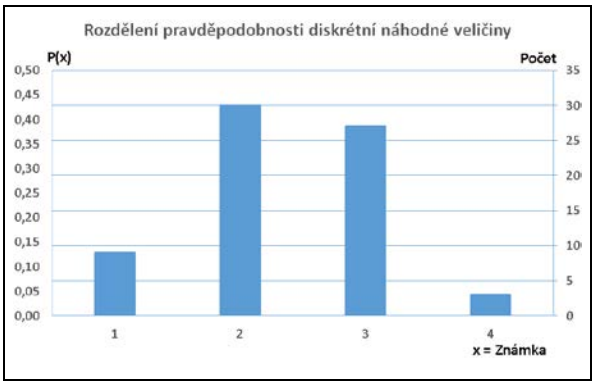
Příklad: Výsledky zkoušek z předmětu Statistika a informatika v zimním semestru 2015 / 2016 jsou uvedeny v prvních dvou sloupcích následující tabulky. Třetí sloupec obsahuje hodnoty pravděpodobnostní funkce P(x), čtvrtý její hodnoty v procentech. Pátý sloupec obsahuje hodnoty distribuční funkce F(x), šestý její hodnoty v procentech. Interpretace dat z tabulky může být např. "S pravděpodobností 13% udělám zkoušku za jedna" nebo "S pravděpodobností 95,65% zkoušku udělám".

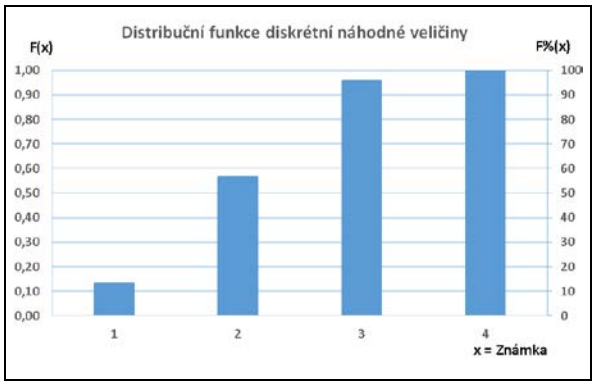
x = Známka	**Počet**	**P(x)**	**P%(x)**	**F(x)**	**F%(x)**
1	9	0,13	13,04	0,13	13,04
2	30	0,43	43,48	0,57	56,52
3	27	0,39	39,13	0,96	95,65
4	3	0,04	4,35	1,00	100,00

Z uvedené tabulky lze zjistit střední hodnotu: μ = 1 . 0,13 + ... + 4 . 0,04 = **2,35**. Rovněž lze zjistit rozptyl:

σ^2 = $(1 - 2{,}35)^2$. 0,13 + ... + $(4 - 2{,}35)^2$. 0,14 = **0,57**.

Grafickým náhledem na pravděpodobnostní a distribuční funkci shora uvedené diskrétní náhodné veličiny mohou být např. tyto dva grafy:





2.3.8 Normální rozdělení

Bezesporu nejdůležitějším rozdělením v teorii pravděpodobnosti a matematické statistice je normální rozdělení. Jeho význam udává např. centrální limitní věta, která za velmi obecných podmínek zaručuje, že součet nezávislých náhodných veličin má přibližně normální rozdělení bez ohledu na rozložení jednotlivých sčítanců. Další význam spočívá v tom, že jím lze aproximovat mnohá jiná rozdělení včetně diskrétních. Normální rozdělený bývá také nazýváno *Gaussovým rozdělením* a graf jeho hustoty *Gaussovou křivkou*.

Nejprve zaveďme normované normální rozdělení:

Definice: *Normované normální rozdělení* je takové, jehož hustota pravděpodobnosti (bývá zvykem ji označovat φ(x) namísto obecného f(x)) má tvar

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

Lze ukázat, že střední hodnota normovaného normálního rozdělení je 0 a jeho rozptyl je 1. Graf hustoty je symetrický kolem nuly, funkce φ má dva inflexní body {-1;+1}.

Distribuční funkce normovaného normálního rozdělení má tvar

$$\Phi(x) = \frac{1}{\sqrt{2 \cdot \pi}} \cdot \int_{-\infty}^{x} e^{\frac{-t^2}{2}} \cdot dt$$

Obecně se pak zavádí normální rozdělení takto:

Definice: *Normální rozdělení* je takové, jehož hustota pravděpodobnosti f(x) má tvar

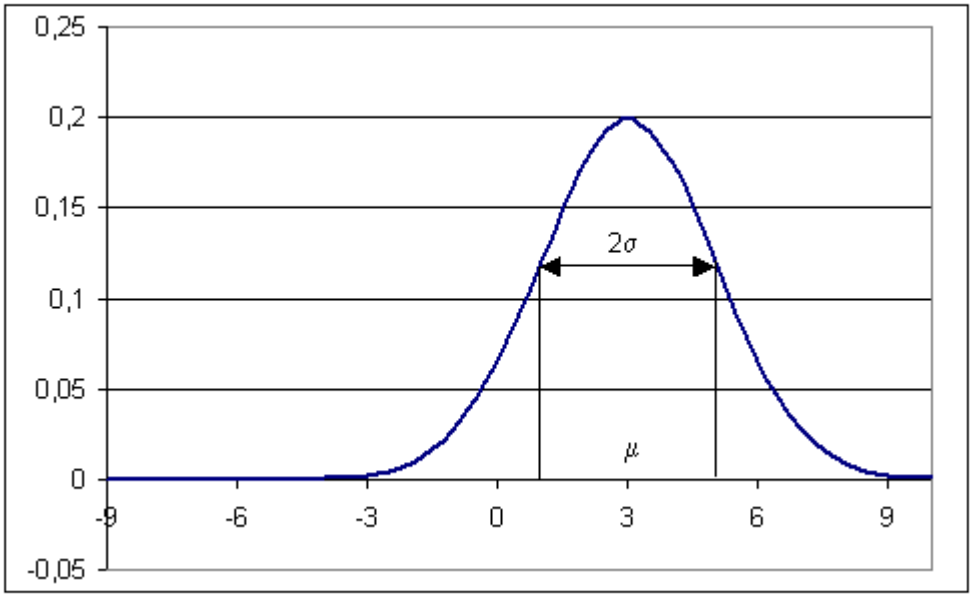
$$f(x) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sigma} \cdot \varphi\left(\frac{x-\mu}{\sigma}\right)$$

kde μ je reálné číslo a σ > 0. Graf hustoty f(x) je symetrický kolem přímky x=μ, přičemž hodnota μ je zároveň střední hodnotou. Funkce má dva inflexní body {μ-σ , μ+σ}, přičemž hodnota σ^2 je zároveň rozptylem.

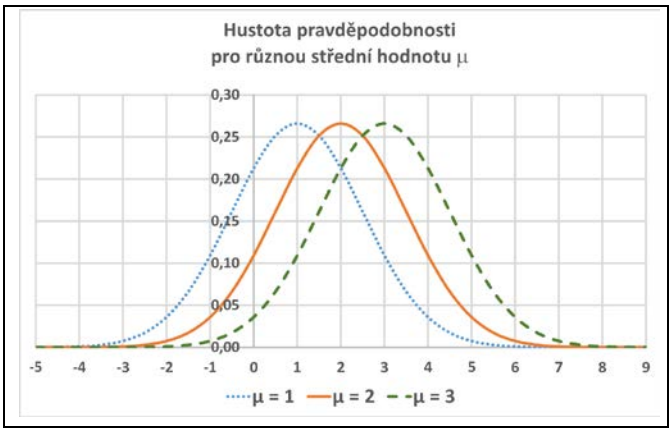
Distribuční funkce obecného normálního rozdělení pak má tvar

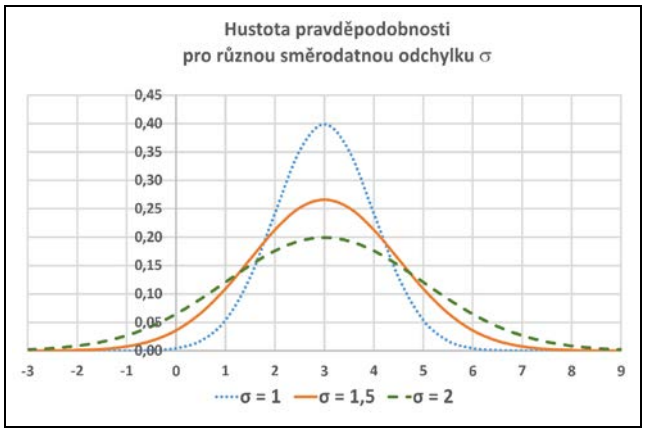
$$F(x) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot \int_{-\infty}^{x} e^{\frac{-(t-\mu)^2}{2\sigma^2}} \cdot dt = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

Příklad: Hustota pravděpodobnosti normálního rozdělení s μ=3 a σ=2 má graf



Význam parametrů μ a σ znázorněné bez animace:





3 Základní pojmy ve statistické analýze dat

Tato kapitola přehledně uvádí základní pojmy, o které se opírají statistické postupy tvořící jádro učebních textů. Následně je rovněž podrobně vysvětlen postup vkládání dat a jejich úprava v prostředí programu Statgraphics.

3.1 Terminologie

Ve statistické analýze dat se běžně používá množství specifických termínů a pojmů. Za účelem správného pochopení výkladu dané problematiky jsou v této podkapitole uvedeny a definovány základní nejčastěji používané statistické pojmy včetně názorných příkladů.

Statistický údaj (data) je hodnota námi měřené sledované veličiny, např. věk jednoho návštěvníka (turisty) dané technické památky.

Statistická jednotka je prvek statistického souboru, u kterého zkoumáme konkrétní vlastnosti (statistické znaky). Statistickou jednotkou může být např. návštěvník (turista), u kterého zjišťujeme více statistických údajů, jako např. pohlaví, věk, stupeň vzdělání, zaměstnání, bydliště apod.

Statistický soubor je konečná množina hodnot, čili souhrn všech námi zjištěných statistických údajů. Statistickým souborem mohou být například stupně vzdělání všech námi dotazovaných návštěvníků dané lokality.

Proměnná (statistický znak) je měřená statistická veličina (vlastnost jednotek), která nabývá ve statistickém (datovém) souboru různých hodnot (např. návštěvnost, délka pobytu, doprava atd.).

Rozsah souboru je počet údajů (dat) jednoho statistického znaku (proměnné) ve statistickém souboru. Obvykle jej označujeme „n“. Rozsahem souboru může být např. počet všech stanovených návštěvností geolokality (např. v jednotlivých dnech) za celé sledované období.

Konstanta je ten případ, kdy proměnná v rámci statistického souboru nabývá pouze jedné hodnoty, tedy v rámci souboru se hodnota nemění.

Nezávisle proměnná je taková sledovaná veličina, která v rámci studovaného souboru dat není ovlivněna změnami hodnot jiné proměnné.

Závisle proměnná je veličina, jež nabývá různých hodnot v závislosti na změnách jiné veličiny.

> Jako příklad můžeme uvést datový soubor, ve kterém je sledována návštěvnost vybrané geolokality a množství spadlých srážek na daném území. V tomto případě je množství srážek nezávisle proměnnou (nabývá různých hodnot nezávisle na návštěvnosti). Závisle proměnnou je návštěvnost (nabývá různých hodnot v závislosti na množství spadlých srážek, tedy návštěvnost je ovlivněna srážkami).

Typy proměnných - na základě způsobu vyjádření se proměnné dělí na kvantitativní a kvalitativní.

Kvantitativní proměnná je hodnota vyjádřená určitými čísly (např. návštěvnost). Číselné proměnné jsou také označovány jako numerické.

Kvalitativní proměnná je hodnota vyjádřená slovy (např. charakter počasí vyjádřený pomocí oblačnosti - jasno, polojasno, oblačno, zataženo). V odborné literatuře se často pro kvalitativní proměnné používá název kategoriální.

Program Statgraphics označuje kvantitativní proměnnou anglickým slovem „Numeric“ a kvalitativní proměnnou anglickým slovem „Character“. Podle typu vztahů mezi hodnotami (podle způsobu vyjádření) rozlišujeme proměnné nominální, ordinální (dle výše uvedeného dělení se jedná o proměnné kvalitativní) a kardinální (proměnná kvantitativní). V současnosti některé statistické programy (např. SAS, SPSS, Statistica) umožňují zadat typ proměnné i podle tohoto typu dělení.

Nominální proměnná se zpravidla vyjadřuje slovem, nebo číselným kódem. U nominální proměnné nemůžeme stanovit pořadí, lze pouze určit, zda jsou stejné, či rozdílné. Příkladem nominální proměnné je např. pohlaví – žena, muž. V tomto případě může nominální proměnná nabývat pouze dvou kategorií a bývá označována jako dichotomická. Dalším příkladem může být rodinné zázemí turistů – svobodný, ženatý, rozvedený. Zde může nominální proměnná nabývat více než dvou kategorií a bývá označována jako vícekategoriální.

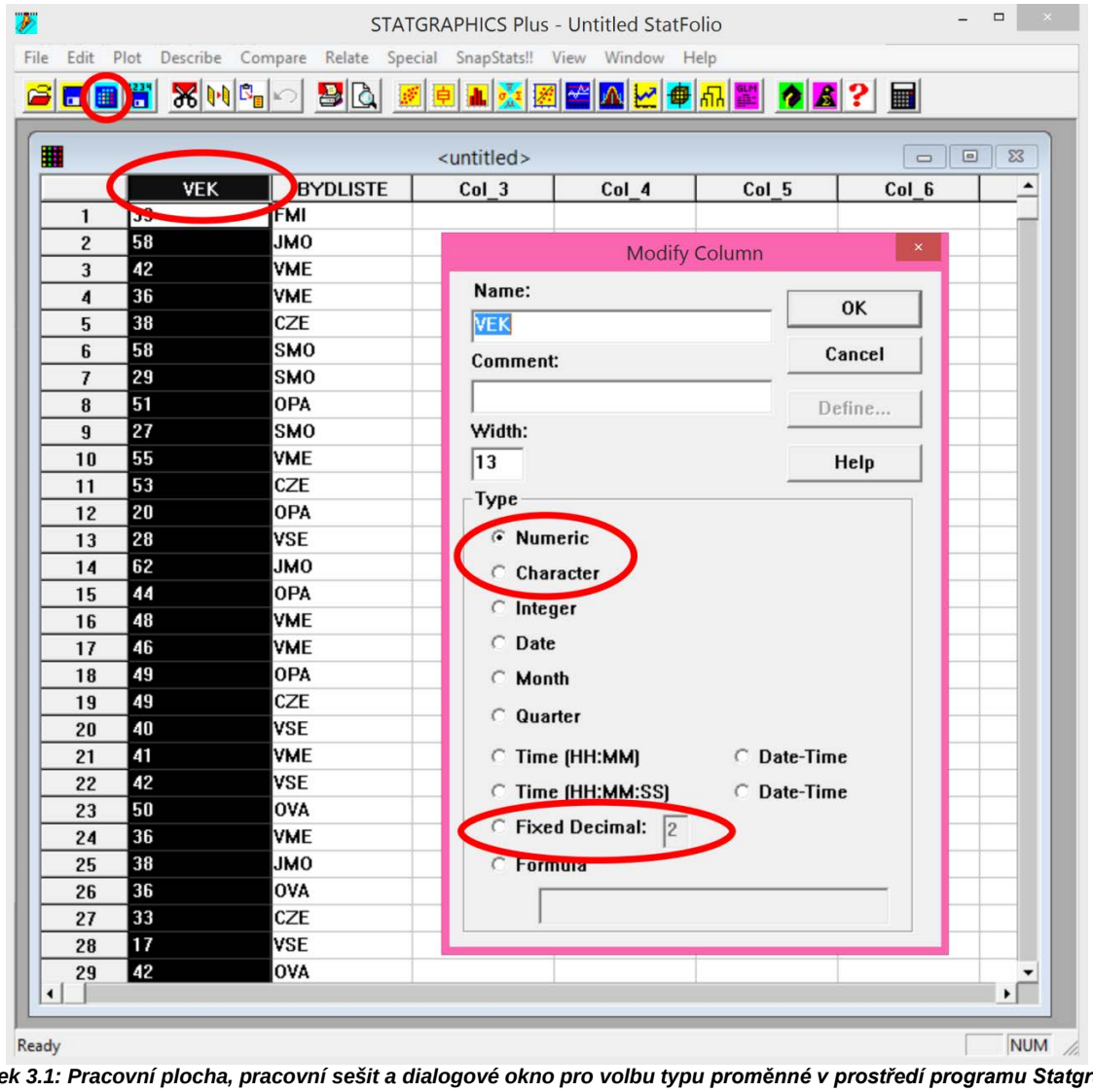
Ordinální proměnná se také zpravidla vyjadřuje slovně, ovšem na rozdíl od nominální proměnné zde můžeme navíc určit pořadí. To znamená, že hodnoty ordinální proměnné lze seřadit od nejmenší po největší podle intenzity sledovaného parametru, avšak nemůžeme určit, o kolik se hodnoty mezi sebou liší. Příkladem ordinální proměnné může být např. úroveň spokojenosti turistů s nabízenými službami.

Kardinální proměnná se vyjadřuje číselně. U kardinální proměnné lze jednoznačně říci o kolik je jedna hodnota větší než druhá (např. věk účastníka zájezdu, cena zájezdu, vzdálenost od místa bydliště apod.). Kardinální proměnné lze dále rozdělit na intervalové (rozdílové) a poměrové (podílové). V případě intervalové proměnné můžeme číselně vyjádřit pouze rozdíl dvou hodnot. U poměrové proměnné můžeme kromě rozdílu dvou hodnot vyjádřit i jejich podíl tzn., že poměrová proměnná nabývá pouze kladných číselných hodnot.

3.2 Postup pro zadání dat a typu proměnné v prostředí programu Statgraphics

Program Statgraphics spustíme dvojklikem levou myší na ikonu programu (modrá ikona s grafem). Následně se automaticky otevře pracovní plocha s jednořádkovým textovým menu, horní lišta s ikonami a prázdný pracovní sešit pro vkládání dat. Data můžeme do pracovního sešitu programu vkládat několika různými způsoby. Prvním způsobem je překopírování dat z pracovního sešitu libovolného tabulkového procesoru (např. MS Excel) pomocí klávesových zkratek Ctrl+C (kopírovat) a Ctrl+V (vložit). Další možností je přímé otevření daného datového souboru v pracovním sešitě. Přímé otevření se provede pomocí textového menu: File → Open → Open Data File. Následně je nutné zvolit typ souboru (jeho příponu, např. xls, sf3, dbf) a zadat název souboru. Přímé otevření datového souboru je také možné provést klikem levou myší na ikonu Open data file (třetí ikona zleva v horní liště, viz obrázek 3.2.1). V případě programu Statgraphics Plus 5.0 je nutné otevírat excelovské soubory pouze s příponou xls (jedná se o starší verzi programu Statgraphics, která neumí přečíst soubory s příponou xlsx). Experimentální data lze vkládat i jejich přímým zápisem do pracovního sešitu.

Po načtení experimentálních dat je nutné zvolit typ proměnné. Volbu typu proměnné provedeme dvojklikem levou myší na záhlaví v pracovním sešitě. Otevře se nám dialogové okno Modify Column (viz obrázek 3.1). Zde si můžeme v řádku Name zvolenou proměnnou libovolně pojmenovat. V nabídce Type zvolíme typ proměnné. V případě, že máme hodnoty uvedeny v číselném formátu, zatrhneme volbu Numeric. Pro slovně vyjádřené hodnoty (ve formě znaků) zatrhneme volbu Character. Dále si u číselných hodnot můžeme zvolit počet desetinných míst. Volbu počtu desetinných míst proměnné provedeme v nabídce Type zatržením volby Fixed Decimal a zadáním číselné hodnoty (viz obrázek 3.1):



Obrázek 3.1: Pracovní plocha, pracovní sešit a dialogové okno pro volbu typu proměnné v prostředí programu Statgraphics

4 Základní statistické charakteristiky

Náplní kapitoly je vysvětlení klasických a robustních odhadů číselných charakteristik, způsob výpočtu a vhodnost jejich použití na experimentálních datech. V rámci kapitoly je také osvětlena problematika odhadů v případě malých souborů dat, tzv. malých výběrů. Počínaje touto kapitolou budou za experimentální data považovány *konečné* množiny hodnot *kvantitativních* veličin - názorně řečeno: několik (třeba hodně, ale ne nekonečně) číselných hodnot.

Základní statistické (číselné) charakteristiky podávají koncentrované souhrnné informace o vlastnostech analyzovaného datového souboru. Základní statistické charakteristiky jsou také velmi často nazývány pojmem popisná statistika, případně deskriptivní statistika. Mezi nejpoužívanější statistické charakteristiky datového souboru patří především odhady míry polohy a variability analyzované proměnné. Míra polohy je střední hodnota, kolem které hodnoty sledované proměnné kolísají. Míra polohy se nejčastěji vyjadřuje prostřednictvím aritmetického průměru, geometrického průměru, mediánu, modusu atd. Míra variability udává, jak jsou hodnoty sledované proměnné rozptýleny (jak kolísají) kolem její střední hodnoty. Míra variability se nejčastěji vyjadřuje rozptylem, směrodatnou odchylkou, variačním koeficientem, interkvartilovým rozpětím atd. Podle konstrukce se číselné charakteristiky se dělí na momentové a kvantilové.

4.1 Klasické odhady míry polohy a variability

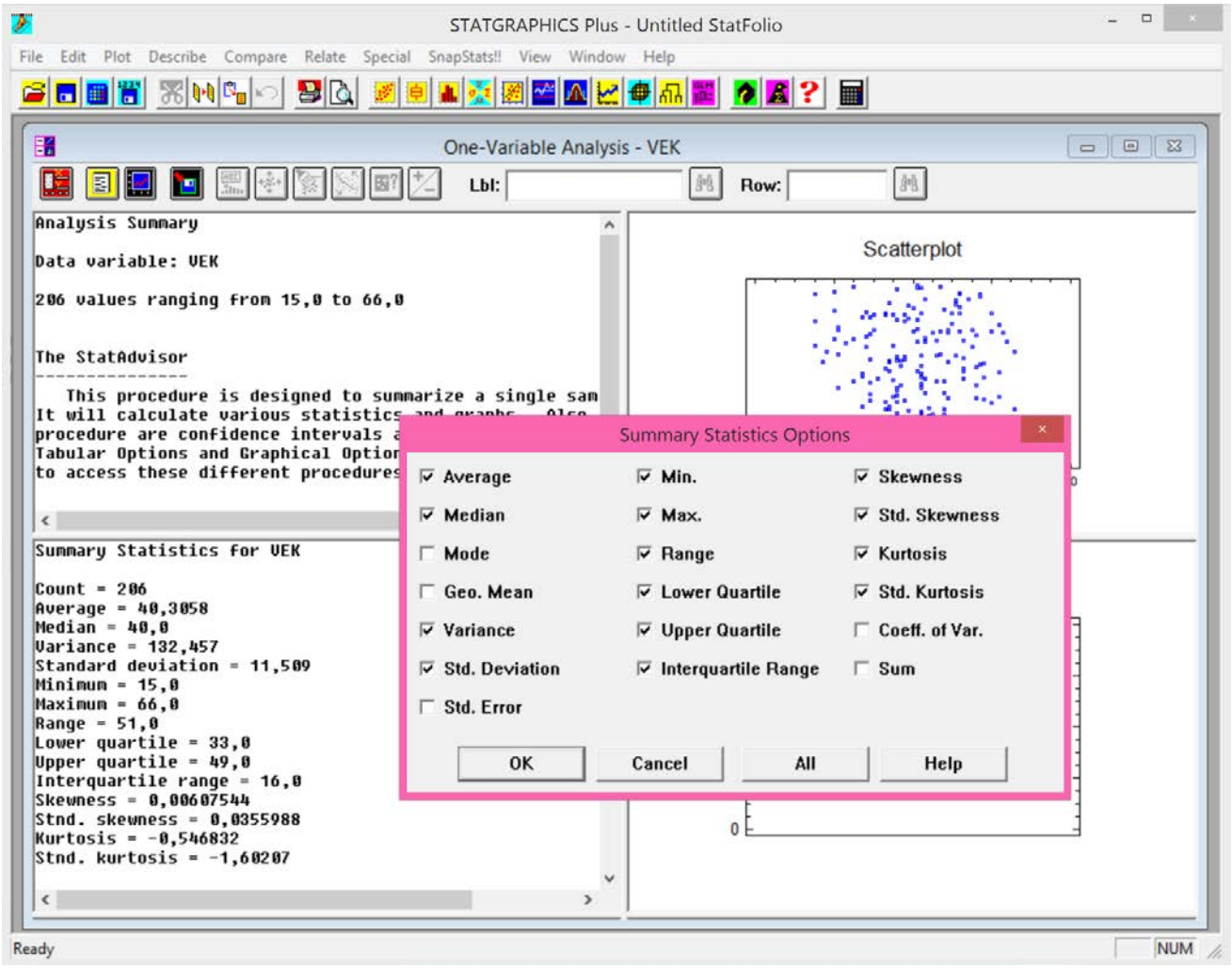
Klasické odhady míry polohy a variability jsou konstruovány na základě tzv. počátečních a centrálních momentů, jedná se tedy o momentové charakteristiky. Míra polohy představující střední hodnotu, je ve statistice označována jako první obecný moment. Míra variability představující rozptyl (disperzi, variaci) je označována jako druhý centrální moment. Matematické pozadí statistických momentů zde nebude blíže rozebíráno, nicméně v případě zájmu je lze nalézt v odborné literatuře, např. Svatošová a Kába (2007), Meloun a Militký (2004), Otyepka et al. (2013), Ott a Longnecker (2016).

Momentové charakteristiky jsou založeny na předpokladu normality rozdělení datového souboru a jsou velmi citlivé na odlehlé hodnoty. Při jejich použití v případě nesplnění uvedených podmínek mohou být výsledné odhady výrazně zkresleny (vypočtené odhady jsou nadhodnoceny, příp. podhodnoceny). Pokud uvedené podmínky splněny nejsou, je nutné pro odhad střední hodnoty použít odhad robustní. Mezi nejznámější momentové charakteristiky patří aritmetický průměr, rozptyl a z něj odvozená směrodatná odchylka.

Aritmetický průměr (mean, average), někdy také nazývaný výběrový průměr, je nejčastěji používaným odhadem míry polohy. Poskytuje maximálně věrohodný odhad střední hodnoty datového souboru při splnění předpokladu normality datového souboru a nepřítomnosti odlehlých hodnot. Výpočet aritmetického průměru je dán podílem sumy hodnot sledované proměnné x_i a rozsahu souboru n:

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

V programu Statgraphics je aritmetický průměr označen jako Average a vypočteme jej následujícím postupem: klikneme levou myší na příkaz Describe v jednořádkovém textovém menu a následně: Numeric Data → One-Variable Analysis. V dialogovém okně One-Variable Analysis zadáme do řádku Data název proměnné, pro kterou chceme aritmetický průměr vypočítat a klikneme na OK. V levém dolním okně se zobrazí souhrn základních popisných charakteristik souboru (Summary Statistics for "název zadané proměnné"). Jestliže chceme zobrazit více popisných charakteristik souboru, než je standardně programem nabízeno, klikneme pravou myší kamkoliv do prostoru levého dolního okna a v nabídce zvolíme Pane Options. Otevře se nám dialogové okno Summary Statistics Options. Ve zobrazené nabídce zatrhneme požadované popisné statistiky (viz obrázek 4.1).



Obrázek 4.1: Výstup analýzy jedné proměnné za účelem stanovení základních statistických charakteristik

Hodnotu aritmetického průměru je možné vypočíst i v tabulkovém procesoru Excel. V pracovním sešitě vypočteme aritmetický průměr pro zvolená data vložením funkce PRŮMĚR.

Geometrický průměr (geometric mean) je dalším parametrem odhadu míry polohy, který lze v programu Statgraphics vypočíst. Zde je označen jako Geo.Mean. Své uplatnění nachází především v oblasti zpracování ekonomických dat (např. cenové indexy, koeficienty růstu, vyčíslení průměrné inflace apod.). Je označován symbolem $\bar{x}_G$ a jeho hodnota se vypočte jako n-tá odmocnina součinu jednotlivých hodnot sledované proměnné x_i:

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

V tabulkovém procesoru Excel vypočteme geometrický průměr pro zvolená data vložením funkce GEOMEAN. Ve statistice se můžeme setkat s dalšími odhady míry polohy, jako např. vážený aritmetický průměr, harmonický průměr, ořezaný průměr, winsorizovaný průměr a další. Jelikož tyto odhady nejsou pro analýzu dat získaných z dotazníkových šetření v rámci turismu příliš vhodné, nebudou zde již blíže popisovány.

Rozptyl (variance), někdy také nazývaný disperze či **populační rozptyl**, je odhadem míry variability pro normálně rozdělená data základního souboru. Hodnota rozptylu nám popisuje rozptýlenost hodnot sledované proměnné kolem její střední hodnoty. Je tedy zřejmé, že čím větší je hodnota rozptylu, tím menší je schopnost stanovené střední hodnoty charakterizovat (reprezentovat) sledovanou proměnnou. Hodnota rozptylu σ^2 (sigma) se vypočte jako podíl sumy čtverců odchylek jednotlivých hodnot proměnné x_i od aritmetického průměru $\bar{x}$ a rozsahu souboru n:

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

V tabulkovém procesoru Excel vypočteme populační rozptyl pro zvolená data vložením funkce VAR.P (u verzí Excelu nižších než 2010 funkce VAR).

Ve statistické analýze dat je však nejčastěji používaným odhadem míry variability dat výběrového souboru tzv. **výběrový rozptyl**. Vzhledem k jeho vysoké frekvenci používání bývá právě tento odhad označován obecně jako rozptyl. Hodnota výběrového rozptylu s^2 se vypočte jako podíl sumy čtverců odchylek jednotlivých hodnot proměnné x_i od aritmetického průměru $\bar{x}$ a rozsahu souboru n sníženého o 1:

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

Snížením rozsahu o jedničku se docílí nevychýleného odhadu rozptylu, což platí především u odhadů pro menší datové soubory ($n < 30$). V případě použití výše uvedeného populačního rozptylu σ^2 dochází k podhodnocení odhadu.

V programu Statgraphics je výběrový rozptyl označen jako Variance a vypočteme jej stejným postupem, jež je uveden výše. V tabulkovém procesoru Excel vypočteme výběrový rozptyl pro zvolená data vložením funkce VAR.S (u verzí Excelu nižších než 2010 funkce VAR.VÝBĚR).

Směrodatná odchylka (standard deviation) - nevýhodou přímého použití odhadu míry variability pomocí rozptylu v praxi je fakt, že výsledná hodnota je druhou mocninou (kvadrátem) jednotky proměnné. Např. je-li denní úhrn spadlých dešťových srážek vyjádřen v jednotkách [mm], pak rozptyl bude vyjádřen v jednotkách [mm^2]. Za účelem odstranění takto vznikajících „podivných" jednotek byla zavedena statistická charakteristika nazývající se směrodatná odchylka. Směrodatná odchylka je druhou odmocninou z rozptylu, čímž je dosaženo shody jednotky s jednotkou proměnné. Stejně jako u rozptylu můžeme vypočítat směrodatnou odchylku pro základní soubor, která je označována symbolem σ:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

V tabulkovém procesoru Excel vypočteme směrodatnou odchylku základního souboru vložením funkce SMODCH.P (u verzí Excelu nižších než 2010 funkcí SMODCH).

Nejčastěji používaným odhadem je, stejně jako v případě rozptylu, směrodatná odchylka pro výběrový soubor (vypočítaná z výběrového rozptylu), jež je označována symbolem s:

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

V programu Statgraphics je směrodatná odchylka pro výběrový soubor označena jako Standard deviation a vypočteme ji stejným postupem, jež je uveden výše. V tabulkovém procesoru Excel vypočteme směrodatnou odchylku pro zvolená data vložením funkce SMODCH.VÝBĚR.S (u verzí Excelu nižších než 2010 funkcí SMODCH.VÝBĚR).

V odborné literatuře, nebo odborných článcích zabývajících se problematikou statistické analýzy dat, jsou hodnoty aritmetického průměru a směrodatné odchylky obecně vyjadřovány ve tvaru $\bar{x} \pm 1{,}96s/\sqrt{n}$ (platí pro hladinu pravděpodobnosti $\alpha = 0{,}05$), nebo ve tvaru $\bar{x} \cdot (s)$.

Variační koeficient (coefficient of variation, CV), nazývaný také relativní směrodatná odchylka (relative standard deviation, RSD), se využívá v případech, kdy potřebujeme srovnat variabilitu dvou, či více souborů majících různé jednotky, nebo hodnoty proměnné v různých řádech (např. [mm/h] vs. [l/den] apod.). Je vlastně mírou relativní variability, která vyjadřuje, jakou procentuální část aritmetického průměru představuje směrodatná odchylka. Variační koeficient se označuje symbolem v_x a jeho hodnota se vypočte jako podíl směrodatné odchylky s a aritmetického průměru $\bar{x}$ vyjádřený v procentech:

$$v_x = \frac{s}{\bar{x}} \cdot 100 \ [\%]$$

V programu Statgraphics je variační koeficient označen jako Coeff. of variation (viz obrázek 4.1). V tabulkovém procesoru Excel vypočteme variační koeficient pro zvolená data pomocí výše uvedeného vzorce s využitím funkcí SMODCH.VÝBĚR.S a PRŮMĚR.

Variační rozpětí (range) není příliš spolehlivým odhadem míry variability analyzovaného souboru, neboť je významně ovlivněn odlehlými hodnotami. Používá se jako rychlá a velmi orientační charakteristika pro předběžné vyhodnocení variability. Variační rozpětí se ve statistice označuje velkým písmenem R a jeho hodnota se vypočte jako rozdíl mezi největší hodnotou x_{max} a nejmenší hodnotou x_{min} analyzované proměnné:

$$R = x_{max} - x_{min}$$

V programu Statgraphics je variační rozpětí označeno anglickým názvem Range (viz obrázek 4.1). V tabulkovém procesoru Excel vypočteme variační rozpětí pro zvolená data pomocí výše uvedeného vzorce s využitím funkcí MAX a MIN.

Retransformovaný průměr je zvláštním typem klasických odhadů míry polohy provedený na transformovaných datech. Pokud se nám podaří v případě datového souboru, který nesplňuje předpoklady pro použití klasických odhadů, nalézt vhodnou transformaci (viz podkapitola **5.4 Transformace**), omezíme tímto vliv asymetrie a odlehlých hodnot a můžeme provést odhad míry polohy pomocí aritmetického průměru. Tuto výslednou statistickou charakteristiku následně pomocí zpětné transformace převedeme do původních hodnot. Odhad střední hodnoty se pak nazývá retransformovaný průměr $\bar{x}_R$.

4.2 Robustní odhady míry polohy a variability

Robustní odhady míry polohy a variability jsou konstruovány na základě tzv. kvantilů, jedná se tedy o kvantilové charakteristiky. Kvantilové charakteristiky jsou používány v případech, kdy datové soubory nesplňují předpoklady pro použití klasických odhadů. Jedná se tedy o soubory dat, které nesplňují předpoklad normality (mají asymetrické rozdělení), nebo jsou v datech přítomny odlehlé hodnoty. Mezi nejznámější kvantilové charakteristiky patří medián, dolní a horní kvartil, a interkvartilové rozpětí.

q-Kvantil: Základní ideou q-kvantilu pro statistické charakteristiky jednorozměrného datového souboru o N prvcích je rozdělení seřazených dat na q zhruba stejně početných podmnožin. Kvantily jsou pak hraniční hodnoty mezi dvěma sousedními podmnožinami. Přesně se **k-tý q-kvantil** náhodné proměnné V definuje jako taková hodnota v, pro kterou je pravděpodobnost, že hodnota náhodné veličiny je

- menší než v - nejvýše rovna k/q; $P(X<v) \leq k/q$,
- větší než v - nejvýše rovna (q-k) / q = 1 - k/q; $P(X>v) \leq 1-k/q$.

Jinak řečeno, k-tý q-kvantil je ta hodnota (dat), kde kumulativní distribuční funkce nabude nebo překročí hodnotu k/q.

Existuje tedy (q-1) q-kvantilů, a to pro každé celé číslo $k \in (0, q)$. Pro množinu seřazených N dat, indexovaných od 1 do N, k-tý q-kvantil je prvek dat s indexem $I_{kq} = \lceil N \,.\, k \,/\, q \rceil$. Z hlediska definice však v případě, že I=N.k/q je celé číslo, pak všechny hodnoty počínaje hodnotou dat s tímto indexem (X_i) až do hodnoty dat s následujícím indexem (X_{I+1}) mohou být kvantilem. V tomto případě bývá zvykem za kvantil považovat střed mezi těmito dvěma hodnotami. Není to však povinnost; za kvantil lze vzít např. menší z obou hodnot nebo mezi nimi interpolovat. Protože z q-kvantilu se odvozuje řada dalších statistik (viz dále), je právě zmíněná situace kritickým místem při jejich určení. Statistici sami se neshodnou už ani na tom, zda kvantily mají být pouze hodnoty datového souboru (viz výše určení jejich indexu), nebo to mohou být i hodnoty zkoumané veličiny v datech však neobsažené (viz výše např. střed mezi dvěma hodnotami).

***Poznámka**: $\lceil X \rceil$ je "horní celá část" X: je-li X celé, výsledkem je X, není-li celé, je výsledkem nejbližší vyšší celé číslo.*

***Poznámka**: Kvantil se často označuje symbolem $\tilde{x}_p$, kde index p udává polohu kvantilu uvnitř datového souboru a pohybuje se v rozsahu 0 až 1. Např. je-li zapsáno $\tilde{x}_{0,30}$, je tím označen 30% kvantil, to znamená, že 30 % hodnot v datovém souboru je menších než tento kvantil a 70 % hodnot je větších.*

Percentil pP je P-tý 100-kvantil. Z definice q-kvantilu pro q=100 se odvozuje i jiná definice: percentil pP je nejmenší hodnota **x** veličiny X, pro níž platí, že pro P% dat $\{x_i\}$ je splněna podmínka $x_i \leq pP$.

Příklad: Pro analýzu vhodnosti zalesnění devastovaných partií geolokalit byl zkoumán sadební materiál tvořený 400 náhodně sebranými semeny jedle. Jejich váhy v [g] se pohybovaly v intervalu <4.001; 4.999>. Percentil p5 je tedy taková váha, pro níž platí, že nejvýše 5% vah je ≤ p5. Protože celkem má datový soubor N = 400 hodnot, q = 100, k = 5, je $I_{kq} = \lceil 400 \,.\, 5 \,/\, 100 \rceil = 20$. Dvacátá hodnota seřazeného souboru dat byla rovna 4.191, je tedy p5 = 4.191.

Pořadí	1.	2.	...	19.	20.	21.	22.	...	399.	400.
Hodnota	4,001	4,014	...	4,190	**4,191**	**4,193**	4,196	...	4,996	4,999

Tabulka pro určení P5

Shora uvedená definice percentilu pomocí kvantilu je sice jednou z velmi často používaných, při jeho určení není však jednoznačná. Viz právě uvedený příklad: 21. hodnota v seřazeném souboru byla rovna 4.193. To ovšem znamená, že všechny hodnoty x < 4.193 splňují shora uvedenou definici percentilu, tedy p5 může být např. i 4.19256.

V Excelu se hodnoty percentilů vypočtou vložením funkce PERCENTIL.INC (u verzí Excelu nižších než 2010 vložením funkce PERCENTIL).

Medián x_m je definován jako 1-ní 2-kvantil, pomocí percentilu jako p50, tj. hodnota, "pod kterou" leží nejvýš polovina hodnot souboru a "nad kterou" leží nejvýš polovina hodnot souboru. Při jeho určení se v praxi postupuje podle definice q-kvantilu: jeho index $i = I_{kq} = \lceil N \,.\, 1 \,/\, 2 \rceil$, což je N/2 pro N sudé, (N+1)/2 pro N liché. Je-li tedy N liché, je mediánem hodnota X_i; je-li N sudé, nejčastěji se za medián přijímá hodnota hodnota $(X_i+X_{i+1})/2$. V jednotlivých krocích:

1. Soubor se uspořádá podle velikosti. Označme takto uspořádaný soubor $Y = \{y_1, y_2, \ldots, y_n\}$; každé y_i je tedy nějaké x_k.
2. Nechť m je celá část podílu n/2: m = [n/2]. Je-li n sudé, je n=2.m, je-li n liché, je n=2.m+1.
3. Je-li n liché, je mediánem hodnota y_{m+1}. Je-li n sudé, je mediánem hodnota $(y_m+y_{m++})/2$.

Příklad: Medián souboru s váhami semen jedle se zjistí postupem popsaným výše. Nejprve se hodnoty seřadí podle velikosti. Získá se následující posloupnost hodnot:

Pořadí	1.	2.	...	199.	200.	201.	202.	...	399.	400.
Hodnota	4,001	4,014	...	4,498	**4,500**	**4,501**	4,501	...	4,996	4,999

Tabulka pro určení mediánu

"Polovinou" souboru je hranice mezi 200.tým a 201.ním prvkem (prvků je sudý počet). Medián je tedy polovina mezi 4,500 a 4,501, tj. **4,5005**.

V programu Statgraphics je vypočtená hodnota mediánu označena jako Median (viz obrázek 4.1). V tabulkovém procesoru Excel vypočteme medián pro zvolená data vložením funkce MEDIAN.

Dolní a horní kvartil q_D a q_H jsou definovány jako 1-ní a 3-tí 4-kvantil, pomocí percentilu jako p25 resp. p75. Označují se proto často také $\tilde{x}_{0,25}$ a $\tilde{x}_{0,75}$. Nepřesně ale názorně řečeno, jsou to hodnoty, "pod kterými" leží nejvýš čtvrtina resp. tři čtvrtiny hodnot souboru a "nad kterými" leží nejvýš tři čtvrtiny resp. čtvrtina hodnot souboru. Indexy dolního resp. horního kvartilu jsou podle definice $I_{kq} = \lceil N \,.\, 1 \,/\, 4 \rceil$ resp. $I_{kq} = \lceil N \,.\, 3 \,/\, 4 \rceil$.

Pro určení hodnot kvartilů však není stanoven žádný jednotný postup. Jednak sami statističtí odborníci používají několik metod, jednak autoři statistického software aplikují různé algoritmy (a pohříchu ani přesně neřeknou jaké). Zhruba lze vidět následující metodiky při určování kvartilů (všechny splňují definici a medián naštěstí všechny určují pro sudý počet dat stejně, jako střed mezi prostředními hodnotami):

A. Soubor se uspořádá podle velikosti. Mediánem se rozdělí na dvě části, ale medián se nezařadí do žádné z nich. Dolní kvartil je pak medián dolní poloviny dat, horní kvartil je medián horní poloviny dat. Metodu popsali např. Moore a McCabe a často se užívá v software statistických kalkulátorů nebo jiné výpočetní techniky.
B. Soubor se uspořádá podle velikosti. Mediánem se rozdělí na dvě části. Je-li mediánem prostřední datová hodnota (rozsah souboru je liché číslo), pak se zařadí do obou polovin. Dolní kvartil je pak medián dolní poloviny dat, horní kvartil je medián horní poloviny dat. Pro sudý počet dat je tedy tato metoda totožná s předchozí. Metodu popsal např. Tukey.
C. Pro sudý počet dat se aplikuje předchozí metoda. Je-li počet N dat lichý, pak existuje celé číslo n takové, že N=4.n+1 nebo N=4.n+3 (4.n+0 a 4.n+2 jsou sudá). Je-li N=4.n+1, je dolní kvartil $q_D = 1/4.X_n+3/4.X_{n+1}$ a horní kvartil $q_H = 3/4.X_{3n+1}+1/4.X_{3n+2}$. Je-li N=4.n+3, je dolní kvartil $q_D = 3/4.X_{n+1}+1/4.X_{n+2}$ a horní kvartil $q_H = 1/4.X_{3n+2}+3/4.X_{3n+3}$.

D. Určí se hodnota i_D = (N+1)/4 a zaokrouhlí se na nejbližší celé číslo. Pokud je i_D přesně mezi dvěmi celými čísly, zaokrouhlí se nahoru. Prvek dat s takto zaokrouhleným indexem i_D je dolním kvartilem. Určí se dále hodnota i_H = 3.(N+1)/4 a zaokrouhlí se na nejbližší celé číslo. Pokud je i_H přesně mezi dvěmi celými čísly, zaokrouhlí se dolů. Prvek dat s takto zaokrouhleným indexem i_H je horním kvartilem. Takto určené kvartily jsou tedy vždy prvky datového souboru. Metodu popsali např. Mendenhall a Sincich.

E. Určí se hodnota i = (N+1)/4. Je-li i celé číslo, je dolním kvartilem prvek dat s indexem i, tj. x_i. Není-li i celé číslo, položí se k = [i] a interpoluje se mezi prvky x_k a x_{k+1}. Dolním kvartilem je pak hodnota $x_k + (x_{k+1} - x_k).(i-k)$. Určí se hodnota j = 3.(N+1)/4. Je-li j celé číslo, je horním kvartilem prvek dat s indexem j, tj. x_j. Není-li j celé číslo, položí se k = [j] a interpoluje se mezi prvky x_k a x_{k+1}. Horním kvartilem je pak hodnota $x_k + (x_{k+1} - x_k).(j-k)$. Metodu používají některé statistické aplikace, např. Minitab.

