Statistika v kinantropologii

Mgr. Martin Sebera, Ph.D.

Úvod

Práce předkládá možnosti využití základního matematicko-statistického aparátu pro potřeby studia v kinantropologickém výzkumu na Fakultě sportovních studií. Možnost statistického zpracování dat je stále velmi dynamicky se rozvíjející disciplína. S rostoucí dostupností výkonné měřící a výpočetní techniky se statistické metody začaly ve stále větším měřítku prosazovat v běžné kinantropologické praxi. Smyslem předloženého materiálu je porozumět mechanismu, na jehož základě jsou základní statistické výpočty prováděny. Přiměřená znalost statistiky pak pomůže studentům lépe chápat zákonitosti naměřených dat. Aplikovat statistické metody a postupy znamená zaznamenávat data o jevech a zpracovávat je, tj. třídit, vyhodnocovat a interpretovat. Statistika se tak nachází v úzkém kontaktu s informačními technologiemi (informatika, výpočetní technika).

Cíl materiálu

Na konci tohoto kurzu bude student schopen:

- porozumět a vysvětlit základní statistické charakteristiky;
- použít testování hypotéz na vybraných případech;
- interpretovat výsledky na reálných příkladech (srovnání dvou skupin, hledání a popis závislostí, zhodnocení statistického modelu);

Text je rozvržen do 7 kapitol. 1 se zabývá základními termíny a odbornými pojmy, kapitola 2 pak rozložením četnosti a grafickým znázorněním dat, 3. základními statistickými charakteristikami, 4. se věnuje obecně testování hypotéz, testy normality a t-testy. 5. Kapitola popisuje základní neparametrické testy, 6. Zmiňuje parametrický i neparametrický výpočet korelačního koeficientu a poslední se věnuje základům lineární regrese.

Pro názorné ukázky příkladu a řešení byl použit sw Statistica 10 CZ, ke kterému mají přístup všichni studenti Masarykovy univerzity. Text označený zelenou barvou nabízí přesný postup při průchodu kontextovým menu v sw Statistica.

http://www.muni.cz/ics/services/software/statistica?lang=cs

Poznámky na úvod

Dvojí zápis dat v sw STATISTICA.

Za drobnou nevýhodu práce v sw STATIATICA lze považovat dva různé zápisy dat, které se používají při jednotlivých výpočtech. Prvním je způsob zápisu **po proměnných**: každý sloupec obsahuje popis proměnné a ve sloupci pod ním pak hodnoty dané proměnné. Druhým způsobem je zápis proměnné společně s tzv. grupovací proměnnou. V prvním sloupci je identifikátor grupovací proměnné a ve druhém sloupci jsou vlastní data – tab. 1.

Tab. 1 Odlišné způsoby zápisu stejných hodnot vybraných proměnných

Zápis po proměnných

délka-skupina 1	**délka-skupina 2**
68	142
199	179
158	138

Zápis s grupovací proměnnou

skupina	**délka**
skupina1	68
skupina1	199
skupina1	158
skupina2	142
skupina2	179
skupina2	138

Ikonky k porozumění a rychlejší orientace v textu (http://openclipart.org/search/?query=smile)

Začátek teorie 

Konec teorie 

Řešený příklad 

Důležitá informace hodna zapamatování 

1 Základní a výběrový soubor

Základní soubor

Představuje soubor všech prvků (osob, objektů, …), které mohou být teoreticky předmětem sledování a u kterých se objevuje sledovaný znak. Počet členů této množiny označujeme N a toto N je obvykle velmi velké, může být i nekonečno (v případě sledování proměnných v čase). Výzkumník zpravidla provádí svá šetření na mnohem menší množině, tzv. výběrovém souboru, který je vybrán některou z metod (výběr náhodný nebo záměrný stratifikovaný) vždy však způsobem, aby výběr co nejlépe reprezentoval základní soubor.

Výběrový soubor

Náhodný výběr představuje takovou techniku, kdy každý prvek má stejnou pravděpodobnost výběru, nebo má stejnou pravděpodobnost dostat se do výběrového souboru. Prakticky se náhodný výběr provádí pomocí losování nebo generátorem náhodných čísel (software), výjimečně tabulkou náhodných čísel.