F. Určí se hodnota i = (N+3)/4. Je-li i celé číslo, je dolním kvartilem prvek dat s indexem i, tj. x_i. Není-li i celé číslo, položí se k = [i] a interpoluje se mezi prvky x_k a x_{k+1}. Dolním kvartilem je pak hodnota $x_k + (x_{k+1} - x_k).(i-k)$. Určí se hodnota j = (3.N+1)/4. Je-li j celé číslo, je horním kvartilem prvek dat s indexem j, tj. x_j. Není-li j celé číslo, položí se k = [j] a interpoluje se mezi prvky x_k a x_{k+1}. Horním kvartilem je pak hodnota $x_k + (x_{k+1} - x_k).(j-k)$. Metodu popsali např. Freund a Perles a používá ji např. Excel.

Příklad - Dolní kvartil souboru s vahami semen jedle metodikou A: Nejprve se hodnoty seřadí podle velikosti. Získá se posloupnost hodnot, jejíž důležitá část je v následující tabulce:

Pořadí	1.	2.	...	100.	101.	...	199.	200.
Hodnota	4,001	4,014	...	**4,378**	**4,379**	...	4,498	4,500

Tabulka pro určení dolního kvartilu

Počet prvků souboru je 400, tedy sudé číslo. Dolní polovinu prvků tvoří prvky s indexy z intervalu <1,200>. Dolní kvartil je roven jejich mediánu, a protože jich je sudý počet, je roven

4,379 - 0,25 . (4,739 - 4,738) = **4,37875**.

Horní kvartil souboru s vahami semen jedle metodikou A: Nejprve se hodnoty seřadí podle velikosti. Získá se posloupnost hodnot, jejíž důležitá část je v následující tabulce:

Pořadí	201.	202.	...	300.	301.	...	399.	400.
Hodnota	4,501	4,501	...	**4,634**	**4,635**	...	4,996	4,999

Tabulka pro určení horního kvartilu

Počet prvků souboru je 400, tedy sudé číslo. Horní polovinu prvků tvoří prvky s indexy z intervalu <201,400>. Horní kvartil je roven jejich mediánu, a protože jich je sudý počet, je roven

4,635 - 0,75 . (4,635 - 4,634) = **4,63425**.

V programu Statgraphics najdeme vypočtenou hodnotu dolního kvartilu pod anglickým pojmem Lower quartile a hodnotu horního kvartilu pod pojmem Upper quartile. V tabulkovém procesoru Excel vypočteme dolní a horní kvartil pro zvolená data vložením funkce QUARTIL.INC (u verzí Excelu nižších než 2010 vložením funkce QUARTIL).

V případě kvantilových charakteristik se také můžeme setkat s pojmem **decily**. Decily dělí datový soubor na deset stejně početných částí a jedná se tedy o 10% kvantily.

Modus (mode) je robustním odhadem míry polohy a ve statistice je označován symbolem $\hat{x}$. Lze jej definovat jako hodnotu, v níž nabývá frekvenční funkce svého maxima. Modus je tedy hodnota, která se v analyzovaném datovém souboru vyskytuje nejčastěji, čili s největší četnosti. Modus nemá příliš velkou vypovídací schopnost a v analýze experimentálních dat je využíván jen zřídka.

V programu Statgraphics je modus označen jako Mode (viz obrázek 4.1). V tabulkovém procesoru Excel lze modus pro zvolená data vypočíst vložením funkce MODE.SNGL (u verzí Excelu nižších než 2010 funkce MODE).

Interkvartilové rozpětí (interquartile range), nazývané také mezikvartilové rozpětí, je robustním odhadem míry variability a označuje se symbolem RF (velmi sporadicky se vyskytuje označení symbolem IQR). Jedná se o odhad založený na kvantilech a používá se pouze tam, kde nelze pro odhad variability datového souboru použít momentové charakteristiky (např. směrodatnou odchylku). Interkvartilové rozpětí se vypočte jako rozdíl horního a dolního kvartilu:

$$R_F = \tilde{x}_{0,75} - \tilde{x}_{0,25}$$

V programu Statgraphics je interkvartilové rozpětí označeno anglickým pojmem Interquartile range (viz obrázek 4.1).

Kvartilová odchylka (quartile deviation, semi-inter quartile range) je taktéž robustním odhadem míry polohy založeném na kvantilech. Označuje se symbolem Q_F a obvykle se vyčísluje spolu s mediánem. Její hodnota se vypočte jako polovina rozdílu horního a dolního kvartilu:

$$Q_F = \frac{\tilde{x}_{0.75} - \tilde{x}_{0.25}}{2}$$

4.3 Odhady míry polohy a variability u malých datových souborů

Pokud máme datový soubor obsahující velmi malé množství dat, stojíme před problémem jak správně odhadnout míru polohy a variability. Jak uvádí Meloun a Militký (2011), v případě datového souboru obsahujícího pouze dvě hodnoty (n = 2) lze použít pro vyčíslení střední hodnoty aritmetický průměr. Ovšem pouze za předpokladu, že jsou si měřené hodnoty blízké. U dvou hodnot velmi rozdílných tento postup použít nelze, jelikož nejsme schopni odhadnout, která hodnota je hodnotou odlehlou. U datového souboru obsahujícího pouze tři hodnoty (n = 3) je určení střední hodnoty opět velmi obtížné. Obvykle se pro výpočet použije aritmetický průměr, avšak pouze ze dvou nejbližších hodnot. Tento postup v praxi přináší výrazně lepší výsledky, než výpočet mediánu ze všech tří hodnot. Pro odhad variability, tedy směrodatné odchylky, lze u souborů obsahujících dvě až tři hodnoty použít pro výpočet rovnici, kterou uvádí ve své publikaci Potts (1992):

$$s = \frac{w}{k}$$

kde w je rozdíl mezi nejvyšší a nejnižší hodnotou souboru; k je faktor závislý na počtu pozorování (počtu hodnot). Faktorem k pro n = 2 je hodnota 1,128, faktorem k pro n = 3 je hodnota 1,693.

Poněkud lepší situace nastává u souborů dat, které obsahují 4 až 20 naměřených experimentálních hodnot. V případě takovýchto souborů je pro odhad míry polohy a variability doporučováno použít tzv. Hornův postup (Horn, 1983). Jak uvádí Meloun a Militký (2011), Hornův postup poskytuje v případě malých výběrů (pro $4 \leq n \leq 20$) nejlepší odhad střední hodnoty (odhad parametru polohy) vyjádřený pivotovou polosumou (P_L) a nejlepší odhad variability (odhad směrodatné odchylky) vyjádřený pivotovým rozpětím (R_L). Výpočetní postup odhadu střední hodnoty a variability metodou Hornova postupu je následující:

Prvním krokem ve výpočetním postupu je seřazení experimentálních dat vzestupně, čili od nejmenší po největší.

Dalším krokem je výpočet hloubky pivotu H, kde n je počet hodnot a $\lfloor X \rfloor$ je celá část čísla X:

$$H = \left\lfloor \frac{n+1}{2} \right\rfloor / 2 \quad \text{(pro n liché)}$$

$$H = \left\lfloor \frac{n+1}{2} + 1 \right\rfloor / 2 \quad \text{(pro n sudé)}$$

V třetím kroku je stanovena hodnota dolního pivotu $x_D = x_{(H)}$ a horního pivotu $x_H = x_{(n+1-H)}$.

Posledním krokem je výpočet pivotové polosumy P_L (odhad střední hodnoty) a pivotového rozpětí R_L (odhad parametru variability, např. směrodatné odchylky):

$$P_L = \frac{x_D + x_H}{2}$$

$$R_L = x_H - x_D$$

V praxi se obecně se pro soubory dat obsahujících nanejvýš 8 hodnot využívá Hornova postupu. Pro soubory dat obsahujících více než 8 hodnot již lze k výpočtům využít statistický software (tzn. použít pro výpočet klasické nebo robustní odhady parametrů).

5 Průzkumová analýza dat

Kapitola je zaměřena na jeden z nejdůležitějších počátečních kroků při statistické analýze dat, kterým je průzkumová analýza dat. Jsou zde podrobně vysvětleny její jednotlivé nástroje, a to grafické diagnostiky, statistické testy, včetně problematiky identifikace odlehlých hodnot. Součástí kapitoly je rovněž řada názorných příkladů, jejich řešení v prostředí programu Statgraphics a podrobná interpretace výsledků.

Průzkumová analýza dat je základním a v současnosti široce využívaným nástrojem při statistickém zpracování dat. Její základy položil v roce 1977 ve své knize "*Exploratory data analysis*" americký matematik John Wilder Tukey (Tukey, 1977). Průzkumová analýza dat bývá velmi často označovaná zkratkou EDA podle anglického názvu Exploratory Data Analysis, v některých publikacích IDA podle anglického názvu Initial Data Analysis. Účelem průzkumové analýzy dat je odhalení zvláštností statistického souboru a ověření předpokladů pro následné statistické zpracování (Meloun a Militký, 2006). Základem EDA je využití řady grafických diagnostik (histogram četnosti, krabicový graf, Q-Q graf atd.) a statistických testů (test normality, test odlehlých hodnot atd.), podle nichž můžeme následně rozhodnout o rozdělení souboru dat (normalita rozdělení) a identifikovat odlehlé hodnoty. Převážná většina metod statistické analýzy je založena na splnění výše uvedených vlastností datových souborů. Identifikace těchto předpokladů je proto nezbytnou podmínkou pro správné vyčíslení základních statistických charakteristik a volbu statistické metody pro analýzu dat.

5.1 Diagnostika grafů

V rámci průzkumové analýzy dat je prvním a nejdůležitějším krokem diagnostika grafických výstupů. Statistické testy bývají často málo přísné a nespolehlivé, proto je vhodné v prvé řadě spoléhat na výsledky grafické diagnostiky a výsledky statistických testů používat pouze pro jejich ověření.

Grafů pro grafickou diagnostiku datových souborů existuje velké množství, ovšem každý statistický software nabízí vykreslení jen některých z nich. V této podkapitole budou popsány pouze grafy, které umožňuje vytvořit program Statgraphics.

5.1.1 Histogram četnosti

Histogram četnosti (frequency histogram) je jedním z nejstarších a nejpoužívanějších grafů v průzkumové analýze dat jedné proměnné. Histogram četnosti je intervalový sloupcový graf, v němž jsou na ose x zobrazeny intervaly (třídy) hodnot proměnné. Jednotlivé intervaly jsou rovny šířce sloupce, přičemž šířky sloupců jsou si navzájem rovny. Hodnoty na ose y odpovídají absolutním, nebo relativním četnostem tříd (intervalů). Absolutní četnost je vyjádřená skutečným počtem zastoupených hodnot v jednotlivých třídách, čili udává, kolik hodnot z datového souboru se vyskytuje v daných intervalech (třídách). Relativní četnost udává zastoupení hodnot v jednotlivých třídách vyjádřenou v procentech. Jednotlivým četnostem odpovídají v grafu výšky sloupce.

Při vytváření grafu svépomocí se může uživatel dostat do svízelné situace při volbě šířky a počtu sloupců. V případě, že si sloupců vytvoříme příliš moc, nebo málo, může dojít k vytvoření histogramu bez potřebné vypovídací schopnosti. Pro tyto účely je vhodné použít ke konstrukci histogramu tzv. Sturgesovo pravidlo, pomocí kterého stanovíme optimální počet a šířku sloupců. V současnosti většina statistických programů, včetně

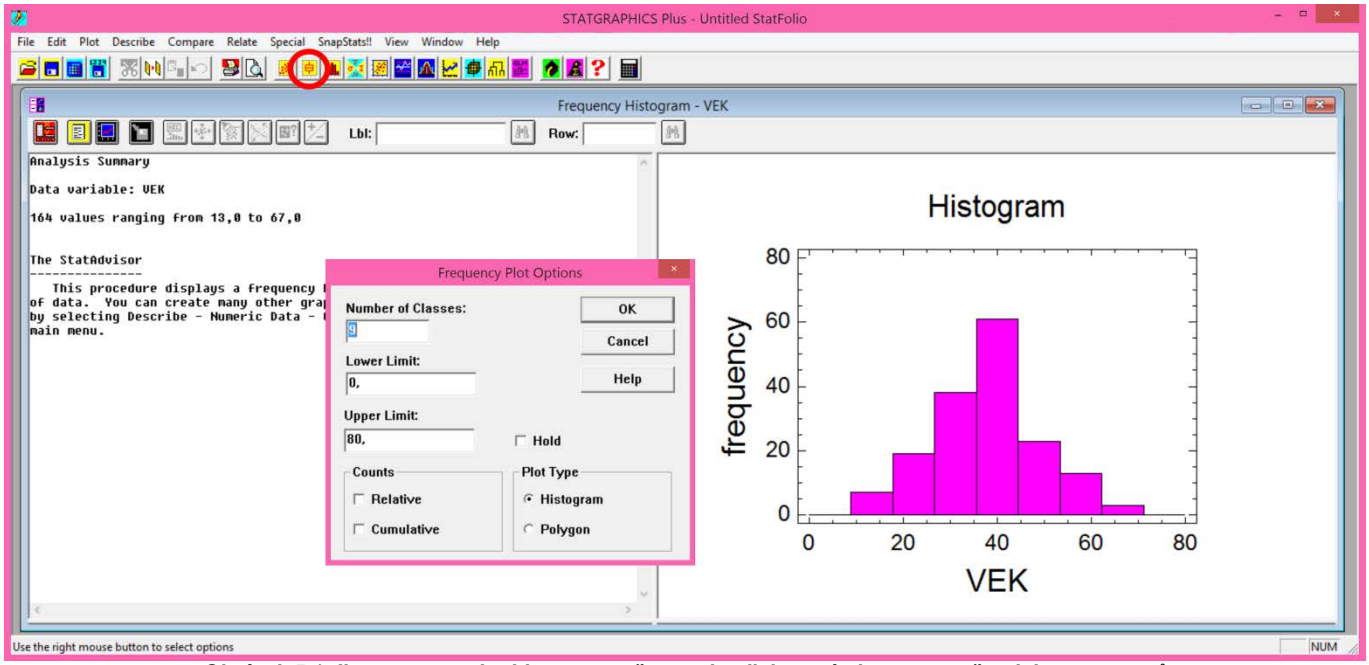
tabulkového procesoru Excel, Sturgesovo pravidlo, nebo jeho modifikaci, ke konstrukci histogramu četnosti využívá.

Postup konstrukce histogramu dle Sturgesova pravidla:

1. Stanovíme počet tříd k dle následující rovnice:
   $k = 1 + [3,3 * \log n]$
   kde n je celkový počet hodnot (rozsah souboru); zápis [X] značí zaokrouhlení hodnoty X nahoru.
2. Určíme šířku třídy (intervalu) T dle následující rovnice:
   $T = (x_{max} - x_{min}) / k$
   kde x_{max} je maximální hodnota v souboru a x_{min} je minimální hodnota v souboru.
3. Určíme meze tříd. Meze první třídy budou $<x_{min}; x_{min} + T>$, meze následující třídy budou $(x_{min} + T; x_{min} + 2T>$, $(x_{min} + 2T; x_{min} + 3T>$ atd. Přičítání násobků šířky třídy provádíme tak dlouho, dokud nemáme stanoveno tolik intervalů, kolik jsme vypočetli tříd k.
4. Určíme četnost zastoupení hodnot v jednotlivých třídách. Čili stanovíme, kolik hodnot z našeho datového souboru odpovídá rozsahu každé jednotlivé třídy.
5. Stanovené hranice tříd vyneseme na osu x, stanovené četnosti vyneseme na osu y.

V prostředí programu Statgraphics lze histogram četnosti vytvořit následujícím postupem: klikneme levou myší na příkaz Plot v jednořádkovém textovém menu a následně Exploratory Plots→Frequency Histogram. Otevře se nám dialogové okno Frequency Histogram, v němž do řádku Data zadáme název proměnné, pro kterou chceme histogram vykreslit a klikneme na OK. Nejrychlejším způsobem jak v programu vytvořit histogram četnosti je kliknutím levou myší na ikonu Histogramu v horní liště s ikonami (viz obrázek 5.1). Tímto postupem program automaticky vytvoří Histogram četnosti s optimální šířkou a počtem sloupců.

Po vytvoření histogramu můžeme počet sloupců a koncové body libovolně měnit. Dále je možné zadat, v jakém formátu chceme mít znázorněnou četnost na ose y (absolutní, nebo relativní četnost). Tyto změny provedeme kliknutím pravou myší do prostoru vytvořeného histogramu. Zobrazí se nabídkové menu, v němž zvolíme příkaz Pane Options a v dialogovém okně Frequency Plot Options zadáme požadované parametry (viz obrázek 5.1).



Obrázek 5.1: Ikona pro tvorbu histogramu četnosti a dialogové okno pro změnu jeho parametrů

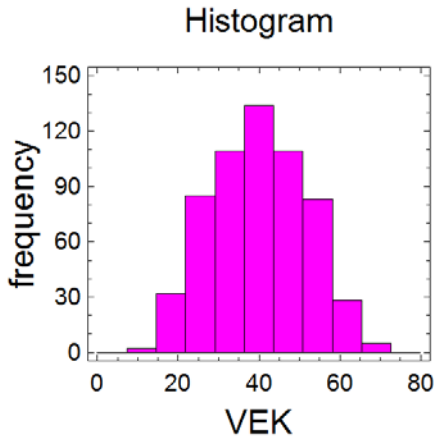
Histogram četnosti lze vytvořit i v tabulkovém procesoru Excel a to následujícím postupem: klikneme myší na záložku Data a v souboru příkazů Analýza vybereme příkaz Analýza dat. V dialogovém okně Analytické nástroje zvolíme příkaz Histogram (Data→Analýza dat→Histogram). V případě, že se v záložce Data nabídka Analýza dat nezobrazuje, je nutné si ji v Excelu aktivovat. Aktivaci provedeme kliknutím levou myší na záložku Soubor→Možnosti→Doplňky. V nabídkovém okně Doplňky v části Neaktivní doplňky aplikací zadáme Analytické

nástroje a klikneme na tlačítko Přejít. Zobrazí se nám nabídkové menu, v němž zatrhneme Analytické nástroje a potvrdíme stiskem tlačítka OK.

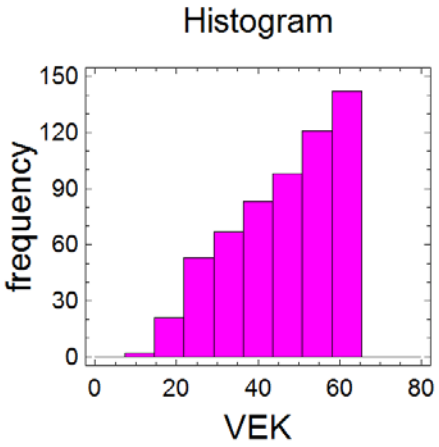
Interpretace histogramu četnosti

Diagnostikou histogramu četnosti můžeme u datového souboru rozhodnout o splnění předpokladu normality (symetrii rozdělení), homogenitě souboru (to znamená, že můžeme určit, zda se soubor rozpadá do dílčích podsouborů, či ne) a identifikovat odlehlé hodnoty (tzv. outliers).

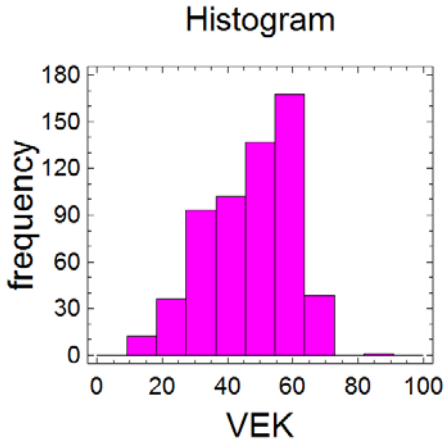
Mějme histogramy četnosti, které byly vytvořeny na základě datových souborů obsahujících věk návštěvníků čtyř různých geolokalit. Histogram četnosti na obrázku 5.2 znázorňuje věk návštěvníků geolokality číslo 1. Z grafu je zřejmé, že se jedná o datový soubor s normálním (symetrickým) rozdělením bez odlehlých hodnot. Dále můžeme říct, že nejvíce navštěvují geolokalitu č. 1 turisté ve věku přibližně 36 až 44 let. Na obrázku 5.3 je uveden histogram četnosti věku návštěvníků geolokality č. 2. Datový soubor má jednoznačně asymetrické rozdělení (není splněn předpoklad normality). Data jsou záporně sešikmené k nižším hodnotám bez přítomnosti odlehlých hodnot. V případě geolokality č. 2 převažují turisté z vyšších věkových kategorií, přičemž největší návštěvnost je ve věkové skupině kolem 58 až 65 let.



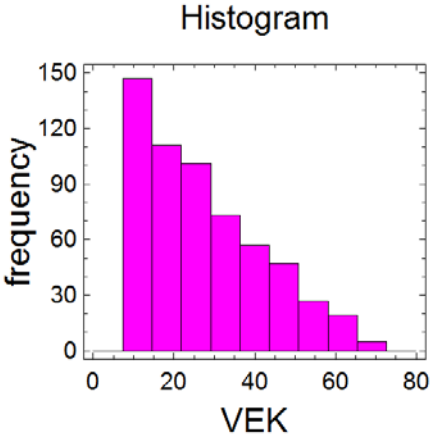
Obrázek 5.2: Histogram četnosti – normální rozdělení dat



Obrázek 5.3: Histogram četnosti – asymetrické rozdělení dat záporně sešikmené



Obrázek 5.4: Histogram četnosti – asymetrické rozdělení dat s odlehlou hodnotou



Obrázek 5.5: Histogram četnosti – asymetrické rozdělení dat kladně sešikmené

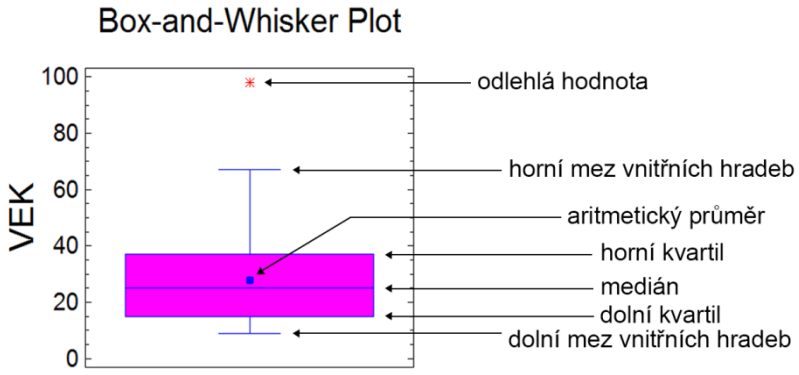
Obrázek 5.4 znázorňuje histogram věku turistů z geolokality č. 3. Zde je opět zřejmé asymetrické rozdělení datového souboru sešikmené k nižším hodnotám. Vpravo lze identifikovat přítomnost jedné odlehlé hodnoty. Lze konstatovat, že geolokalitu č. 3, stejně jako v případě geolokality č. 2, navštěvují převážně turisté vyšších věkových kategorií, přičemž nejčastěji turisté ve věkové skupině 54 až 63 let. Jedna odlehlá hodnota vpravo značí,

že geolokalitu č. 3 zřejmě navštívil jeden turista ve věkové skupině 82 a 91 let. Může se ovšem také jednat o hrubou chybu, tzn., že došlo ke špatnému zápisu hodnoty (výšky věku) během dotazníkového průzkumu (problematika odlehlých hodnot viz podkapitola 5.3 Odlehlé hodnoty a jejich identifikace). Obrázek 5.5 znázorňuje histogram věku návštěvníků geolokality č. 4, ze kterého je patrné, že data jsou asymetricky rozdělena a kladně sešikmená k vyšším hodnotám. V souboru dat se nevyskytují odlehlé hodnoty. Lze konstatovat, že geolokalitu navštěvují především turisté nižších věkových skupin, s převahou turistů ve věku 7 až 14 let.

5.1.2 Krabicový graf

Krabicový graf je v současnosti nejpoužívanějším grafem průzkumové analýzy dat a je součástí téměř všech statistických programů. Velmi často se můžeme setkat i s názvem krabička s vousy, nebo vousatá krabička. Tyto názvy byly převzaty z anglického názvu Box and Whisker Plot. Některé statistické programy používají název Boxplot.

Konstrukce krabicového grafu je založena na pěti základních statistických charakteristikách a to na mediánu, horním kvartilu, dolním kvartilu, minimální a maximální hodnotě souboru (viz obrázek 5.6). Základem grafu je obdélník (krabice), jehož dolní vodorovná linie (dno) znázorňuje dolní kvartil ($\tilde{x}_{0,25}$) a horní vodorovná linie (víko) horní kvartil ($\tilde{x}_{0,75}$). Obdélník je rozdělen další vodorovnou linií, která představuje medián datového souboru. Výška obdélníku je dána rozdílem mezi dolním a horním kvartilem a nazývá se interkvartilové rozpětí. Interkvartilové rozpětí obsahuje 50 % dat souboru. Z obdélníku vybíhají úsečky (vousy). Konec dolní úsečky se nazývá dolní mez vnitřních hradeb a udává minimální hodnotu souboru bez odlehlých hodnot. Konec horní úsečky se nazývá horní mez vnitřních hradeb a udává maximální hodnotu souboru bez odlehlých hodnot. Odlehlými hodnotami jsou v krabicovém grafu ta data, která jsou větší, nebo menší, než 1,5 násobek interkvartilového rozpětí. Odlehlé hodnoty leží v grafu mimo rozsah úseček a jsou znázorněny jako izolované body pomocí čtverečku, kroužku, hvězdičky apod. (viz obrázek 5.6). Některé statistické programy umožňují nastavit délku úseček (vousů) libovolně, např. od minima do maxima včetně odlehlých bodů a vyznačit hodnotu aritmetického průměru datového souboru (pomocí křížku, kroužku, čtverečku apod.).

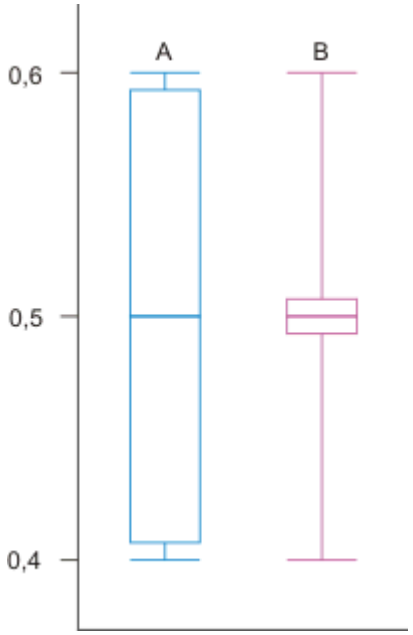


Obrázek 5.6: Krabicový graf

Vypovídací schopnost krabicového grafu je poměrně značná, pokud je správně chápána jeho podstata. Právě pro pochopení vypovídací schopnosti krabicového grafu lze ve výuce celkem s úspěchem použít tento příklad:

V dnešní bio- době, kdy výrobci označují s oblibou své potraviny právě předponou bio-, máme bio- snad všechny druhy potravin. Krkonošské, Beskydské a jiné bio-krávy žerou jen bio-trávu, produkují bio-hnůj - ale také bio-mléko. Konkurence je veliká, bio-farmáři soupeří o bio-zákazníka nejrůznějším způsobem: od mléčných (samozřejmě bio) automatů na mléko až po osobní odběr snad rovnou od bio-vemene. Máme-li v rozumném perimetru několik takových producentů mléka, jak vybrat toho nejlepšího? Běžný zákazník rozhodně nebude porovnávat kvalitu, protože už jen fyzikální a mikrobiologické analýzy by ho finančně zruinovaly - navíc celkem oprávněně tuší, že výsledky u všech určitě budou splňovat přísné normy EU.

Zákazník se tedy zaměří na poctivost prodejce. Začne odebírat a pečlivě měřit půllitrové dávky a hodnotit je metodami, kterým se naučil přečtením tohoto článku. Vybere dva jemu nejbližší prodejce A a B, odebere od každého 100 dávek deklarovaných jako půllitrové a z naměřených skutečných objemů nechá sestavit krabicové grafy. Získá následující (obrázek 5.7):

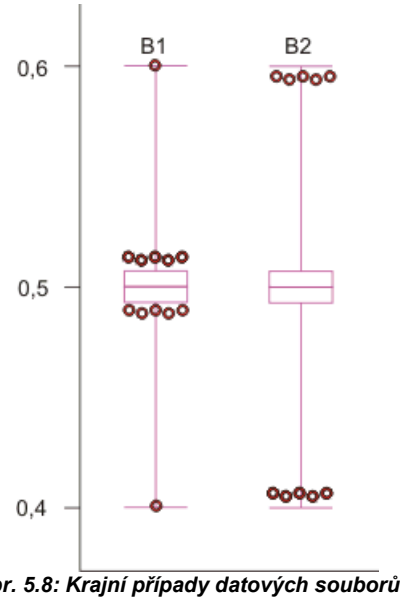


Obr. 5.7: Krabicový graf půllitrových dávek bio-mléka

Otázkou tedy je: který z obou je poctivější, ke kterému raději chodit?

Stejnou otázku klade autor tohoto článku svým studentům po vyslechnutí přednášky o kvartilech, mediánu a krabicovém grafu. Přibližně 20 % se jich přiklání v prodejci A, aniž však dovedou popsat důvod. 80 % prohlásí za lepšího prodejce B a zdůvodňují to tím, že mnoho jeho prodejů je téměř přesně půl litru (to, že oba nám alespoň jednou deci ubrali, ale na druhé straně alespoň jednou nám deci přidali - to obě skupiny studentů shodně potvrzují). Ovšem na otázku druhé skupině - kolik nám tedy ze 100 dávek prodali skoro přesně - se ozývají tipy od 90, 94 - a co třeba 96? Třeba i to!

Teprve po několikerém připomenutí konstrukce krabicového grafu začnou posluchači chápat, že ony dva nízoučké obdélníky v případě B reprezentují 25 % + 25 % = 50 naměřených hodnot. Tedy krabicový graf sděluje, že prodejce B nám nejméně z poloviny naměřil skoro úplně přesně. Dále zobrazuje skutečnost, že nejméně jednou nám prodal o deci méně a nejméně jednou o deci více. Jak je to ovšem se zbývajícími 24 prodeji pod správnou míru a 24 prodeji nad správnou míru - o tom už tento tvar krabicového grafu nevypovídá. Mohou nastat krajní případy, z nich dva jsou na obrázku 5.8 schematicky znázorněny jako B1 a B2:



Obr. 5.8: Krajní případy datových souborů B

V případě B1 jen jediná hodnota je extrémně odlehlá směrem k minimu, zatímco zbývajících 24 je velmi blízko "středním" 50 hodnotám; analogicky směrem k maximu. V případě B2 je tomu naopak: všech 24 nízkých hodnot je téměř u minima a všech 24 vysokých hodnot je téměř u maxima. Příklady dat odpovídajících grafům B1 a B2 jsou v následujících dvou četnostních tabulkách:

0.600	1x
0.510	24x
0.500	50x
0.490	24x
0.400	1x

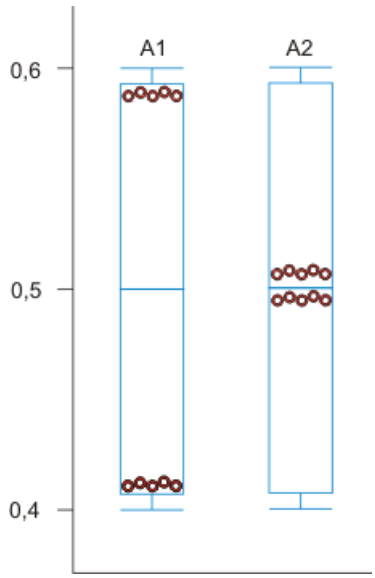
Data, jejichž krabicový graf je B1

0.600	1x
0.595	23x
0.510	1x
0.500	50x
0.490	1x
0.405	23x
0.400	1x

Data, jejichž krabicový graf je B2

Pokud v praxi dávají data krabicový graf podobný případu B, pak následujícím krokem by měl být rozbor datového souboru. Nejčastěji se zjistí, že jen nepatrné množství dat tvoří odlehlé hodnoty (případ B1: stačí zjistit příčinu a tyto hodnoty ze souboru vyloučit), nebo že došlo ke smíchání dat třech různých souborů do jednoho (případ B2: stačí zpracovat tři soubory samostatně).

Ovšem i případ A stojí za pozornost. Krabicový graf A totiž vypovídá o tom, že nejméně 25x nás téměř o deci ošidili - ovšem na druhé straně nám nejméně 25x téměř deci přidali. Jak je to však ve zbývajících 50 případech, o tom krabicový graf už nevypovídá. Mohou nastat krajní případy, z nich dva jsou na obrázku 5.9 schematicky znázorněny jako A1 a A2:



Obr. 5.9: Krajní případy datových souborů A

V případě A1 je dalších 25 hodnot téměř rovných velmi malým hodnotám poblíž minima, a zbývajících 25 hodnot téměř rovných velmi velkým hodnotám poblíž maxima. V případě A2 je všech zbývajících 50 hodnot velmi podobných někde v okolí prostřední hodnoty.

Příklady dat odpovídajících grafům A1 a A2 jsou v následujících dvou četnostních tabulkách:

0.600	1x
0.595	49x
0.405	49x
0.400	1x

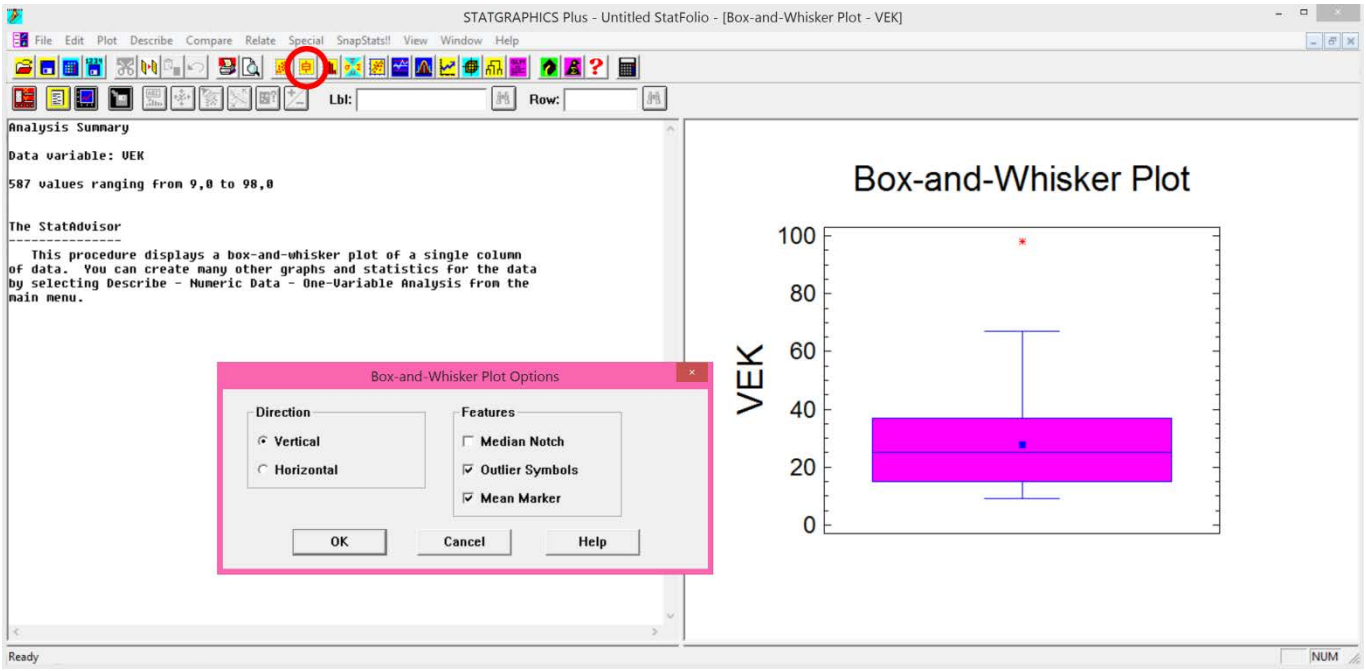
Data, jejichž krabicový graf je A1

0.600	25x
0.590	1x
0.500	48x
0.410	1x
0.400	25x

Data, jejichž krabicový graf je A2

Pokud v praxi dávají data krabicový graf podobný případu A, pak následujícím krokem by měl být rozbor datového souboru. Nejčastěji se zjistí, že došlo ke smíchání dat dvou různých souborů do jednoho (případ A1: stačí zpracovat dva soubory samostatně), nebo že došlo ke smíchání dat třech různých souborů do jednoho (případ A2: stačí zpracovat tři soubory samostatně).

V programu Statgraphics se krabicový graf vytvoří následujícím postupem: Plot→Exploratory Plots→Box-and-Whisker Plot. V dialogovém okně Box-and-Whisker Plot zadáme do řádku Data název proměnné, pro kterou chceme krabicový graf vykreslit a klikneme na OK. Nejrychlejším způsobem jak vytvořit krabicový graf je kliknutím levou myší v horní liště s ikonami na ikonu Boxplot (viz obrázek 5.10).



Obrázek 5.10: Dialogové okno pro změnu parametrů krabicového grafu

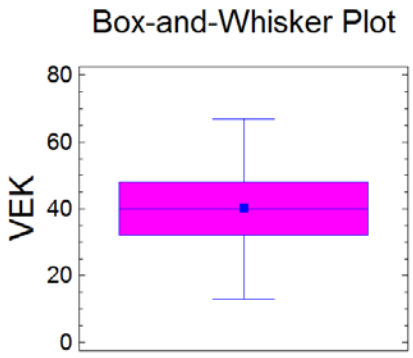
Jakmile máme krabicový graf vytvořen, můžeme si nastavit, zda jej chceme mít zobrazen horizontálně, či vertikálně. Dále si můžeme nastavit zobrazení aritmetického průměru a zobrazení odlehlých hodnot. Jestliže zadáme zobrazení bez odlehlých hodnot, délky úseček (vousů) budou zobrazeny od minima do maxima včetně odlehlých hodnot. Tyto změny provedeme kliknutím pravou myší do prostoru vytvořeného krabicového grafu. V nabídkovém menu zvolíme příkaz Pane Options a v dialogovém okně Box-and-Whisker Plot Options vyklikneme příkaz Outlier Symbols (viz obrázek 5.10).

Interpretace krabicového grafu

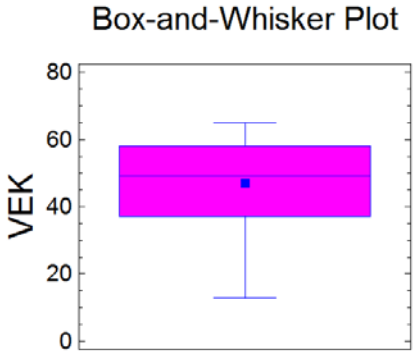
Diagnostikou krabicového grafu můžeme rozhodnout o symetrii rozdělení (splnění předpokladu normality) datového souboru, identifikovat odlehlé hodnoty, určit hodnotu mediánu, případně aritmetického průměru.

Mějme krabicové grafy, které byly stejně jako histogramy četnosti uvedené výše, vytvořeny na základě výšky věku návštěvníků čtyř různých geolokalit. Na obrázku 5.11 je uveden krabicový graf pro geolokalitu č. 1, z něhož je patrné, že soubor dat má normální rozdělení (je symetrický). Normální rozdělení poznáme podle toho, že délky úseček (vousů) jsou stejné, linie znázorňující medián dělí obdélník na přibližně stejně velké poloviny a hodnota mediánu se rovná hodnotě aritmetického průměru (40 let). Dále se v souboru nevyskytují odlehlé hodnoty. Na základě uvedeného grafu můžeme konstatovat, že nejvíce geolokalitu č. 1 navštěvují turisté ve věkovém rozmezí mezi 30 až 50 lety, přičemž se zvyšujícím a snižujícím věkem návštěvnost postupně klesá. Na obrázku 5.12 je uveden krabicový graf pro geolokalitu č. 2. Dle grafu mají data jednoznačně asymetrické rozdělení (není splněn předpoklad normality), záporně sešikmené k nižším hodnotám bez přítomnosti odlehlých hodnot. Sešikmení dat k nižším hodnotám indikuje úsečka pod dolní hranou obdélníku (krabice), která je výrazně delší než úsečka nad jeho horní hranou. Sešikmené rozdělení dat nám také indikuje poloha mediánu. Pokud linie znázorňující medián leží blíže k jedné z krajních linií obdélníku (víku, nebo dnu krabice) jsou data sešikmené v opačném směru.

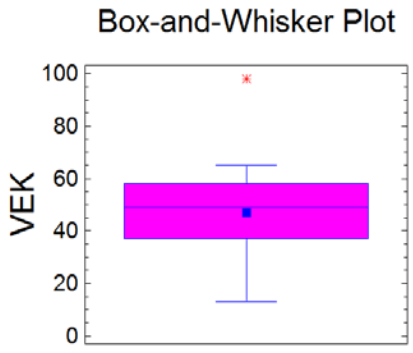
Hodnota mediánu (49 let) se v tomto případě nerovná aritmetickému průměru (47 let). Lze konstatovat, že v případě geolokality č. 2 převažují turisté z vyšších věkových kategorií, přičemž s klesajícím věkem klesá zároveň i návštěvnost. Obrázek 5.13 znázorňuje krabicový graf věku turistů z geolokality č. 3. Jedná se opět o asymetrické rozdělení sešikmené k nižším hodnotám, ovšem v tomto případě zde lze v horní části grafu identifikovat jednu odlehlou hodnotu. Z grafu vyplývá, že geolokalita č. 3 je navštěvována převážně turisty vyššího věku. Odlehlá hodnota nám indikuje, že lokalitu zřejmě navštívil jeden turista velmi vysokého věku (98 let). Obrázek 5.14 uvádí krabicový graf věku návštěvníků geolokality č. 4. Horní úsečka grafu (vous) je výrazně delší než dolní, medián je mírně blíž k dolní straně obdélníka (krabice) a hodnota mediánu (25 let) se nerovná aritmetickému průměru 28 let). Z uvedeného vyplývá, že data jsou asymetricky rozdělena s kladným zešikmením k vyšším hodnotám bez odlehlých hodnot. Z grafu je zřejmé, že geolokalitu č. 4 navštěvují především turisté nižších věkových skupin, přičemž návštěvnost geolokality klesá se vzrůstajícím věkem.



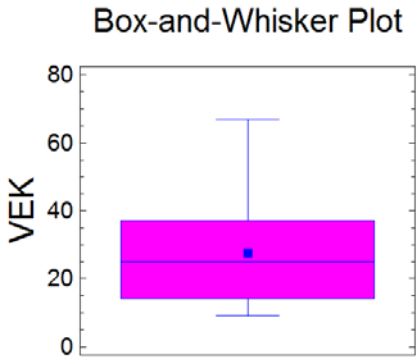
Obrázek 5.11: Krabicový graf s normálním rozdělením dat



Obrázek 5.12: Krabicový graf s asymetrickým záporně sešikmeným rozdělením dat



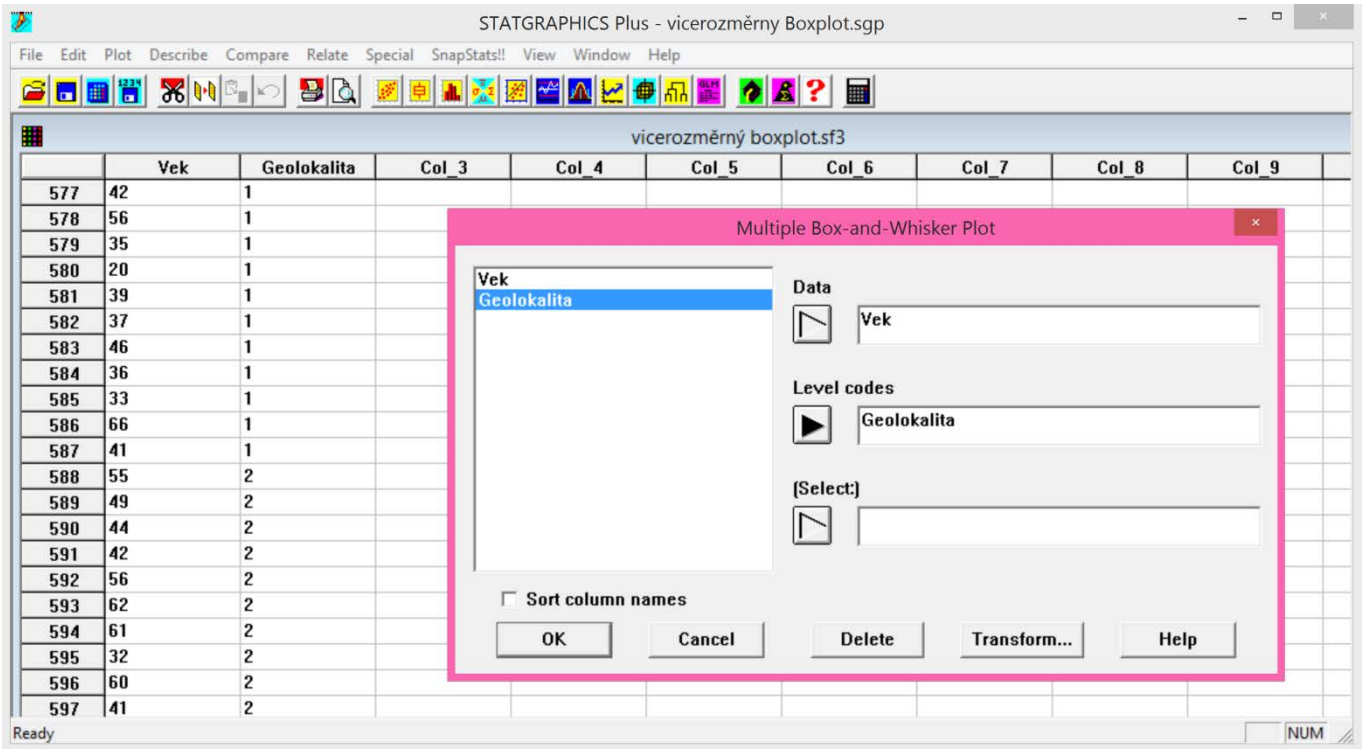
Obrázek 5.13: Krabicový graf s asymetrickým záporně sešikmeným rozdělením dat s jednou odlehlou hodnotou



Obrázek 5.14: Krabicový graf s asymetrickým kladně sešikmeným rozdělením dat

V programu Statgraphics lze také vytvořit krabicové grafy pro více datových souborů v rámci jednoho obrázku. Jedná se o tzv. vícenásobný krabicový graf (Multiple Box and Whisker Plot), který umožňuje provést průzkumovou analýzu pro jednotlivé datové soubory a zároveň porovnat datové soubory mezi sebou. Vzájemným srovnáním vytvořených krabicových grafů můžeme odhadnout, zda jsou jednotlivé analyzované proměnné podobné, či rozdílné atd.

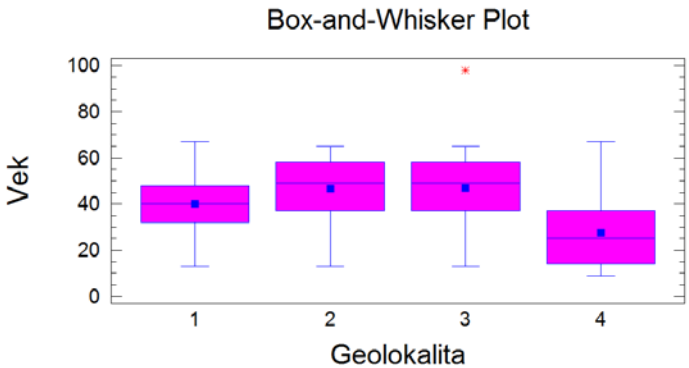
Pokud chceme vícenásobný krabicový graf vytvořit v programu Statgraphics, je nutné nejprve specifickým způsobem připravit data v pracovním sešitě. Data jednotlivých proměnných musí být vloženy pouze do jednoho sloupce pod sebou. V druhém sloupci musí být jednotlivým proměnným přiřazen číselný kód, neboli faktor (např. 1, 2, 3 atd.), za účelem jejich rozlišení. Na obrázku 5.15 je znázorněn pracovní sešit, ve kterém jsou v prvním sloupci vloženy hodnoty zjištěných věků návštěvníků na čtyřech různých geolokalitách (data jsou stejná, jako v předchozích čtyřech vzorových příkladech krabicových grafů). Ve druhém sloupci mají data první proměnné, tedy hodnoty věků návštěvníků zjištěné na první geolokalitě, přiřazen faktor 1. Data z druhé geolokality mají přiřazen faktor 2, data z třetí geolokality faktor 3 a data ze čtvrté geolokality faktor 4. Sloupec s faktory si můžeme nazvat libovolně, např. "Geolokalita".



Obrázek 5.15: Pracovní sešit s daty a dialogové okno pro vytvoření vícenásobného krabicového grafu

Nyní můžeme následujícím postupem vytvořit vícenásobný krabicový graf: Plot→Exploratory Plots→Multiple Box-and-Whisker Plot. Otevře se dialogové okno Multiple Box-and-Whisker Plot. Zde si do řádku Data zadáme název sloupce, ve kterém máme zadány hodnoty všech proměnných (zjištěný věk návštěvníků na všech čtyřech lokalitách). Do řádku Level codes zadáme název sloupce, ve kterém máme zadány faktory (číselné kódy pro jednotlivé geolokality) a klikneme na OK (viz obrázek 5.15).

Obrázek 5.16 uvádí výsledný vícenásobný krabicový graf. V uvedeném grafu jsou patrné významné rozdíly zjištěného věku návštěvníků na jednotlivých geolokalitách. Druhá a třetí geolokalita se od ostatních liší vyšším věkem návštěvníků a čtvrtou geolokalitu navštěvují výrazně mladší návštěvníci než je tomu u ostatních sledovaných geolokalit. Dále je z grafu patrné, že výška obdélníku krabicového grafu první geolokality je zřetelně menší, než u ostatních. To značí, že data z první sledované geolokality vykazují nižší stupeň variability (proměnlivosti) než ostatní soubory dat. Interpretace jednotlivých krabicových grafů je totožná s interpretací uvedenou v předchozím textu pro obrázky 5.11 až 5.14.



Obrázek 5.16: Vícenásobný krabicový graf

5.1.3 Normální pravděpodobnostní graf

Normální pravděpodobnostní graf (normal probability plot), nazývaný také jako rankitový graf, umožňuje velmi přesně rozhodnout, zda data souboru pocházejí z normálního rozdělení. Graf je založen na kumulativní distribuční funkci normovaného normálního rozdělení (Forthofer et al., 2007). Při jeho konstrukci jsou na osu x vynášeny

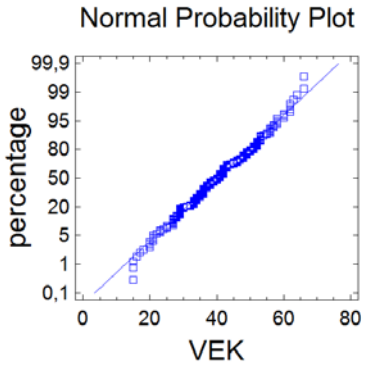
skutečně naměřené hodnoty, které jsou vzestupně seřazeny. Na osu y jsou vynášeny kumulativní procenta kvantilů normovaného normálního rozdělení. V některých statistických programech (včetně programu Statgraphics) je možné v grafu vykreslit ideální přímku, na které by v případě normálního rozdělení měly zobrazené body ležet.

Normální pravděpodobnostní graf vytvoříme v programu Statgraphics následujícím postupem: Plot→Exploratory Plots→Normal Probability Plot. V dialogovém okně Normal Probability Plot zadáme do řádku Data název proměnné, pro kterou chceme graf vykreslit a klikneme na OK. V grafu se ideální přímka normálního rozdělení zobrazuje automaticky. Jestliže chceme mít graf vykreslený bez této přímky, klikneme pravou myší do prostoru grafu. Zobrazí se nabídkové menu, zvolíme Pane Options a v dialogovém okně Normal Probability Plot Options zatrhneme v nabídce Fitted Line položku None. Dále zde v nabídce Direction můžeme změnit uspořádání os (tzn. zaměnit pozici osy x za y).

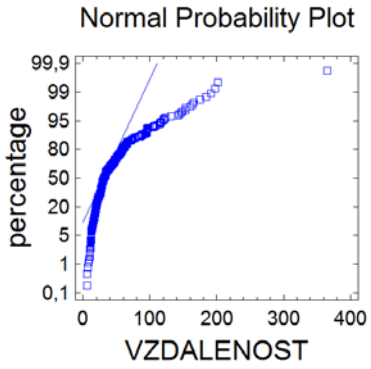
Interpretace normálního pravděpodobnostního grafu

Diagnostikou normálního pravděpodobnostního grafu můžeme rozhodnout o normalitě datového souboru a identifikovat přítomnost významně odlehlých hodnot. V případě, že zobrazená data vytváří přímku, je rozdělení dat normální. Pokud se data seskupují do konkávní křivky, jedná se o asymetricky rozdělená data s kladným zešikmením. Asymetricky rozdělená data se záporným zešikmením budou v grafu vytvářet konvexní křivku. Silně odlehlé hodnoty se zobrazí jako body významně vzdálené od ostatních.

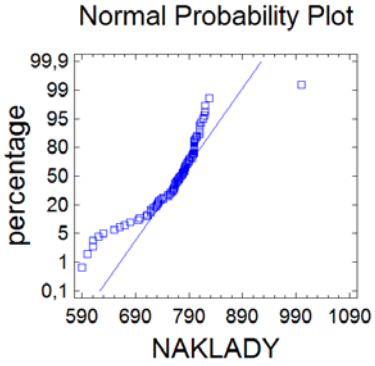
Mějme normální pravděpodobností grafy, které byly vytvořeny na základě datových souborů obsahujících věk návštěvníků, vzdálenost od místa bydliště a celkových nákladů na návštěvu vybrané technické památky. Obrázek 5.17 uvádí graf věku návštěvníků, ve kterém se vykreslené body velmi těsně přimykají ideální přímce. Rozložení bodů v grafu je typické pro soubor dat s normálním rozdělením. Na obrázku 5.18, který zobrazuje vzdálenost od místa bydliště návštěvníků, body v grafu vykreslují konkávní křivku. Jedná se tedy o asymetrické rozdělení dat kladně sešikmené k vyšším hodnotám. V grafu lze identifikovat jednu extrémně odlehlou hodnotu vpravo (bod je významně vzdálen od ostatních), která je způsobena velmi vzdáleným bydlištěm jednoho z návštěvníků.



Obrázek 5.17: Normální pravděpodobnostní graf – normální rozdělení



Obrázek 5.18: Normální pravděpodobnostní graf – asymetrické rozdělení kladně sešikmené



Obrázek 5.19: Normální pravděpodobnostní graf – asymetrické rozdělení záporně sešikmené

Na obrázku 5.19, jež znázorňuje graf celkových nákladů na návštěvu, jsou body seřazeny do konvexní křivky. Konvexní tvar křivky značí, že soubor dat je asymetricky rozdělen se záporným zešikmením k nižším hodnotám. Graf jednoznačně indikuje přítomnost jedné extrémně odlehlé hodnoty vpravo. Odlehlá hodnota je v tomto případě způsobena velmi vysokými náklady jednoho z dotazovaných návštěvníků.

5.1.4 Jednorozměrný bodový graf

Jednorozměrný bodový graf (scatterplot) je jednorozměrnou projekcí zkoumaných hodnot do osy x. Přestože se jedná o poměrně jednoduchý graf, má velkou vypovídací schopnost. Při jeho konstrukci jsou na osu x vynášeny reálně naměřené hodnoty. Ve směru osy y jsou jednotlivé hodnoty náhodně rozprostřeny (rozmítnuty) prostřednictvím generátoru náhodných čísel. Rozprostření zamezuje překrývání jednotlivých dat v grafu a tím zvyšuje jeho vypovídací schopnost.

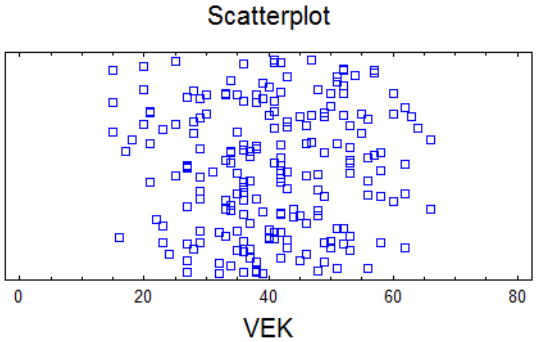
V prostředí programu Statgraphics vytvoříme jednorozměrný bodový graf klikem levou myší na příkaz: Plot→Scatterplots→Univariate Plot. V dialogovém okně Univariate Plot zadáme do řádku Data název proměnné, pro kterou chceme graf vykreslit a klikneme na OK.

Interpretace jednorozměrného bodového grafu

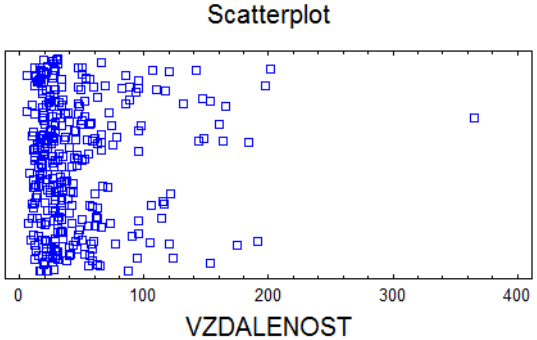
Diagnostikou tohoto grafu můžeme velmi jednoduše rozhodnout o symetrii rozdělení, identifikovat lokální koncentrace dat (shluky) a odhalit přítomnost odlehlých hodnot.

Mějme jednorozměrné bodové grafy, které byly vytvořeny na základě stejných datových souborů jako v případě normálních pravděpodobnostních grafů interpretovaných výše. Z grafu na obrázku 5.20 je patrné, že zobrazená data tvoří mrak bodů bez zahuštění a vzdálených bodů. Můžeme říci, že soubor dat má normální rozdělení bez přítomnosti odlehlých hodnot. Z grafu jednoznačně vyplývá, že věkové rozmezí návštěvníků hodnocené technické památky je přibližně 20 až 60 let. V grafu na obrázku 5.21 můžeme vidět mrak bodů se zahuštěním v jeho levé části (mezi hodnotami 5 až 65), přičemž v pravé části se nachází jeden významně vzdálený bod. Jedná se tedy o soubor dat, který má asymetrické rozdělení kladně sešikmené s přítomností jedné odlehlé hodnoty. Z grafu vyplývá, že převážná většina návštěvníků byla ze vzdálenosti 5 až 65 km od místa bydliště, přičemž jeden návštěvník byl od místa svého bydliště vzdálen přibližně 362 km.

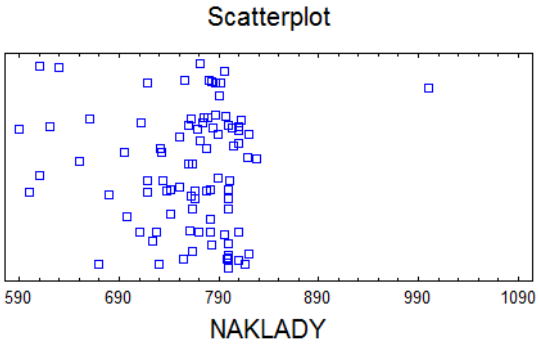
Na obrázku 5.22 můžeme vidět mrak bodů se zahuštěním v jeho pravé části (cca mezi hodnotami 730 až 830) a jedním vzdáleným bodem napravo od mraku. Zde se jedná o asymetrické rozdělení záporně sešikmené s jednou odlehlou hodnotou. Z uvedeného grafu vyplývá, že převážná část návštěvníků utratila za návštěvu dané technické památky přibližně 730 až 830 Kč. Jeden návštěvník zřejmě utrácel více než ostatní a návštěva ho stála cca 1000 Kč (viz odlehlá hodnota v pravé části grafu). V případě, že bychom v grafu měli dva, či více oddělených mraků bodů, jednalo by se bimodální, případně vícemodální rozdělení (více oddělených shluků bodů). U tohoto rozdělení je nejlepší cestou pro správné statistické vyhodnocení rozdělit data do samostatných souborů a následně analyzovat každý soubor samostatně.



Obrázek 5.20: Jednorozměrný bodový graf – normální rozdělení



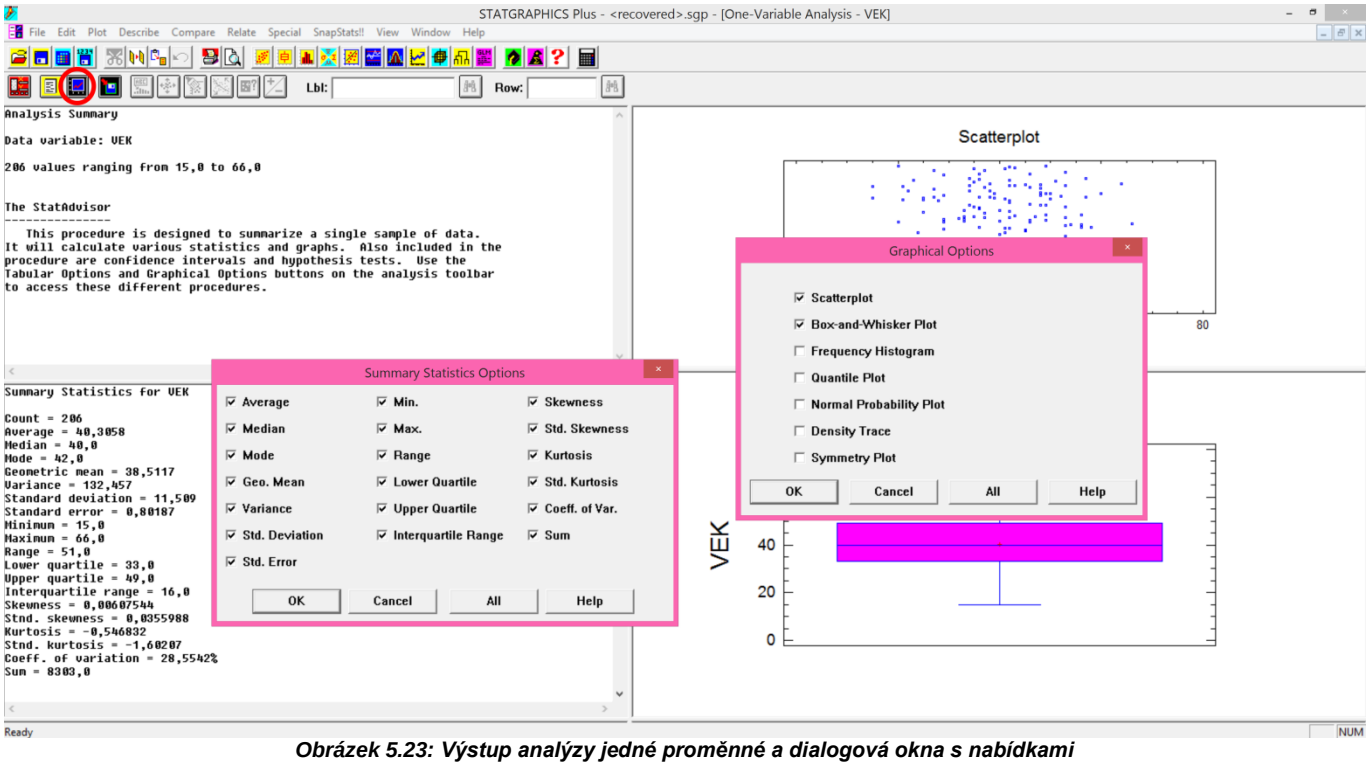
5.21: Jednorozměrný bodový graf – asymetrické rozdělení kladně sešikmené



Obrázek 5.22: Jednorozměrný bodový graf – asymetrické rozdělení záporně sešikmené

Veškeré výše uvedené typy grafů lze v programu Statgraphics zobrazit najednou pomocí analýzy jedné proměnné. Postup je následující: klikneme levou myší na příkaz Describe v jednořádkovém textovém menu a následně: Numeric Data→One-Variable Analysis. V dialogovém okně One-Variable Analysis zadáme do řádku Data název proměnné, pro kterou chceme grafy vykreslit a klikneme na OK.

Výstup analýzy se zobrazí ve čtyřech různých oknech (viz obrázek 5.23). V horním pravém okně se zobrazí jednorozměrný bodový graf (Scatterplot), v dolním pravém okně krabicový graf (Box-and-Whisker Plot). V levém horním okně se zobrazí souhrnný popis provedené analýzy (Analysis Summary) a v levém dolním okně souhrn základních popisných charakteristik souboru (Summary Statistics). V levých oknech se vždy nachází odstavec se slovním komentářem (The StatAdvisor). Komentář lze s výhodou využít při následné interpretaci provedených analýz.



Obrázek 5.23: Výstup analýzy jedné proměnné a dialogová okna s nabídkami

Pokud chceme zobrazit další typy grafů (např. histogram četnosti apod.) klikneme levou myší na ikonu Graphical options. Následně v zobrazeném dialogovém okně zatrhneme typy grafů, které chceme vykreslit (viz obrázek 5.23). Jestliže chceme zobrazit více popisných charakteristik souboru, než je standardně programem nabízeno, klikneme pravou myší do prostoru levého dolního okna (Summary Statistics) a v nabídce zvolíme Pane Options. V dialogovém okně Summary Statistics Options zatrhneme požadované popisné statistiky.

Program Statgraphics umožňuje výsledné grafy dále graficky upravovat. Můžeme si změnit název grafu, popis os, velikost a styl písma, barvy čar, grafické symboly apod. Úpravu provedeme kliknutím pravou myší do prostoru

vybraného grafu a v nabídce zvolíme Graphics Options. Zobrazí se dialogové menu se záložkami, pomocí kterých graf upravíme dle našich požadavků a představ. Následné uložení provedeme kliknutím pravou myší do prostoru vybraného grafu. Zobrazí se nám nabídka, v jejíž dolní části najdeme příkaz Save Graph. Po zvolení tohoto příkazu můžeme daný graf uložit pod námi vybraným názvem v jednom z nabízených formátů, např. wmf, JPG, png a další.

5.1.5 Kvantilový graf

Kvantilový graf (Quantile Plot) slouží především ke zjištění, zda data pochází z určitého typu rozdělení. Na osu x se vynáší pořádková statistika sledované proměnné a na osu y kumulativní pravděpodobnost teoretického rozdělení. Ve většině statistických programů je v grafu defaultně znázorněná teoretická křivka odpovídající normálnímu rozdělení. V případě, že se vykreslené body sledované proměnné těsně přimykají teoretické křivce normálního rozdělení, můžeme rozdělení analyzovaného datového souboru považovat za normální.

V programu Statgraphics vytvoříme kvantilový graf klikem levou myší na příkaz: Describe→Distributions→ Distribution Fitting (Uncensored Data). V dialogovém okně Distribution Fitting (Uncensored Data) zadáme do řádku Data název proměnné, pro kterou chceme graf vykreslit a klikneme na OK. Zobrazí se nám výstup analýzy ve standardním nastavení obsahující čtyři okna. Nyní klikneme na ikonu Graphical options. Následně v dialogovém okně Graphical Options zatrhneme nabídku Quantile plot.

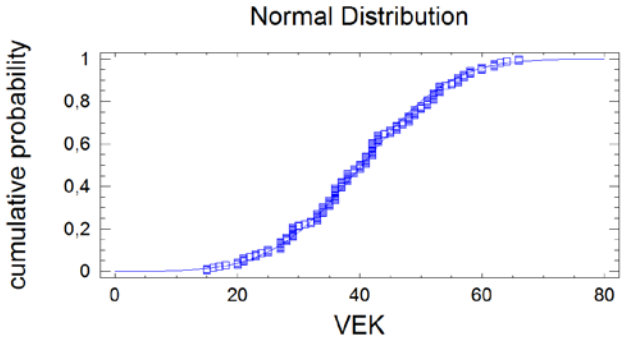
Program Statgraphics umožňuje vykreslit teoretickou křivku i pro jiná rozdělení (lognormální, exponenciální atd.). V případě, že chceme rozdělení souboru dat srovnat s jiným rozdělením než normálním, klikneme pravou myší do prostoru grafu. V nabídce zvolíme Analysis Options. Zobrazí se nabídkové menu Probability Distributions Options, ve kterém si zatrhnutím zvolíme požadovaný typ rozdělení.

Interpretace kvantilového grafu

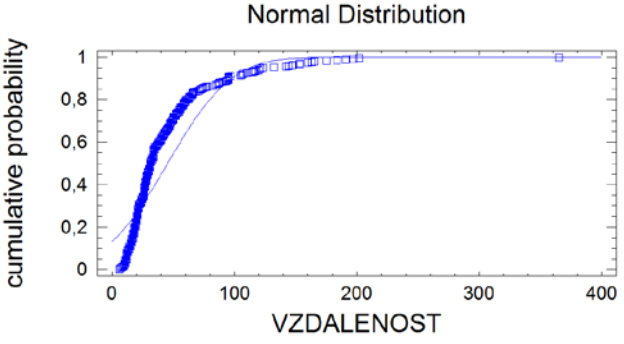
Diagnostikou kvantilového grafu můžeme rozhodnout především o charakteru rozdělení dat, homogenitě a přítomnosti odlehlých hodnot.

Mějme kvantilové grafy, jež byly vytvořeny (stejně jako v předchozích příkladech) na datových souborů obsahujících věk návštěvníků, vzdálenost od místa bydliště a celkových nákladů na návštěvu vybrané technické památky.

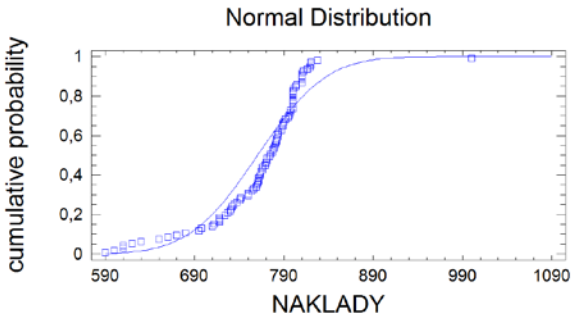
V kvantilovém grafu na obrázku 5.24 je zřetelně vidět, že zobrazená data nevytvářejí oddělené shluky a dobře se přimykají teoretické křivce normálního rozdělení. Taktéž zde není přítomna hodnota výrazně vzdálená od ostatních. Analyzovaný soubor dat má tedy normální rozdělení, je homogenní bez přítomnosti odlehlých hodnot. Na obrázku 5.25 se data teoretické křivce nepřimykají a jsou seřazeny do konkávního tvaru. V pravé části grafu lze identifikovat jednu významně vzdálenou hodnotu. Zde má soubor dat jednoznačně asymetrické rozdělení s kladným zešikmením k vyšším hodnotám s jednou odlehlou hodnotou. Kvantilový graf na obrázku 5.26 znázorňuje data s asymetrickým rozdělením, jež jsou záporně zešikmena k nižším hodnotám (data se nepřimykají teoretické křivce normálního rozdělení a jsou seřazeny do konvexní křivky). Dále je zde přítomna jedna odlehlá hodnota (jeden vzdálený bod v pravé části grafu).



Obrázek 5.24: Kvantilový graf – normální rozdělení



5.25: Kvantilový graf – asymetrické rozdělení kladně sešikmené



Obrázek 5.26: kvantilový graf – asymetrické rozdělení záporně sešikmené

5.1.6 Kvantil-kvantilový graf

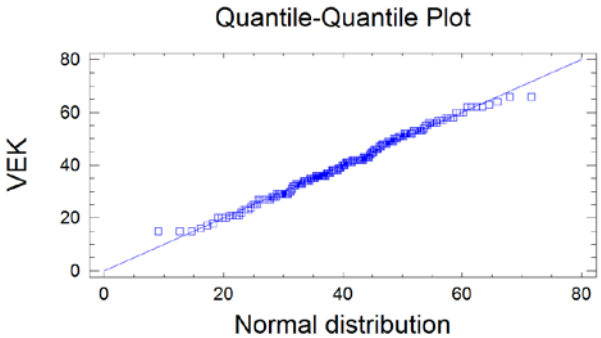
Kvantil-kvantilový graf (Quantile-Quantile Plot), velmi často označován také jako Q-Q graf (Q-Q Plot), slouží především k posouzení shody rozdělení souboru dat s vybraným teoretickým rozdělením. Zpravidla se pro teoretické rozdělení volí rozdělení normální. Při jeho konstrukci se na osu x vynáší kvantily zvolené distribuční funkce. Na osu y se vynáší kvantily pozorované distribuce sledované proměnné. Grafem prochází teoretická křivka zvolené distribuční funkce. V případě, že se vykreslené body sledované proměnné těsně přimykají teoretické křivce, rozdělení souboru dat odpovídá zvolenému teoretickému rozdělení.

V programu Statgraphics vytvoříme kvantil-kvantilový graf stejným postupem jako v případě kvantilového grafu: Describe→Distributions→Distribution Fitting (Uncensored Data). Po zobrazení dialogového okna Graphical Options zatrhneme nabídku Quantile-Quantile Plot. Volbu rozdělení pro teoretickou křivku provedeme stejným postupem, jenž je uveden u kvantilového grafu.

Interpretace kvantil-kvantilového grafu

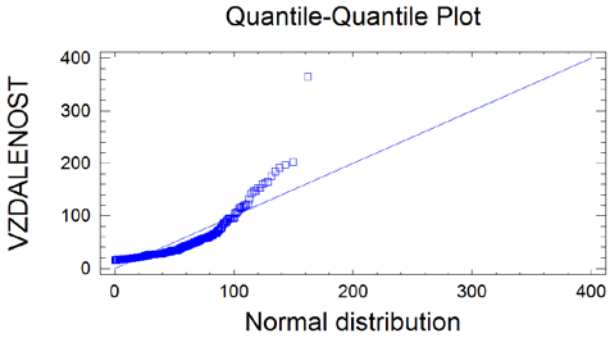
Tímto grafem diagnostikujeme především charakter rozdělení dat souboru. Dále můžeme rozhodnout o homogenitě souboru a přítomnosti odlehlých hodnot.

Mějme kvantil-kvantilové grafy, jež byly vytvořeny na základě stejných datových souborů jako v případě kvantilových grafů. Na obrázku 5.27 je uveden kvantil-kvantilový graf pro soubor dat s normálním (symetrickým) rozdělením, homogenní, bez přítomnosti odlehlých hodnot.

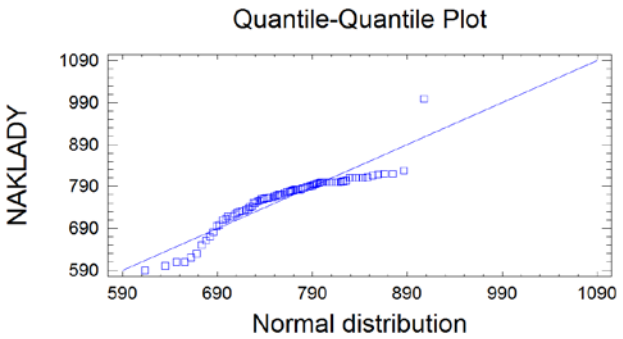


Obrázek 5.27: Kvantil-kvantilový graf – normální rozdělení

Obrázek 5.28 znázorňuje graf pro soubor dat s asymetrickým rozdělením kladně sešikmeným. V horní části grafu lze identifikovat jednu odlehlou hodnotu. Obrázek 5.29 znázorňuje graf pro soubor dat s asymetrickým rozdělením záporně sešikmeným, přičemž v horní části grafu lze opět identifikovat jednu odlehlou hodnotu.



5.28: Kvantil-kvantilový graf – asymetrické rozdělení kladně sešikmené



5.29: Kvantil-kvantilový graf – asymetrické rozdělení záporně sešikmené

5.1.7 Numerická metoda ověření normality

O normalitě (symetrii) datového souboru lze rozhodnout na základě vypočtených hodnot koeficientů šikmosti (skewness) a špičatosti (kurtosis). **Koeficient šikmosti** udává, do jaké míry jsou hodnoty souboru rovnoměrně rozptýleny kolem středu. Symetrické rozdělení má hodnotu koeficientu šikmosti přibližně rovnou nule. V případě kladné hodnoty koeficientu je datový soubor rozdělen asymetricky s kladným zešikmením k vyšším hodnotám. V případě záporné hodnoty koeficientu má soubor asymetrické rozdělení se záporným zešikmením k nižším hodnotám. **Koeficient špičatosti** udává, s jakou hustotou jsou hodnoty souboru rozptýleny kolem středu. Kladná hodnota koeficientu indikuje rozdělení dat špičatější a záporná hodnota rozdělení dat plošší než je normální rozdělení. Normální rozdělení dat má tedy hodnoty koeficientů šikmosti a špičatosti přibližně rovny nule.

V prostředí programu Statgraphics hodnoty uvedených koeficientů získáme následujícím postupem: Describe→Numeric Data→One-Variable Analysis. V dialogovém okně One-Variable Analysis zadáme do řádku Data název proměnné, pro kterou chceme hodnoty koeficientů vypočíst a klikneme na OK. V levém dolním okně Summary Statistics nalezneme vypočtenou hodnotu koeficientu šikmosti pod pojmem Skewness a hodnotu koeficientu špičatosti pod pojem Kurtosis. Pokud se námi požadované hodnoty nezobrazí, klikneme pravou myší do prostoru okna a v nabídce zvolíme Pane Options. V dialogovém okně Summary Statistics Options zatrhneme požadované popisné statistiky (viz obrázek 5.23).

Hodnoty uvedených koeficientů je možné vypočíst i v tabulkovém procesoru Excel. V pracovním sešitě vypočteme koeficient šikmosti pro zvolená data vložením funkce SKEW a koeficient špičatosti vložením funkce KURT.

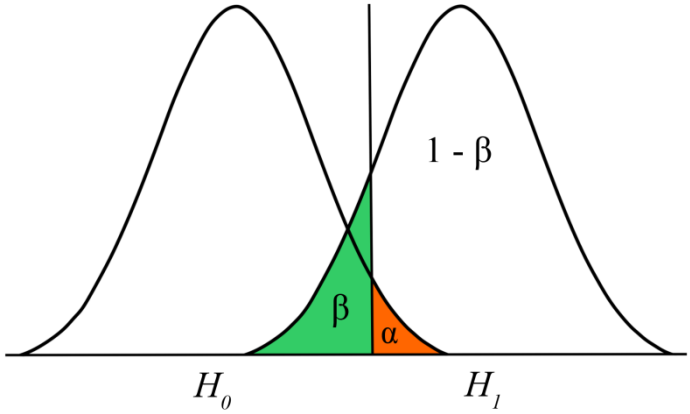
5.2 Statistické testy

Statistickými testy ověřujeme základní předpoklady o datovém souboru. Jedná se především o testování normality datového souboru, nezávislosti, testy odlehlých hodnot apod. Klasické metody statistické analýzy jsou obvykle založeny na předpokladu normality rozdělení datového souboru, proto mezi nejvýznamnější z uvedených testů patří právě testy normality. Statistické testy jsou zpravidla méně citlivé na porušení předpokladů normality než diagnostické grafy. Proto je vždy důležité nejprve provést diagnostiku grafů a tyto závěry následně ověřit pomocí statistických testů.

Statistické testy se provádějí pomocí testování statistických hypotéz. Základem je statistická hypotéza, která předpokládá např. určité rozdělení dat, homogenitu, nezávislost atd. Nejprve vyslovíme hypotézu H_0, která formuluje náš předpoklad - např. soubor dat má normální rozdělení. Hypotéza H_0 se nazývá nulová hypotéza, případně testovaná hypotéza. Následně proti hypotéze H_0 vyslovíme hypotézu H_A, která je opačná - např. soubor dat nemá normální rozdělení. Hypotéza H_A se nazývá alternativní hypotéza. Nyní testujeme platnost hypotézy H_0 proti H_A. V případě, že přijmeme hypotézu H_0 zamítáme H_A, čili H_0 je pravdivá. V případě, že zamítneme H_0 přijímáme H_A, čili H_0 je nepravdivá.

Při testování hypotéz se můžeme dopustit dvou chyb: chyby 1. druhu a chyby 2. druhu. Chyba 1. druhu znamená, že zamítneme hypotézu H_0 i když je pravdivá. Chyba 1. druhu se označuje řeckým písmenem α a nazývá

se hladina významnosti. Chyba 2. druhu znamená, že přijmeme hypotézu H_0 i když je nepravdivá. Chyba 2. druhu se označuje řeckým písmenem β a nazývá se síla testu a je dána číslem 1 - β. Při statistickém testování se standardně postupuje tak, že si zvolíme dostatečně nízkou hodnotu hladiny významnosti α. Hladina významnosti α se obvykle volí rovna 0,05 (případně 0,01) a zahrnuje tak 5% (1%) pravděpodobnost výskytu chyby 1. druhu. Jelikož obě chyby spolu vzájemně souvisí, určíme takto i chybu 2. druhu β. Vzájemnou závislost obou uvedených chyb znázorňuje obrázek 5.30.



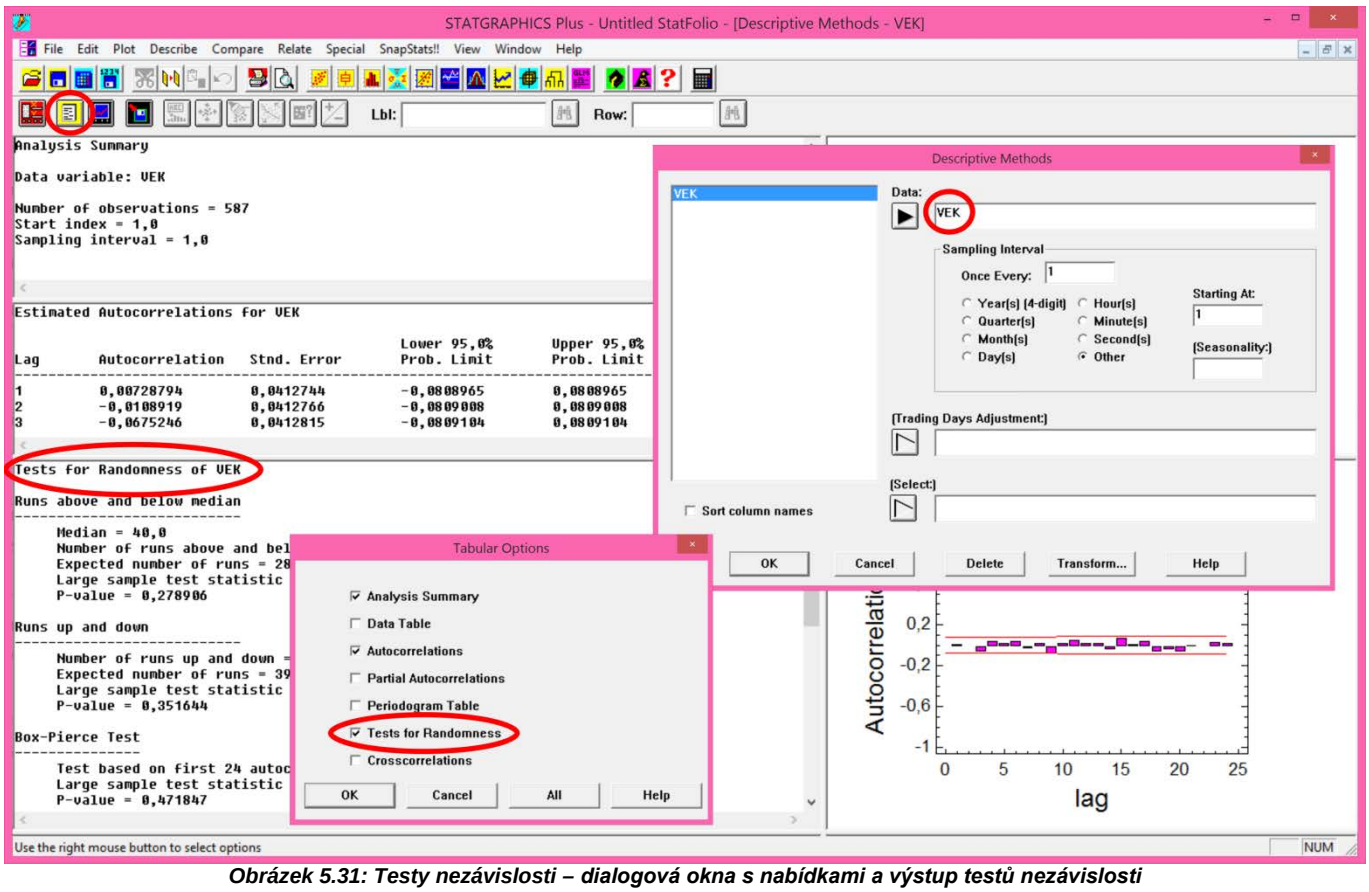
Obrázek 5.30: Vzájemná závislost hladiny významnosti α a síly testu β

Testování hypotézy H_0 vs. H_A můžeme provést dvěma způsoby. První způsob (evropské testování) využívá pro určení platnosti hypotézy testovací kritérium. Testování hypotéz se provede srovnáním vypočtené hodnoty (testovací kritérium) s kritickou tabelární hodnotou. V případě, že vypočtená hodnota je nižší než kritická tabelární hodnota, hypotézu H_0 přijímáme na dané hladině významnosti. Pokud je vypočtená hodnota vyšší než kritická tabelární hodnota, hypotézu H_0 zamítáme a přijímáme hypotézu H_A na dané hladině významnosti. Druhý způsob (americké testování) využívá pro testování *p*-hodnotu (*p*-value). *p*-hodnotu lze definovat, jako spočtenou hladinu významnosti, na které bylo ještě možné při daných naměřených hodnotách zamítnout nulovou hypotézu H_0 (Pavlík, 2005). Je-li vypočtená *p*-hodnota větší než zvolená hladina významnosti α hypotézu H_0 přijímáme. V případě, že *p*-hodnota je menší než α, hypotézu H_0 zamítáme a přijímáme alternativní hypotézu H_A.

5.2.1 Testy nezávislosti

Při statistické analýze dat se předpokládá, že data ve zpracovávaném souboru jsou náhodným výběrem a tedy nejsou na sobě vzájemně závislá. Příčinou vzniku závislosti (tzv. autokorelace) může být například nenáhodný výběr dotazovaných turistů, nestabilita měřicího přístroje, vliv vnějších faktorů (teplota, tlak, specifická období – prázdniny, víkendy atd.). Závislost dat v souboru může při statistické analýze vést k nadhodnocení odhadů a zkreslení interpretovaných výsledků. V případě, že se závislost v datech prokáže, je nutné zkontrolovat celý postup vedoucí k jejich získání a hlouběji analyzovat její příčiny. Test nezávislosti se nejčastěji provádí pomocí testování autokorelačních koeficientů (Otyepka et al., 2013).

Program Statgraphics nabízí pro testování nezávislosti tři druhy testů: mediánový test (Runs above and below median), test založený na bodech zvratu (Runs up and down) a Box-Pierceův test (Box-Pierce Test). Uvedené testy lze provést následujícím postupem: klikneme levou myší na příkaz Special v jednořádkovém textovém menu a následně Time-Series Analysis→Descriptive Methods. V dialogovém okně Descriptive Methods zadáme do řádku Data název proměnné, pro kterou chceme testy provést a klikneme na OK. Zobrazí se nám výstup analýzy obsahující čtyři okna. Nyní klikneme levou myší na ikonu Tabular options (druhá ikona zleva v horní liště výstupu). Zobrazí se dialogové okno Tabular Options, ve kterém v nabídce zatrhneme Tests for Randomness. Po těchto úkonech se v levé části zobrazí další okno s výsledky testů (viz obrázek 5.31). Dvojklikem levou myší do prostoru okna se výsledky zobrazí přes celou levou část pracovního prostředí.



Obrázek 5.31: Testy nezávislosti – dialogová okna s nabídkami a výstup testů nezávislosti

Interpretace testů nezávislosti

Příklad: Pro testování byl zadán datový soubor obsahující věk 588 návštěvníků sledované geolokality, získaný v rámci dotazníkového průzkumu.

Výsledky testování (Tests for Randomness of VEK):

1) Runs above and below median

Median = 40,0
Number of runs above and below median = 270
Expected number of runs = 283,35
Large sample test statistic z = -1,08278
P-value = 0,278906

2) Runs up and down

Number of runs up and down = 381
Expected number of runs = 391,0
Large sample test statistic z = -0,931402
P-value = 0,351644

3) Box-Pierce Test

Test based on first 24 autocorrelations
Large sample test statistic = 23,8214
P-value = 0,471847

Vyhodnocení testu: Program Statgraphics vyjadřuje výsledky testů pomocí p-hodnoty. Všechny tři testy mají vypočtenou p-hodnotu vyšší než 0,05, což znamená, že přijímáme nulovou hypotézu H_0 (H_0 = data jsou nezávislá). Provedenými testy bylo s 95% statistickou jistotou prokázáno, že data jsou nezávislá.

Na rozdíl od jiných statistických programů program Statgraphics přímo nabízí v odstavci The StatAsvisor vlastní interpretaci provedených analýz a testů, které lze při zpracovávání experimentálních dat s výhodou využít.

Výsledky pro výše uvedený příklad:

„Three tests have been run to determine whether or not VEK is a random sequence of numbers. A time series of random numbers is often called white noise, since it contains equal contributions at many frequencies. The first test counts the number of times the sequence was above or below the median. The number of such runs equals 270, as compared to an expected value of 283,35 if the sequence were random. Since the P-value for this test is greater than or equal to 0.10, we cannot reject the hypothesis that the series is random at the 90% or higher confidence level. The second test counts the number of times the sequence rose or fell. The number of such runs equals 381, as compared to an expected value of 391,0 if the sequence were random. Since the P-value for this test is greater than or equal to 0.10, we cannot reject the hypothesis that the series is random at the 90% or higher confidence level. The third test is based on the sum of squares of the first 24 autocorrelation coefficients. Since the P-value for this test is greater than or equal to 0.10, we cannot reject the hypothesis that the series is random at the 90% or higher confidence level."