Příklad 1

Základní soubor všech atletů v ČR je definován svým registrem, který je dostupný na webu Českého atletického svazu. Výběrový soubor můžeme vytvořit 3 způsoby:

1. *náhodný systematický výběr*. Do výběrového souboru vložíme každého k-tého člena z pořadí všech atletů, kteří byli předtím seřazení podle svého příjmení. Hodnota *k* je určena předem.
2. *náhodný vícestupňový výběr*. Typicky ho provádíme ve více krocích (dva či více). Např. nejprve seřadíme seznam atletů podle příslušnosti ke krajům, poté v každé takové skupině provedeme náhodný výběr.
3. *náhodný stratifikovaný výběr.* Základní soubor rozdělíme podle jasného kritéria a jedinci jsou vybíráni do vzorku náhodně z těchto skupin (např. vybereme disciplíny a věkovou kategorii a poté provedeme náhodný výběr).

Statistické jednotky

Jsou prvky statistického souboru, které mají alespoň jednu společnou vlastnost. U statistických jednotek zjišťujeme statistické znaky.

Statistický znak

Je označením určité vlastnosti jednotky, kterou zkoumáme. Hodnota statistického znaku představuje míru sledované vlastnosti

Znaky, které nabývají více než jedné varianty, označujeme a nazýváme **proměnnými**.

Typy proměnných

Při statistické analýze potřebujeme u každé proměnné určit její typ. Můžeme se setkat s několika způsoby klasifikace proměnných, v našem textu popisujeme přístup, který za hlavní kritérium považuje *typy vztahů mezi hodnotami*. Podle Řezankové (2005) u tohoto hlediska rozlišujeme proměnné:

- **Nominální.** Hodnotou je číslo nebo text. U těchto proměnných můžeme provádět jen rozdělení četností, případně operaci porovnání. ***Příklad***: student absolvoval motorický test „běh na 50 m“ s výkonem 7,4 s a motorický test „leh-sed s výsledkem 50 opakování za minutu. Číselné hodnoty 7,4 a 50 určují jen odlišné výsledků motorických testů, nic jiného se vyčíst nedá
- **Ordinální znaky** umožňuje provádět srovnání a tím určit pořadí. V případě textových proměnných je nutné tyto převést na čísla. ***Příklad***: v dotaznících vyjadřujeme míru souhlasu s daným tvrzením. Svou kondicí hodnotím jako: *vynikající – velmi dobrou – dobrou – slabou – špatnou*. Výroky respondentů můžeme určit pořadí, jak který respondent souhlasí s tvrzením. Však netvrdíme, že rozdíl mezi odpověďmi *vynikající* a *velmi dobrou* je stejný jako mezi *slabou* a *špatnou*.
- **Intervalové** kromě porovnání můžeme provádět operaci součtu a rozdílu. ***Příklad***: výška a hmotnost jedince. Naměříme-li u batolete výšku v cm po čtyřech měsících hodnoty 60, 62, 64, 66, znamená to, že každým měsícem dítě vyrostlo o 2 cm.
- **Poměrové znaky** umožňují interpretovat kromě operace rovnosti, uspořádání a rozdílu ještě operace podílu a součinu. ***Příklad***: zaběhne-li atlet 100 m za 11 s a druhý atlet za 22 s, je možné prohlásit, že první je dvakrát rychlejší než druhý.

Nominální a ordinální proměnné jsou souhrnně označovány jako *kvalitativní*; intervalové a poměrové proměnné jsou souhrnně označovány jako *kvantitativní* (numerické, kardinální). Kvantitativní proměnné můžeme podle jiného hlediska dělit na

- *diskrétní*, které nabývají pouze celočíselných obměn (*počet permanentek do posilovny*) a
- *spojité* (*metrické*), jež mohou nabývat libovolných hodnot z určitého intervalu (věk respondenta, výkon ve vrhu koulí).

Nominální, ordinální a kvantitativní diskrétní proměnné můžeme souhrnně označit jako **kategoriální** (obměny těchto proměnných nazýváme kategoriemi).

- *dichotomické* (*alternativní*), které nabývají pouze dvou kategorií (ekonomicky aktivní a neaktivní, kuřák a nekuřák), a
- *vícekategoriální* (*množné*), jež nabývají více než dvou kategorií (rodinný stav, obor).

Kontrolní otázky a úkoly

1. Uveďte 3 základní a výběrové soubory ze sportovního prostředí
2. Definujte typy proměnných
3. U každého typu proměnných uveďte 3 ze sportovního prostředí
4. Uveďte proměnné, která jsou kategoriální, ale nejsou ordinální
5. Jakého typu je proměnná „pohlaví“?