5.2.2 Testy normality

Řada statistických analýz a testů je založena na předpokladu normality datového souboru. Nesplnění předpokladů normality může být způsobeno jak přítomností odlehlých hodnot tak vlastním rozdělením datového souboru. V případě, že je normalita splněna, můžeme použít klasické metody statistické analýzy a parametrické testy. Pokud však normalita splněna není, je nutné použít buď neparametrické metody statistické analýzy a neparametrické testy, vyloučit odlehlé hodnoty (viz podkapitola 5.3 Odlehlé hodnoty a jejich identifikace), nebo se pokusit dosáhnout normality datového souboru pomocí vhodné transformace dat (viz podkapitola 5.4 Transformace).

O splnění normality datového souboru rozhodujeme především na základě grafických diagnostik. Následně pak za účelem potvrzení těchto závěrů provádíme testování pomocí testů normality. Testů normality existuje velké množství např. Kolmogorov-Smirnovův test, Lillieforsův test, Shapiro-Wilkův test, Anderson-Darlingův test a další. V následujícím textu budou popsány a vysvětleny pouze testy, které lze provést v programu Statgraphics. Veškeré níže uvedené testy testují hypotézu H_0 = soubor dat má normální rozdělení proti H_A = soubor dat nemá normální rozdělení.

Kolmogorov-Smirnovův test ((Kolmogorov-Smirnov Test; K-S test) - je neznámějším a nejpoužívanějším testem normality. Je založen na testování největšího rozdílu (v absolutní hodnotě) mezi pozorovanou a teoretickou kumulativní distribuční funkcí. K-S test se obvykle používá v případě malých datových souborů, které obsahují méně než 50 hodnot (n < 50) (Miller a Miller, 2010).

V programu Statgraphics provedeme K-S test následovně: Describe→Distributions→Distribution Fitting (Uncensored Data). V dialogovém okně Distribution Fitting (Uncensored Data) zadáme do řádku název proměnné a klikneme na OK. V dolní části levého okna se následně zobrazí výsledek K-S testu.

Příklad: Pro testování byl zadán datový soubor obsahující vzdálenost od místa bydliště 48 návštěvníků sledované geolokality, získaný v rámci dotazníkového průzkumu.

Výsledky testování:

EDF Statistic	Value	Modified Form	P-Value
Kolmogorov-Smirnov D	0,178964	3,22305	<0,01*
Anderson-Darling A^2	23,7551	23,8108	0,0000*

Vyhodnocení testu: Z výsledku K-S testu vyplývá, že vypočtená *p*-hodnota (<0,01) je menší než 0,05. Čili zamítáme hypotézu H_0 ve prospěch alternativní hypotézy H_A.

Závěr: Provedením Kolmogorov-Smirnovova testu bylo s 95% statistickou jistotou prokázáno, že data nelze považovat za výběr s normálním rozdělením.

Anderson-Darlingův test (Anderson-Darling Test; A-D test) - je modifikací Kolmogorov-Smirnovova testu. Zásadní rozdíl mezi těmito dvěma testy spočívá v tom, že A-D test používá specifické kritické hodnoty pro všechny distribuce. A-D test klade větší důraz na konce rozdělení datového souboru a je přísnější (citlivější na odchylky od normality) než K-S test. Test je vhodný pro datové soubory, které obsahují více než 50 hodnot (n > 50) (Razali a Wah, 2011).

V programu Statgraphics provedeme A-D test totožným postupem jako u K-S testu, přičemž výsledky se zobrazí přímo pod výsledky K-S testu (viz výsledky testování K-S testu uvedené výše). Z výše uvedeného výsledku testu je zřejmé, že p-hodnota 0,0000 je menší než 0,05 a my zamítáme hypotézu H_0 ve prospěch hypotézy H_A. Závěr testu zní: Provedením Anderson-Darlingova testu bylo s 95% statistickou jistotou prokázáno, že datový soubor nelze považovat za výběr s normálním rozdělením.

Chí-kvadrát test dobré shody (Chi-Square Test; χ^2-test) - je založen na testování shody mezi teoretickými a pozorovanými četnostmi pomocí testačního kritéria χ^2 (chí-kvadrát). Chí-kvadrát test je vhodný především pro velké datové soubory, které obsahují více než 50 hodnot (n > 50) (Miller a Miller, 2010).

V programu Statgraphics provedeme chí-kvadrát test opět stejným postupem jako u předchozích dvou testů, čili Describe→Distributions→Distribution Fitting (Uncensored Data). Výsledek testu se zobrazí v dolní části levého okna.

Příklad: Pro testování byl zadán datový soubor obsahující počty návštěvníků-cyklistů sledované geolokality, získaný v rámci dotazníkového průzkumu v průběhu 200 po sobě následujících dnů.

Výsledky testování:

Chi-Square Test

	Lower Limit	Upper Limit	Observed Frequency	Expected Frequency	Chi-Square
at or below	-7,3901	0	22,22	22,22	
	-7,3901	13,116	19	22,22	0,47
	13,116	28,1374	70	22,22	102,72
	28,1374	41,2263	38	22,22	11,20
	41,2263	53,7937	17	22,22	1,23
	53,7937	66,8826	17	22,22	1,23
	66,8826	81,904	7	22,22	10,43
	81,904	102,41	11	22,22	5,67
above	102,41		21	22,22	0,07

Chi-Square = 155,231 with 6 d.f. **P-Value = 0,0**

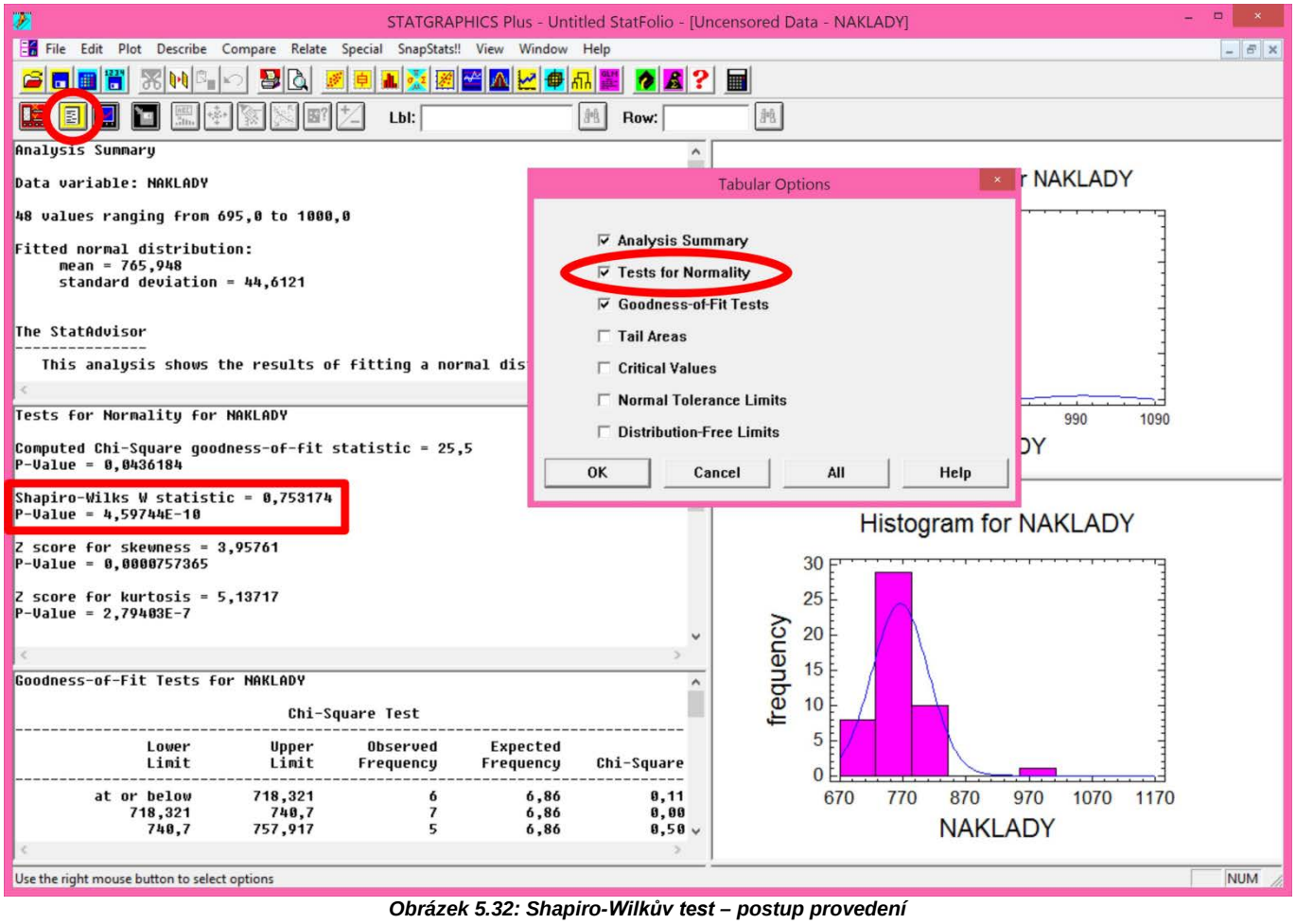
Vyhodnocení testu: Z výsledku chí-kvadrát testu vyplývá, že vypočtená *p*-hodnota 0,0 je menší než 0,05.

Závěr: Provedením chí-kvadrát testu bylo na hladině významnosti 0,05 prokázáno, že data nelze považovat za výběr s normálním rozdělením.

Shapiro-Wilkův test (Shapiro-Wilk Test; S-W test) - patří mezi nejpřísnější testy normality. Je založen na testování kvantilů normální distribuční funkce s kvantily experimentálních dat pomocí testovacího kritéria *W*.

S-W test je vhodný pro testování normality souborů obsahujících méně než 50 hodnot (n < 50) (Razali a Wah, 2011). S-W test patří mezi nekvalitnější a nejpoužívanější testy normality pro malé soubory dat.

V programu Statgraphics provedeme S-W test následujícím postupem: Describe→Distributions→Distribution Fitting (Uncensored Data). V dialogovém okně Distribution Fitting (Uncensored Data) zadáme do řádku Data název proměnné a klikneme na OK. Zobrazí se nám výstup analýzy ve čtyřech různých oknech. Nyní klikneme levou myší na ikonu Tabular options a v nabídce Tabular Options zatrhneme Tests for Normality. V levé části se zobrazí nové okno s výsledkem S-W testu (viz obrázek 5.32).



Obrázek 5.32: Shapiro-Wilkův test – postup provedení

Příklad: Pro testování byl zadán datový soubor obsahující náklady na výlet 48 návštěvníků sledované geolokality, získaný v rámci dotazníkového průzkumu.

Výsledek S-W testu:

```
Shapiro-Wilks W statistic = 0,753174
P-Value = 4,59744E-10
```

Vyhodnocení testu: Z výsledku S-W testu je zřejmé, že vypočtená p-hodnota $4{,}59744.10^{-10}$ je nižší než 0,05 - zamítáme hypotézu H_0 ve prospěch alternativní hypotézy H_A.

Závěr: Provedením Shapiro-Wilkova testu bylo s 95% statistickou jistotou prokázáno, že data nelze považovat za výběr s normálním rozdělením.

5.3 Odlehlé hodnoty a jejich identifikace

Odlehlé hodnoty (tzv. outliers), neboli odlehlá pozorování, jsou specifická data, jež se svou numerickou hodnotou významně liší od ostatních dat v analyzovaném datovém souboru. Přítomnost odlehlých hodnot v datovém souboru

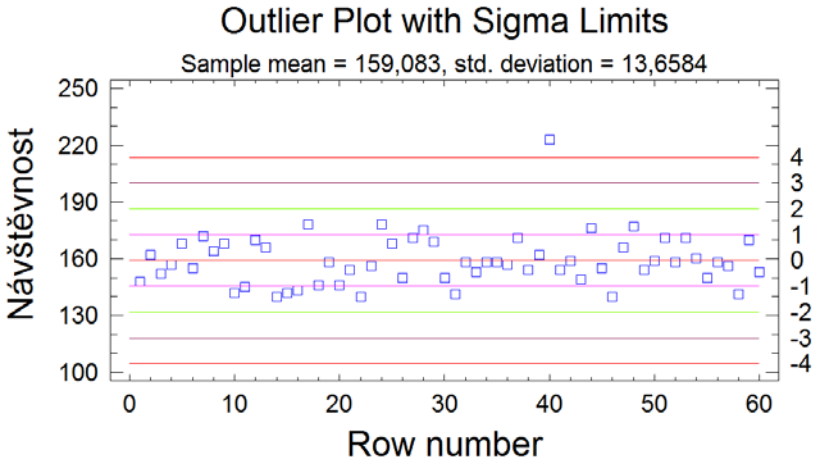
může být způsobena mnoha příčinami. Jejich výskyt může být důsledkem hrubých chyb (např. špatný zápis dat, chybná měření), mohou pocházet z jiného základního souboru než ostatní analyzovaná data, nebo mohou být výsledkem specifického jevu (např. vliv specifických období (prázdniny, víkendy), vlivy počasí, exkurze apod.). Odlehlé hodnoty výrazně ovlivňují klasické odhady míry polohy a rozptýlení (aritmetický průměr a směrodatná odchylka) a proto je nutné jim věnovat zvláštní pozornost.

Odlehlé hodnoty v datovém souboru můžeme identifikovat pomocí grafických diagnostik (viz podkapitola 5.1 Diagnostika grafů) a specifických testů. V případě, že jsou odlehlé hodnoty v souboru identifikovány, vyvstává otázka co s nimi? První možností je odlehlé hodnoty z datového souboru vyloučit. Vyloučení odlehlých hodnot nelze provádět mechanicky ihned po jejich identifikaci. Vždy je nutné nejprve příčiny jejich výskytu hlouběji analyzovat. U experimentálních dat se obecně nevylučují žádné odlehlé hodnoty v případě, že máme malý rozsah datového souboru, nebo k jejich dosažení bylo vynaloženo velké množství finančních prostředků (např. nákladná měření). K vyloučení přistupujeme pouze tehdy, kdy s jistotou víme, že odlehlá hodnota je způsobena hrubou chybou. Pokud máme dostatečný rozsah datového souboru, můžeme vyloučit jednu, v nezbytném případě maximálně dvě odlehlé hodnoty. Druhou možností je odlehlé hodnoty v datovém souboru ponechat. V tomto případě však musíme přistoupit k využití robustních metod a odhadů, které nejsou citlivé na přítomnost odlehlých hodnot (např. odhady založené na kvantilech - medián).

5.3.1 Identifikace odlehlých hodnot pomocí grafu

Program Statgraphics nabízí speciální graf pro identifikaci odlehlých hodnot s názvem Outlier Plot with Sigma Limits. Graf je založen na stanovení násobků směrodatné odchylky (±sigma, neboli ±σ) datového souboru. Jestliže je některá hodnota datového souboru vyšší (nižší) než aritmetický průměr plus minus trojnásobek směrodatné odchylky (±3σ), můžeme ji považovat za odlehlou. Konstrukce grafu spočívá ve vynesení pořadí dat v datovém souboru (číslo řádku) na osu x. Na levou osu y jsou vynášeny hodnoty sledované proměnné (např. počty návštěvníků, jejich věk apod.) a na pravou osu y násobky směrodatné odchylky.

V programu Statgraphics vytvoříme graf pro identifikaci odlehlých hodnot následujícím postupem: Describe→Numeric Data→Outlier Identification. V dialogovém okně Outlier Identification zadáme do řádku Data název analyzované proměnné a klikneme na OK. Zobrazí se výstup analýzy s výše uvedeným grafem v pravém horním okně. Výsledný graf pro identifikaci odlehlých hodnot je uveden na obrázku 5.33.



Obrázek 5.33: Graf pro identifikaci odlehlých hodnot

Graf uvedený na obrázku 5.33 byl vytvořen na základě datového souboru obsahujícího celkové počty návštěvníků vybrané technické památky v jednotlivých dnech pracovního týdne v průběhu třech měsíců. V grafu lze v jeho pravé části nahoře identifikovat jeden odlehlý bod, který má jednoznačně hodnotu vyšší než je trojnásobek směrodatné odchylky (dle pravé osy y, hodnota je dokonce vyšší než čtyřnásobek směrodatné odchylky). Pokud chceme zjistit, o kterou hodnotu se v analyzovaném datovém souboru jedná, postupujeme následovně: pro zobrazení grafu na celou plochu dvakrát klikneme levou myší do prostoru grafu. Poté klikneme levou myší přímo na identifikovanou odlehlou hodnotu. Nad grafem se vedle ikon v řádku označeném Row objeví číslo řádku datového souboru, ve kterém se odlehlá hodnota nachází. Jestliže klik levou myší v grafu na odlehlé hodnotě

podržíme, vpravo nahoře se zobrazí hodnoty souřadnice osy x a levé osy y. Po těchto úkonech se v pracovním sešitě řádek s odlehlou hodnotou podkreslí modře. Pokud chceme řádek vymazat, klikneme pravou myší na pořadové číslo řádku (řádek se podkreslí oranžově) a na klávesnici zmáčkneme Delete. V uvedeném grafu se jedná o odlehlou hodnotu nacházející se na řádku 40, přičemž celkový počet návštěvníků v daném dnu činil 223 osob.

5.3.2 Identifikace odlehlých hodnot pomocí mediánových souřadnic

Program Statgraphics dále nabízí identifikaci odlehlých hodnot na základě mediánových souřadnic. Mediánové souřadnice jsou vypočteny jako modifikované z-skóre na základě mediánové absolutní odchylky (MAD - median absolute deviation). Metoda je robustní a je tedy vhodná i pro soubory dat s asymetrickým rozdělením. V případě, že vypočtená hodnota mediánové souřadnice v absolutní hodnotě je větší než tři, jedná se o odlehlou hodnotu. V zahraniční literatuře je pro identifikaci odlehlých hodnot často doporučována kritická hodnota 3,5 (Iglewicz a Hoaglin, 1993).

Postup pro výpočet v programu Statgraphics: Describe→Numeric Data→Outlier Identification. V dialogovém okně Outlier Identification zadáme do řádku Data název analyzované proměnné a klikneme na OK. Výstup analýzy se zobrazí v levém okně ve formě tabulky. Statgraphics vždy hodnotí prvních pět nejvyšších a prvních pět nejnižších hodnot nacházejících se v datovém souboru. Ve výsledné tabulce je v prvním sloupci uvedeno číslo řádku, ve kterém se hodnoty nacházejí, v druhém sloupci jsou uvedeny přímo analyzované hodnoty. Hodnoty mediánových souřadnic jsou v tabulce uvedeny v posledním sloupci s názvem Modified MAD Z-Score.

Příklad: Pro testování byl zadán datový soubor obsahující počty návštěvníků vybrané technické památky v jednotlivých dnech pracovního týdne v průběhu třech měsíců.

Výsledky testování:

Sorted Values

Studentized Values Row	Value	Modified Without Deletion	With Deletion	MAD Z-Score
22	140,0	-1,39719	-1,43309	-1,42835
46	140,0	-1,39719	-1,43309	-1,42835
14	140,0	-1,39719	-1,43309	-1,42835
31	141,0	-1,32397	-1,35559	-1,349
58	141,0	-1,32397	-1,35559	-1,349
...				
44	176,0	1,23855	1,26567	1,42835
48	177,0	1,31177	1,34271	1,50771
24	178,0	1,38498	1,42014	1,58706
17	178,0	1,38498	1,42014	1,58706
40	**223,0**	**4,67966**	**5,98024**	**5,15794**

Vyhodnocení testu: Z výsledků vyplývá, že hodnota 223,0 nacházející se na 40. řádku datového souboru má mediánovou souřadnici větší než tři (5,15794) a můžeme ji tedy považovat za odlehlou hodnotu.

5.3.3 Grubbsův test (Grubbs Test)

Grubbsův test se používá k testování odlehlých hodnot u souborů dat, které splňují předpoklad normálního rozdělení. Test je založen na výpočtu testovacího kritéria na bázi aritmetického průměru a směrodatné odchylky. Testuje se vždy nejodlehlejší hodnota datového souboru na zvolené hladině významnosti α.

V prostředí programu Statgraphics provedeme Grubbsův test stejným postupem jako v případě identifikace odlehlých hodnot pomocí mediánových souřadnic (Describe→Numeric Data→Outlier Identification). Výsledek testu se zobrazí v levém okně pod výslednou tabulkou mediánových souřadnic. Test vždy vybírá pro testování hodnotu, která je vzdálená nejvyšším násobkem směrodatné odchylky od aritmetického průměru. Tyto hodnoty jsou uvedeny ve třetím sloupci tabulky pod názvem Studentized Values Without Deletion.

Příklad: Pro testování byl zadán datový soubor obsahující počty návštěvníků vybrané technické památky v jednotlivých dnech pracovního týdne v průběhu třech měsíců.

Výsledek Grubbsova testu:

```
Grubbs' Test (assumes normality)
----------------------------------------
Test statistic = 4,67966
P-Value = 0,0000106565
```

Vyhodnocení testu: Z výsledku testu je zřejmé, že byla testována hodnota 223,0 v řádku 40 (viz tabulka výše), která je vzdálená od aritmetického průměru o 4,67966 násobek směrodatné odchylky. Z výsledku Grubbsova testu vyplývá, že vypočtená *p*-hodnota 0,0000106565 je menší než 0,05. Čili zamítáme hypotézu H_0 ve prospěch alternativní hypotézy H_A.

Závěr: Provedením Grubbsova testu bylo s 95% statistickou jistotou prokázáno, že hodnotu 223,0 lze považovat za odlehlou.

5.3.4 Hoaglinova modifikace vnitřních hradeb (Hoaglin's modification of inner bounds)

Jedná se o jednoduchý numerický postup pro identifikaci odlehlé hodnoty. Metoda je založena na předpokladu, že datový soubor má bez odlehlých hodnot normální rozdělení.

Postup identifikace odlehlých hodnot uvedenou metodou (Meloun a Militký, 2011):

$$BD = \tilde{x}_{0,25} - K\,(\tilde{x}_{0,75} - \tilde{x}_{0,25})$$

$$BH = \tilde{x}_{0,75} + K\,(\tilde{x}_{0,75} - \tilde{x}_{0,25})$$

kde BD je modifikace dolní vnitřní hradby, BH je modifikace horní vnitřní hradby, $\tilde{x}_{0,25}$ je dolní kvartil, $\tilde{x}_{0,75}$ je horní kvartil. Parametr K je volen tak, aby pravděpodobnost $P(n, K)$, že žádný prvek pocházející z normálního rozdělení nebude ležet mimo vnitřní hradby, byla dostatečně vysoká (např. 0,95). Pro rozsah datového souboru $8 \le n \le 100$ a $P(n, K) = 0{,}95$ lze využít aproximaci hodnoty K:

$$K \approx 2{,}25 - 3{,}6/n$$

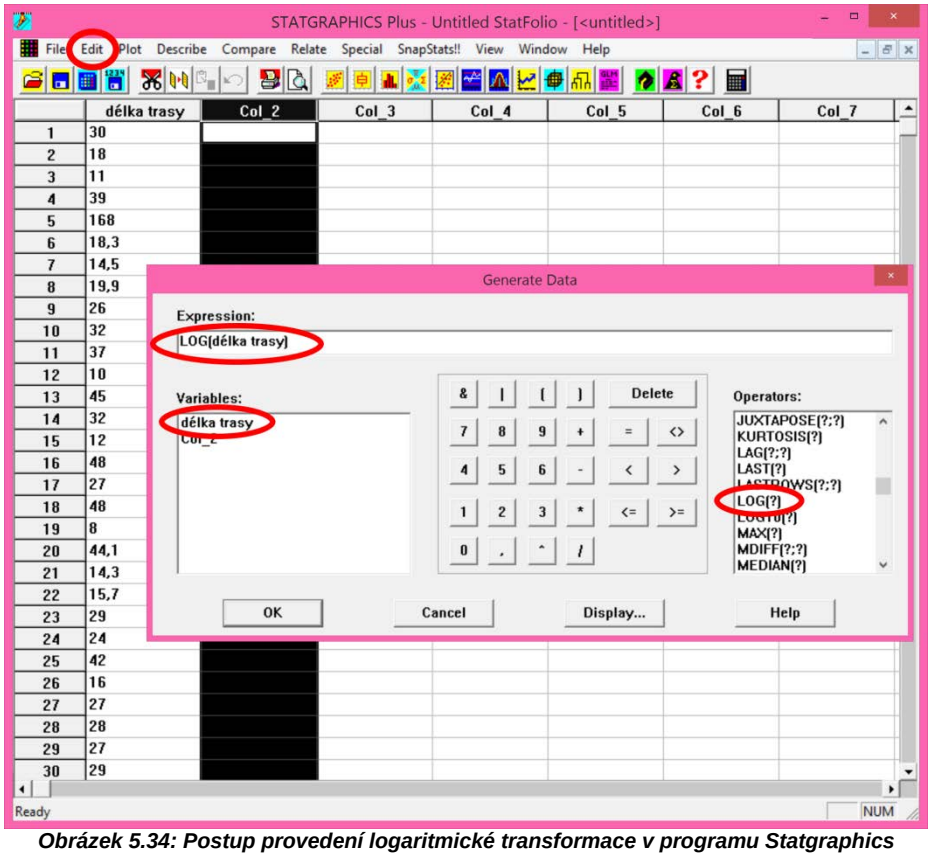
Za odlehlé hodnoty jsou považovány všechny prvky datového souboru, které leží mimo rozsah modifikovaných vnitřních hradeb [BD, BH].

5.4 Transformace

V případě, že jsme průzkumovou analýzou dat (grafickými diagnostikami a statistickými testy) dospěli k závěru, že analyzovaná experimentální data nesplňují předpoklad normálního rozdělení (nehomogenita, asymetrie, přítomnost odlehlých hodnot, které nelze vyloučit), stojíme před otázkou, jaké statistické metody a popisné charakteristiky následně použít pro jejich správné vyhodnocení. Jednou z možných cest je provedení transformace dat. Vhodnou transformací dat lze v mnoha případech dosáhnout stabilizace rozptylu a přiblížení dat k normálnímu

rozdělení. Povede-li se nám vhodnou transformaci nalézt, můžeme data vyhodnotit klasickými metodami statistické analýzy, jež vyžadují splnění předpokladu normality (např. vyčíslit odhad střední hodnoty pomocí aritmetického průměru atd.). Veškeré statistické analýzy se provedou s transformovanými daty a nakonec se výsledky vyčíslí pomocí zpětné transformace opačnou funkcí (tzv. retransformací). Ve statistické analýze dat patří mezi obecně používané transformace mocninná, exponenciální, Box-Coxova, přičemž nejpoužívanější je transformace logaritmická.

Logaritmická transformace - užití logaritmické transformace je vhodné pouze pro asymetricky rozdělená kladná data. Transformace se provádí obvykle pomocí přirozeného logaritmu ln *x*. Dekadický logaritmus je používán pro transformaci pouze ojediněle. Zpětná transformace do původních dat se provede opačnou funkcí (retransformací), čili e^{xln}.



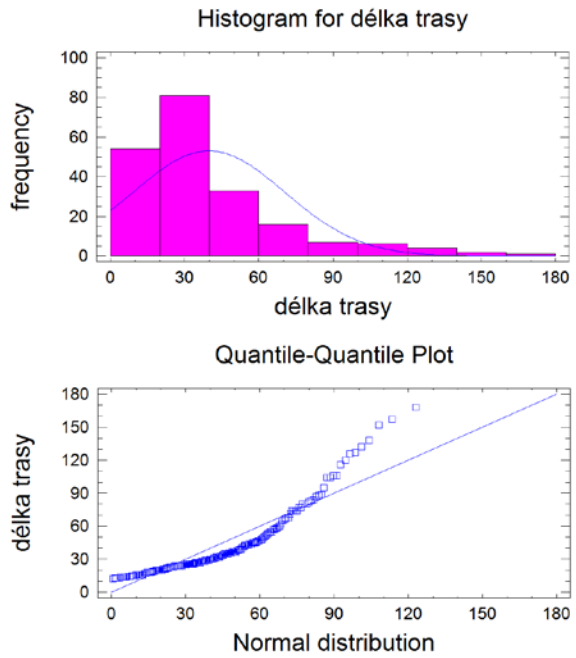
Obrázek 5.34: Postup provedení logaritmické transformace v programu Statgraphics

V programu Statgraphics provedeme logaritmickou transformaci následujícím postupem: nejprve si klikem levou myší na záhlaví označíme sloupec, do kterého chceme transformovaná data umístit (označený řádek se začerní) a v jednořádkovém textovém menu zvolíme Edit→Generate Data (viz obrázek 5.34). Následně se zobrazí dialogové okno Generate Data v němž v řádku Operators vyhledáme příkaz LOG(?) a dvakrát na něj klikneme levou myší (pozn. program Statgraphics používá na rozdíl od běžného označení ln pro přirozený logaritmus označení LOG a pro dekadický logaritmus označení LOG10). Poté klikneme do řádku Expression kde se zobrazil příkaz LOG(?). Vymažeme otazník a dvakrát klikneme levou myší ve sloupci Variables na analyzovanou proměnnou, kterou chceme transformovat. Finální výraz v řádku Expression má tvar: LOG(název analyzované proměnné) (Viz obrázek 5.34). Klikneme na OK a transformované hodnoty původní proměnné se v pracovním sešitě zobrazí v námi předem definovaném sloupci. Dalším krokem je opětovná průzkumová analýza dat (grafická diagnostika a testování) s cílem ověřit, zda transformace byla úspěšná a vedla k normalitě dat.

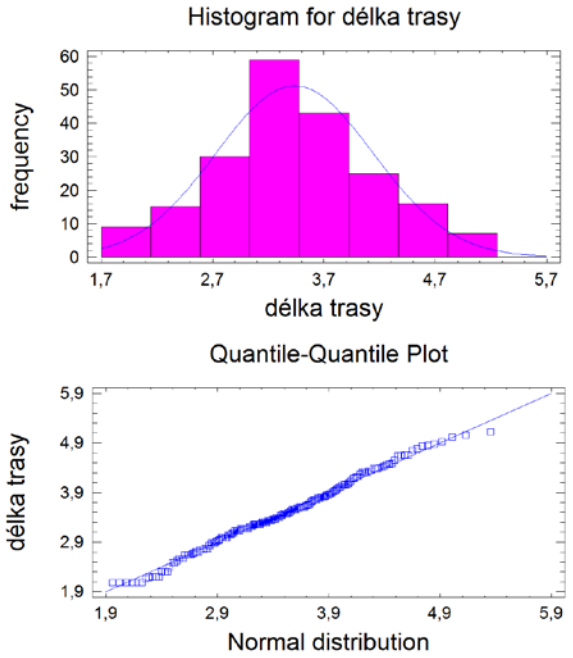
Pro názornost a lepší pochopení problematiky transformace dat je na obrázku 5.35 uveden histogram četnosti a kvantil-kvantilový graf pro vybranou proměnnou (jedná se o délku trasy, kterou dotazovaní cyklisté v rámci výletu ujeli) před logaritmickou transformací a na obrázku 5.36 po transformaci. Ze srovnání grafů pro původní data a data po transformaci je zřejmé, že transformace vedla k zesymetričtění rozdělení a přiblížení normalitě.

Pro názornost vlivu logaritmické transformace na rozdělení dat byl také proveden chí-kvadrát test. Provedením testu datového souboru před transformací bylo zjištěno, že vypočtená *p*-hodnota $5{,}55112.10^{-16}$ je významně menší

než 0,05. Hypotéza H_0 se tedy zamítá a na hladině významnosti 0,05 je prokázáno, že data nelze považovat za výběr z normálního rozdělení. Pro transformovaná data však chí-kvadrát test vypočetl *p*-hodnotu 0,133581, která je větší než 0,05, čili hypotéza H_0 se přijímá a data lze považovat za výběr z normálního rozdělení. Závěr: Logaritmická transformace byla úspěšná, transformovaná data splňují předpoklad normality a lze je dále analyzovat klasickými statistickými metodami.



Obrázek 5.35: Histogram četnosti a Q-Q graf pro data před transformací



Obrázek 5.36: Histogram četnosti a Q-Q graf pro data po transformaci

6 Testování statistických hypotéz

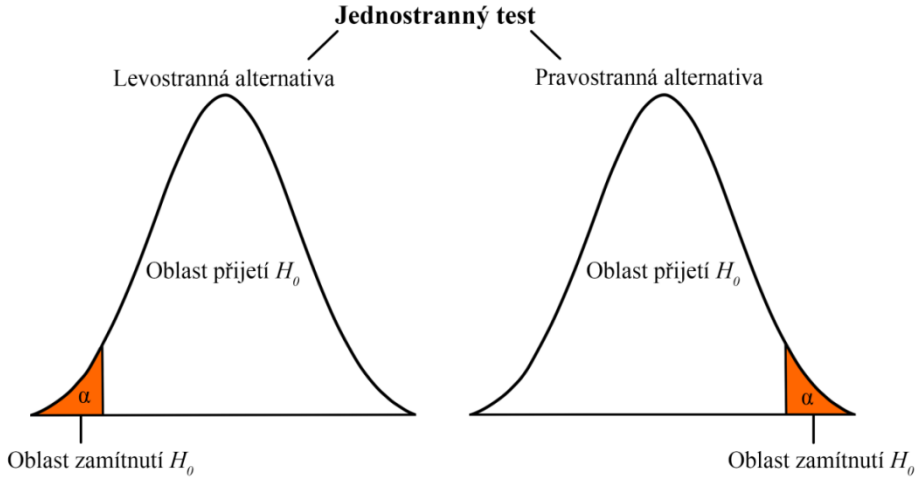
Kapitola je zaměřena na osvětlení problematiky týkající se metod testování statistických hypotéz. V rámci kapitoly je názorně vysvětlen test správnosti, test shodnosti, párový test, včetně vhodnosti jejich použití, interpretace výsledků a provedení v prostředí programu Statgraphics. Závěr kapitoly je věnován neparametrickým testům.

Při vyhodnocování experimentálních dat velmi často potřebujeme vyřešit otázky typu: liší se zjištěné hodnoty dvou různých souborů (např. soubory dat zjištěné na dvou různých geolokalitách, hodnoty naměřené dvěma různými meteostanicemi apod.), liší se námi naměřené experimentální hodnoty od hodnot deklarovaných výrobcem (např. koncentrace soli v potravině) atd.? K vyřešení těchto otázek se ve statistické analýze používají metody testování statistických hypotéz. Podstata testování statistických hypotéz (co je to statistická hypotéza, hladina významnosti α, způsoby testování) je v těchto výukových textech již popsána v podkapitole 5.2 Statistické testy.

Základem testování statistických hypotéz je vytvoření nulové hypotézy H_0 a k ní alternativní hypotézy H_A. Statistické hypotézy se podle charakteru řešené úlohy formulují jako jednostranné (v angličtině nazývané one-tailed test, nebo one-sided test), nebo oboustranné (v angličtině nazývané two-tailed test, nebo two-sided test). V případě jednostranného testu formulujeme hypotézy takto:

$H_0 : \mu = \mu_0$; $H_A : \mu < \mu_0$ (nebo $H_A : \mu > \mu_0$)

kde μ je skutečná hodnota parametru (např. střední hodnota) a μ_0 je předpokládaná (zvolená) hodnota. V odborné literatuře bývá často jednostranný test blíže specifikován a to pojmem levostranná alternativa (angl. left-tailed test) pro $H_A : \mu < \mu_0$ a pravostranná alternativa (angl. right-tailed test) pro $H_A : \mu > \mu_0$. Takto definovanými hypotézami testujeme, zda naměřené hodnoty na zvolené hladině významnosti α jsou menší (příp. větší) než předpokládaná hodnota. Grafické znázornění přijetí či zamítnutí hypotéz a rozdělení hladiny významnosti α pro jednostranný test znázorňuje obrázek 6.1.



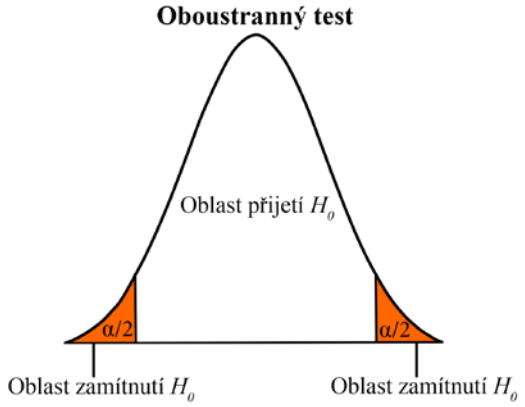
Obrázek 6.1: Kritický obor u jednostranného testu

Jako příklad využití jednostranného testu (levostranná alternativa $H_A : \mu < \mu_0$) můžeme uvést následující problém: deklarovaná váha balení dodávané potraviny je 100 g (μ_0) a nás zajímá, zda výrobce nepodvádí a dodržuje uvedenou hmotnost balení. Vybereme náhodně 20 dodaných balení, zvážíme je (čímž získáme soubor experimentálních dat), vypočteme střední hodnotu (μ) a provedeme jednostranný test s výše uvedenými formulacemi hypotéz. V případě, že testem prokážeme platnost hypotézy H_0 jsme spokojení, výrobce nás nepodvádí. Ovšem pokud testem zamítneme hypotézu H_0 ve prospěch hypotézy H_A, prokážeme, že dodávaná balení mají nižší než deklarovanou váhu a výrobce nás zřejmě podvádí.

V případě oboustranného testu formulujeme hypotézy následovně:

$$H_0 : \mu = \mu_0; H_A : \mu \neq \mu_0$$

Takto definovanými hypotézami testujeme, zda naměřené hodnoty na zvolené hladině významnosti α jsou shodné či se od předpokládané hodnoty významně liší. Grafické znázornění přijetí či zamítnutí hypotéz a rozdělení hladiny významnosti α pro oboustranný test znázorňuje obrázek 6.2.



Obrázek 6.2: Kritický obor u oboustranného testu

Jako příklad pro využití oboustranného testu můžeme uvést následující příklad: výrobce dodává šunku s deklarovaným obsahem masa (μ_0) a my jsme se rozhodli kvalitu dodávané šunky ověřit. K dispozici máme 20 vzorků. Stanovíme obsah masa ve vzorcích (experimentální data) a z těchto dat vypočteme střední hodnotu (μ). Následně provedeme oboustranný test s použitím výše uvedených formulací hypotéz. Pokud testem prokážeme

platnost hypotézy H_0 dodávaná šunka je kvalitní. Jestliže však testem zamítneme hypotézu H_0 ve prospěch hypotézy H_A, prokážeme, že obsah masa v šunce neodpovídá hodnotě uváděné výrobcem, čili je nižší, nebo vyšší.

6.1 Parametrické testy

6.1.1 Test správnosti

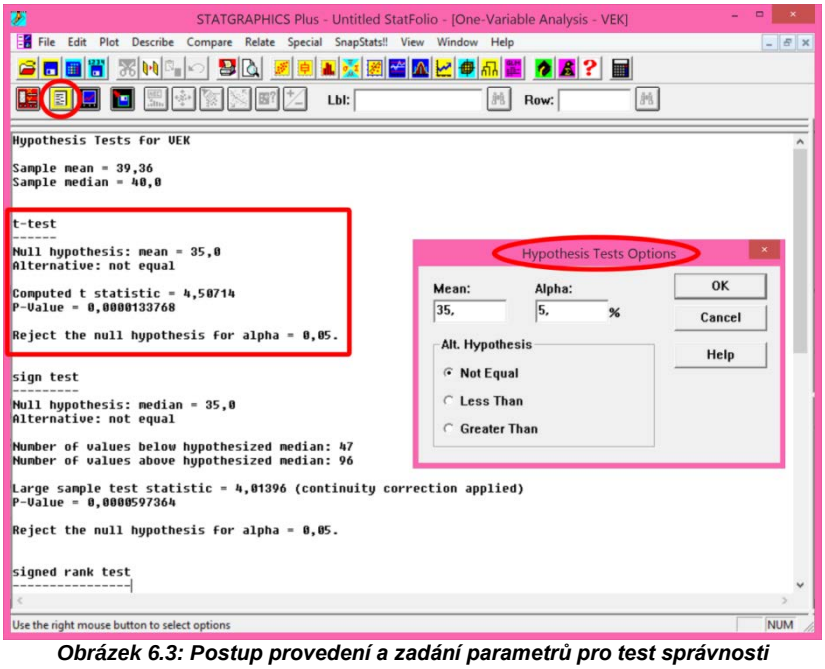
Test správnosti bývá velmi často v odborné literatuře označován také jako „test střední hodnoty". Test správnosti je parametrickým testem, čili podmínkou jeho použití je splnění předpokladu normality datového souboru. Pokud datový soubor předpoklad normality nesplňuje, je nutné použít neparametrický test (viz podkapitola "6.2 Neparametrické testy"). Test správnosti se využívá především u statistických úloh, kdy potřebujeme rozhodnout, zda se zvolený parametr datového souboru (nejčastěji se jedná o střední hodnotu, nebo rozptyl) rovná (příp. je menší, či větší) předem známé hodnotě.

Provedením testu správnosti můžeme odpovědět např. na otázky typu: Je obsah soli v masném výrobku ve shodě s deklarovanou hodnotou? Měří naše zařízení ve shodě s hodnotou etalonu? Liší se námi naměřené hodnoty množství srážek v dané geolokalitě od střední hodnoty uváděné ČHMÚ pro předešlý rok? Test správnosti se provádí pomocí jednovýběrového t-testu (Studentův test). Testovací kritérium t je definováno následující rovnicí:

$$t = \frac{\bar{x} - \mu_0}{s} \cdot \sqrt{n}$$

kde $\bar{x}$ je střední hodnota (aritmetický průměr) testovaného souboru dat, μ_0 je konstanta (zvolená hodnota), s je směrodatná odchylka a n je rozsah souboru.

Postup testování v prostředí programu Statgraphics: Describe→Numeric Data→One-Variable Analysis. V dialogovém okně One-Variable Analysis zadáme do řádku Data název proměnné, kterou chceme testovat a klikneme na OK. Následně klikneme na ikonu Tabular Options a v nabídce zatrhneme Hypothesis Tests. V levém dolním okně Hypothesis Tests nalezneme výsledky provedeného t-testu pro defaultně nastavené hodnoty. Dvojklikem levou myší do prostoru okna si výsledky zvětšíme přes celou levou plochu. Pro zadání zvolené hodnoty μ_0 a formulaci alternativní hypotézy H_A klikneme pravou myší do prostoru okna a v zobrazené nabídce zadáme Pane Options. Zobrazí se dialogové okno Hypothesis Tests Options. Zde do řádku Mean zadáme zvolenou střední hodnotu pro nulovou hypotézu a do řádku Alpha zadáme zvolenou hladinu významnosti. V nabídce Alt. Hypothesis zformulujeme alternativní hypotézu - pro oboustranný test zatrhneme v nabídce Not Equal, pro jednostranný test levostrannou alternativu Less Than, pravostrannou alternativu Greater Than (viz obrázek 6.3).



Obrázek 6.3: Postup provedení a zadání parametrů pro test správnosti

Příklad: Pro testování byl zadán datový soubor obsahující věk 150 dotázaných návštěvníků technické památky v průběhu roku 2016. Za rok 2015 uvádí správce této památky průměrný věk návštěvníka 35 let. Je věk dotázaných návštěvníků v roce 2016 shodný s průměrným věkem zjištěným v roce 2015?

Výsledek t-testu:

```
t-test
------------------------------------------------------
Null hypothesis: mean = 35,0
Alternative: not equal
Computed t statistic = 4,50714
P-Value = 0,0000133768
Reject the null hypothesis for alpha = 0,05.
```

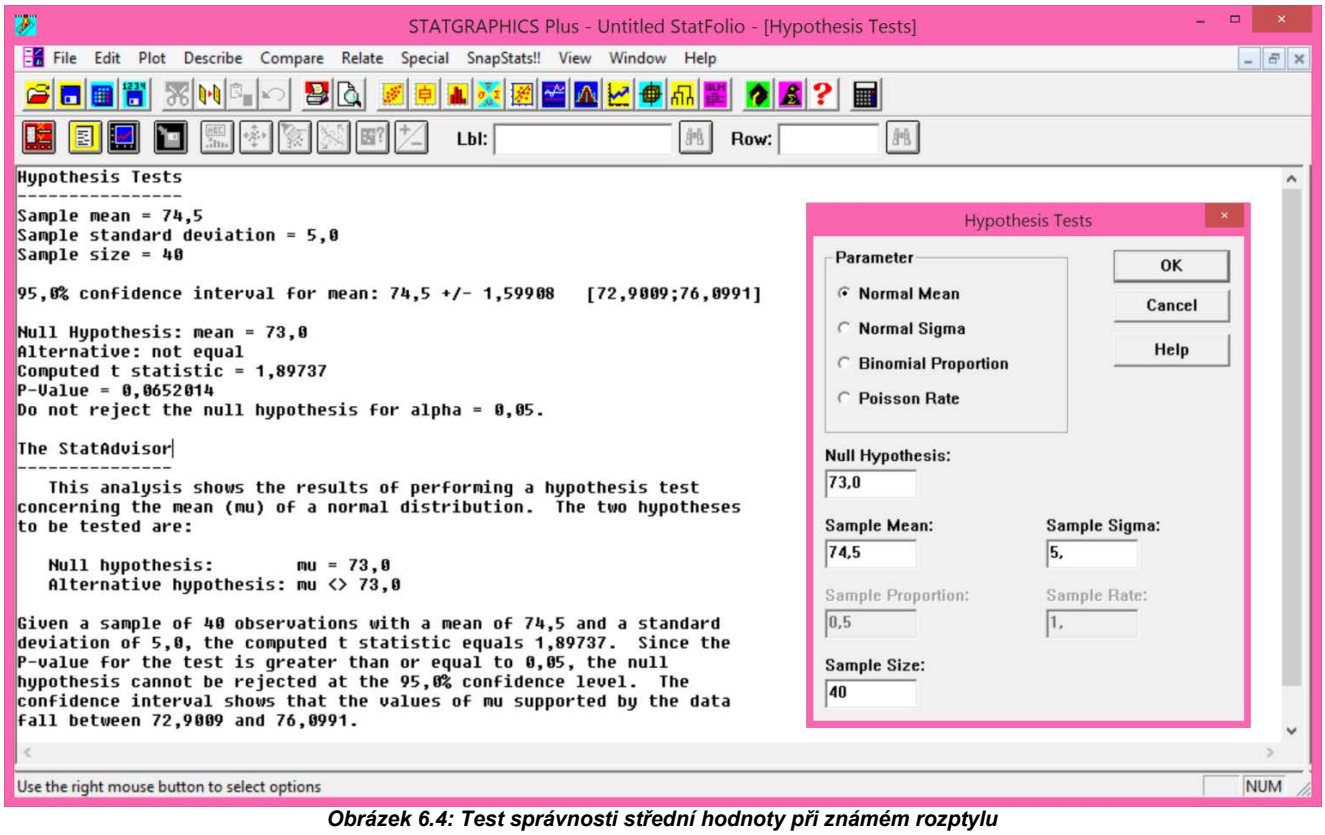
Vyhodnocení testu: Z výsledku t-testu vyplývá, že vypočtená *p*-hodnota 0,0000133768 je menší než 0,05, což znamená, že zamítáme nulovou hypotézu H_0 (H_0 = věk je shodný) ve prospěch alternativní hypotézy H_A (H_A = věk není shodný).

Závěr: Provedeným testem bylo s 95% statistickou jistotou prokázáno, že věk 150 návštěvníků technické památky v roce 2016 se významně liší od průměrného věku návštěvníků zjištěného v roce 2015.

Další variantou testu správnosti je testování střední hodnoty při známém rozptylu. Využíváme jej v případech, kdy dopředu známe nejenom střední hodnotu μ_0, ale i odhad variability (rozptyl, či směrodatnou odchylku).

Test správnosti střední hodnoty při známém rozptylu provedeme v programu Statgraphics následujícím postupem: Describe→Hypothesis Tests. Zobrazí se dialogové okno Hypothesis Tests, kde v nabídce Parameter zvolíme Normal Mean, do řádku Null Hypothesis zadáme zvolenou střední hodnotu μ_0, do řádku Sample Mean zadáme vypočtenou střední hodnotu testovaného souboru (tedy aritmetický průměr $\bar{x}$), do řádku Sample Size zadáme počet hodnot vyskytujících se v testovaném souboru a do řádku Sample Sigma zadáme zvolenou hodnotu směrodatné odchylky (viz obrázek 6.4).

Pro volbu hladiny významnosti α a formulaci alternativní hypotézy H_A klikneme pravou myší do prostoru levého okna Hypothesis Tests a zadáme Analysis Options. Zobrazí se dialogové okno Hypothesis Tests Options, ve kterém zadáme námi požadované parametry testu.



Obrázek 6.4: Test správnosti střední hodnoty při známém rozptylu

Příklad: V restauraci na sledované geolokalitě bylo odebráno 40 vzorků čepované limonády, přičemž měřeným parametrem byl obsah cukru. Průměrný obsah cukru ve vzorcích činil 74,5 g/l. V minulosti byl již podobný průzkum proveden a restaurace uvádí průměrný obsah cukru 73 g/l se směrodatnou odchylkou 5 g/l. Liší se aktuálně stanovený obsah cukru v limonádě od průměrné hodnoty uváděné restaurací dle předchozího průzkumu?

Výsledek t-testu:

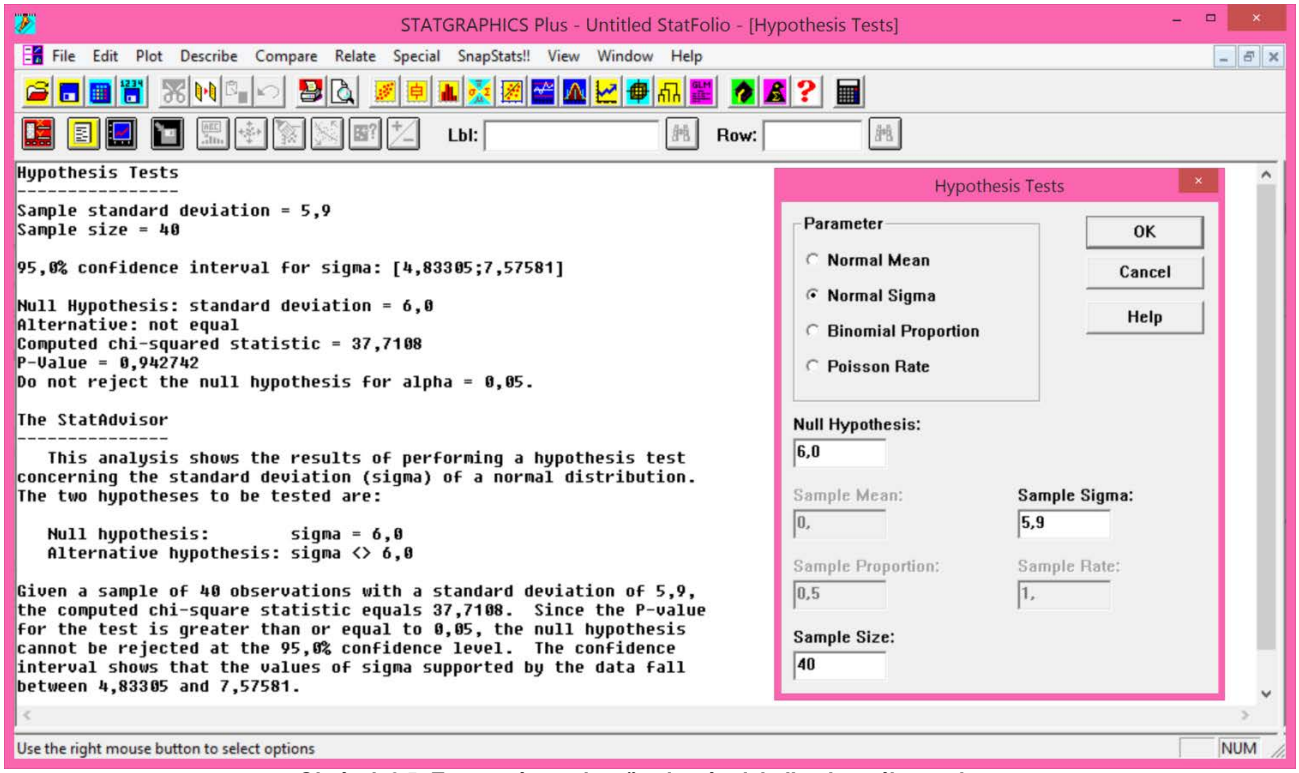
```
Hypothesis Tests
------------------------
Sample mean = 74,5
Sample standard deviation = 5,0
Sample size = 40
95,0% confidence interval for mean: 74,5 +/- 1,59908   [72,9009; 76,0991]
Null Hypothesis: mean = 73,0
Alternative: not equal
Computed t statistic = 1,89737
P-Value = 0,0652014
Do not reject the null hypothesis for alpha = 0,05.
```

Vyhodnocení testu: Z výsledku t-testu vyplývá, že vypočtená p-hodnota 0,0652014 je větší než 0,05, což znamená, že přijímáme nulovou hypotézu H_0.

Závěr: Provedeným testem bylo s 95% statistickou jistotou prokázáno, že stanovené obsahy cukru v čepované limonádě jsou srovnatelné s výsledky uváděnými restaurací dle předchozího průzkumu.

Test správnosti můžeme použít i v případech, kdy potřebujeme pouze otestovat variabilitu datového souboru vůči předem dané hodnotě. Testování rozptylu, respektive směrodatné odchylky, se provádí pomocí χ^2-testu (chí-kvadrát testu).

Postup testování v programu Statgraphics je obdobný jako v případě testu správnosti střední hodnoty při známé hodnotě rozptylu: Describe→Hypothesis Tests. Zobrazí se dialogové okno Hypothesis Tests, kde v nabídce Parameter zvolíme Normal Sigma, do řádku Null Hypothesis zadáme zvolenou hodnotu směrodatné odchylky, do řádku Sample Size zadáme počet hodnot vyskytujících se v testovaném souboru a do řádku Sample Sigma zadáme vypočtenou hodnotu směrodatné odchylky analyzovaného souboru dat (viz obrázek 6.5).



Obrázek 6.5: Test správnosti směrodatné odchylky datového souboru

Vyhodnocení testu následně provedeme pomocí výsledné *p*-hodnoty stejným způsobem jako v předchozích příkladech. Pro vyhodnocení výsledku můžeme využít i interpretaci, kterou nabízí přímo program Statgraphics v části The StatAdvisor.

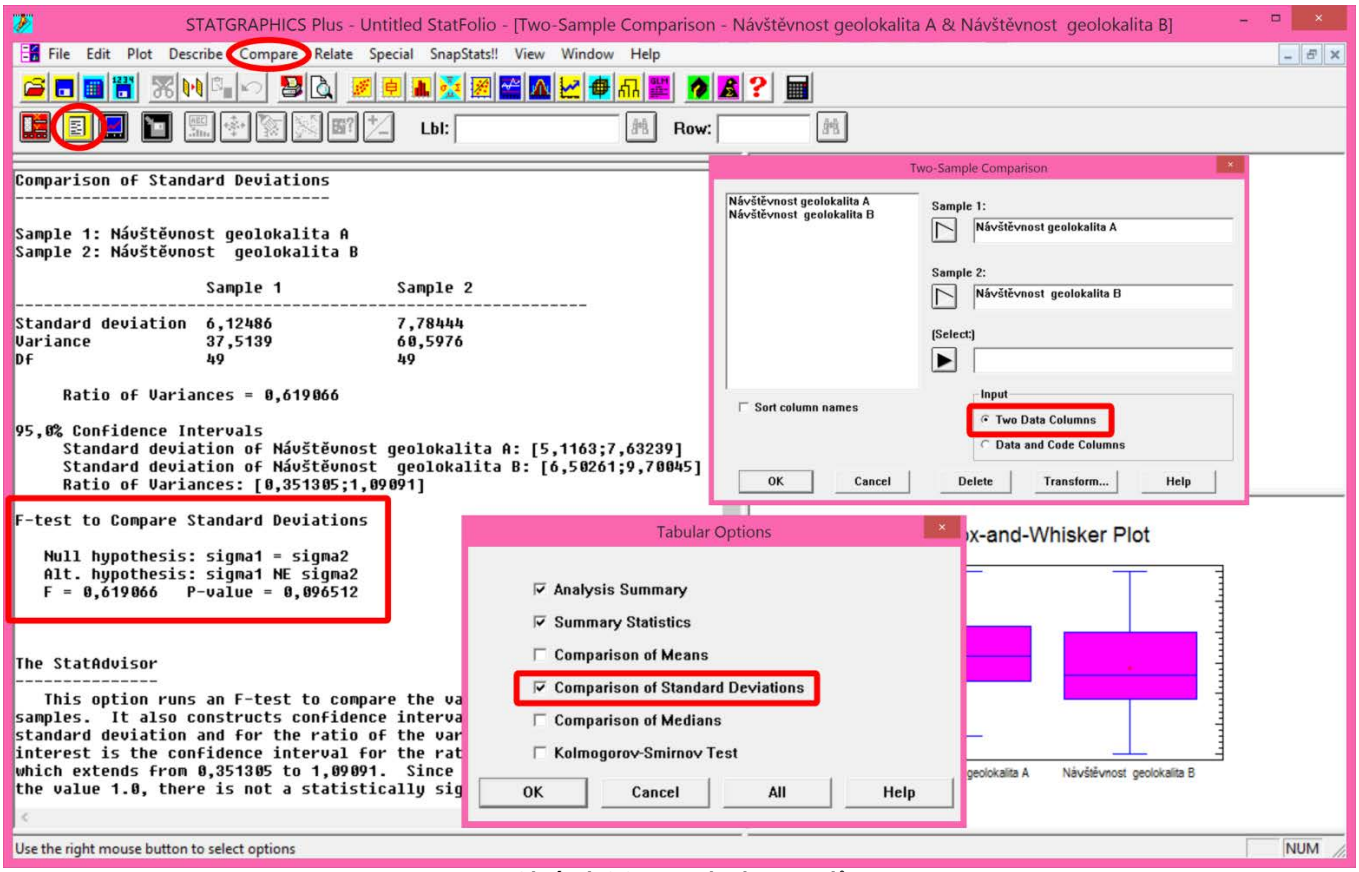
6.1.2 Test shodnosti

Test shodnosti je velmi často v odborné literatuře označován jako "test shody středních hodnot". Jedná se o parametrický dvouvýběrový test, který předpokládá normalitu a nezávislost náhodných výběrů. V případě, že náhodné výběry jsou závislé (např. koncentrace soli v masném výrobku byly měřeny dvěma různými přístroji na témže vzorku, čili je mezi nimi párová souvztažnost), je nutné pro testování shody středních hodnot použít párový test (viz odstavec 6.1.3 Párový test). Pokud datové soubory nesplňují předpoklad normality, je nutné použít neparametrický test (viz podkapitola 6.2 Neparametrické testy).

Test shodnosti nachází uplatnění u statistických úloh, kde potřebujeme rozhodnout, zda jsou střední hodnoty dvou datových souborů shodné, nebo se liší. Provedením testu shodnosti můžeme odpovědět např. na otázky typu: Jsou střední hodnoty věku návštěvníků zjištěné v rámci dotazníkového průzkumu na geolokalitě A a B shodné? Je hmotnost výrobku dodavatele A a B stejná? Liší se návštěvnost geolokality v létě a v zimě?

Test shodnosti se provádí pomocí klasického Studentova dvouvýběrového t-testu. Pro správnou volbu postupu testování je nutné nejprve ověřit, zda rozptyly náhodných výběrů jsou shodné (vykazují homoskedasticitu), či rozdílné (vykazují heteroskedasticitu). K testování shody rozptylů se používá Fisher-Snedecorův test (F-test).

Test shody rozptylů provedeme v programu Statgraphics následovně: Compare→Two Samples→Two-Sample Comparison. V dialogovém okně Two-Sample Comparison zadáme do řádku Sample 1 název prvního datového souboru (výběru), do řádku Sample 2 název druhého datového souboru, v nabídce Input zatrhneme Two Data Columns a klikneme na OK. Zobrazí se nám výstup analýzy obsahující čtyři okna. Nyní klikneme levou myší na ikonu Tabular options (druhá ikona zleva v horní liště výstupu). Zobrazí se dialogové okno Tabular Options. Nejprve musíme provést test shody rozptylů, proto v nabídce zatrhneme Comparison of Standard Deviations. Po těchto úkonech se v levé části zobrazí další okno s výsledkem F-testu. Dvojklikem levou myší do prostoru okna se výsledky zobrazí přes celou levou část výstupů (viz obrázek 6.6).



Obrázek 6.6: Test shody rozptylů

Příklad: Na dvou vybraných geolokalitách (A a B) byla v rámci dotazníkového výzkumu sledovaná jejich návštěvnost v průběhu 50ti dnů. Jsou rozptyly výsledných datových souborů (náhodných výběrů) shodné, či rozdílné?

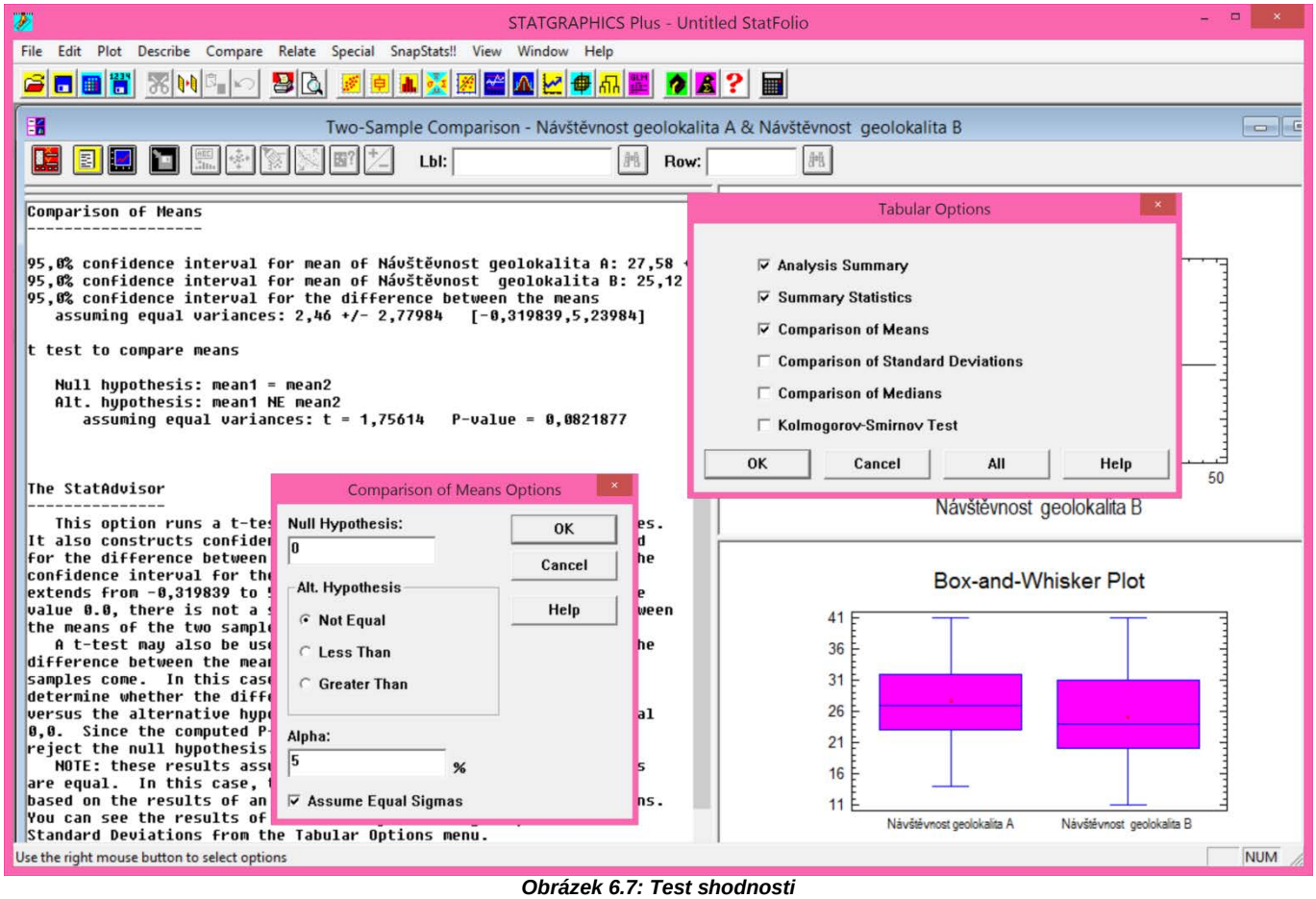
Výsledek F-testu:

```
F-test to Compare Standard Deviations:
Null hypothesis: sigma1 = sigma2
Alt. hypothesis: sigma1 NE sigma2
F = 0,619066   P-value = 0,096512
```

Vyhodnocení testu: Z výsledku F-testu vyplývá, že vypočtená *p*-hodnota 0,096512 je větší než 0,05, což znamená, že přijímáme nulovou hypotézu H_0.

Závěr: Provedeným F-testem bylo s 95% statistickou jistotou prokázáno, že rozptyly (resp. směrodatné odchylky) datových souborů jsou shodné.

Nyní můžeme provést test shodnosti: znova klikneme levou myší na ikonu Tabular options a v dialogovém okně zatrhneme Comparison of Means. V levé části výstupu se zobrazí další okno s výsledkem dvouvýběrového t-testu. Následně musíme programu zadat, zda má t-test být proveden pro shodné, či rozdílné rozptyly. Klikneme pravou myší do okna s výsledkem testu a zvolíme Pane Options. Zobrazí se dialogové okno Comparison of Means Options, v jehož dolní části se nachází volba Assume Equal Sigmas. Jelikož nám F-test prokázal, že rozptyly jsou shodné, volbu zatrneme (viz obrázek 6.7). Pokud bychom F-testem došli k závěru, že rozptyly shodné nejsou, je nutné tuto volbu zrušit (políčko u volby musí být prázdné). Dále v dialogovém okně zadáme požadovanou hladinu významnosti α a klikneme na OK.



Obrázek 6.7: Test shodnosti

Příklad – pokračování výše uvedeného: Na dvou vybraných geolokalitách (A a B) byla v rámci dotazníkového výzkumu sledovaná jejich návštěvnost v průběhu 50ti dnů. Je střední hodnota věku návštěvníků geolokality A shodná se střední hodnotou věku návštěvníků geolokality B?

Výsledek dvouvýběrového t-testu:

```
t test to compare means:
Null hypothesis: mean1 = mean2
Alt. hypothesis: mean1 NE mean2
assuming equal variances: t = 1,75614   P-value = 0,0821877
```

Vyhodnocení testu: Z výsledku dvouvýběrového t-testu vyplývá, že vypočtená *p*-hodnota 0,0821877 je větší než 0,05, což znamená, že přijímáme nulovou hypotézu H_0.

Závěr: Provedeným testem shodnosti bylo prokázáno, že na 95% hladině významnosti neexistuje statisticky významný rozdíl mezi středními hodnotami věku návštěvníků geolokality A a geolokality B.

6.1.3 Párový test

Párový test je parametrickým testem, který předpokládá normalitu náhodných výběrů a nepřítomnost odlehlých hodnot. V případě nedodržení uvedených předpokladů je nutné použít neparametrickou modifikaci tohoto testu. Dalším předpokladem užití párového testu je vzájemná závislost náhodných výběrů. Je tedy vhodný pro párová data, čili pro data mezi kterými existuje určitá souvstažnost (logická vazba). Testem lze např. řešit statistické úlohy typu: Poskytují dvě různé meteostanice (rozdílného typu, rozdílných výrobců) stejné výsledky? Mělo zlepšení poskytovaných služeb na geolokalitě významný vliv na spokojenost návštěvníků (porovnáváme spokojenost před a po zlepšení služeb)? Mělo snížení vstupného vliv na návštěvnost technické památky?

Vzhledem k vzájemné závislosti náhodných výběrů nelze pro testování použít dvouvýběrový t-test. Pro testování se používá párový t-test, jehož testovací kritérium je založeno na podílu střední hodnoty a směrodatné odchylky vypočtených z rozdílů mezi párovými daty. Tímto je vlastně dvourozměrná úloha převedena na jednorozměrnou, čili jednovýběrový t-test.

Postup provedení párového testu v programu Statgraphics: Compare→Two Samples→ Paired-Sample Comparison. V dialogovém okně Paired-Sample Comparison zadáme do řádku Sample 1 název prvního datového souboru (výběru), do řádku Sample 2 název druhého datového souboru a klikneme na OK. Zobrazí se nám výstup analýzy obsahující čtyři okna. Nyní klikneme levou myší na ikonu Tabular options (druhá ikona zleva v horní liště výstupu). V nabídce dialogového okna Tabular Options zvolíme Hypothesis Tests. Po těchto úkonech se v levé části zobrazí další okno s výsledkem jednovýběrového t-testu. (viz obrázek 6.8).

Příklad: V průběhu července 2015 byla sledována návštěvnost technické památky. Na jaře roku 2016 bylo sníženo vstupné a v průběhu července 2016 opět sledována návštěvnost. Mělo snížení vstupného vliv na návštěvnost technické památky?

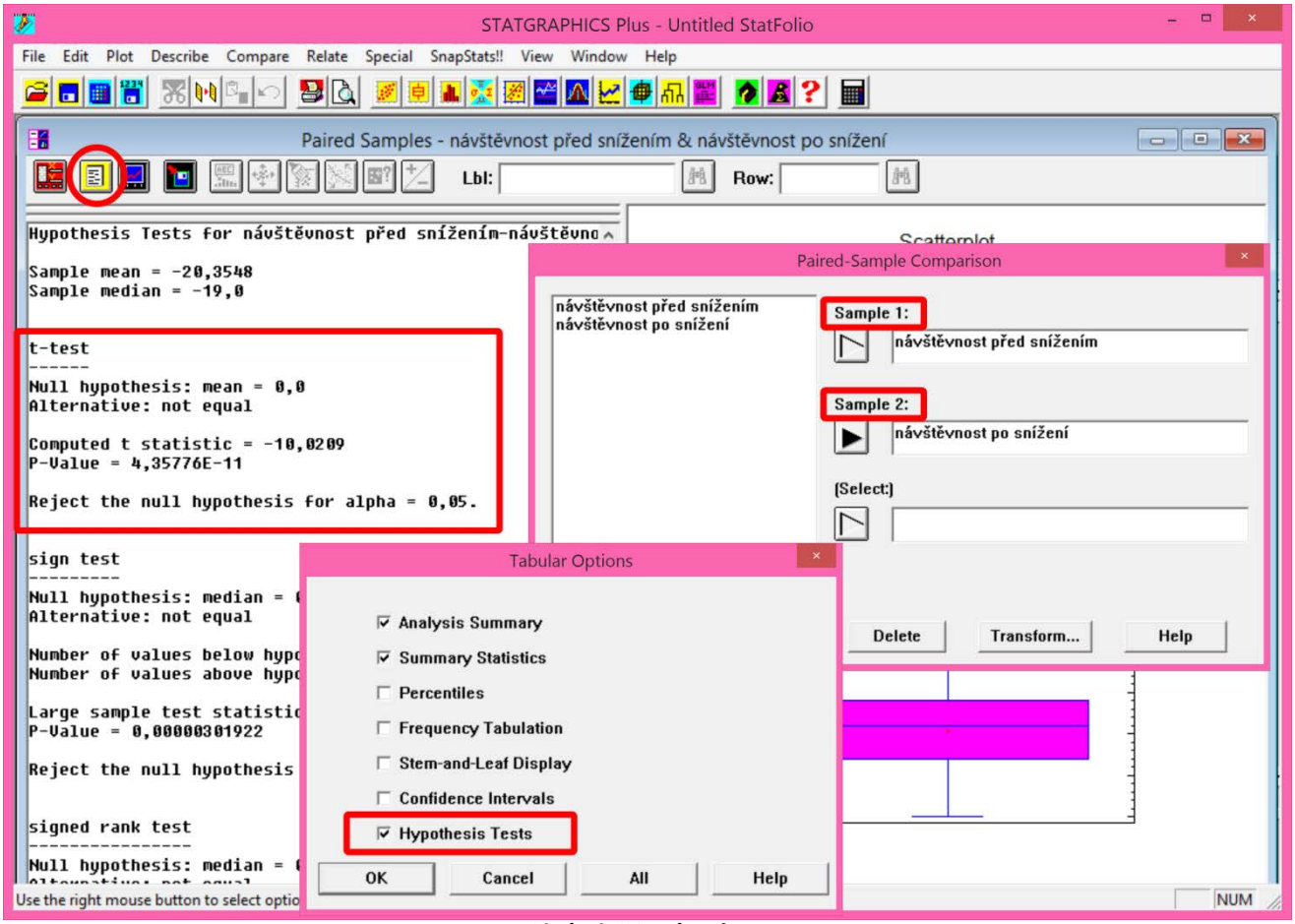
Výsledek párového testu:

```
t-test
--------
Null hypothesis: mean = 0,0
Alternative: not equal
Computed t statistic = -10,0209
P-Value = 4,35776E-11
Reject the null hypothesis for alpha = 0,05.
```

Vyhodnocení testu: Z výsledku párového testu (jednovýběrového t-testu) vyplývá, že vypočtená *p*-hodnota $4{,}35776.10^{-11}$ je menší než 0,05, což znamená, že zamítáme nulovou hypotézu H_0.

Závěr: Párovým testem bylo na 95% hladině významnosti prokázáno, že snížení vstupného mělo významný vliv na návštěvnost technické památky.

Jestliže datové soubory nesplňují předpoklad normality pro provedení testů uvedených v podkapitole 6.1, pokusíme se nejprve provést transformaci dat. Pouze v případech kdy transformace není úspěšná, přistupujeme k použití neparametrických testů.



Obrázek 6.8: Párový test

6.2 Neparametrické testy

V předchozích podkapitolách byly podrobně popsány parametrické testy, které jsou založeny na předpokladu normálního rozdělení analyzovaného datového souboru. V praxi se však často stává, že experimentální data předpoklad normality nesplňují, případně rozdělení dat neznáme vůbec. Zejména u přírodních dat se normální rozdělení vyskytuje spíše výjimečně. Pro testování těchto datových souborů se používají neparametrické testy.

Neparametrické testy nevyžadují předpoklad specifického rozdělení dat a jsou robustní na odlehlé hodnoty. Výpočty testovacích kritérií jsou založeny na pořádkových statistikách (mediánech), proto jsou běžně nazývány pořadovými testy. Na rozdíl od parametrických testů jsou však slabší, tzn., že v případě jejich použití dochází s větší pravděpodobností k přijetí nepravdivé hypotézy.

Neparametrických testů existuje velké množství jako např. Wicoxonův test, Mann-Whitneyův test, znaménkový test, Wald-Wolfowitzův test, neparametrický Kolmogorov-Smirnovův test, mediánový test atd. V následujících odstavcích budou vysvětleny pouze ty neparametrické testy, které lze provést v programu Statgraphics.

6.2.1 Jednovýběrový znaménkový test

Jednovýběrový znaménkový test (One Sample Sign Test) je neparametrickou obdobou jednovýběrového t-testu (testu správnosti). Používáme jej pro statistické úlohy, kdy potřebujeme provést test správnosti, avšak datový soubor nesplňuje předpoklady použití parametrického testu.

V programu Statgraphics provedeme znaménkový test stejným postupem jako v případě testu správnosti, jež je uveden v odstavci 6.1.1 Test správnosti. Výsledek testu se nám zobrazí v levé části výstupu analýzy pod výsledkem t-testu (viz obrázek 6.9).

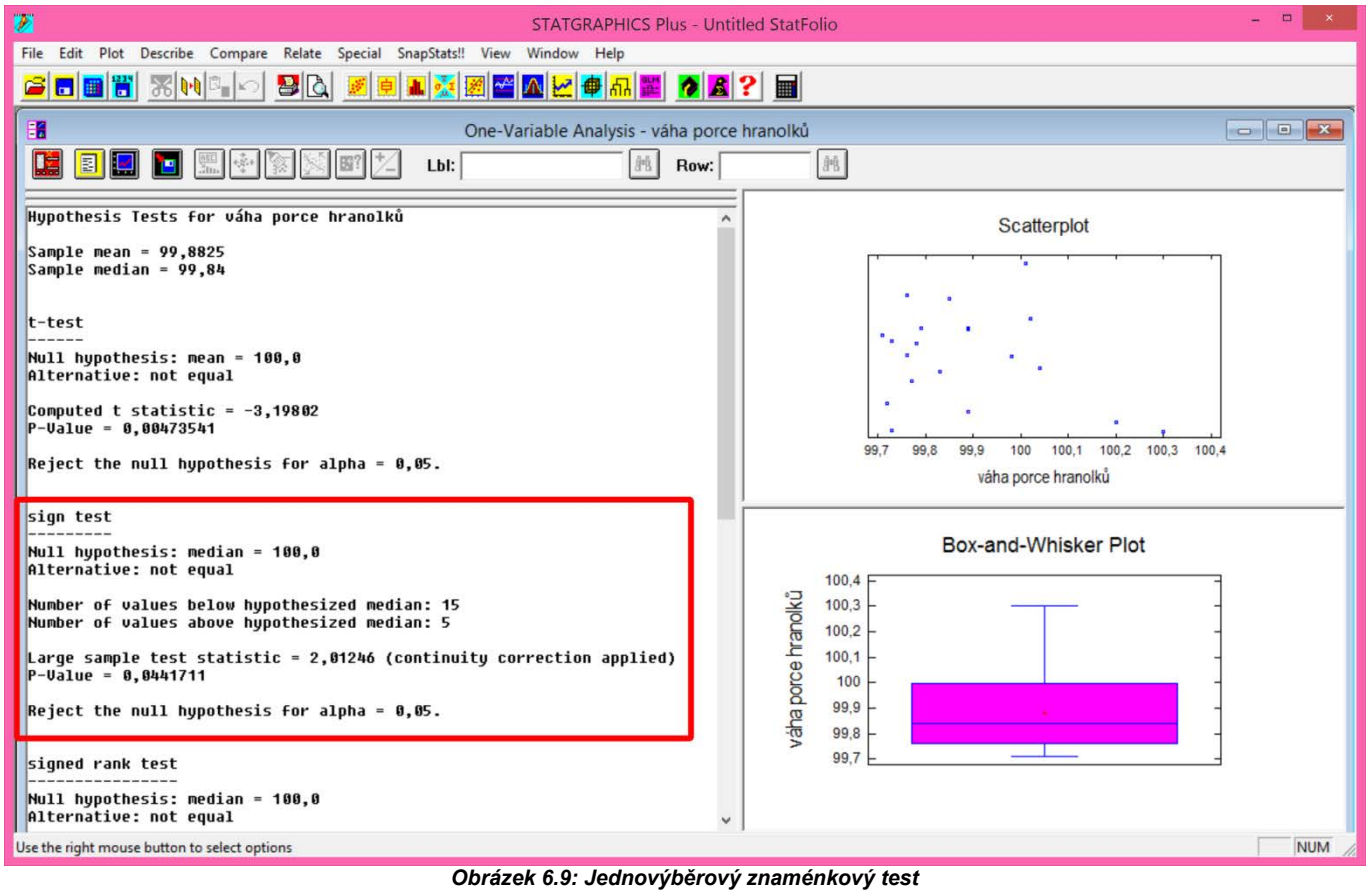
Příklad: Restaurace na dané geolokalitě prodává smažené hranolky, u nichž uvádí hmotnost 100 g. Náhodně bylo zváženo 20 prodaných porcí. Rozhodněte, zda střední hodnota (medián) hmotností odpovídá uváděné hmotnosti.

Výsledek jednovýběrového znaménkového testu (sign test):

```
sign test
------------
Null hypothesis: median = 100,0
Alternative: not equal
Number of values below hypothesized median: 15
Number of values above hypothesized median: 5
Large sample test statistic = 2,01246 (continuity correction applied)
P-Value = 0,0441711
Reject the null hypothesis for alpha = 0,05.
```

Vyhodnocení testu: Z výsledku jednovýběrového znaménkového testu vyplývá, že vypočtená *p*-hodnota 0,0441711 je menší než 0,05, což znamená, že zamítáme nulovou hypotézu H_0 ve prospěch alternativní hypotézy H_A.

Závěr: Jednovýběrovým znaménkovým testem bylo s 95% statistickou jistotou prokázáno, že hmotnost prodávaných porcí hranolek se významně liší od hmotnosti uváděné restaurací.



Obrázek 6.9: Jednovýběrový znaménkový test

6.2.2 Mann-Whitneyův test

Mann-Whitneyův test (Mann-Whitney W test) je neparametrickou obdobou testu shodnosti (dvouvýběrového t-testu). Jeho použití je vhodné v případech, kdy alespoň jeden z analyzovaných náhodných výběrů nesplňuje předpoklad normálního rozdělení.

Mann-Whitneyův test provedeme v programu Statgraphics následujícím postupem: Compare→Two Samples→ Two-Sample Comparison. V dialogovém okně Two-Sample Comparison zadáme do řádku Sample 1 název prvního datového souboru (výběru), do řádku Sample 2 název druhého datového souboru, v nabídce Input zatrhneme Two Data Columns a klikneme na OK. Zobrazí se nám výstup analýzy obsahující čtyři okna. Nyní klikneme levou myší na ikonu Tabular options (druhá ikona zleva v horní liště výstupu). Zobrazí se dialogové okno Tabular Options v němž zatrhneme Comparison of Medians. V levé části výstupu se zobrazí další okno s výsledkem Mann-Whitneyova testu (viz obrázek 6.10).

Příklad: V průběhu měsíce bylo měřeno množství srážek na dvou geolokalitách vzdálených od sebe pouze 5 km. Rozhodněte, zda jsou střední hodnoty množství srážek na geolokalitách shodné, nebo se liší?

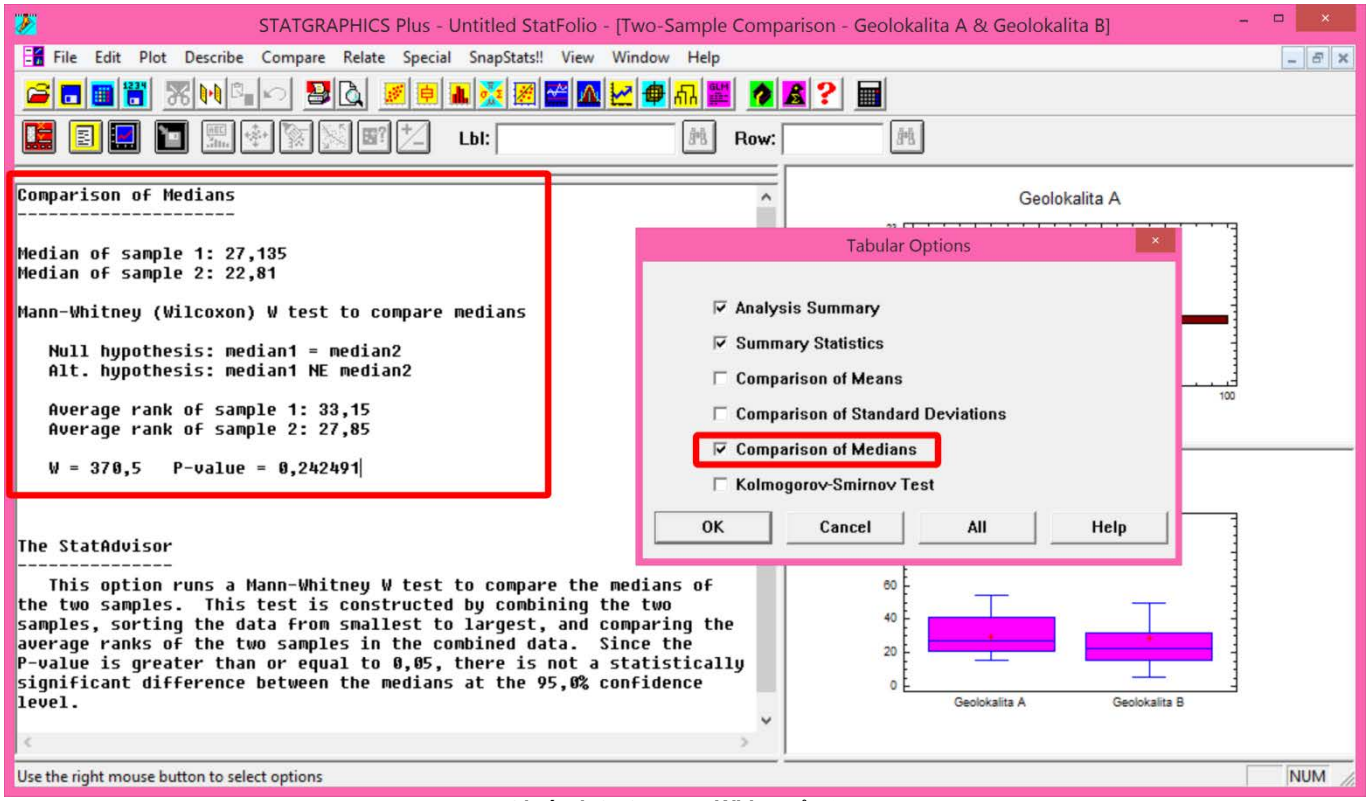
Výsledek Mann-Whitneyova testu:

```
Comparison of Medians
------------------------------
Median of sample 1: 27,135
Median of sample 2: 22,81

Mann-Whitney (Wilcoxon) W test to compare medians
Null hypothesis: median1 = median2
Alt. hypothesis: median1 NE median2
Average rank of sample 1: 33,15
Average rank of sample 2: 27,85
W = 370,5    P-value = 0,242491
```

Vyhodnocení testu: Z výsledku Mann-Whitneyova testu vyplývá, že vypočtená *p*-hodnota 0,242491 je větší než 0,05, což znamená, že přijímáme nulovou hypotézu H_0.

Závěr: Provedeným Mann-Whitneyovým testem bylo s 95% statistickou jistotou prokázáno, že střední hodnoty (mediány) množství srážek na geolokalitách jsou ve sledovaném období shodné.



Obrázek 6.10: Mann-Whitneyův test

6.2.3 Znaménkový test pro párová data

Tento test je neparametrickou obdobou klasického parametrického párového testu. Jeho užití je vhodné v případech, že máme závislé náhodné výběry a nejméně jeden z nich nesplňuje předpoklad normálního rozdělení.

Postup provedení v programu Statgraphics je stejný jako v případě parametrického párového testu, jež je uveden v odstavci 6.1.3 Párový test. Výsledek testu se nám zobrazí v levé části výstupu analýzy pod výsledkem t-testu (viz obrázek 6.11).

Příklad: Ve dvaceti vzorcích kofoly byl stanovován obsah kofeinu ve dvou různých laboratořích. Proveďte porovnání analýz pomocí znaménkového testu pro párová data a zjistěte, zda obě laboratoře vykazují stejné výsledky.

Výsledek testu:

```
sign test
------------
Null hypothesis: median = 0,0
Alternative: not equal

Number of values below hypothesized median: 9
Number of values above hypothesized median: 11
Large sample test statistic = 0,223607 (continuity correction applied)
P-Value = 0,823059

Do not reject the null hypothesis for alpha = 0,05
```

Vyhodnocení testu: Z výsledku znaménkového testu pro párová data vyplývá, že vypočtená *p*-hodnota 0,823059 je větší než 0,05, což znamená, že přijímáme nulovou hypotézu H_0.

Závěr: Provedeným testem bylo s 95% statistickou jistotou prokázáno, že obě laboratoře poskytují stejné výsledky.

Obrázek 6.11: Znaménkový test pro párová data

7 Regresní a korelační analýza

Náplní této kapitoly je osvětlení principů statistických metod používaných ve statistické analýze dat k identifikaci vzájemné závislosti dvou, případně více sledovaných proměnných. V rámci kapitoly jsou dále podrobně vysvětleny způsoby výpočtu a správné použití jednotlivých metod na experimentálních datech. Pro názornost a lepší pochopení problematiky regresní a korelační analýzy je kapitola doplněna názornými příklady, včetně jejich podrobné interpretace.

7.1 Regresní analýza

Regresní analýzou se rozumí metody zkoumání jednostranné závislosti nějaké veličiny (nikoliv nutně náhodné - tzv. závisle proměnné) na jiné veličině nebo jiných veličinách (tzv. nezávisle proměnných). Při zkoumání regresní závislosti se pozornost zaměřuje na nalezení (střední kvadratické) **regresní funkce**, která vystihuje průběh této závislosti, a výpočet **reziduálního rozptylu**, který vystihuje těsnost této závislosti.

Pro jednoduchost a z důvodu zaměření této publikace je následující text zaměřen převážně na **diskrétní** náhodné veličiny, které jsou pro praktické aplikace nejdůležitější.

7.1.1 Regresní funkce

Jsou-li **X** a **Y** dvě náhodné veličiny mající střední hodnoty, pak regresní funkcí náhodné veličiny **Y** na náhodné veličině **X** je podmíněná střední hodnota E(Y|X) jako funkce podmínky:

$$E(Y|X = x_i) = \frac{\sum_j y_j p_{ij}}{\sum_j p_{ij}}$$

kde $p_{ij} = P(X=x_i, Y=y_j)$ a

$$\sum_{ij} p_{ij} = 1$$

Regresní funkce mají důležitou *minimalizační vlastnost*. Platí pro ně totiž

$$E(Y - E(Y|X))^2 = \min_g E(Y - g(X))^2$$

kde se minimum bere přes všechny měřitelné funkce **g** jedné proměnné.

Jestliže náhodné veličiny X a Y jsou nezávislé, je E(Y|X) = EY, takže regresní funkce je konstantou. Jestliže sdružené rozložení náhodných veličin X a Y je normální, pak E(Y|X=x) = EY + β.(x-EX), kde

$$\beta = \frac{cov(X,Y)}{DX}$$

je tzv. **regresní koeficient Y na X**, takže regresní funkce je lineární. DX je rozptyl veličiny X, *cov* je kovariance.

7.1.2 Střední kvadratická regresní funkce

Regresní funkce mohou být značně složité. Proto se často aproximují jednoduššími funkcemi; přitom se vychází z minimalizační vlastnosti regresní funkce. Funkci **g**, která minimalizuje střední kvadratickou odchylku

S_g=E(Y-g(X))^2, hledáme nikoliv ve třídě všech měřitelných funkcí jedné proměnné, ale jen v nějaké podtřídě všech měřitelných funkcí jedné proměnné (např. mezi všemi lineárními funkcemi, polynomy apod).

Funkce získaná touto cestou se nazývá **střední kvadratická regresní funkce**. Není-li možná záměna se shora definovanou regresní funkcí, pak se přívlastek *střední kvadratická* vynechává.