2 Průzkumová analýza dat

Účelem průzkumové**Chyba! Záložka není definována.** analýzy**Chyba! Záložka není definována.** dat**Chyba! Záložka není definována.Chyba! Záložka není definována.** (exploratory data analysis) je odhalit jejich zvláštnosti a ověřit předpoklady pro následné statistické zpracování. O výběru předpokládáme, že jeho rozdělení je normální a data splňují předpoklady nezávislosti**Chyba! Záložka není definována.Chyba! Záložka není definována.** a homogenity**Chyba! Záložka není definována.Chyba! Záložka není definována.**. V této kapitole se omezíme pouze na **ověřování normality**Chyba! Záložka není definována. **a homogenity**Chyba! Záložka není definována.Chyba! Záložka není definována. **dat**.

K ověření těchto předpokladů budeme používat **především grafické metody,** které vděčí za svůj vznik především rozvoji počítačové grafiky. Tak zvaná **vizualizace dat**Chyba! Záložka není definována. je důležitým nástrojem při analýze dat, regresní analýze, analýze časových řad a dalších.

Statistické programy nabízejí více grafů pro identifikaci statistických zvláštností dat. Následuje stručný popis nejpoužívanějších grafů průzkumové**Chyba! Záložka není definována.** analýzy**Chyba! Záložka není definována.**.

2.1 Frekvenční grafy

Jedním ze základních typů grafického zobrazení dat jsou tzv. **frekvenční (četnostní) grafy,** které zobrazují informace obsažené ve frekvenční tabulce (tabulka rozdělení četností). Mezi nejznámější frekvenční grafy**Chyba! Záložka není definována.** patří **histogram**Chyba! Záložka není definována.**y a polygony četností.**

Před konstrukcí těchto grafů musíme výběr roztřídit, tj. **rozdělit n hodnot do *k* intervalů (tříd).** Statistické programy počet tříd a jejich šířku většinou nabízí. Nabízené hodnoty můžeme ve většině případů měnit. Při volbě intervalů můžeme například postupovat tak, že vhodně zaokrouhlíme dolů nejmenší hodnotu a nahoru největší hodnotu. Takto vzniklý rozsah hodnot pak dělíme na počet intervalů**Chyba! Záložka není definována.**, který bývá blízký deseti. Praktické hledisko je zde to, aby středy nebo hranicemi intervalů byla „rozumná“ celá čísla.

Ať kreslíme histogram**Chyba! Záložka není definována.** sami nebo pomocí počítače, musíme vždy zvolit z několika možností. Vzhledem k tomu, že často nevíme a priori, co můžeme v datech čekat, jsou tato rozhodnutí založena na zkušenosti.

Sturgesovo pravidlo**Chyba! Záložka není definována.** doporučuje volit počet intervalů**Chyba! Záložka není definována.** ***k*** podle vztahu

$$k \approx 1 + 3{,}3 \cdot \log_{10}(n)$$

kde n je rozsah výběru.

Pro přibližně symetrická rozdělení výběru lze počet intervalů**Chyba! Záložka není definována.** (tříd) ***k*** počítat podle vztahu

$$k \approx \left\lfloor 2\sqrt{n} \right\rfloor$$

kde n je rozsah výběru a výraz $\left\lfloor 2\sqrt{n} \right\rfloor$ označuje celočíselnou část čísla v závorce.

Výhodou uvedených vzorců je sice standardizace, ale nevýhodou např. to, že hranice a středy tříd mohou být „nehezká“ čísla. Nutno dělat kompromisy mezi navrhovaným počtem

intervalů a „rozumnými“ hranicemi a středy intervalů. Neexistuje universální postup, vždy se musí vycházet z konkrétních dat.

Rovněž **šířka intervalů**Chyba! Záložka není definována. závisí na povaze zkoumaných dat. Většinou pracujeme s **intervaly stejné délky** (mluvíme o ekvidistantním**Chyba! Záložka není definována.** dělení). Nestejnou šířku tříd (neekvidistantní**Chyba! Záložka není definována.** dělení) volíme jen ve zdůvodnitelných případech (např. při velmi zešikmeném rozdělení výběru), neboť komplikuje další práci s rozdělením četností.

Tabulku rozdělení četností znázorňujeme graficky pomocí histogram**Chyba! Záložka není definována.**u**Chyba! Záložka není definována.**. V **histogramu** každé třídě odpovídá obdélník, jehož výška je rovna četnosti a šířka třídnímu rozpětí.