7.1.3 Lineární střední kvadratická regresní funkce

Nejjednodušší střední kvadratická regresní funkce je **lineární střední kvadratická regresní funkce**. V tomto případě je obecný tvar minimalizační funkce g(x)=a.x+b, ovšem jen pro jedinou dvojici koeficientů [a,b] nabývá střední kvadratická odchylka S_g minima. Jinými slovy jen pro jedinou dvojici koeficientů [a,b] je střední hodnota E(Y-a.X-b)^2 minimální. Tato dvojice koeficientů určuje **nejlepší lineární odhad** náhodné veličiny Y prostřednictvím náhodné veličiny X. Proces nalezení zmíněných koeficientů se nazývá často **regrese přímkou** a používá se při něm **metody nejmenších čtverců**.

Je-li v tomto případě DX $\neq$ 0, pak koeficienty [a,b] jsou jednoznačně určeny a platí

$$b = \frac{\mathrm{cov}(X,Y)}{DX} \qquad a = EY - b \cdot EX$$

Je tedy vidět, že v případě dvourozměrného normálního rozložení je lineární střední kvadratická regresní funkce rovna regresní funkci, a koeficient b=β se nazývá **regresní koeficient Y na X**.

Jsou-li náhodné veličiny X a Y nezávislé, pak β = 0, takže nejlepším lineárním odhadem náhodné veličiny Y je její střední hodnota EY. Obráceně, může být β = 0 i když náhodné veličiny X a Y jsou závislé. V tomto případě může být absence náhodné veličiny X v odhadu způsobena předpokladem lineárnosti odhadu.

7.2 Metoda nejmenších čtverců

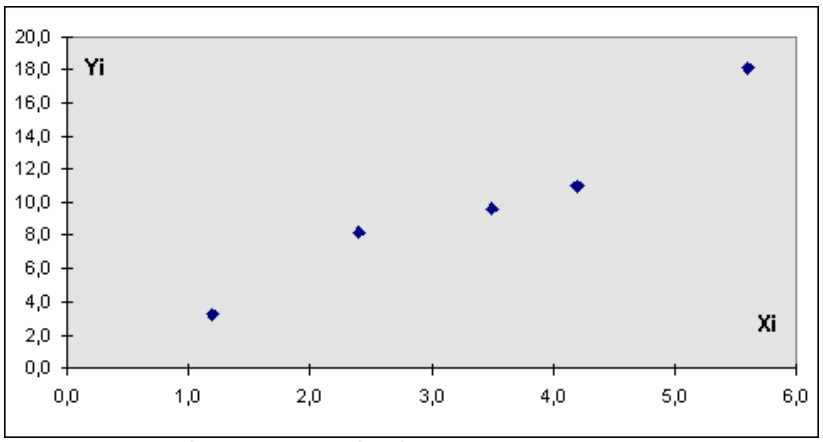
7.2.1 Princip metody

Pojmy zavedené v předchozím odstavci budeme aplikovat na konečnou množinu náhodných veličin X a Y. Nechť tedy je dána množina dvojic { [x_i, y_i] } pro i=1, 2, .., n. V praxi je nejčastěji reprezentována tabulkou, např.:

i	x_i	y_i
1	1,2	3,2
2	2,4	8,2
3	3,5	9,6
4	4,2	11,0
5	5,6	18,1

Velmi dobrou názornou pomůckou je geometrická interpretace. Dvojice [x_i,y_i] je možno chápat jako souřadnice bodů v rovině; pokud nebude moci dojít k záměně, budou dvojice dat nazývány *body* a lineární funkce y=a.x+b *přímkou*. Za těchto předpokladů lze data ze shora uvedené tabulky zobrazit v rovině známým způsobem (viz obrázek 7.1).

Zde je na místě uvést, jaký je nejčastější případ praktického použití popisovaného aparátu. Data jsou sice (z hlediska matematické teorie) náhodné veličiny; jsou však zároveň hodnotami, které nesou informaci o stavu nějakého procesu, objektu, fyzikální nebo chemické veličiny apod. Dvojice dat tedy nese informaci o současném stavu dvou objektů, veličin apod. Označíme-li tyto objekty, veličiny apod. X a Y, pak zápis dvojice [x_i, y_i] do tabulky vznikl jako záznam této skutečnosti: v okamžiku, kdy byla veličina X ve stavu x_i, byla veličina Y ve stavu y_i. Už v tom je implicitně vyjádřeno, že Y je pokládáno za veličinu závislou na veličině X: změní-li se X na nějakou hodnotu, změní se i Y na nějakou hodnotu. Tedy první závěr: vytvoří-li se např. shora uvedená tabulka, už v tom okamžiku se data pokládají za závislá.



Obrázek 7.1: Grafické znázornění dat z tabulky shora

Za druhé: závislost Y na X je

a) známa přesně nebo
b) známa na úrovni analytického předpisu nebo
c) neznáma.

V případě ad a) není z hlediska popisovaných metod co řešit. Je-li tou známou závislostí např. y=3.x+4, pak tabulka dat může sloužit např. ke zkoumání přesnosti měření apod.

Případ ad c) je častým úvodem řešení vědeckých nebo technických problémů; snahou je tento případ převést na případ uvedený pod bodem b) nalezením vhodného tvaru obecné funkce.

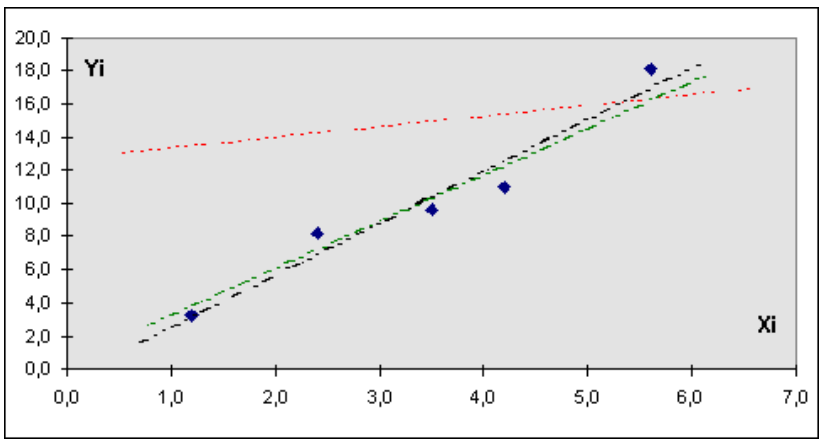
Případ ad b) je nejčastější a k jeho řešení směřuje tato kapitola. Je známa funkční závislost daná svým obecným funkčním předpisem (např. y=a.x+b). Pro některé hodnoty $\underline{x}$ byly zjištěny příslušné hodnoty $\underline{y}$. Sledovaný konkrétní děj se zcela jistě řídil nějakým konkrétním funkčním předpisem (např. y=3.x+4), ale v okamžiku zjišťování $[x_i, y_i]$ ještě skutečnost a=3 a b=4 nebyla známa. Tyto hodnoty jsou ale většinou cílem výzkumu.

Situace je tedy tato: ví se, že Y=f(X) a předpis pro $\underline{f}$ je obecně znám; ví se, že pro některá $\underline{x_j}$ byla naměřena nějaká $\underline{y_i}$; cílem je nalézt takové hodnoty koeficientů ve funkčním předpise $\underline{f}$, aby pro všechna $\underline{i}$ bylo $y_i=f(x_i)$.

Problém je však v tom, že v praxi téměř nikdy neplatí **přesně** $y_i=f(x_i)$, ale jen $y_i \approx f(x_i)$. Vždy dochází k nepřesnostem při získávání hodnot, např. díky nepřesnosti měřidla, nedokonalosti lidského vnímání, ale i díky nezaznamenávaným vlivům okolí apod. Proto je vhodné předchozí odstavec přeformulovat takto:

Ví se, že Y=f(X) a předpis pro $\underline{f}$ je obecně znám; ví se, že pro některá $\underline{x_j}$ byla naměřena nějaká $\underline{y_i}$; ví se, že toto měření nebylo zcela přesné; cílem je nalézt takové hodnoty koeficientů ve funkčním předpise $\underline{f}$, ze kterého - podrobíce se jistým chybám - původně y_i vzešly, tedy aby pro všechna $\underline{i}$ bylo y_i pokud možno co nejblíže $f(x_i)$.

Na dokreslení uveďme následující obrázek 7.2 a příklad pro lineární závislost:



Obrázek 7.2: Grafické znázornění možných lineárních závislostí

Ví se, že původní funkční závislostí byla lineární závislost (přímka) y=a.x+b. Přímek v rovině je však nekonečně mnoho; každá je jednoznačně dána konkrétní dvojicí koeficientů [a,b] (např. [3,4]). Z obrázku 7.2 je zřejmé, že vyznačené body nejsou "moc blízko" tečkované přímce. Obě čerchované přímky jsou na tom v tomto ohledu daleko lépe. Ale která z nich?

Především je nutno přesně definovat poněkud vágní pojem "být blízko" resp. "co nejblíže". Obecně je přesná definice podána v popisu minimalizační vlastnosti regresní funkce. Zde se zaměřme pouze na lineární závislost (přímku).

Jako míru přesnosti bývá zvykem chápat *odchylku* bodu od přímky. Za tuto odchylku je možno brát vzdálenost bodu od přímky - rozumí se běžnou, *kolmou* vzdálenost. To však by působilo při zjišťování koeficientů "nejlepší" přímky značné potíže: musely by se spouštět kolmice na neznámou přímku, neznámé koeficienty by byly (Pythagorova věta) pod odmocninou aj. Proto se v tomto případě chápla" odchylka - ve směru osy y, tj. hodnota y_i-f(x_i), pro přímku y_i-a.x_i-b.

Při posuzování, zda body jsou "dost blízko" funkci (např. přímky), se hodnotí ne odchylka každého bodu zvlášť, ale je nutno jakýmsi způsobem zohlednit všechny body najednou. Jako kriterium se nabízí *součet* odchylek všech bodů. Jsou-li body "hodně rozházené okolo" funkce (např. přímky), zdá se být součet jejich odchylek velký, jsou-li "málo rozházené", zdá se být menší. Pokud je však odchylka skutečně definována jako y_i-f(x_i) (pro přímku y_i-a.x_i-b), jsou některé odchylky kladné a některé záporné. Ve svém důsledku to znamená, že čtyři body mající odchylky po řadě (100, -100, 100, -100) perfektně vyhovují, protože součet jejich odchylek je nula - nejmenší možný!

Tato závada by šla obejít zavedením ne součtu odchylek, ale součtu *absolutních hodnot* odchylek. To by však při určování konkrétních hodnot koeficientů a, b přímky působilo obdobné problémy jako odmocnina u shora zmíněné kolmé odchylky. Proto se jako kriterium (jehož nejmenší hodnota se hledá) přijímá součet *kvadrátů* odchylek, tj. výraz

$$S = \sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2$$

7.2.2 Lineární regrese přímkou

Je-li funkcí f(x) lineární funkce y = a.x + b, pak má poslední vzorec předchozího odstavce tvar

$$S = \sum_{i=1}^{n} \left(y_i - a \cdot x_i - b\right)^2$$

Hodnota **S** se liší přímka od přímky; závisí na koeficientech a a b a v tomto smyslu je tedy **S** funkcí a, b:

$$S = S\,(a,b) = \sum_{i=1}^{n} \left(y_i - a \cdot x_i - b\right)^2$$

Základní úlohou je pak nalézt taková a_m, b_m, aby

$$S\,(a_m, b_m) = \min_{a,b \in R} S\,(a,b)$$

Hledání extrémů funkcí více proměnných je jednou z úloh diferenciálního počtu (viz); v našem případě se redukuje na řešení soustavy

$$\frac{\partial S}{\partial a} = 0$$
$$\frac{\partial S}{\partial b} = 0$$

Je tedy

$$\frac{\partial S}{\partial a} = \partial \frac{\sum (y_i - a \cdot x_i - b)^2}{\partial a} = \sum \partial \frac{(y_i - a \cdot x_i - b)^2}{\partial a} = \sum 2 \cdot (y_i - a \cdot x_i - b) \cdot (- x_i) = 0$$

$$\frac{\partial S}{\partial b} = \partial \frac{\sum (y_i - a \cdot x_i - b)^2}{\partial b} = \sum \partial \frac{(y_i - a \cdot x_i - b)^2}{\partial b} = \sum 2 \cdot (y_i - a \cdot x_i - b) \cdot (- 1) = 0$$

po vytknutí konstant

$$-2 \cdot \sum \left(x_i \cdot y_i - a \cdot x_i^2 - b \cdot x_i\right) = 0$$

$$-2 \cdot \sum \left(y_i - a \cdot x_i - b\right) = 0$$

a proto také

$$\sum \left(x_i \cdot y_i - a \cdot x_i^2 - b \cdot x_i\right) = 0$$

$$\sum \left(y_i - a \cdot x_i - b\right) = 0$$

po úpravě

$$\sum x_i \cdot y_i = \sum a \cdot x_i^2 + \sum b \cdot x_i$$

$$\sum y_i = \sum a \cdot x_i + \sum b$$

Sčítá se přes i (do n); na něm a ani b nezávisí, lze je tedy vytknout. Po prohození stran je

$$a \cdot \sum x_i^2 + b \cdot \sum x_i = \sum x_i \cdot y_i$$

$$a \cdot \sum x_i + b \cdot n = \sum y_i$$

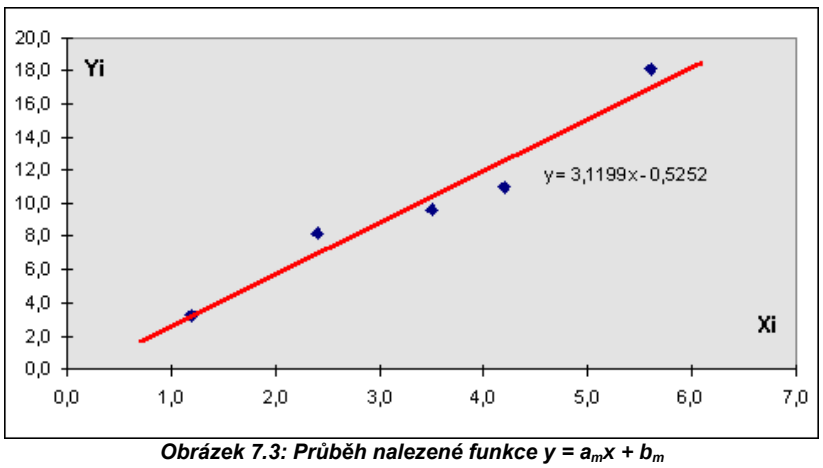
To je konečná tzv. **soustava normálních rovnic**. Protože všechna x_i i y_i jsou známa, jsou známy i všechny Σ a soustava normálních rovnic je tedy běžná soustava dvou lineárních rovnic o dvou neznámých a a b Řešením této soustavy (pokud existuje) jsou dvě hodnoty a_m a b_m takové, že S (a_m,b_m) = min.

Postup demonstrujme na datech ve shora uvedené tabulce. Je $\Sigma\ x_i = 16.9$, $\Sigma\ y_i = 50.1$, $\Sigma\ x_i^2 = 68.45$, $\Sigma\ x_i y_i = 204.68$. Protože bodů je 5, má soustava normálních rovnic tvar

$$68.45\ a_m + 16.9\ b_m = 204.68$$

$$16.9\ a_m + 5\ b_m = 50.1$$

Řešením je $a_m = 3.1199$, $b_m = -0.5252$. Průběh funkce $y = a_m x + b_m$ vzhledem k zadaným $[x_i,y_i]$ je na následujícím obrázku 7.3:



Obrázek 7.3: Průběh nalezené funkce $y = a_m x + b_m$

7.3 Korelační analýza

Pod pojmem korelace se obecně rozumí vzájemný vztah mezi dvěma veličinami. Mění-li se jedna, mění se korelativně i druhá. Ve statistice se v zúženém slova smyslu korelační analýza zabývá zkoumáním vzájemné *lineární* závislosti dvou analyzovaných proměnných. V předchozím odstavci byl uveden postup kvantifikace koeficientů takové lineární závislosti, což lze provést pro jakákoliv data (má-li soustava normálních rovnic řešení). Ovšem pro praktické využití nalezené závislosti je třeba posoudit, jak dobře nalezená závislost "vyhovuje" daným datům. Tato míra lineární korelační závislosti se obecně vyjadřuje prostřednictvím koeficientu determinace R^2. Koeficient determinace udává, jaká část variability náhodné proměnné *y* je způsobena závislostí na náhodné proměnné *x* a jaká část je způsobena náhodnými vlivy. V praxi se však k vyjádření těsnosti vzájemné lineární závislosti nejčastěji používá korelační koeficient obecně označovaný R (koeficient determinace je jeho druhou mocninou). U konkrétních koeficientů korelace se často místo označení R používá malé r.

Hodnota korelačního koeficientu se pohybuje v rozmezí -1 až 1. Čím více se hodnota korelačního koeficientu blíží číslu ±1 tím silnější je lineární závislost mezi analyzovanými proměnnými. Pokud se hodnota korelačního koeficientu rovná nule, jsou analyzované proměnné nekorelované. V tomto případě je však nutné si uvědomit, že mezi proměnnými pouze neexistuje lineární závislost, avšak muže mezi nimi existovat závislost nelineární (např. kvadratická). Pokud se hodnota korelačního koeficientu pohybuje v rozmezí -1 až 0, je korelace negativní (záporná). Negativní korelace znamená, že s rostoucí hodnotou sledované proměnné *x* klesá hodnota proměnné *y* (mezi proměnnými platí nepřímá závislost). Korelační koeficient, jehož hodnota se nachází v rozsahu 0 až +1 značí pozitivní (kladnou) korelaci. Pozitivní korelace znamená, že s rostoucí hodnotou proměnné *x* zároveň roste i hodnota proměnné *y* (mezi proměnnými platí přímá závislost). Mezi nejčastěji používané korelační koeficienty patří Pearsonův korelační koeficient a Spearmanův korelační koeficient.

Grafickým znázorněním lineární závislosti je přímka (y = a.x + b). Přímka je dána dvěma parametry a to směrnicí (a, angl. slope) a úsekem (b, angl. intercept, také y-intercept). Aby byly výsledky jakékoliv statistické analýzy statisticky věrohodné, je nutné mít na každý parametr minimálně pět experimentálních bodů. Proto v případě korelační analýzy je pro rigoróznost výsledku nutné mít pro obě proměnné minimálně deset experimentálních hodnot.

U korelační analýzy je důležité mít vždy na paměti, že korelace vyjadřuje pouze kvantitativní vztah mezi dvěma proměnnými, ale nevysvětluje příčinu (korelace neimplikuje kauzalitu). To znamená, že korelaci nelze vysvětlovat jako poznání reálných příčinných vztahů (Meloun a Militký, 2011)!

Po vyčíslení korelačního koeficientu před námi stojí rozhodnutí, zda je lineární závislost mezi sledovanými proměnnými významná, či není. Rozhodnutí o významnosti korelačního koeficientu můžeme provést několika způsoby. Prvním z nich je využití vypočtené *p*-hodnoty, kdy pomocí hypotézy testujeme, zda je korelace mezi proměnnými nulová. Pokud je výsledná *p*-hodnota větší než námi zvolená hladina významnosti α, přijímáme hypotézu H_0, která říká, že mezi proměnnými neexistuje žádný lineární vztah. V opačném případě, čili když vypočtená *p*-hodnota bude menší než zvolená hladina významnosti α, hypotézu H_0 zamítneme ve prospěch hypotézy alternativní H_A, která říká, že mezi proměnnými existuje významná lineární závislost. Další možností jak rozhodnout o významnosti korelačního koeficientu je srovnání s jeho kritickou tabelární hodnotou (kritické tabelární hodnoty pro Pearsonův a Spearmanův korelační koeficient jsou uvedeny v příloze 1 a 2). Korelaci mezi proměnnými považujeme za významnou tehdy, jestliže je vypočtená absolutní hodnota korelačního koeficientu vyšší než příslušná kritická tabelární hodnota. Třetí možností je posouzení významnosti korelace pomocí testu významnosti korelačního koeficientu. Testovací kritérium pro provedení testu vypočteme dle následující rovnice:

$$t_r = \frac{|r| \times \sqrt{n-2}}{\sqrt{1-r^2}}$$

Vypočtenou absolutní hodnotu t_r následně srovnáme s příslušnou hodnotou kvantilu studentova rozdělení $t_{\alpha,n-2}$ (kritické tabelární hodnoty Studentova t rozdělení jsou uvedeny v příloze 3). V případě, že vypočtená hodnota t_r je vyšší než tabelární hodnota kvantilu studentova rozdělení $t_{\alpha,n-2}$, zamítáme na zvolené hladině významnosti α hypotézu H_0, že sledované proměnné jsou nekorelované. Testem jsme tedy prokázali, že mezi sledovanými proměnnými existuje významná korelační závislost. Jestliže je hodnota t_r nižší než tabelární hodnota $t_{\alpha,n-2}$ hypotézu H_0 přijímáme, což znamená, že korelační závislost je nevýznamná.

7.3.1 Pearsonův korelační koeficient

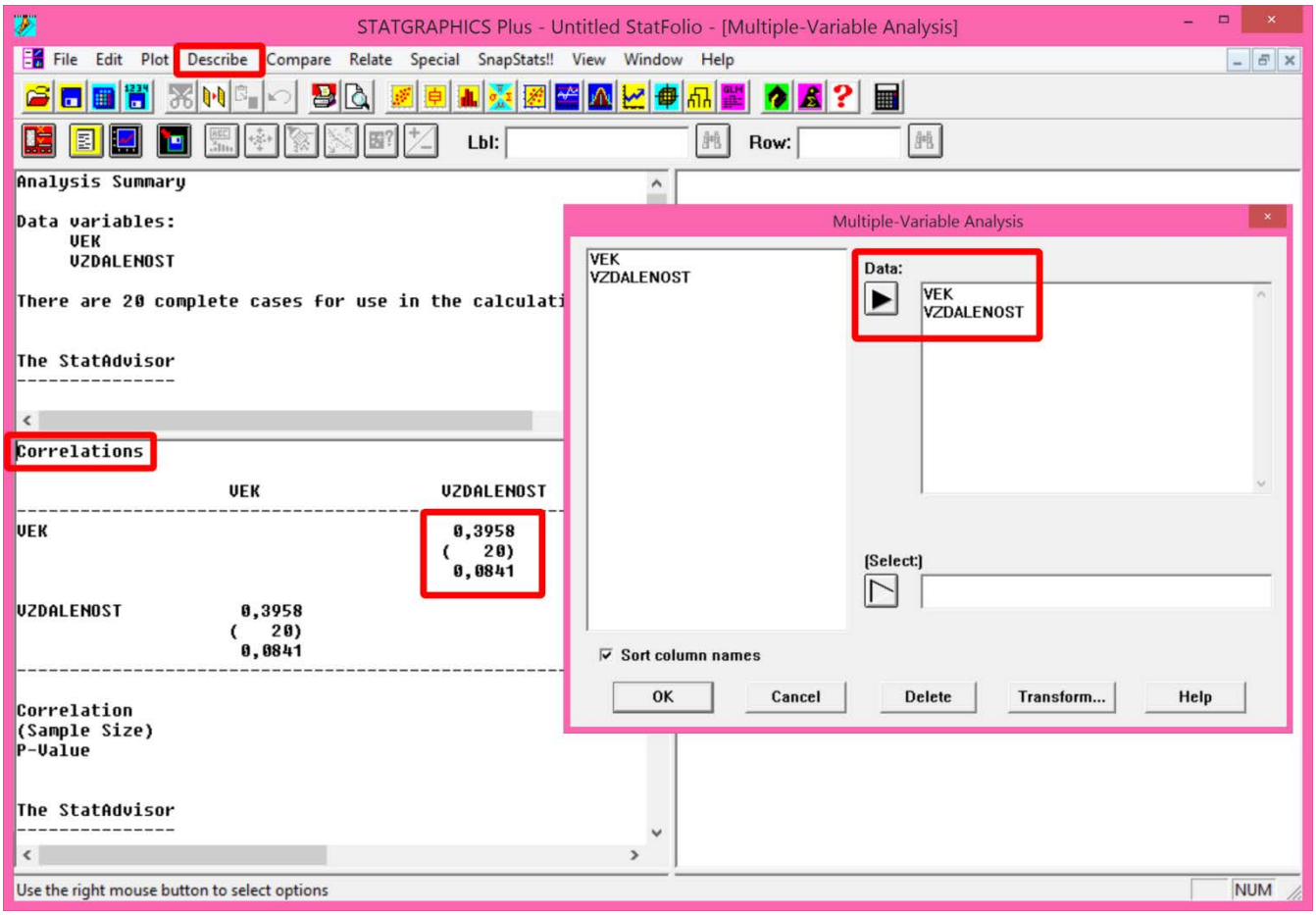
Pearsonův korelační koeficient, v literatuře někdy také označován jako výběrový korelační koeficient, je parametrickým párovým koeficientem. Jelikož je jeho výpočet založen na momentových odhadech míry polohy a variability (viz podkapitola 4.1 Klasické odhady míry polohy a variability), je předpokladem jeho použití normalita datových souborů bez odlehlých hodnot.

Pearsonův korelační koeficient lze vypočíst dle následující rovnice:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

kde $\bar{x}$ a $\bar{y}$ jsou střední hodnoty obou datových souborů.

V programu Statgraphics provedeme výpočet Pearsonova korelačního koeficientu následovně: Describe→Numeric Data→Multiple-Variable Analysis. V dialogovém okně Multiple-Variable Analysis zadáme do řádku data názvy dvou datových souborů proměnných, které chceme analyzovat a klikneme na OK. Zobrazí se nám výstup analýzy, přičemž výsledek korelační analýzy se nachází v levém dolním okně s názvem Correlations. Zde jsou pod sebou ve sloupci uvedeny tři hodnoty. První z nich představuje vypočtenou hodnotu Pearsonova korelačního koeficientu, druhá hodnota je uvedena v závorce a představuje počet párových hodnot a třetí je vypočtená *p*-hodnota (viz obrázek 7.4).



Obrázek 7.4: Pearsonův korelační koeficient - postup pro výpočet v programu Statgraphics

Hodnotu Pearsonova korelačního koeficientu je možné vypočíst i v tabulkovém procesoru Excel. V pracovním sešitě jej pro zvolená data vypočteme vložením funkce PEARSON, nebo vložením funkce CORREL (obě funkce poskytují shodné výsledky).

7.3.2 Spearmanův korelační koeficient

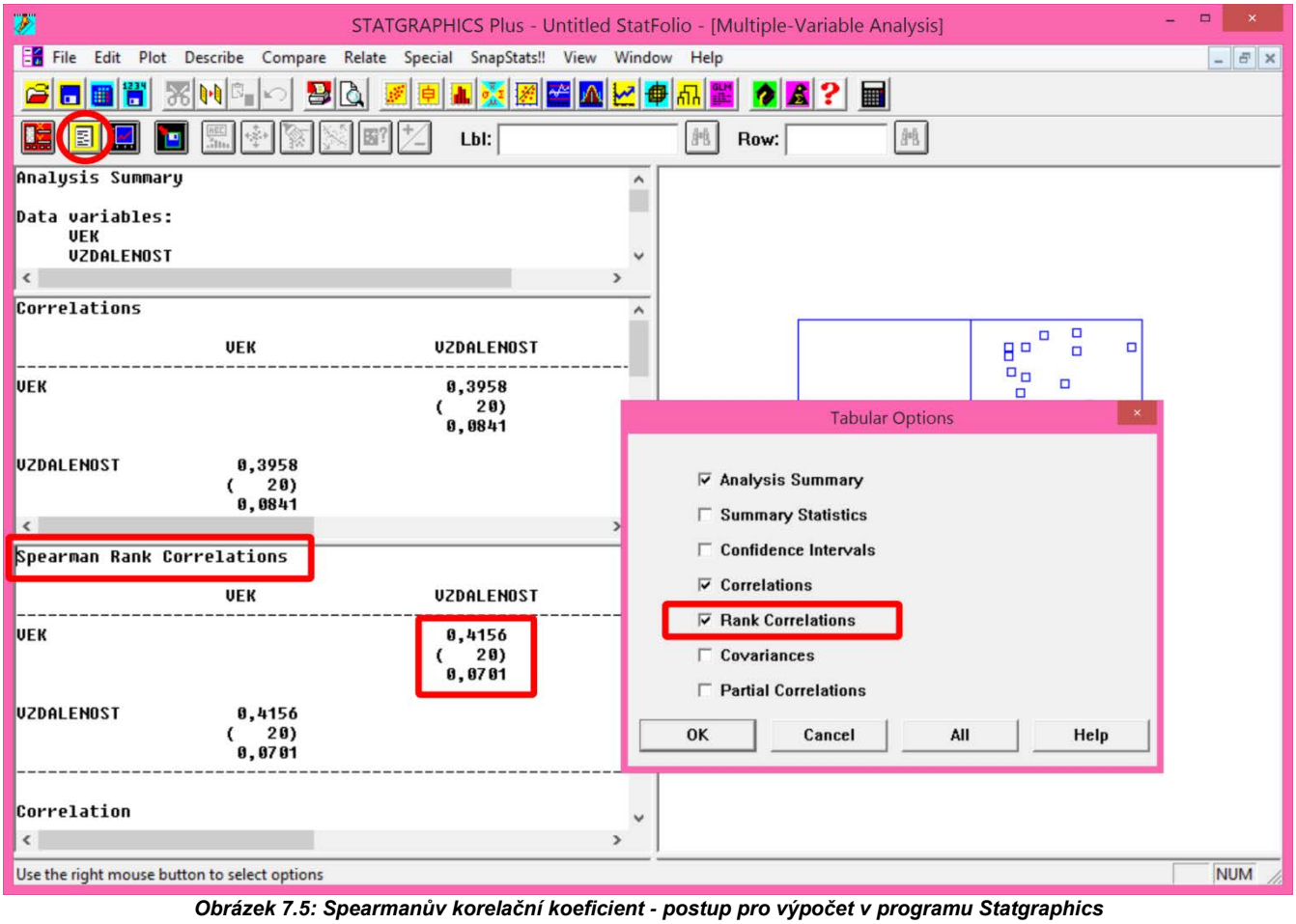
Spearmanův korelační koeficient je neparametrickým pořadovým koeficientem a označuje se symbolem r_s. Jeho výpočet není založen na výpočtech odhadů z experimentálních hodnot, ale na jejich pořadí (podobně jako v případě kvantilových charakteristik a neparametrických testů). Užití Spearmanova korelačního koeficientu je vhodné zejména v případech, kdy není splněn předpoklad normality datových souborů, nebo je prokázána přítomnost odlehlých hodnot.

Spearmanův korelační koeficient lze vypočíst dle následující rovnice:

$$r_s = \frac{\sum_{i=1}^{n} x_{ri} \cdot y_{ri} - n \cdot \bar{x}_r \bar{y}_r}{(n-1) \cdot s_{xr} \cdot s_{yr}}$$

kde x_{ri}, y_{ri} jsou pořadím hodnoty x_i a y_i v rámci vzestupně uspořádaných hodnot, $\bar{x}_r$ a $\bar{y}_r$ jsou aritmetické průměry hodnot pořadí proměnné x a y, s_{xr} a s_{yr} jsou výběrové směrodatné odchylky hodnot pořadí proměnné x a y.

Postup výpočtu v prostředí programu Statgraphics: provedeme stejné úkony jako v případě výpočtu Pearsonova korelačního koeficientu, jež je uveden v předchozí podkapitole. Po zobrazení výstupu analýzy klikneme levou myší na ikonu Tabular options a ve zobrazené nabídce zatrhneme příkaz Rank Correlations. V levé části výstupu analýzy se objeví nové okno s názvem Spearman Rank Correlations s výsledným Spearmanovým korelačním koeficientem (viz obrázek 7.5):



Obrázek 7.5: Spearmanův korelační koeficient - postup pro výpočet v programu Statgraphics

8 Řešené příklady statistického zpracování dat

Za účelem názorné ukázky praktického použití a možnosti procvičení výše uvedeného výkladu problematiky statistické analýzy je v této kapitole uvedena řada řešených příkladů včetně vstupních experimentálních dat a finálních interpretací výsledků.

8.1 Statistická analýza velkých výběrů

Zadání:

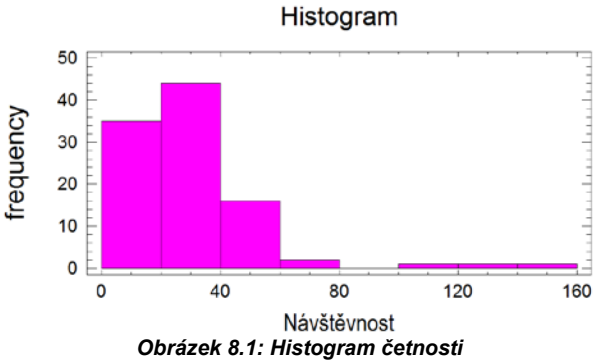
V průběhu let 2014 – 2016 byla sledována návštěvnost vybrané technické památky. Proveďte průzkumovou analýzu dat a stanovte nejlepší odhad polohy s následujícími parametry: n = 100, α = 0,05, veličina="Návštěvnost"

Návštěvnost				
26	14	38	25	52
14	22	14	38	60
17	25	19	25	27
26	41	25	19	51
22	20	29	20	19
32	41	12	21	48
144	26	14	20	14
32	15	9	136	15
33	16	35	108	21
38	20	25	24	52
21	11	24	37	25
14	42	15	27	21
22	29	21	9	62
25	15	20	30	46
21	15	9	33	16
14	29	20	41	19
11	7	30	31	63
33	14	21	51	55
30	32	27	53	36
46	16	57	55	11

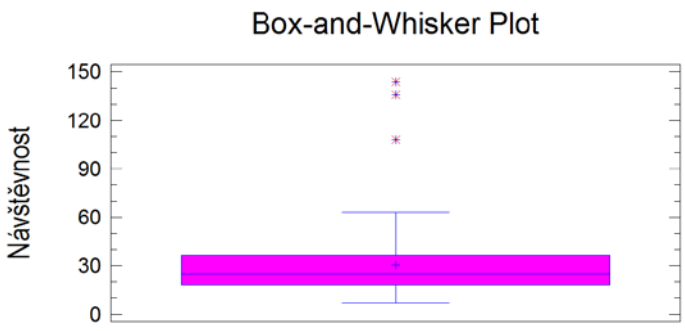
Řešení:

1. Průzkumová analýza dat

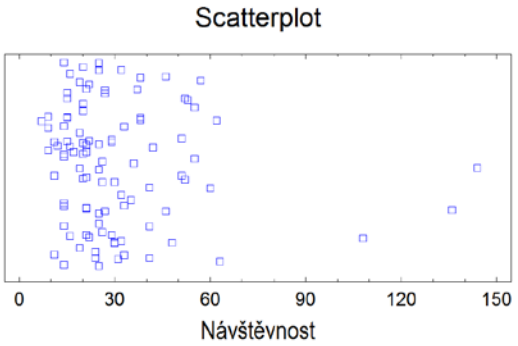
1.1. Diagnostika grafů:



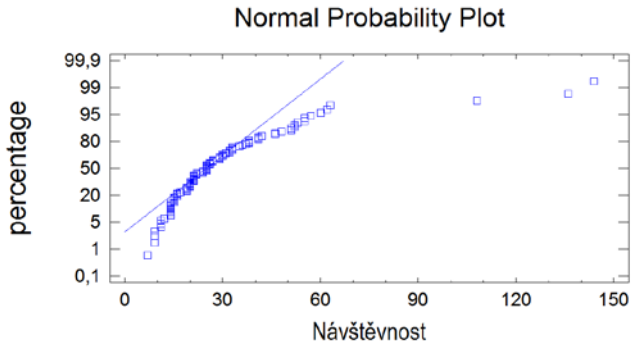
Obrázek 8.1: Histogram četnosti



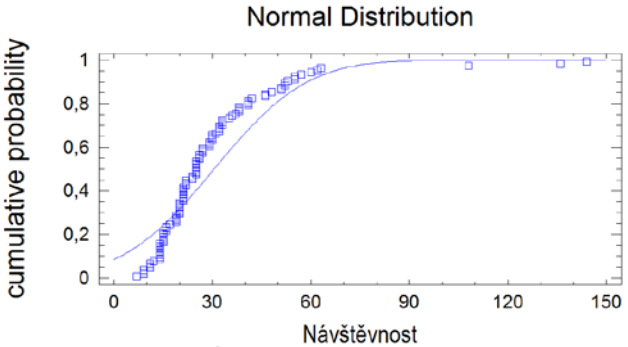
Obrázek 8.2: Krabicový graf



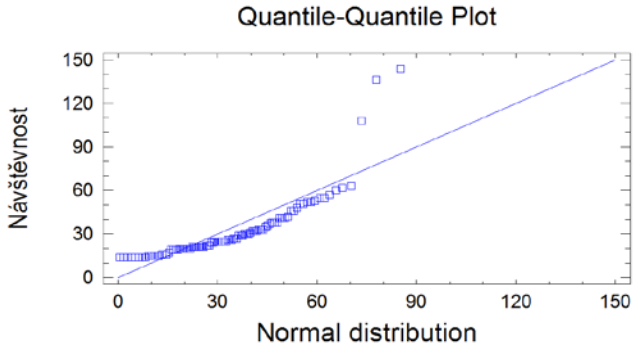
Obrázek 8.3: Jednorozměrný bodový graf



Obrázek 8.4: Normální pravděpodobnostní graf



Obrázek 8.5: Kvantilový graf



Obrázek 8.6: Kvantil-kvantilový graf

Z histogramu četnosti na obrázku 8.1 vyplývá, že data mají silně asymetrické rozdělení sešikmené k vyšším hodnotám, přičemž vpravo lze identifikovat tři odlehlé hodnoty. Krabicový graf na obrázku 8.2 potvrzuje výraznou asymetrii s kladným zešikmením k vyšším hodnotám. V horní části grafu je patrná přítomnost tří odlehlých hodnot. V jednorozměrném bodovém grafu na obrázku 8.3 je patrný mrak bodů se zahuštěním v levé části, což potvrzuje asymetrické rozdělení kladně sešikmené. Vpravo jsou přítomny tři odlehlé hodnoty. Na obrázku 8.4 body v normálním pravděpodobnostním grafu vykreslují konkávní křivku, což indikuje asymetrické rozdělení dat s kladným zešikmením. V pravé části grafu se nacházejí tři odlehlé hodnoty. V kvantilovém grafu (obrázek 8.5) se data teoretické křivce normálního rozdělení nepřimykají a jsou seřazeny do konkávního tvaru - soubor dat má jednoznačně asymetrické rozdělení s kladným zešikměním s třemi odlehlými hodnotami. V kvantil-kvantilovém grafu (obrázek 8.6) většina bodů neleží na přímce, což indikuje asymetrické rozdělení. Nahoře lze opět identifikovat přítomnost tří odlehlých hodnot.

1.2. Ověření normality:

Výstup

```
Skewness = 2,97301
Kurtosis = 11,7155
```

Hodnota koeficientu šikmosti je kladná a je významně vyšší než nula, což znamená, že datový soubor je rozdělen asymetricky s kladným zešikmením. Kladná hodnota koeficientu špičatosti indikuje rozdělení dat špičatější než je normální rozdělení.

Závěr: Z grafických diagnostik vyplývá, že soubor dat má výrazně asymetrické rozdělení sešikmené k vyšším hodnotám se třemi odlehlými body nahoře. Vzhledem k charakteru dat, nelze tyto odlehlé body vyloučit. Pro další statistické hodnocení bude zřejmě nutná transformace dat.

1.3. Statistické testy:

Testy nezávislosti – výstup:

```
Tests for Randomness of Cd
Runs above and below median
```

```
-----------------------------------
Median = 25,0
Number of runs above and below median = 41
Expected number of runs = 47,4946
Large sample test statistic z = -1,25019
P-value = 0,211231
Runs up and down
----------------------
Number of runs up and down = 63
Expected number of runs = 66,3333
Large sample test statistic z = -0,678158
P-value = 0,497669
Box-Pierce Test
-------------------
Test based on first 24 autocorrelations
Large sample test statistic = 17,9638
P-value = 0,804761
```

Závěr: Všechny tři uvedené testy mají vypočtenou *p*-hodnotu vyšší než 0,05, což znamená, že přijímáme nulovou hypotézu H_0 (H_0 = data jsou nezávislá). Provedenými testy bylo s 95% statistickou jistotou prokázáno, že data jsou nezávislá.

Testy normality – výstup:

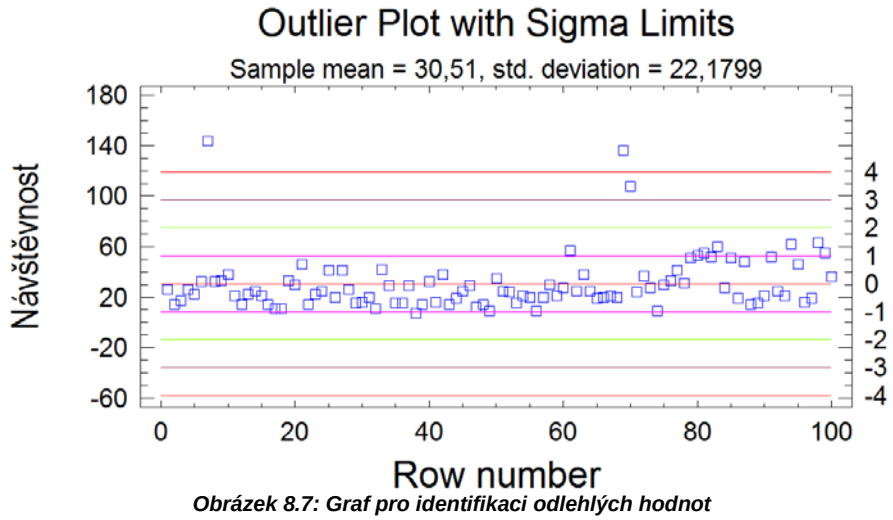
Vzhledem k velkému rozsahu datového souboru (n > 50) byl pro testování normality použit Anderson-Darlingův test (A-D test) a Chí-kvadrát test dobré shody (χ^2-test).

```
Anderson-Darling A^2: P-Value = 0,0000
Chi-Square Test: P-Value = 1,19507E-7
```

Závěr: Výsledná *p*-hodnota provedeného A-D testu 0,0000 je menší než 0,05. Zamítáme hypotézu H_0 ve prospěch hypotézy H_A. Provedením A-D testu bylo s 95% statistickou jistotou prokázáno, že datový soubor nelze považovat za výběr s normálním rozdělením. Z výsledku chí-kvadrát testu dobré shody vyplývá, že vypočtená *p*-hodnota 1,19507E-7 je menší než 0,05. Provedením chí-kvadrát testu bylo na hladině významnosti 0,05 prokázáno, že data nelze považovat za výběr z normálního rozdělení.

1.4. Identifikace odlehlých hodnot:

Identifikace odlehlých hodnot byla provedena pomocí diagnostiky grafu odlehlých hodnot a pomocí mediánových souřadnic.



Obrázek 8.7: Graf pro identifikaci odlehlých hodnot

Identifikace pomocí mediánových souřadnic – výstup:

Sorted Values

Row	Value	Studentized Values		Modified MAD Z-Score
		Without Deletion	With Deletion	
38	7,0	-1,05997	-1,07141	-1,349
56	9,0	-0,969796	-0,979342	-1,19911
74	9,0	-0,969796	-0,979342	-1,19911
49	9,0	-0,969796	-0,979342	-1,19911
17	11,0	-0,879624	-0,88752	-1,04922
...				
94	62,0	1,41975	1,44173	2,77294
98	63,0	1,46484	1,48853	2,84789
70	108,0	3,4937	3,75255	6,22039
69	136,0	4,7561	5,44993	8,31883
7	144,0	5,11679	6,00682	8,91839

V grafu odlehlých hodnot na obrázku 8.7 lze identifikovat tři odlehlé hodnoty, jejichž hodnota je vyšší než trojnásobek směrodatné odchylky datového souboru. Z výsledných hodnot mediánových souřadnic vyplývá, že hodnoty 144, 136 a 108 nacházející se na 7., 69. a 70. řádku datového souboru mají mediánovou souřadnici větší než tři (8,91839; 8.31883; 6.22039) a můžeme je tedy považovat za odlehlé hodnoty.

Závěr: Předpoklad o normalitě výběru byl zamítnut, což je ve shodě se závěry uvedenými v části 1.1) Diagnostika grafů a části 1.2) Ověření normality. Předpoklad o nezávislosti hodnot byl přijat. Grafem odlehlých hodnot a kritériem mediánových souřadnic byly identifikovány 3 odlehlé body, což je opět ve shodě se závěry grafických diagnostik. Vzhledem k charakteru dat nelze odlehlé body z další analýzy vyloučit. Jejich vyloučení by mohlo následně vést ke ztrátě důležité informace. Vzhledem k asymetrickému rozdělení dat a přítomnosti odlehlých hodnot se jeví jako nejvhodnější pro následné vyhodnocení studovaného souboru provedení transformace dat.

1.5. Transformace dat:

Pro transformaci dat byla použita logaritmická transformace pomocí přirozeného logaritmu xln = ln (x). Následně byla úspěšnost transformace ověřena grafickými diagnostikami, testy normality a testováním odlehlých hodnot.