Histogramy možno v jednotlivých programech **doplnit křivkou hustoty** (v případě absolutních nebo relativních četností) **nebo distribuční funkcí** (v případě kumulativních četností**Chyba! Záložka není definována.**) některého teoretického rozdělení. Pomocí takto doplněných grafů rozhodujeme o vhodnosti proložení těmito funkcemi.

Vzhledem k subjektivní podstatě třídění histogram**Chyba! Záložka není definována.**y umožňují pouze odhady (někdy dosti nepřesné) tvaru funkce hustoty pravděpodobnostního rozdělení, typických vlastností a individuálních zvláštností dat.

2.2 Krabicové grafy

Krabicový graf**Chyba! Záložka není definována.** (Box and whiskers plot) zobrazuje data ve formě **obdélníku (krabice)**, z něhož vybíhají úsečky (vousy). V nejjednodušší variantě představuje krabicový graf obdélník o délce rovné kvartilovému rozpětí**Chyba! Záložka není definována.** a s vhodně zvolenou šířkou, která je úměrná $\sqrt{n}$. **V místě mediánu je vertikální čára**.

Z obou protilehlých stran tohoto obdélníku vycházejí **vodorovné úsečky (vousy)** ukončené hodnotami $B_D^{'}, B_H^{'}$, které leží uvnitř tzv. vnitřních hradeb B_D, B_H, kde

$$B_D = \tilde{x}_{0.25} - 1{,}5.(\tilde{x}_{0.75} - \tilde{x}_{0.25}) \quad (1)$$

$$B_H = \tilde{x}_{0.75} + 1{,}5.(\tilde{x}_{0.75} - \tilde{x}_{0.25}) \quad (2)$$

Hodnoty vně intervalu $(B_D^{'}, B_H^{'})$ jsou tzv. **vybočující hodnoty**Chyba! Záložka není definována.Chyba! Záložka není definována.**.**

Vybočující pozorování mohou být chybná nebo odlehlá. Chybné hodnoty**Chyba! Záložka není definována.**, pokud identifikujeme zdroj chyb, je nutné vždy vyloučit nebo opravit. U odlehlých hodnot závisí rozhodnutí na tom, zda nalezneme nějaký důvod, proč daná hodnota není projevem přirozené variability.

Krabicové grafy tedy dávají kondenzovanou vizuální informaci o úrovni (medián) a variabilitě (kvartilové rozpětí) dat. Umožňují orientační posouzení symetrie dat a přítomnosti vybočujících měření.

2.3 Grafy pro ověření normalityChyba! Záložka není definována. výběru

Pro zjišťování odchylek rozdělení výběru od normálního rozdělení slouží jednak **porovnání histogramChyba! Záložka není definována.u a křivky hustoty normálního rozdělení,** jednak různé **grafy pro ověření normalityChyba! Záložka není definována. výběru**. V těchto grafech jsou vyneseny hodnoty empirické (vyznačené body) a teoretické (vyznačené plnou čarou) distribuční funkce. Na základě velikosti odchylek od lineárního průběhu můžeme posoudit míru porušení předpokladu normality**Chyba! Záložka není definována.**.

2.4 Příklad

Data uvedená v tabulce 1 (získané body na přijímací zkoušce) podrobte průzkumové**Chyba! Záložka není definována.** (exploratorní) analýze**Chyba! Záložka není definována.**.

Tab. 1 Body získané v přijímacím řízení

17 12 15 16 18 17 18 9 12 15 16 11 16 17 18 12 14 20 21 17 11 14 15 20 14 12 16 15 14 15 16 15

Průzkumová analýza ve STATISTICE

a) Frekvenční grafy

Tabulku intervalového rozdělení četností nabízí program v modulu Statistika. Postup příkazů je následující:

Potup: Statistika – Základní statistiky a tabulky – Popisné statistiky – Tabulky četností

Výsledkem této procedury je textový výstup (tab. 2), který obsahuje frekvenční tabulku