Data po transformaci:

Transformovaná data xln				
3,258097	3,828641	2,772589	4,043051	4,007333
2,639057	2,639057	3,637586	3,218876	3,951244
2,833213	3,091042	2,639057	3,637586	4,094345
3,258097	3,218876	2,944439	3,218876	3,295837
3,091042	3,713572	3,218876	2,944439	3,931826
3,465736	2,995732	3,367296	2,995732	2,944439
4,969813	3,713572	2,484907	3,044522	3,871201
3,465736	3,258097	2,639057	2,995732	2,639057
3,496508	2,70805	2,197225	4,912655	2,70805
3,637586	2,772589	3,555348	4,682131	3,044522
3,044522	2,995732	3,218876	3,178054	3,951244
2,639057	2,397895	3,178054	3,610918	3,218876
3,091042	3,73767	2,70805	3,295837	3,044522
3,218876	3,367296	3,044522	2,197225	4,127134
3,044522	2,70805	2,995732	3,401197	3,828641
2,639057	2,70805	2,197225	3,496508	2,772589

Transformovaná data xln				
2,397895	3,367296	2,995732	3,713572	2,944439
2,397895	1,94591	3,401197	3,433987	4,143135
3,496508	2,639057	3,044522	3,931826	4,007333
3,401197	3,465736	3,295837	3,970292	3,583519

Transformovaná data byla opět podrobena grafické diagnostice. Z výsledných grafů vyplývá, že transformace dat vedla k zesymetričtění rozdělení a přiblížení k normalitě. Přestože lze v grafech stále identifikovat mírně odlehlé hodnoty, je zřejmé, transformací došlo k jejich výraznému přiblížení k ostatním hodnotám v datovém souboru. Hodnoty koeficientu šikmosti a špičatosti se po transformaci výrazně přiblížily nulové hodnotě (šikmost - 0,478559; špičatost - 0,652687). Testy normality (Anderson-Darlingův test a Chí-kvadrát test dobré shody) bylo s 95% statistickou jistotou prokázáno, že transformovaná data lze považovat za výběr z normálního rozdělení.

Závěr: Ze statistického hlediska bylo provedení transformace oprávněné a úspěšné.

2. Odhad míry polohy

Vzhledem k úspěšné transformaci je odhad polohy proveden pomocí retransformovaného průměru $\overline{x}_R$.

Postup:

- Nejprve vypočteme z transformovaných dat aritmetický průměr: $\overline{x} = 3{,}24326$
- Nyní tuto výslednou hodnotu pomocí zpětné transformace e^{xln} převedeme do původních hodnot, čímž vyčíslíme hodnotu retransformovaného průměru: $\overline{x}_R = 25{,}62 \approx 26$.

Souhrnný závěr: Průzkumovou analýzou dat bylo zjištěno, že soubor dat má výrazně asymetrické rozdělení kladně sešikmené k vyšším hodnotám se třemi odlehlými body nahoře. Ověřením předpokladů o výběru dat byl předpoklad o normalitě výběru zamítnut, což je ve shodě se závěry EDA. Předpoklad o nezávislosti hodnot byl přijat. Diagnostika grafu odlehlých hodnot a kritérium mediánových souřadnic určilo tři odlehlé body (horní), což je opět ve shodě se závěry EDA. Vzhledem k charakteru dat, nelze tyto odlehlé body vyloučit. Jejich vyloučení by mohlo následně vést ke ztrátě důležité informace. Bylo prokázáno, že transformace dat je nutná. Provedená logaritmická transformace byla úspěšná a poskytuje nejlepší odhad střední hodnoty a to $\overline{x}_R = 26$.

Pozn: *V případě, že by transformace dat úspěšná nebyla, bylo by nutné pro odhad střední hodnoty použít medián.*

8.2 Statistická analýza malých výběrů metodou Hornova postupu

Zadání:

V průběhu 7 dnů byly na vybrané geolokalitě sledovány celkové denní tržby stánku prodávajícího zmrzlinu. Určete parametr polohy a rozptýlení s použitím Hornovy metody pivotů s následujícími parametry: n = 7, veličina="Celkové denní tržby (Kč)". Výsledky porovnejte s klasickými a robustními odhady parametrů polohy a rozptýlení pomocí programu Statgraphics.

Celkové denní tržby (Kč)						
1240	810	949	1170	894	1904	1287

Řešení:

1. Seřazení dat vzestupně

Celkové denní tržby (Kč)						
810	894	949	1170	1203	1287	1904

2. Hornova metoda pivotů

Hloubka pivotu	$H = (int((n+1)/2))/2 = (int((7+1)/2))/2 = 2$
Dolní pivot	$X_D = X_{(H)} = X_{(2)} = 894$
Horní pivot	$X_H = X_{(n+1-H)} = X_{(7+1-2)} = X_{(6)} = 1287$
Pivotová polosuma	$P_L = (X_D + X_H)/2 = (894 + 1287)/2 \approx 1091$
Pivotové rozpětí	$R_L = X_H - X_D = 1287 - 894 = 393$

Závěr: Bodový odhad parametru polohy je přibližně 1091 Kč a parametru rozptýlení 393 Kč.

3. Klasické a robustní odhady parametru polohy a rozptýlení

Výstup:

```
Summary Statistics for denní tržby
Count = 7
Average = 1173,86
Median = 1170,00
Standard deviation = 366,98
Interquartile range= 393,00
```

Závěr: Klasické odhady parametru polohy a rozptýlení činí: aritmetický průměr je přibližně 1174 Kč se směrodatnou odchylkou 367 Kč. Robustní odhad těchto parametrů: medián 1170 Kč a interkvartilové rozpětí 393 Kč.

Souhrnný závěr: Při porovnání výsledků získaných Hornovou metodou pivotů s klasickými a robustními odhady bylo zjištěno, že tyto odhady parametru polohy poskytují mírně nadhodnocené výsledky. Je zřejmé, že v případě malých výběrů je nejvhodnější aplikace Hornovy metody pivotů.

8.3 Test správnosti

Zadání:

Na sledované geolokalitě stánkový prodejce párků deklaruje obsah soli v tomto masném výrobku 2,16 % se směrodatnou odchylkou 0,1 %. V rámci kontroly bylo náhodně odebráno 36 vzorků prodávaných párků. Rozhodněte, zda jsou prodávané párky kvalitní a nepřekračují deklarovaný obsah soli. Parametry: n = 36, veličina="Obsah soli (%)"

Obsah soli (%)		
2,26	1,84	2,12
2,08	1,94	2,22
2,13	1,89	2,22
2,15	2,24	1,97
2,01	2,07	2,02
2,22	1,97	2,21
2,16	2,24	2,31
2,15	1,97	2,28
2,01	2,26	2,19
1,96	2,30	2,13
2,09	2,29	2,05
1,83	2,06	2,12

Řešení:

1. Průzkumová analýza dat

Průzkumovou analýzou dat bylo zjištěno, že data mají normální rozdělení bez odlehlých hodnot. Nezávislost výběru byla přijata. Transformace dat není nutná.

2. Odhad střední hodnoty datového souboru

Jelikož datový soubor splňuje předpoklad normality, byla střední hodnota vypočítána pomocí aritmetického průměru $\bar{x}$ = 2,11 %.

3. Test správnosti při známém rozptylu

Formulace hypotézy H_0:	$H_0 \equiv \mu = \mu_0$;
Formulace hypotézy H_A:	$H_A \equiv \mu > \mu_0$
Nulová hypotéza:	2,16
Střední hodnota datového souboru:	2,11
Zvolená hodnota směrodatné odchylky:	0,1
Počet hodnot v datovém souboru:	36

Chceme zjistit, jestli obsah soli nepřekračuje deklarovanou hodnotu => pro testování zvolíme jednostranný test pravostrannou alternativu (Greater Than).

Výstup testu:

```
Hypothesis Tests
----------------
Sample mean = 2,11
Sample standard deviation = 0,1
Sample size = 36
95,0% lower confidence bound for mean: 2,11 - 0,0281596   [2,08184]
Null Hypothesis: mean = 2,16
Alternative: greater than
Computed t statistic = -3,0
P-Value = 0,997525
Do not reject the null hypothesis for alpha = 0,05.
```

Závěr: Z výsledku t-testu vyplývá, že vypočtená *p*-hodnota 0,997525 je větší než 0,05, což znamená, že přijímáme nulovou hypotézu H_0.

Souhrnný závěr: Na základě provedeného testu bylo s 95% statistickou jistotou prokázáno, že obsah soli v prodávaných párcích nepřekračuje hodnotu deklarovanou prodejcem.

8.4 Test shodnosti

Zadání:

V rámci dotazníkového průzkumu byl zjišťován věk padesáti návštěvníků na geolokalitě A a věk padesáti návštěvníků na geolokalitě B. Rozhodněte, zda je střední hodnota věku návštěvníků na geolokalitě A shodná

se střední hodnotou věku návštěvníků na geolokalitě B. Parametry: n_1 = 50, n_2 = 50, α = 0,05, veličina="Věk návštěvníků (rok)"

Geolokalita A

Věk návštěvníků A				
32	55	44	53	34
25	41	39	17	21
44	42	29	60	44
42	40	53	15	34
36	39	42	53	48
26	41	55	46	30
27	25	41	34	48
55	30	40	35	58
25	31	18	42	57
41	65	29	42	54

Geolokalita B

Věk návštěvníků B				
42	35	45	47	34
29	44	39	39	30
35	67	51	35	36
20	55	27	37	45
62	22	30	32	50
48	19	46	16	22
52	43	48	48	17
59	36	48	53	49
45	39	22	37	41
34	22	35	59	45

Řešení:

1. Průzkumová analýza dat

Soubor A: Z provedené průzkumové analýzy vyplývá, že data jsou nezávislá, homogenní, bez odlehlých hodnot. Předpoklad normality byl přijat. Z porovnání rozdělení vyplývá, že data mají normální rozdělení. Transformace dat není nutná.

Soubor B: Z provedené průzkumové analýzy vyplývá, že data jsou nezávislá, homogenní, bez odlehlých hodnot. Předpoklad normality byl přijat. Z porovnání rozdělení vyplývá, že data mají normální rozdělení. Transformace dat není nutná.

2. Test shody rozptylů

Výstup:

```
F-test to Compare Standard Deviations:
Null hypothesis: sigma1 = sigma2
Alt. hypothesis: sigma1 NE sigma2
F = 0,968034   P-value = 0,909929
```

Závěr: Z výsledku F-testu vyplývá, že vypočtená *p*-hodnota 0,909929 je větší než 0,05, což znamená, že přijímáme nulovou hypotézu H_0. Testováním bylo s 95% statistickou jistotou zjištěno, že rozptyly se považují za shodné (homoskedasticita).

3. Test shodnosti

Výstup:

```
t test to compare means:
Null hypothesis: mean1 = mean2
Alt. hypothesis: mean1 NE mean2
assuming equal variances: t = 0,0497249 P-value = 0,960443
```

Závěr: Na základě zjištění, že oba výběry mají shodný rozptyl (vykazují homoskedasticitu), byl pro test shodnosti použit dvouvýběrový t-test pro shodné rozptyly. Z výsledku t-testu vyplývá, že vypočtená *p*-hodnota 0,960443 je větší než 0,05, což znamená, že přijímáme nulovou hypotézu H_0.

Souhrnný závěr: Testem shodnosti bylo prokázáno, že na 95% hladině významnosti neexistuje statisticky významný rozdíl mezi středními hodnotami věku návštěvníků na geolokalitě A a geolokalitě B. Střední hodnoty věku návštěvníků na geolokalitách tedy lze považovat za shodné.

8.5 Párový test

Zadání:

Za účelem zhodnocení vlivu nově poskytovaných služeb byla sledována návštěvnost dané geolokality. Návštěvnost byla sledována nejprve 16 dnů před zavedením služeb a následně 16 dnů po jejich zavedení. Rozhodněte, zda zavedení nových služeb mělo významný vliv na návštěvnost geolokality (porovnáváme návštěvnost před a po zavedení nově poskytovaných služeb). Parametry: $n_1 = 16$, $n_2 = 16$, $\alpha = 0{,}05$, veličina="Návštěvnost"

Návštěvnost před změnami							
41	45	44	48	37	32	44	47
29	37	16	28	43	48	58	49

Návštěvnost po změnách							
69	65	76	52	60	85	64	60
54	87	70	62	72	61	71	69

Řešení:

1. Průzkumová analýza dat

Soubor dat před změnami: Z provedené průzkumové analýzy vyplývá, že data jsou nezávislá bez odlehlých hodnot. Předpoklad normality byl přijat.

Soubor dat po změnách: Z provedené průzkumové analýzy vyplývá, že data jsou nezávislá bez odlehlých hodnot. Předpoklad normality byl přijat.

2. Párový test

Výstup:

```
t-test
------
Null hypothesis: mean = 0,0
Alternative: not equal

Computed t statistic = -7,26788
```

```
P-Value = 0,00000275003
Reject the null hypothesis for alpha = 0,05.
```

Závěr: Z výsledku párového testu (jednovýběrového t-testu) vyplývá, že vypočtená *p*-hodnota 0,00000275003 je menší než 0,05, což znamená, že zamítáme nulovou hypotézu H_0 ve prospěch alternativní hypotézy H_A.

Souhrnný závěr: Párový test zamítl na hladině významnosti α = 0,05 hypotézu o shodě návštěvnosti před a po zavedení nových služeb. Z výsledků vyplývá, že zavedení nově poskytovaných služeb mělo významný vliv na výslednou návštěvnost dané geolokality.

8.6 Jednovýběrový znaménkový test

Zadání:

Na sledované geolokalitě má stánkový prodejce piva povoleno prodávat pouze desetistupňové pivo (10°) s obsahem alkoholu max. 4,3 %. V rámci kontroly bylo odebráno 30 vzorků prodávaného piva. Rozhodněte, zda střední hodnota alkoholu v odebraných vzorcích splňuje výše uvedený limit. Parametry: n = 30, veličina="Obsah alkoholu (%)"

Obsah alkoholu (%)		
4,3	5,6	2,3
5,2	2,3	2,3
2,4	1,7	1,6
3,4	5,3	3,2
2,7	5,0	1,8
3,4	2,7	1,1
5,5	1,4	2,0
4,1	1,1	1,8
3,5	1,7	2,7
1,1	1,9	3,6

Řešení:

1. Průzkumová analýza dat

Průzkumovou analýzou dat bylo zjištěno, že data jsou nezávislá bez odlehlých hodnot. Data nemají normální rozdělení.

2. Jednovýběrový znaménkový test

Formulace hypotéz: $H_0 : \mu = \mu_0$; $H_A : \mu > \mu_0$

Nulová hypotéza: 4,3

Výstup:

```
sign test
---------
Null hypothesis: median = 4,3
Alternative: greater than
Number of values below hypothesized median: 17,5
Number of values above hypothesized median: 8,8
Large sample test statistic = -3,88896 (continuity correction applied)
P-Value = 0,99995
```

```
Do not reject the null hypothesis for alpha = 0,05.
```

Závěr: Z výsledku jednovýběrového znaménkového testu vyplývá, že vypočtená *p*-hodnota 0,99995 je větší než 0,05, což znamená, že přijímáme nulovou hypotézu H_0.

Souhrnný závěr: Na základě provedeného testu bylo s 95% statistickou jistotou prokázáno, že střední hodnota alkoholu v odebraných vzorcích prodávaného desetistupňového piva splňuje povolený limit.

8.7 Mann-Whitneyův test

Zadání:

V zájmové oblasti byla na dvou geolokalitách sledována prodejnost suvenýrů. Prodejnost byla sledována v průběhu jednoho měsíce, přičemž vždy na konci dne byla zaznamenána výše tržeb. Rozhodněte, zda je střední hodnota výše tržeb za prodané suvenýry na geolokalitě A shodná se střední hodnotou výše tržeb na geolokalitě B. Parametry: $n_1 = 30$, $n_2 = 30$, $\alpha = 0,05$, veličina="Prodejnost suvenýrů (tržby v Kč)"

Geolokalita A

Prodejnost suvenýrů (Kč)		
2422	1498	1386
1722	1498	980
1414	1554	714
658	1302	1806
644	1176	896
812	966	924
1316	434	1274
1092	994	882
1694	588	644
1708	854	1204

Geolokalita B

Prodejnost suvenýrů (Kč)		
700	406	70
504	812	140
588	168	182
868	224	322
714	448	630
826	336	882
504	84	532
406	168	1050
462	266	168
336	168	140

Řešení:

1. Průzkumová analýza dat

Soubor A: Z provedené průzkumové analýzy vyplývá, že data jsou nezávislá, bez odlehlých hodnot. Předpoklad normality byl přijat.

Soubor B: Z provedené průzkumové analýzy vyplývá, že data jsou nezávislá, bez odlehlých hodnot. Data nesplňují předpoklad normality.

2. Mann-Whitneyův test

Výstup:

```
Comparison of Medians
---------------------
Median of sample 1: 1134,0
Median of sample 2: 406,0
Mann-Whitney (Wilcoxon) W test to compare medians
Null hypothesis: median1 = median2
Alt. hypothesis: median1 NE median2
Average rank of sample 1: 43,4
Average rank of sample 2: 17,6
W = 63,0   P-value = 1,09528E-8
```

Závěr: Z výsledku Mann-Whitney W testu vyplývá, že vypočtená *p*-hodnota 1,09528E-8 je významně menší než 0,05, což znamená, že zamítáme nulovou hypotézu H_0 ve prospěch alternativní hypotézy H_A.

Souhrnný závěr: Mann-Whitneyovým testem bylo prokázáno, že na 95% hladině významnosti existuje statisticky významný rozdíl mezi středními hodnotami výše tržeb na geolokalitě A a geolokalitě B. Z výsledku testu je zřejmé, že prodejnost suvenýrů na dvou sledovaných geolokalitách je významně rozdílná.

8.8 Znaménkový test pro párová data

Zadání:

Za účelem zhodnocení vlivu rekonstrukce odpočinkových zón na dané geolokalitě byla sledována spokojenost návštěvníků. Návštěvníci byli dotazováni nejprve 20 dnů před rekonstrukcí a následně 20 dnů po rekonstrukci. Rozhodněte, zda rekonstrukce odpočinkových zón měla významný vliv na spokojenost návštěvníků (porovnáváme procentuální podíl kladných odpovědí v jednotlivých dnech před a po rekonstrukci). Parametry: $n_1 = 20$, $n_2 = 20$, $\alpha = 0{,}05$, veličina="Spokojenost (%)"

Spokojenost před rekonstrukcí (%)									
39	40	46	46	45	46	46	43	43	45
44	44	43	47	45	44	43	44	41	12

Spokojenost po rekonstrukci (%)									
71	74	71	89	74	65	70	76	69	70
74	75	73	72	75	66	72	73	70	73

Řešení:

1. Průzkumová analýza dat

Datový soubor – spokojenost před rekonstrukcí: Z provedené průzkumové analýzy vyplývá, že data jsou nezávislá s jednou odlehlou hodnotou. Data nesplňují předpoklad normality.

Datový soubor – spokojenost po rekonstrukci: Z provedené průzkumové analýzy vyplývá, že data jsou nezávislá s jednou odlehlou hodnotou. Data nesplňují předpoklad normality.

2. Znaménkový test pro párová data

Výstup:

```
sign test
---------
Null hypothesis: median = 0,0
Alternative: not equal
```

```
Number of values below hypothesized median: 20
Number of values above hypothesized median: 0
Large sample test statistic = 4,24853 (continuity correction applied)
P-Value = 0,0000215322
Reject the null hypothesis for alpha = 0,05.
```

Závěr: Z výsledku znaménkového testu pro párová data vyplývá, že vypočtená *p*-hodnota 0,0000215322 je významně menší než 0,05, což znamená, že zamítáme nulovou hypotézu H_0 ve prospěch alternativní hypotézy H_A.

Souhrnný závěr: Znaménkovým testem pro párová data bylo prokázáno, že na 95% hladině významnosti existuje statisticky významný rozdíl mezi spokojeností návštěvníků před a po rekonstrukci odpočinkových zón. Je zřejmé, že provedená rekonstrukce měla významný vliv na spokojenost návštěvníků.

8.9 Pearsonův korelační koeficient

Zadání:

V rámci dotazníkového průzkumu byl sledován věk návštěvníků dané geolokality. Návštěvníci byli dále dotazováni, jakou délku turistické trasy preferují (km). Určete Pearsonův korelační koeficient a testujte na hladině významnosti α = 0,05, zda mezi oběma proměnnými existuje statisticky významná korelace (lineární závislost). Parametry: n_1 = 30, n_2 = 30, α = 0,05, veličiny="Věk (rok)", "Délka trasy (km)"

Věk (rok)									
39	58	42	36	38	58	29	51	27	55
53	20	23	62	44	48	46	49	49	40
41	42	50	36	38	36	33	17	42	51

Délka trasy (km)									
5,5	3,5	9,0	8,5	5,0	2,0	9,0	3,0	9,0	2,5
1,0	13,0	12,0	3,5	6,0	4,0	4,5	3,5	6,5	7,5
5,5	8,5	4,5	9,5	5,0	10,0	12,0	14,0	8,0	7,0

Řešení:

1. Průzkumová analýza dat

Soubor dat – věk: Z provedené průzkumové analýzy vyplývá, že data jsou nezávislá, bez odlehlých hodnot. Data splňují předpoklad normality.

Soubor dat – délka trasy: Z provedené průzkumové analýzy vyplývá, že data jsou nezávislá, bez odlehlých hodnot. Data splňují předpoklad normality.

2. Výpočet Pearsonova korelačního koeficientu

Výstup:

```
Correlations
                    Délka trasy         Věk
                                        0,8734
Délka trasy                             (30)
                                        0,0000
```

		0,8734
Věk		(30)
		0,0000

Závěr: Hodnota Pearsonova korelačního koeficientu mezi věkem návštěvníků a preferovanou délkou turistické trasy je 0,8734. Analýza byla provedena pro 30 experimentálních párových dat. Vypočtená p-hodnota je 0,0000.

3. Významnost vypočteného korelačního koeficientu

3.1. Dle vypočtené *p*-hodnoty:

Z výsledku korelační analýzy vyplývá, že vypočtená *p*-hodnota 0,0000 je významně menší než 0,05. Zamítáme nulovou hypotézu H_0 ve prospěch alternativní hypotézy H_A, což znamená, že mezi proměnnými existuje významná pozitivní korelace (hodnota korelačního koeficientu je v rozsahu 0 až +1).

3.2. Srovnáním s kritickou tabelární hodnotou Pearsonova korelačního koeficientu:

Vypočtená hodnota r = 0,8734. Kritická tabelární hodnota pro 30 hodnot je 0,361. Jelikož je vypočtená absolutní hodnota korelačního koeficientu vyšší než tabelární kritická hodnota, můžeme s 95% statistickou jistotou tvrdit, že mezi proměnnými existuje významná pozitivní korelace.

3.3. Testem významnosti korelačního koeficientu:

$$t_r = \frac{|0{,}8734|\sqrt{30-2}}{\sqrt{1-0{,}8734^2}} \approx 9{,}490$$

Vypočtená hodnota $t_r \approx 9{,}490$. Kritická tabelární hodnota Studentova rozdělení $t_{(\alpha = 0{,}05,\ n-2 = 28)} = 2{,}048$. Jelikož je vypočtená hodnota t_r vyšší než kritická tabelární hodnota Studentova rozdělení, zamítáme na hladině významnosti $\alpha = 0{,}05$ hypotézu H_0, že sledované proměnné jsou nekorelované.

Souhrnný závěr: Provedenou průzkumovou analýzou bylo zjištěno, že oba datové soubory analyzovaných proměnných splňují předpoklad normality a v datech nejsou přítomny odlehlé hodnoty. Na základě těchto výsledků byl pro výpočet použit Pearsonův korelační koeficient, jehož hodnota je 0,8734. Významnost korelačního koeficientu byla posouzena třemi metodami. Z výše uvedených výsledků vyplývá, že v rámci sledované geolokality existuje významná pozitivní korelace mezi věkem návštěvníků a preferovanou délkou turistické trasy.

8.10 Spearmanův korelační koeficient

Zadání:

V období 10. – 20. srpna 2016 byla sledována denní návštěvnost geolokality a množství spadlých srážek (mm/den). Určete Spearmanův korelační koeficient a testujte na hladině významnosti α = 0,05, zda mezi oběma proměnnými existuje statisticky významná korelace (lineární závislost). Parametry: $n_1 = 10$, $n_2 = 10$, α = 0,05, veličiny="Návštěvnost", "Množství srážek (mm/den)"

Návštěvnost									
248	210	173	322	497	370	285	68	136	24

Množství srážek (mm/den)									
7,4	8,9	12	1,2	0	0	6,3	27,4	13,8	54,6

Řešení:

1. Průzkumová analýza dat

Soubor dat – návštěvnost: Z provedené průzkumové analýzy vyplývá, že data jsou nezávislá, bez odlehlých hodnot. Data splňují předpoklad normality.

Soubor dat – množství srážek: Z provedené průzkumové analýzy vyplývá, že data jsou nezávislá s jednou odlehlou hodnotou. Data nesplňují předpoklad normality.

2. Výpočet Spearmanova korelačního koeficientu

Výstup:

Spearman Rank Correlations

	návštěvnost	srážky
návštěvnost		-0,9970 (10) 0,0028
srážky	-0,9970 (10) 0,0028	

Závěr: Hodnota Spearmanova korelačního koeficientu mezi návštěvností a množstvím spadlých srážek je -0,9970. Analýza byla provedena pro 10 experimentálních párových dat. Vypočtená p-hodnota je 0,0028.

3. Významnost vypočteného korelačního koeficientu

3.1. Dle vypočtené *p*-hodnoty:

Z výsledku korelační analýzy vyplývá, že vypočtená *p*-hodnota 0,0028 je významně menší než 0,05. Zamítáme nulovou hypotézu H_0 ve prospěch alternativní hypotézy H_A, což znamená, že mezi proměnnými existuje významná negativní korelace (hodnota korelačního koeficientu je v rozsahu -1 až 0).

3.2. Srovnáním s kritickou tabelární hodnotou Spearmanova korelačního koeficientu:

Vypočtená hodnota r_s = -0,9970. Kritická tabelární hodnota pro 10 hodnot je 0,648. Jelikož je vypočtená absolutní hodnota korelačního koeficientu vyšší než tabelární kritická hodnota, můžeme s 95% statistickou jistotou tvrdit, že mezi proměnnými existuje významná negativní korelace.

3.3. Testem významnosti korelačního koeficientu:

$$t_r = \frac{|-0,9970|\sqrt{10-2}}{\sqrt{1-(-0,9970)^2}} \approx 36,433$$

Vypočtená hodnota $t_r \approx 36,433$. Kritická tabelární hodnota Studentova rozdělení $t_{(\alpha = 0,05,\ n-2 = 8)} = 2,306$. Jelikož je vypočtená hodnota t_r vyšší než kritická tabelární hodnota Studentova rozdělení, zamítáme na hladině významnosti $\alpha = 0,05$ hypotézu H_0, že sledované proměnné jsou nekorelované.

Souhrnný závěr: Provedenou průzkumovou analýzou dat bylo zjištěno, že jeden datový soubor nesplňuje předpoklad normality a v datech je přítomna jedna odlehlá hodnota. Na základě těchto výsledků byl pro výpočet použit Spearmanův korelační koeficient, jehož výsledná hodnota je -0,9970. Významnost korelačního koeficientu byla posouzena třemi metodami. Z výše uvedených výsledků vyplývá, že v rámci sledované geolokality existuje významná negativní korelace mezi denní návštěvností a množstvím spadlých srážek.

4. Výpočet Spearmanova korelačního koeficientu v tabulkovém procesoru Excel

Výpočet provedeme dle vzorce (viz odstavec 7.3.2 Spearmanův korelační koeficient):

$$r_s = \frac{\sum_{i=1}^{n} x_{ri} \cdot y_{ri} - n \cdot \bar{x}_r \bar{y}_r}{(n-1) \cdot s_{xr} \cdot s_{yr}}$$

Postupovat budeme následovně:

1. Vytvoříme tabulku, která bude obsahovat 11 řádků (máme 10 experimentálních párových hodnot, 1 řádek pro záhlaví) a 5 sloupců. V prvním sloupci bude uvedena hodnota proměnné x => x, ve druhém sloupci pořadí hodnoty x_i v rámci vzestupně uspořádaných hodnot => x_{ri}, ve třetím sloupci hodnota proměnné y => y, ve čtvrtém sloupci pořadí hodnoty y_i v rámci vzestupně uspořádaných hodnot => y_{ri}, v pátém sloupci uvedeme pro každý pár dat hodnotu x_i vynásobenou hodnotou y_i. => x_{ri} . y_{ri}

2. Nejprve do připravené tabulky vložíme hodnoty proměnné x a y (1. a 3. sloupec). Následně do sloupce x_{ri} napíšeme číslo pořadí proměnné x dle její hodnoty vzestupně (2. sloupec). Totéž provedeme pro sloupec y_{ri} (4. sloupec). V případě, že máme v některých řádcích hodnoty proměnné shodné, zadáme průměrnou hodnotu daných pořadí. Do posledního sloupce zadáme hodnotu rovnající se součinu příslušné hodnoty x_{ri} a y_{ri}.

x	x_{ri}	y	y_{ri}	x_{ri} . y_{ri}
248	6	7,4	5	30
210	5	8,9	6	30
173	4	12	7	28
322	8	1,2	3	24
497	10	0	1,5	15
370	9	0	1,5	13,5
285	7	6,3	4	28
68	2	27,4	9	18
136	3	13,8	8	24
24	1	54,6	10	10

3. Vypočteme sumu hodnot v 5. sloupci (x_{ri} . y_{ri}) => $\sum_{i=1}^{n} x_{ri} * y_{ri} = 220,5$

4. Vypočteme aritmetické průměry z hodnot ve 2. a 4. sloupci (z pořadových hodnot x_{ri} a y_{ri}) => $\bar{x}_r$ = **5,5**; $\bar{y}_r$ = **5,5**

5. Vypočteme výběrové směrodatné odchylky z hodnot ve 2. a 4. sloupci (z pořadových hodnot x_{ri} a y_{ri}) => s_{xr} = **3,02765**; s_{yr} = **3,018462**

6. Nakonec dosadíme veškeré vypočtené hodnoty do výše uvedeného vzorce => $r_s = \frac{220,5-(10*(5,5*5,5))}{(10-1)*(3,02765*3,018462)} \approx -0,9970$

7. Výsledná vypočtená hodnota Spearmanova korelačního koeficientu je -0,9970.

8.11 Souhrnný příklad

Zadání:

V průběhu července a srpna roku 2016 byla cestovní kanceláří sledována návštěvnost dvou zahraničních destinací – Egypt a Španělsko. Celkem bylo získáno 46 hodnot pro každou destinaci (údajů). Dále byla cestovní kanceláří v průběhu prosince 2016 sledována návštěvnost a průměrná cena zájezdu pouze pro Egypt. Celkem bylo získáno 12 hodnot (údajů). Stanovte střední hodnoty návštěvnosti pro Egypt a Španělsko v létě 2016, střední hodnotu návštěvnosti pro Egypt v zimě 2016 a střední hodnotu průměrné ceny zájezdu do Egypta v zimě 2016. Rozhodněte, zda byla v létě 2016 návštěvnost Egypta významně nižší než návštěvnost Španělska? Existuje v zimě 2016 významná lineární závislost mezi návštěvností Egypta a průměrnou cenou zájezdu? Cestovní kancelář si dále pro Egypt v prosinci 2016 stanovila minimální limit průměrné ceny prodaných zájezdů na osobu 12 000 Kč. Splnila cestovní kancelář v případě Egypta v zimě 2016 stanovený minimální limit průměrné ceny prodaných zájezdů?

Parametry:

α = 0,05

Léto 2016: Egypt (návštěvnost) n_1 = 46; Španělsko (návštěvnost) n_2 = 46

Zima 2016: Egypt (návštěvnost) n_3 = 12; Egypt (průměrná cena zájezdu v Kč) n_4 = 12

Návštěvnost – Léto 2016

Egypt (počet osob)									
391	69	682	253	131	117	87	105	115	2171
169	228	166	172	51	420	240	209	270	121
171	163	148	115	32	216	165	142	219	225
203	103	154	501	135	140	147	236	100	129
132	141	144	176	136	98				

Návštěvnost – Léto 2016

Španělsko (počet osob)									
2764	308	3932	8480	2648	2488	4360	1056	1040	1268
1876	2020	2704	1972	1660	1812	3412	2332	11680	1612
1828	2788	1356	1440	2036	2492	708	1976	2052	3540
3700	780	1464	3456	708	924	876	956	960	2572
2748	2648	2844	3776	1452	760				

Návštěvnost – Zima 2016

Egypt (počet osob)											
21	28	257	15	101	10	46	81	88	8	29	49

Průměrná cena zájezdu – Zima 2016

Egypt (Kč)											
17690	17220	9190	18690	11690	19950	14590	12890	12690	22890	16900	14200

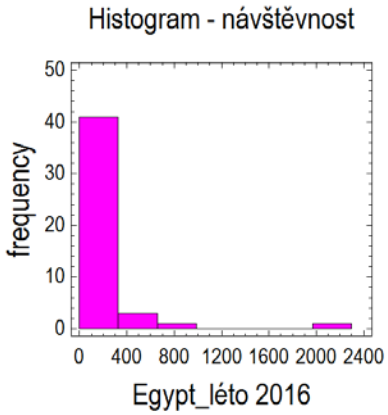
Řešení:

Pro správný výběr výpočtu středních hodnot datových souborů a následné testování musíme nejprve jednotlivě pro každý datový soubor provést průzkumovou analýzu dat.

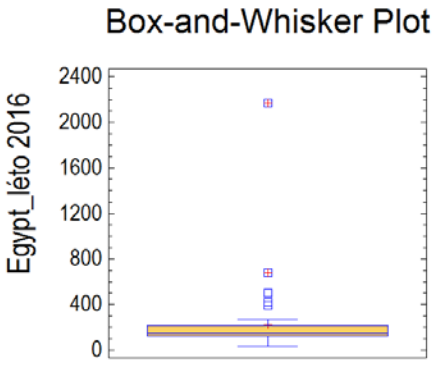
1. Průzkumová analýza dat

1.1. Léto 2016 - Egypt návštěvnost

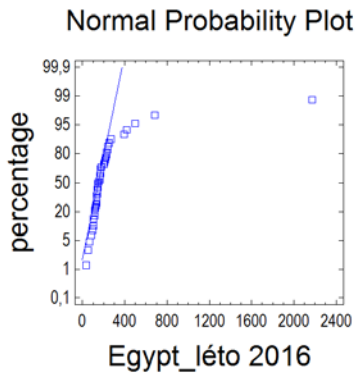
1.1.1. Diagnostika grafů:



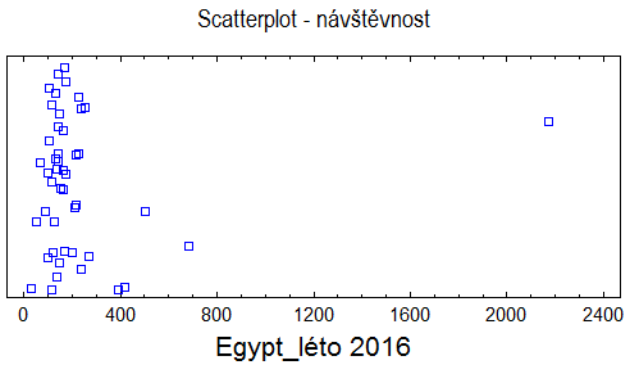
Obrázek 8.8: Histogram četnosti



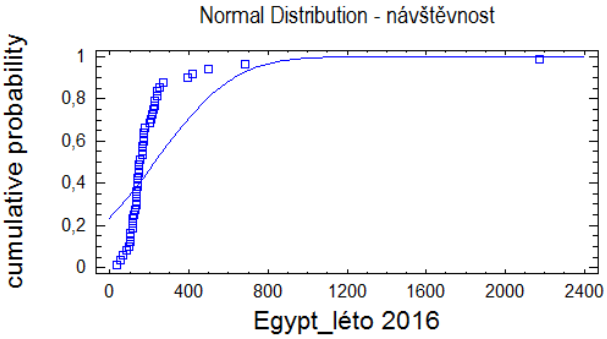
Obrázek 8.9: Krabicový graf



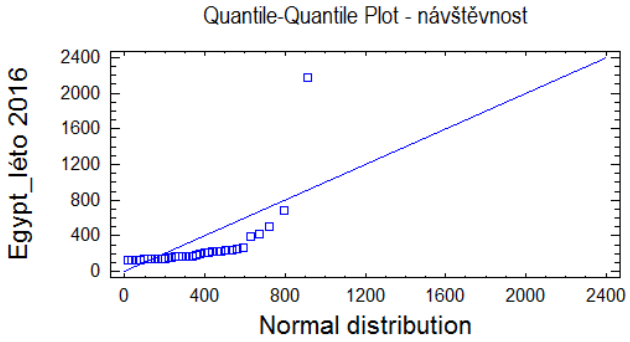
Obrázek 8.10: Normální pravděpodobnostní graf



Obrázek 8.11: Jednorozměrný bodový graf



Obrázek 8.12: Kvantilový graf



Obrázek 8.13: Kvantil-kvantilový graf

Z histogramu četnosti na obrázku 8.8 je zřejmé, že data mají silně asymetrické rozdělení sešikmené k vyšším hodnotám, vpravo lze identifikovat jednu odlehlou hodnotu. Krabicový graf na obrázku 8.9 potvrzuje výraznou asymetrii s kladným zešikmením. V grafu je patrná přítomnost pěti odlehlých hodnot. V normálním pravděpodobnostním grafu (obrázek 8.10) body vykreslují konkávní křivku, což indikuje asymetrické rozdělení dat s kladným zešikmením. V pravé části grafu lze identifikovat jednu odlehlou hodnotu. Od skupiny bodů se nepatrně oddělují další čtyři podezřelé hodnoty, které lze taktéž podezřívat z odlehlosti. V jednorozměrném bodovém grafu na obrázku 8.11 je patrný mrak bodů se zahuštěním v levé části, což potvrzuje asymetrické rozdělení kladně sešikmené. Opět jsou zde patrné čtyři mírně odlehlé hodnoty a napravo jedna významně odlehlá hodnota. V kvantilovém grafu (obrázek 8.12) se data teoretické křivce normálního rozdělení nepřimykají a jsou seřazeny do konkávního tvaru - soubor dat má jednoznačně asymetrické rozdělení s kladným zešikměním s pěti odlehlými hodnotami. V kvantil-kvantilovém grafu (obrázek 8.13) většina bodů neleží na přímce, což indikuje asymetrické rozdělení. Nahoře je opět zřejmá přítomnost jedné odlehlé hodnoty. Dále se od skupiny bodů nepatrně oddělují čtyři hodnoty, které lze podezřívat z odlehlosti.

1.1.2. Ověření normality:

Výstup:

```
Skewness = 5,51567
Kurtosis = 33,7011
```

Hodnota koeficientu šikmosti i špičatosti je kladná a je významně vyšší než nula, což znamená, že datový soubor je rozdělen asymetricky s kladným zešikmením.

Závěr: Z grafických diagnostik vyplývá, že soubor dat má výrazně asymetrické rozdělení sešikmené k vyšším hodnotám s jednou významně odlehlou hodnotou a dalšími čtyřmi podezřelými hodnotami nahoře.

1.1.3. Statistické testy:

Testy nezávislosti – výstup:

```
Runs above and below median:
```

P-value = 0,296597

Runs up and down:

P-value = 0,0846208

Box-Pierce Test

P-value = 0,993937

Závěr: Všechny tři testy nezávislosti mají vypočtenou *p*-hodnotu vyšší než 0,05, což znamená, že přijímáme nulovou hypotézu H_0 (H_0 = data jsou nezávislá). Provedenými testy bylo s 95% statistickou jistotou prokázáno, že data jsou nezávislá.

Testy normality – výstup:

Na základě velikosti datového souboru (obsahuje méně než 50 hodnot) byl pro testování normality použit Kolmogorov-Smirnovův test (K-S test) a Shapiro-Wilkův test (S-W test).

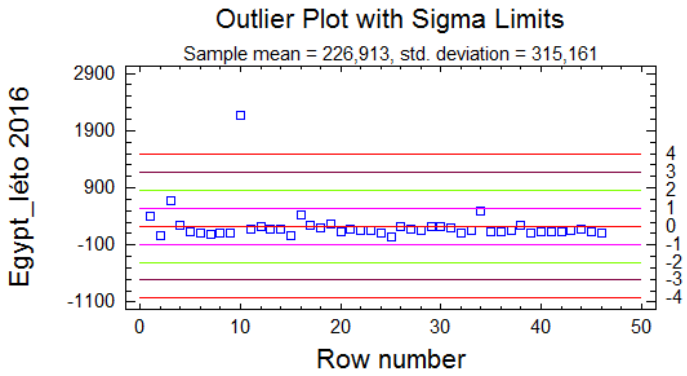
Kolmogorov-Smirnov Test: P-Value = <0,01

Shapiro-Wilks Test: P-Value = 0,0

Závěr: Výsledná *p*-hodnota provedeného K-S testu je <0,01 a S-W testu je 0,0. Jelikož jsou výsledné *p*-hodnoty významně menší než 0,05, zamítáme hypotézu H_0 ve prospěch hypotézy H_A. Provedením K-S testu a S-W testu bylo s 95% statistickou jistotou prokázáno, že datový soubor nelze považovat za výběr s normálním rozdělením.

1.1.4. Identifikace odlehlých hodnot:

Identifikace odlehlých hodnot byla provedena pomocí diagnostiky grafu odlehlých hodnot a pomocí mediánových souřadnic.



Obrázek 8.14: Graf pro identifikaci odlehlých hodnot

V grafu odlehlých hodnot (obrázek 8.14) lze identifikovat jednu odlehlou hodnotu, jejíž hodnota je vyšší než trojnásobek směrodatné odchylky datového souboru. Pomocí mediánových souřadnic bylo identifikováno 5 odlehlých hodnot. Tyto hodnoty mají mediánovou souřadnici větší než tři a můžeme je tedy považovat za odlehlé hodnoty.

Závěr: Předpoklad o normalitě výběru byl zamítnut, což je ve shodě s diagnostikou grafů i hodnotami koeficientů šikmosti a špičatosti. Předpoklad o nezávislosti hodnot byl přijat. Grafem odlehlých hodnot byla identifikována 1 odlehlá hodnota (pozn.: jelikož je identifikace odlehlých hodnot v grafu založena na výpočtu směrodatné odchylky, může být tato diagnostika v případě silně asymetrických dat málo věrohodná). Kritériem mediánových souřadnic bylo identifikováno 5 odlehlých hodnot, což je opět ve shodě se závěry většiny grafických diagnostik. S ohledem na charakter dat nelze odlehlé body z další analýzy vyloučit (možná ztráta důležité informace). Vzhledem k asymetrickému rozdělení dat a přítomnosti odlehlých hodnot se jeví jako nejvhodnější pro následné vyhodnocení studovaného souboru provedení transformace dat.

1.1.5. Transformace dat:

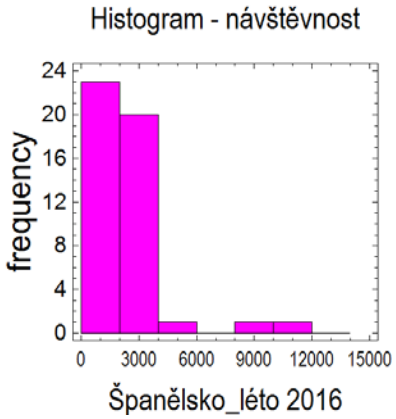
Pro transformaci dat byla použita logaritmická transformace pomocí přirozeného logaritmu xln = ln (x). Následně byla úspěšnost transformace ověřena grafickými diagnostikami, testy normality a testováním odlehlých hodnot.

Z výsledných grafů vyplývá, že transformací dat nebylo dosaženo zesymetričtění rozdělení a přiblížení k normalitě. V grafech lze stále identifikovat významně odlehlé hodnoty. Hodnoty koeficientu šikmosti a špičatosti se po transformaci nepřiblížily nulové hodnotě. Taktéž testy normality (Kolmogorov-Smirnovův test a Shapiro-Wilkův test) bylo s 95% statistickou jistotou prokázáno, že transformovaná data nelze považovat za výběr z normálního rozdělení.

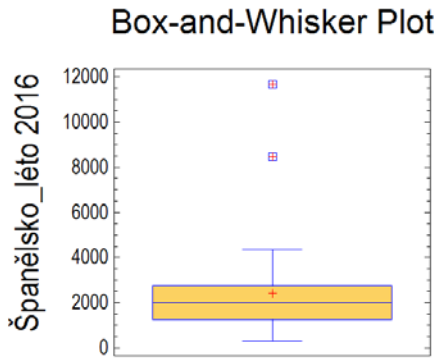
Závěr: Ze statistického hlediska nebyla provedená transformace dat úspěšná a pro další odhady a testy je nutné využít robustních a neparametrických metod statistické analýzy.

1.2. Léto 2016 - Španělsko návštěvnost

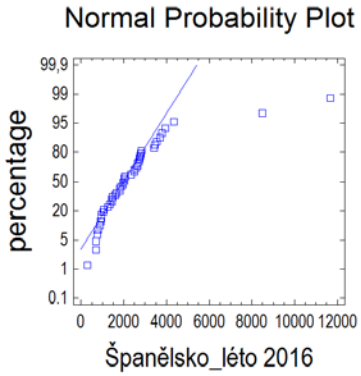
1.2.1. Diagnostika grafů:



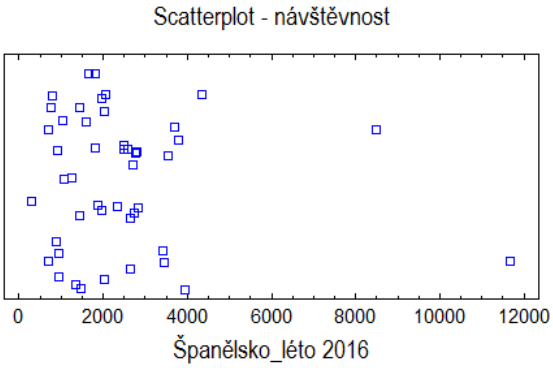
Obrázek 8.15: Histogram četnosti



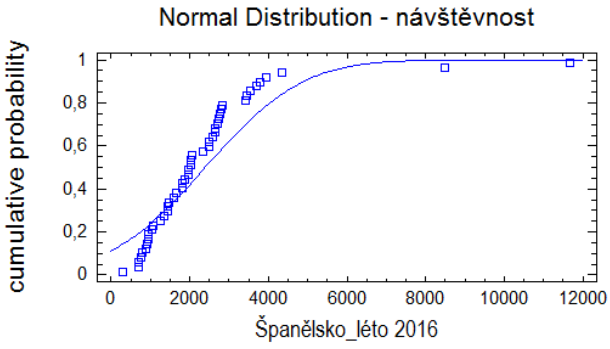
Obrázek 8.16: Krabicový graf



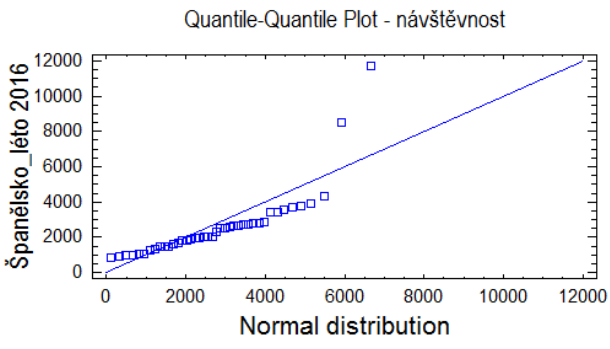
Obrázek 8.17: Normální pravděpodobnostní graf



Obrázek 8.18: Jednorozměrný bodový graf



Obrázek 8.19: Kvantilový graf



Obrázek 8.20: Kvantil-kvantilový graf

Z histogramu četnosti a krabicového grafu (obrázek 8.15 a 8.16) vyplývá, že data mají silně asymetrické rozdělení sešikmené k vyšším hodnotám se dvěma odlehlými hodnotami. V normálním pravděpodobnostním

grafu (obrázek 8.17) jsou body seřazeny do konkávní křivky, což indikuje asymetrické rozdělení dat s kladným zešikmením. V pravé části grafu lze identifikovat dvě odlehlé hodnoty. V jednorozměrném bodovém grafu na obrázku 8.18 je patrný mrak bodů se zahuštěním v levé části, což potvrzuje asymetrické rozdělení kladně sešikmené. V pravé části grafu jsou přítomny dvě odlehlé hodnoty. V kvantilovém grafu (obrázek 8.19) se data teoretické křivce normálního rozdělení nepřimykají a seřazení bodů do konkávního tvaru indikuje asymetrické rozdělení s kladným zešikměním se dvěma odlehlými hodnotami. Body v kvantil-kvantilovém grafu (obrázek 8.20) neleží na přímce, což potvrzuje asymetrické rozdělení. Nahoře je opět zřejmá přítomnost dvou odlehlých hodnot.

1.2.2. Ověření normality:

Výstup:

```
Skewness = 3,06574
Kurtosis = 12,1208
```

Hodnota koeficientu šikmosti i špičatosti je kladná a je významně vyšší než nula, což znamená, že datový soubor je rozdělen asymetricky s kladným zešikmením.

Závěr: Z grafických diagnostik vyplývá, že soubor dat má výrazně asymetrické rozdělení sešikmené k vyšším hodnotám se dvěma odlehlými hodnotami nahoře.

1.2.3. Statistické testy:

Testy nezávislosti – výstup:

```
Runs above and below median:
P-value = 0,100966
Runs up and down:
P-value = 0,0846208
Box-Pierce Test
P-value = 0,682736
```

Závěr: U všech provedených testů nezávislosti je vypočtená *p*-hodnota vyšší než 0,05, což znamená, žepřijímáme nulovou hypotézu H_0 (H_0 = data jsou nezávislá). Testováním bylo s 95% statistickou jistotou prokázáno, že data jsou nezávislá.

Testy normality – výstup:

Na základě velikosti datového souboru (obsahuje méně než 50 hodnot) byl pro testování normality použit Kolmogorov-Smirnovův test (K-S test) a Shapiro-Wilkův test (S-W test).

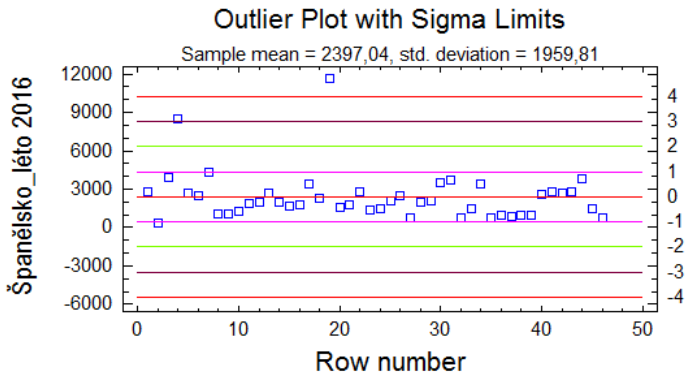
```
Kolmogorov-Smirnov Test: P-Value = <0,01
Shapiro-Wilks Test: P-Value = 4,1706E-11
```

Závěr: Výsledná *p*-hodnota provedeného K-S testu je <0,01 a S-W testu je 4,1706E-11. Jelikož jsou výsledné *p*-hodnoty významně menší než 0,05, zamítáme hypotézu H_0 ve prospěch hypotézy H_A. Provedením K-S testu a S-W testu bylo s 95% statistickou jistotou prokázáno, že datový soubor nelze považovat za výběr s normálním rozdělením.

1.2.4. Identifikace odlehlých hodnot:

Identifikace odlehlých hodnot byla provedena pomocí diagnostiky grafu odlehlých hodnot a pomocí mediánových souřadnic.

V grafu odlehlých hodnot (obrázek 8.21) lze identifikovat dvě odlehlé hodnoty, jejichž hodnota je vyšší než trojnásobek směrodatné odchylky datového souboru. Pomocí mediánových souřadnic byly rovněž identifikovány 2 odlehlé hodnoty.



Obrázek 8.21: Graf pro identifikaci odlehlých hodnot

Závěr: Předpoklad o normalitě výběru byl zamítnut, což je ve shodě s diagnostikou grafů i hodnotami koeficientů šikmosti a špičatosti. Předpoklad o nezávislosti hodnot byl přijat. Grafem odlehlých hodnot byly identifikovány 2 odlehlé hodnoty. Kritériem mediánových souřadnic byly taktéž identifikovány 2 odlehlé hodnoty, což je opět ve shodě se závěry většiny grafických diagnostik. S ohledem na charakter dat nelze odlehlé body z další analýzy vyloučit (možná ztráta důležité informace). Vzhledem k asymetrickému rozdělení dat a přítomnosti odlehlých hodnot se jeví jako nejvhodnější pro následné vyhodnocení studovaného souboru provedení transformace dat.

1.2.5. Transformace dat:

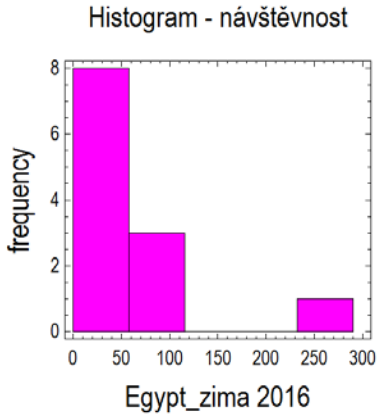
Pro transformaci dat byla použita logaritmická transformace pomocí přirozeného logaritmu xln = ln (x). Následně byla úspěšnost transformace ověřena grafickými diagnostikami, testy normality a testováním odlehlých hodnot.

Z výsledných grafů vyplývá, že transformací dat bylo dosaženo zesymetričtění rozdělení a přiblížení k normalitě. Některými grafickými diagnostikami lze stále identifikovat dvě velmi mírně odlehlé hodnoty. Hodnoty koeficientu šikmosti a špičatosti se po transformaci výrazně přiblížily nulové hodnotě. Taktéž testy normality (Kolmogorov-Smirnovův test a Shapiro-Wilkův test) bylo s 95% statistickou jistotou prokázáno, že transformovaná data lze považovat za výběr z normálního rozdělení.

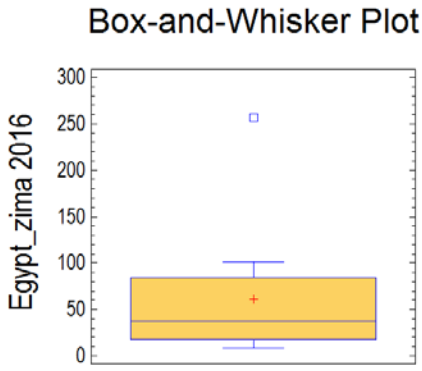
Závěr: Ze statistického hlediska byla provedená transformace dat úspěšná a pro další odhady a testy je možné při použití transformovaných dat využít klasických metod statistické analýzy.

1.3. Zima 2016 - Egypt návštěvnost

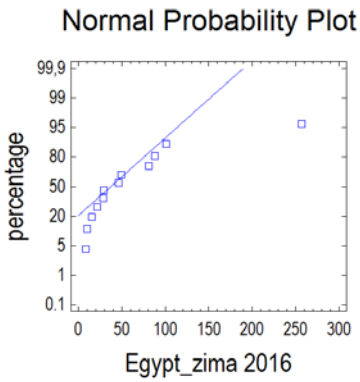
1.3.1. Diagnostika grafů:



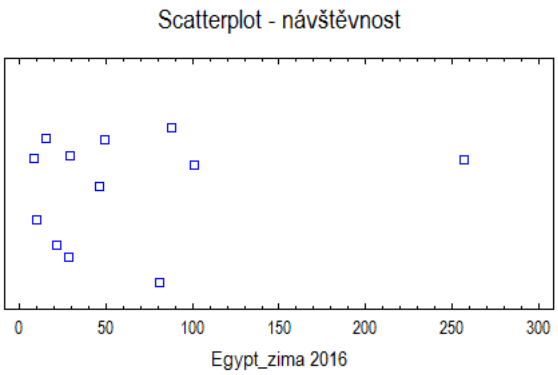
Obrázek 8.22: Histogram četnosti



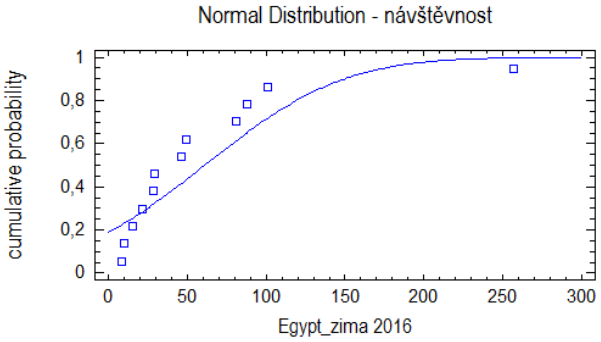
Obrázek 8.23: Krabicový graf



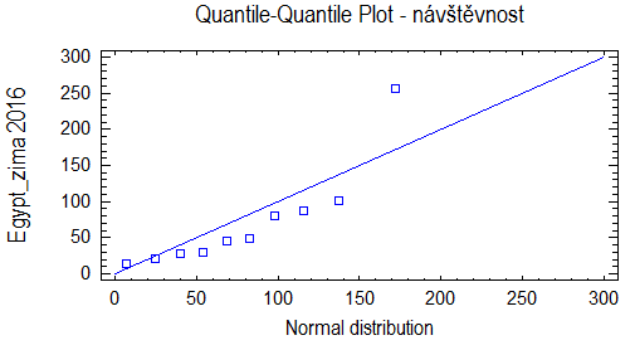
Obrázek 8.24: Normální pravděpodobnostní graf



Obrázek 8.25: Jednorozměrný bodový graf



Obrázek 8.26: Kvantilový graf



Obrázek 8.27: Kvantil-kvantilový graf

Dle histogramu a krabicového grafu (obrázek 8.22 a 8.23) je rozdělení dat asymetrické, silně sešikmené. V grafech lze identifikovat jednu odlehlou hodnotu. V normálním pravděpodobnostním grafu (obrázek 8.24) body seřazené do konkávní křivky potvrzují asymetrické rozdělení dat s kladným zešikmením s jednou odlehlou hodnotou vpravo. V jednorozměrném bodovém grafu na obrázku 8.25 je zřejmý mrak bodů se zahuštěním v levé části, což opět potvrzuje asymetrické rozdělení kladně sešikmené. V pravé části grafu je patrná jedna odlehlá hodnota. V kvantilovém grafu (obrázek 8.26) se data teoretické křivce normálního rozdělení nepřimykají a seřazení bodů do konkávního tvaru indikuje asymetrické rozdělení s kladným zešikměním s jednou odlehlou hodnotou. Body v kvantil-kvantilovém grafu (obrázek 8.27) neleží na přímce, což potvrzuje asymetrické rozdělení. Nahoře je opět zřejmá přítomnost jedné odlehlé hodnoty.

1.3.2. Ověření normality:

Výstup:

```
Skewness = 2,32936
Kurtosis = 6,24844
```

Vypočtený koeficient šikmosti i špičatosti je kladný a významně vyšší než nula, což znamená, že datový soubor je rozdělen asymetricky s kladným zešikmením.

Závěr: Z grafických diagnostik vyplývá, že soubor dat má výrazně asymetrické rozdělení sešikmené k vyšším hodnotám s jednou odlehlou hodnotou nahoře.

1.3.3. Statistické testy:

Testy nezávislosti – výstup:

```
Runs above and below median:
P-value = 0,762065
Runs up and down:
P-value = 0,901433
```

Box-Pierce Test

P-value = 0,817367

Závěr: U všech provedených testů nezávislosti je vypočtená *p*-hodnota vyšší než 0,05, což znamená, že přijímáme nulovou hypotézu H_0 (H_0 = data jsou nezávislá). Testováním bylo s 95% statistickou jistotou prokázáno, že data jsou nezávislá.

Testy normality – výstup:

Na základě velikosti datového souboru (obsahuje méně než 50 hodnot) byl pro testování normality použit Kolmogorov-Smirnovův test (K-S test) a Shapiro-Wilkův test (S-W test).

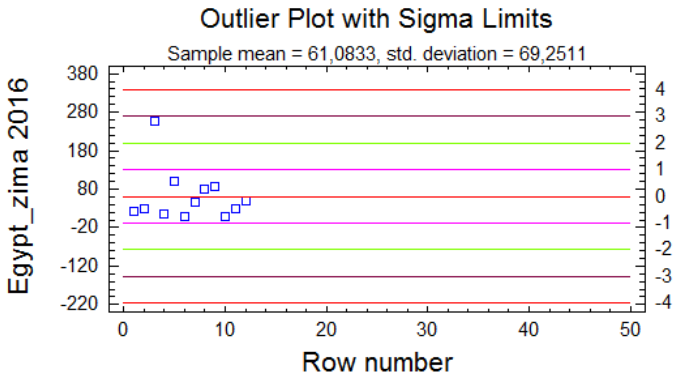
Kolmogorov-Smirnov Test: P-Value = <0,01

Shapiro-Wilks Test: P-Value = 0,000934327

Závěr: Výsledná *p*-hodnota provedeného K-S testu je <0,01 a S-W testu je 0,000934327. Výsledné *p*-hodnoty jsou významně menší než 0,05, proto zamítáme hypotézu H_0 ve prospěch hypotézy H_A. Provedením K-S testu a S-W testu bylo s 95% statistickou jistotou prokázáno, že datový soubor nelze považovat za výběr s normálním rozdělením.

1.3.4. Identifikace odlehlých hodnot:

Identifikace odlehlých hodnot byla provedena pomocí diagnostiky grafu odlehlých hodnot a pomocí mediánových souřadnic.



Obrázek 8.28: Graf pro identifikaci odlehlých hodnot

V grafu odlehlých hodnot (obrázek 8.28) lze identifikovat možnou jednu odlehlou hodnotu (její hodnota je velmi blízko hodnotě trojnásobku směrodatné odchylky). Pomocí mediánových souřadnic byla identifikována 1 odlehlá hodnota.

Závěr: Předpoklad o normalitě výběru byl zamítnut, což je ve shodě s diagnostikou grafů i hodnotami koeficientů šikmosti a špičatosti. Předpoklad o nezávislosti hodnot byl přijat. Grafem odlehlých hodnot i kritériem mediánových souřadnic byla identifikována 1 odlehlá hodnota, což je opět ve shodě se závěry grafických diagnostik. Odlehlé body nelze vzhledem k charakteru dat vyloučit. Vzhledem k asymetrickému rozdělení dat a přítomnosti odlehlé hodnoty se jeví jako nejvhodnější pro následné vyhodnocení studovaného souboru provedení transformace dat.

1.3.5. Transformace dat:

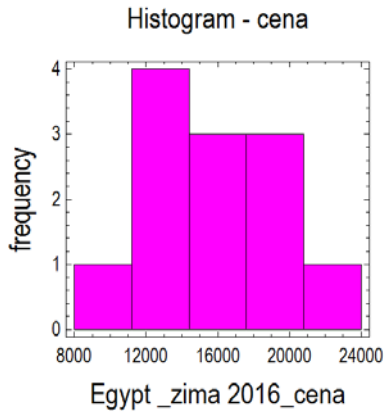
Pro transformaci dat byla použita logaritmická transformace pomocí přirozeného logaritmu xln = ln (x). Následně byla úspěšnost transformace ověřena grafickými diagnostikami, testy normality a testováním odlehlých hodnot.

Transformovaná data byla opět podrobena grafické diagnostice. Z výsledných grafů vyplývá, že transformací dat bylo dosaženo zesymetričtění rozdělení a přiblížení k normalitě. Odlehlé hodnoty v transformovaných datech již nebyly identifikovány. Hodnoty koeficientu šikmosti a špičatosti se po transformaci výrazně přiblížily nulové hodnotě. Taktéž testy normality (Kolmogorov-Smirnovův test a Shapiro-Wilkův test) bylo s 95% statistickou jistotou prokázáno, že transformovaná data lze považovat za výběr z normálního rozdělení.

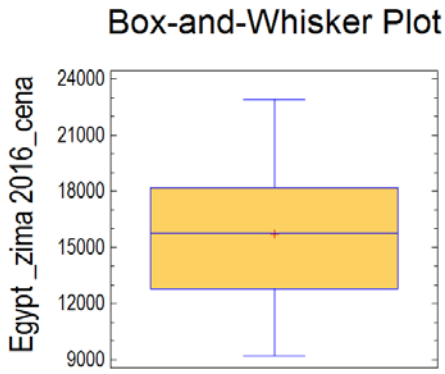
Závěr: Ze statistického hlediska byla provedená transformace dat úspěšná a pro další odhady a testy je možné při použití transformovaných dat využít klasických metod statistické analýzy.

1.4. Zima 2016 – Egypt cena

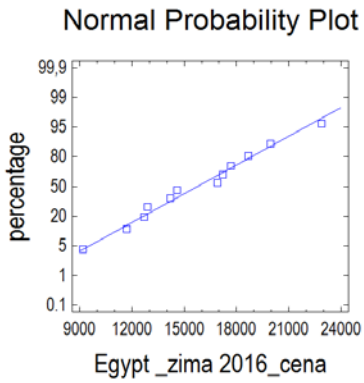
1.4.1. Diagnostika grafů:



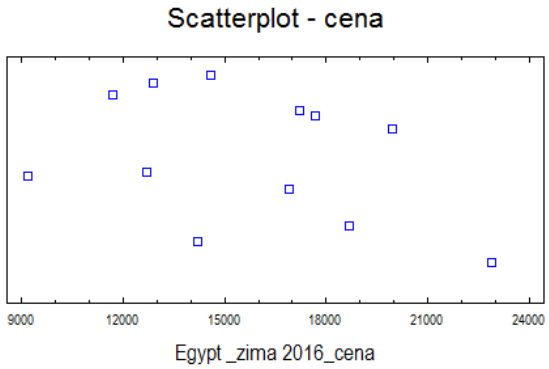
Obrázek 8.29: Histogram četnosti



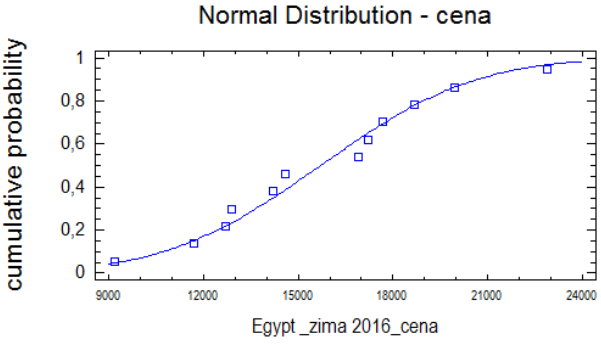
Obrázek 8.30: Krabicový graf



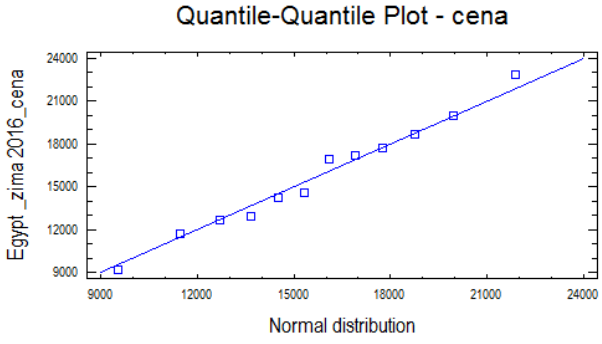
Obrázek 8.31: Normální pravděpodobnostní graf



Obrázek 8.32: Jednorozměrný bodový graf



Obrázek 8.33: Kvantilový graf



Obrázek 8.34: Kvantil-kvantilový graf

Z histogramu četnosti na obrázku 8.29 je zřejmé, že data mají normální rozdělení bez odlehlých hodnot. Krabicový graf na obrázku 8.30 potvrzuje normalitu datového souboru bez přítomnosti odlehlých hodnot. V normálním pravděpodobnostním grafu (obrázek 8.31) se body přimykají teoretické přímce, což indikuje normální rozdělení. V grafu nelze identifikovat odlehlé hodnoty. V jednorozměrném bodovém grafu na obrázku 8.32 je patrný mrak bodů bez zahuštění a odlehlých bodů. V kvantilovém grafu (obrázek 8.33) se data teoretické křivce normálního rozdělení velmi dobře přimykají - soubor dat má jednoznačně normální rozdělení bez odlehlých hodnot. V kvantil-kvantilovém grafu (obrázek 8.34) většina bodů leží na přímce, což indikuje symetrické rozdělení. V grafu nelze identifikovat odlehlé hodnoty.

1.4.2. Ověření normality:

Výstup:

```
Skewness = 0,160826
Kurtosis = 0,334013
```

Hodnota koeficientu šikmosti i špičatosti je velmi blízká nule, což znamená, že datový soubor splňuje předpoklad normálního rozdělení.

Závěr: Z grafických diagnostik vyplývá, že soubor dat má normální rozdělení bez odlehlých hodnot.

1.4.3. Statistické testy:

Testy nezávislosti – výstup:

```
Runs above and below median:
P-value = 0,762065
Runs up and down:
P-value = 0,901433
Box-Pierce Test
P-value = 0,511953
```

Závěr: Všechny tři testy nezávislosti mají vypočtenou *p*-hodnotu vyšší než 0,05, což znamená, že přijímáme nulovou hypotézu H_0 (H_0 = data jsou nezávislá). Provedenými testy bylo s 95% statistickou jistotou prokázáno, že data jsou nezávislá.

Testy normality – výstup:

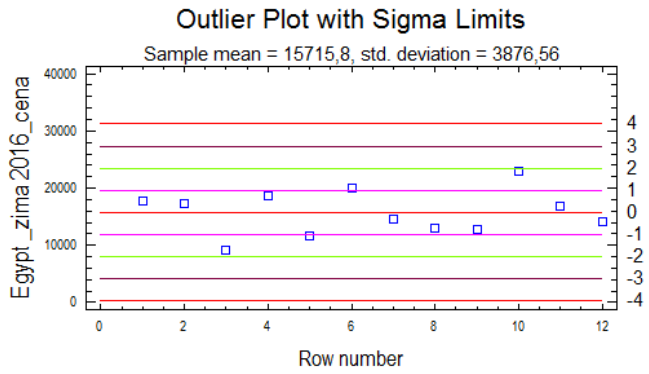
Na základě velikosti datového souboru (obsahuje méně než 50 hodnot) byl pro testování normality použit Kolmogorov-Smirnovův test (K-S test) a Shapiro-Wilkův test (S-W test).

```
Kolmogorov-Smirnov Test: P-Value = >=0,10
Shapiro-Wilks Test: P-Value = 0,989116
```

Závěr: Výsledná *p*-hodnota provedeného K-S testu je >=0,10 a S-W testu je 0,989116. Jelikož jsou výsledné *p*-hodnoty významně větší než 0,05, přijímáme hypotézu H_0. Provedením K-S testu a S-W testu bylo s 95% statistickou jistotou prokázáno, že datový soubor lze považovat za výběr s normálním rozdělením.

1.4.4, Identifikace odlehlých hodnot:

Identifikace odlehlých hodnot byla provedena pomocí diagnostiky grafu odlehlých hodnot a pomocí mediánových souřadnic.



Obrázek 8.35: Graf pro identifikaci odlehlých hodnot

Dle grafu odlehlých hodnot (obrázek 8.35) je zřejmé, že se v datech nevyskytují odlehlé hodnoty. Pomocí mediánových souřadnic byla nepřítomnost odlehlých hodnot potvrzena.

Závěr: Předpoklad o normalitě výběru byl přijat, což je ve shodě s diagnostikou grafů i hodnotami koeficientů šikmosti a špičatosti. Předpoklad o nezávislosti hodnot byl přijat. Grafem odlehlých hodnot i kritériem mediánových souřadnic nebyly identifikovány žádné odlehlé hodnoty, což je opět ve shodě se závěry grafických

diagnostik. Transformace dat není nutná. Pro další odhady a testy je možné využít klasických metod statistické analýzy.

2. Odhad míry polohy

Průzkumovou analýzou dat jsme zjistili, že první analyzovaný datový soubor (Léto 2016 – Egypt - návštěvnost) má výrazně asymetrické rozdělení sešikmené k vyšším hodnotám s přítomností jedné odlehlé hodnoty. Transformace dat nebyla úspěšná. Z uvedených důvodů je nutné pro odhad míry polohy (odhad střední hodnoty) použít robustní metody. Odhad střední hodnoty tedy provedeme pomocí mediánu. V případě druhého a třetího datového souboru (Léto 2016 – Španělsko – návštěvnost; Zima 2016 – Egypt – návštěvnost) byla opět zjištěna výrazná asymetrie s přítomností odlehlých hodnot. Transformace dat však byla úspěšná a pro odhad míry polohy (odhad střední hodnoty) lze použít na tato transformovaná data klasické metody. Odhad střední hodnoty tedy provedeme pomocí retransformovaného průměru. Čtvrtý analyzovaný datový soubor (Zima 2016 – Egypt - cena) má normální rozdělení bez přítomnosti odlehlých hodnot. Odhad střední hodnoty lze provést pomocí aritmetického průměru.

Výsledné střední hodnoty:

2.1. Léto 2016 – Egypt návštěvnost

$\tilde{x}_{0,5} = \mathbf{151}$

2.2. Léto 2016 – Španělsko návštěvnost

$\bar{x}_R = \mathbf{1913}$

2.3. Zima 2016 – Egypt návštěvnost

$\bar{x}_R = \mathbf{38}$

2.4. Zima 2016 – Egypt cena

$\bar{x} = 15715{,}8 \approx \mathbf{15715\ Kč}$

Závěr: Střední hodnota návštěvnosti Egypta v létě 2016 je 151 osob, střední hodnota návštěvnosti Španělska v létě 2016 je 1913 osob, střední hodnota návštěvnosti Egypta v zimě 2016 je 38 osob a střední hodnota průměrné ceny zájezdu do Egypta v zimě 2016 je 15715 Kč.

3. Mann-Whitneyův test

Nyní musíme rozhodnout, zda byla v létě 2016 návštěvnost Egypta významně nižší než návštěvnost Španělska. Pro analýzu použijeme datový soubor návštěvnosti Egypta v létě 2016 a datový soubor návštěvnosti Španělska v létě 2016. K testování použijeme, vzhledem k asymetrickému rozdělení a odlehlým hodnotám v datových souborech, neparametrickou obdobu testu shodnosti, tedy Mann-Whitneyův test. Pro testování zvolíme jednostranný test levostrannou alternativu - Less Than (potřebujeme zjistit, zda byla v létě 2016 návštěvnost Egypta významně nižší než návštěvnost Španělska).

Mann-Whitneyův test - výstup:

```
Comparison of Medians
---------------------
Median of sample 1: 151,0
Median of sample 2: 1998,0
Mann-Whitney (Wilcoxon) W test to compare medians
Null hypothesis: median1 = median2
Alt. hypothesis: median1 < median2
Average rank of sample 1: 24,1522
Average rank of sample 2: 68,8478
W = 2086,0   P-value = 0,0
```

Závěr: Z výsledku Mann-Whitneyova testu vyplývá, že vypočtená *p*-hodnota 0,0 je významně menší než 0,05, což znamená, že zamítáme nulovou hypotézu H_0 ve prospěch alternativní hypotézy H_A. Mann-Whitneyovým testem bylo na 95% hladině významnosti prokázáno, že návštěvnost Egypta v létě 2016 byla významně nižší než návštěvnost Španělska.

4. Korelační analýza

Další otázkou, kterou musíme dle zadání pomocí statistické analýzy vyřešit je, zda existuje v zimě 2016 významná lineární závislost mezi návštěvností Egypta a průměrnou cenou zájezdu. Tento problém vyřešíme pomocí korelační analýzy. Průzkumovou analýzou dat jsme zjistili, že datový soubor pro návštěvnost Egypta v zimě 2016 má asymetrické rozdělení dat s přítomností jedné odlehlé hodnoty, kterou nelze vyloučit. Proto pro vyčíslení těsnosti vzájemné lineární závislosti použijeme Spearmanův korelační koeficient.

Pozn.: *Pokud bychom použili Pearsonův korelační koeficient, který je vhodný pouze pro data splňující podmínku normality bez odlehlých hodnot, měli bychom výsledek nesprávný - výrazně podhodnocený, příp. nadhodnocený. Dále nelze korelovat jeden soubor dat s normálním rozdělením a druhý soubor s transformovanými daty.*

Výsledná hodnota Spearmanova korelačního koeficientu:

```
Zima 2016 Egypt návštěvnost a Zima 2016 Egypt cena
r_s = -1,0000
```

5. Významnost vypočteného korelačního koeficientu

Nyní musíme rozhodnout, zda tato lineární závislost je významná, či není.

Vypočtená *p*-hodnota 0,0000 je významně menší než zvolená hladina pravděpodobnosti 0,05. Zamítáme nulovou hypotézu H_0 ve prospěch alternativní hypotézy H_A, což znamená, že mezi proměnnými existuje významná negativní korelace (hodnota korelačního koeficientu je v rozsahu -1 až 0). Ze srovnání vypočtené hodnoty korelačního koeficientu s kritickou tabelární hodnotou Spearmanova korelačního koeficientu, která pro n = 12 činí 0,587, vyplývá, že absolutní hodnota r_s = -1,0000 je významně vyšší. Můžeme tedy s 95% statistickou jistotou tvrdit, že mezi že mezi proměnnými existuje významná negativní korelace.

Závěr: Na základě výsledků průzkumové analýzy dat byl pro výpočet použit Spearmanův korelační koeficient, jehož výsledná hodnota je -1,0000. Významnost korelačního koeficientu byla posouzena s použitím dvou metod. Z výše uvedených výsledků vyplývá, že existuje významná negativní lineární závislost mezi návštěvností Egypta a průměrnou cenou zájezdu v zimě 2016. Je tedy zřejmé, že v uvedeném případě s narůstající cenou zájezdu návštěvnost dané destinace klesá.

6. Test správnosti

Posledním úkolem vyplývajícím ze zadání je rozhodnout, zda v případě v případě Egypta v zimě 2016 splnila cestovní kancelář stanovený minimální limit průměrné ceny prodaných zájezdů, který činil 12 000 Kč na osobu. K testování použijeme vzhledem k normalitě datového souboru test správnosti. Pro testování zvolíme jednostranný test levostrannou alternativu – Less Than (potřebujeme zjistit, zda nejsou průměrné ceny prodaných zájezdů významně nižší, než udává stanovený minimální limit).

Test správnosti - výstup:

```
t-test
----------------
Null Hypothesis: mean = 12000,0
Alternative: less than
Computed t statistic = 3,32048
P-Value = 0,996587
Do not reject the null hypothesis for alpha = 0,05.
```

Závěr: Z výsledku testu správnosti vyplývá, že vypočtená *p*-hodnota 0,996587 je větší než 0,05, což znamená, že přijímáme nulovou hypotézu H_0. Na základě provedeného testu bylo s 95% statistickou jistotou prokázáno,

že cestovní kancelář v případě Egypta v zimě 2016 stanovený minimální limit průměrné ceny prodaných zájezdů splnila.

Souhrnný závěr: Průzkumovou analýzou dat bylo zjištěno, že první analyzovaný datový soubor (Léto 2016 – Egypt - návštěvnost) má výrazně asymetrické rozdělení sešikmené k vyšším hodnotám s přítomností jedné odlehlé hodnoty. Transformace dat nebyla úspěšná. Na základě těchto výsledků byla pro odhad míry polohy (odhad střední hodnoty) použita robustní metoda, a to medián. V případě druhého a třetího datového souboru (Léto 2016 – Španělsko – návštěvnost; Zima 2016 – Egypt – návštěvnost) byla opět zjištěna výrazná asymetrie s přítomností odlehlých hodnot. Transformace dat však byla úspěšná a pro odhad míry polohy (odhad střední hodnoty) byl použit retransformovaný průměr. Čtvrtý analyzovaný datový soubor (Zima 2016 – Egypt - cena) splnil předpoklad normality, přičemž v datech nebyly identifikovány odlehlé hodnoty. Odhad střední hodnoty byl u tohoto datového souboru proveden pomocí aritmetického průměru. V létě 2016 činila střední hodnota návštěvnosti Egypta 151 osob a střední hodnota návštěvnosti Španělska 1913 osob. V zimě 2016 činila střední hodnota návštěvnosti Egypta 28 osob a střední hodnota ceny zájezdu 15 715 Kč. Mann-Whitneyovým testem bylo na 95% hladině významnosti prokázáno, že návštěvnost Egypta v létě 2016 byla významně nižší než návštěvnost Španělska. Pomocí korelační analýzy byla s 95% statistickou jistotou prokázána významná negativní lineární závislost mezi návštěvností Egypta a průměrnou cenou zájezdu v zimě 2016. Vypočtená hodnota korelačního koeficientu je -1,0000. Ke stanovení vzájemné lineární závislosti byl použit Spearmanův korelační koeficient. Z výsledku vyplývá, že v případě Egypta v zimě 2016 s narůstající cenou zájezdu návštěvnost klesala. Z výsledku testu správnosti vyplývá, že vypočtená *p*-hodnota 0,996587 je větší než 0,05, což znamená, že přijímáme nulovou hypotézu H_0. Na základě provedeného testu bylo s 95% statistickou jistotou prokázáno, že cestovní kancelář v případě Egypta v zimě 2016 stanovený minimální limit průměrné ceny prodaných zájezdů splnila. Testem správnosti bylo s 95% statistickou jistotou prokázáno, že cestovní kancelář v případě Egypta v zimě 2016 stanovený minimální limit 12 000 Kč průměrné ceny prodaných zájezdů splnila.

Literatura

[1] FORTHOFER, R. N., E. S. LEE a M. HERNANDEZ. *Biostatistics: a guide to design, analysis, and discovery*. 2nd ed. Burlington, MA: Elsevier Academic Press, 2007, 502 s. ISBN 01-236-9492-2.

[2] HORN, Paul S. Some Easy t Statistics. *Journal of the American Statistical Association*. 1983, **78**(384), 930-936. DOI: 10.1080/01621459.1983.10477042. ISSN 0162-1459.

[3] IGLEWICZ, B. a D. C. HOAGLIN. *How to Detect and Handle Outliers*. Milwaukee, Wis.: ASQC Quality Press, 1993, 87 s. ISBN 08-738-9247-X.

[4] MILLER, J. a J. MILLER. *Statistics and chemometrics for analytical chemistry*. 6th ed. Harlow: Pearson Education Limited, 2010, 278 s. ISBN 978-027-3730-422.

[5] MELOUN, Milan a Jiří MILITKÝ. *Interaktivní statistická analýza dat*. 3. vyd. Praha: Karolinum, 2012, 953 s. ISBN 978-80-246-2173-9.

[6] MELOUN, Milan a Jiří MILITKÝ. *Statistical data analysis: A practical guide with 1250 exercises and answer key on CD*. Philadelphia: Woodhead Publishing India Pvt Ltd, 2011, 1600 s. ISBN 978-93-80308-11-1.

[7] MELOUN, Milan a Jiří MILITKÝ. *Kompedium statistického zpracování dat: Metody a řešené úlohy.* 2. vyd. Praha: Academia, 2006, 982 s. ISBN 80-200-1396-2.

[8] OTT, Lyman a Michael LONGNECKER. *An introduction to statistical methods & data analysis*. Seventh edition. Australia: Cengage Learning, 2016, 1157 s. ISBN 978-1-305-26947-7.

[9] OTYEPKA, M., P. BANÁŠ a E. OTYEPKOVÁ. *Základy zpracování dat.* Olomouc: Univerzita Palackého v Olomouci, 2013, 90 s. ISBN 978-80-244-3636-4.

[10] PAVLÍK, J. *Aplikovaná statistika*. Praha: Vysoká škola chemicko-technologická, 2005, 172 s. ISBN 80-708-0569-2.

[11] POTTS, P. J. *A Handbook of Silicate Rock Analysis*. Boston, MA: Springer US, 1992, 622 s. ISBN 978-0-216-93209-8.

[12] RAZALI, N. M. a Y. B. WAH. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*. 2011, roč. 2, č. 1, s. 21-33.

[13] SVATOŠOVÁ, Libuše a Bohumil KÁBA. *Statistické metody I*. Vyd. 1. V Praze: Česká zemědělská univerzita, Provozně ekonomická fakulta, 2007, 134 s. ISBN 978-80-213-1672-0.

[14] TUKEY, John Wilder. *Exploratory data analysis*. Reading, Mass.: Addison-Wesley Pub. Co., 1977, 688 s. ISBN 02-010-7616-0.

[15] WIKIPEDIE: *Veličina*. Dostupné na https://cs.wikipedia.org/wiki/Veli%C4%8Dina (cit. 23. 9. 2016).

Přílohy

Příloha 1: Kritické hodnoty Pearsonova korelačního koeficientu

N	Hladina významnosti α - oboustranná alternativa			
	0,20	0,10	0,05	0,01
3	0,951	0,988	0,997	1,000
4	0,800	0,900	0,950	0,99
5	0,687	0,805	0,878	0,959
6	0,608	0,729	0,811	0,917
7	0,551	0,669	0,754	0,875
8	0,507	0,621	0,707	0,834
9	0,472	0,582	0,666	0,798
10	0,443	0,549	0,632	0,765
11	0,419	0,521	0,602	0,735
12	0,398	0,497	0,576	0,708
13	0,380	0,476	0,553	0,684
14	0,365	0,458	0,532	0,661
15	0,351	0,441	0,514	0,641
16	0,338	0,426	0,497	0,623
17	0,327	0,412	0,482	0,606
18	0,317	0,400	0,468	0,590
19	0,308	0,389	0,456	0,575
20	0,299	0,378	0,444	0,561
21	0,291	0,369	0,433	0,549
22	0,284	0,360	0,423	0,537
23	0,277	0,352	0,413	0,526
24	0,271	0,344	0,404	0,515
25	0,265	0,337	0,396	0,505
26	0,260	0,330	0,388	0,496
27	0,255	0,323	0,381	0,487
28	0,250	0,317	0,374	0,479
29	0,245	0,311	0,367	0,471
30	0,241	0,306	0,361	0,463
40	0,207	0,264	0,312	0,403
50	0,184	0,235	0,279	0,361
60	0,168	0,214	0,254	0,330
70	0,155	0,198	0,235	0,306
80	0,145	0,185	0,220	0,286
90	0,136	0,174	0,207	0,270
100	0,129	0,165	0,197	0,256

Příloha 2: Kritické hodnoty Spearmanova korelačního koeficientu

N	Hladina významnosti α - oboustranná alternativa			
	0,20	0,10	0,05	0,01
5	0,800	0,900	1,000	---
6	0,657	0,829	0,886	1,000
7	0,571	0,714	0,786	0,929
8	0,524	0,643	0,738	0,881
9	0,483	0,600	0,700	0,833
10	0,455	0,564	0,648	0,794
11	0,427	0,536	0,618	0,755
12	0,406	0,503	0,587	0,727
13	0,385	0,484	0,560	0,703
14	0,367	0,464	0,538	0,679
15	0,354	0,446	0,521	0,654
16	0,341	0,429	0,503	0,635
17	0,328	0,414	0,488	0,618
18	0,317	0,401	0,472	0,600
19	0,309	0,391	0,460	0,584
20	0,299	0,380	0,447	0,570
21	0,292	0,370	0,436	0,556
22	0,284	0,361	0,425	0,544
23	0,278	0,353	0,416	0,532
24	0,271	0,344	0,407	0,521
25	0,265	0,337	0,398	0,511
26	0,259	0,331	0,390	0,501
27	0,255	0,324	0,383	0,492
28	0,250	0,318	0,375	0,483
29	0,245	0,312	0,368	0,475
30	0,240	0,306	0,362	0,467
35	0,222	0,283	0,335	0,433
40	0,207	0,264	0,313	0,405
45	0,194	0,248	0,294	0,382
50	0,184	0,235	0,279	0,363
55	0,175	0,224	0,266	0,346
60	0,168	0,214	0,255	0,331
70	0,155	0,198	0,235	0,307
80	0,145	0,185	0,220	0,287
90	0,136	0,174	0,207	0,271
100	0,129	0,165	0,197	0,257

Příloha 3: Kritické hodnoty Studentova t rozdělení

df	Hladina významnosti α - oboustranná alternativa			
	0,20	0,10	0,05	0,01
1	3,078	6,314	12,710	63,660
2	1,886	2,920	4,303	9,925
3	1,638	2,353	3,182	5,841
4	1,533	2,132	2,776	4,604
5	1,476	2,015	2,571	4,032
6	1,440	1,943	2,447	3,707
7	1,415	1,895	2,365	3,499
8	1,397	1,860	2,306	3,355
9	1,383	1,833	2,262	3,250
10	1,372	1,812	2,228	3,169
11	1,363	1,796	2,201	3,106
12	1,356	1,782	2,179	3,055
13	1,350	1,771	2,160	3,012
14	1,345	1,761	2,145	2,977
15	1,341	1,753	2,131	2,947
16	1,337	1,746	2,120	2,921
17	1,333	1,740	2,110	2,898
18	1,330	1,734	2,101	2,878
19	1,328	1,729	2,093	2,861
20	1,325	1,725	2,086	2,845
21	1,323	1,721	2,080	2,831
22	1,321	1,717	2,074	2,819
23	1,319	1,714	2,069	2,807
24	1,318	1,711	2,064	2,797
25	1,316	1,708	2,060	2,787
26	1,315	1,706	2,056	2,779
27	1,314	1,703	2,052	2,771
28	1,313	1,701	2,048	2,763
29	1,311	1,699	2,045	2,756
30	1,310	1,697	2,042	2,750
40	1,303	1,684	2,021	2,705
50	1,299	1,676	2,009	2,378
60	1,296	1,671	2,000	2,660
70	1,294	1,667	1,994	2,648
80	1,292	1,664	1,990	2,639
90	1,291	1,662	1,987	2,632
100	1,290	1,660	1,984	2,626

Autor: Ing. Jarmila Drozdová, Ph.D.
doc. Dr. Vladimír Homola, Ph.D.

Katedra, institut: Institut geologického inženýrství

Název: Statistika pro Geovědní a montánní turismus (Učební texty předmětu Statistika a informatika - část Statistika)

Místo, rok, vydání: Ostrava, 2017, 1. vydání

Počet stran: 108

Vydala: Vysoká škola báňská-Technická univerzita Ostrava

Tisk: Institut geologického inženýrství 541

Náklad: 30 ks

Neprodejné

ISBN 978-80-248-4067-3