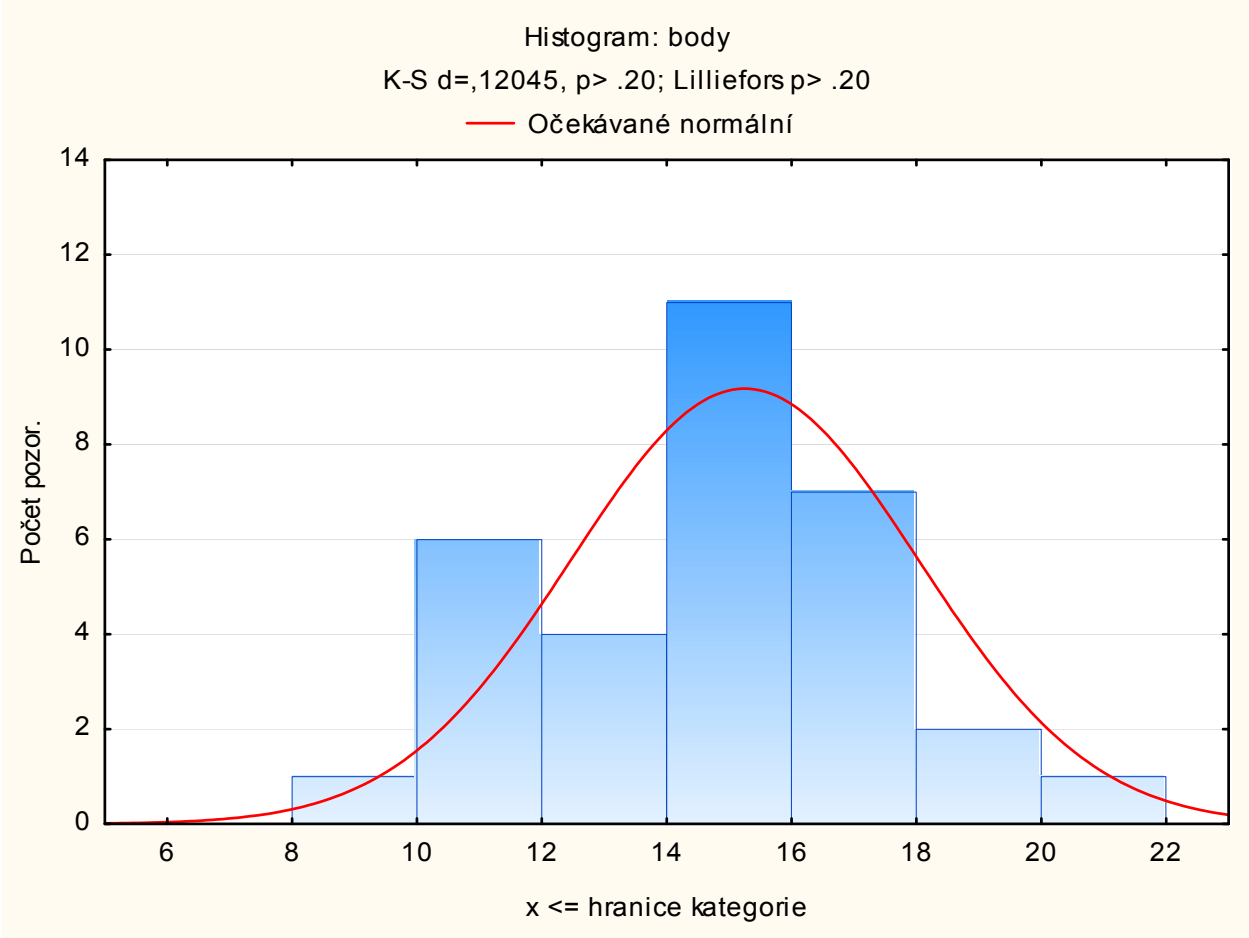
Tab. 2 Tabulka četností: body

	Četnost	Kumulativní četnost	Rel.četn. (platných)	Kumul. % (platných)	Rel.četn. všech	Kumul. % všech
6<x<=8	0	0	0,00000	0,0000	0,00000	0,0000
8<x<=10	1	1	3,12500	3,1250	3,12500	3,1250
10<x<=12	6	7	18,75000	21,8750	18,75000	21,8750
12<x<=14	4	11	12,50000	34,3750	12,50000	34,3750
14<x<=16	11	22	34,37500	68,7500	34,37500	68,7500
16<x<=18	7	29	21,87500	90,6250	21,87500	90,6250
18<x<=20	2	31	6,25000	96,8750	6,25000	96,8750
20<x<=22	1	32	3,12500	100,0000	3,12500	100,0000

ChD	0	32	0,00000		0,00000	100,0000

Graf histogram**Chyba! Záložka není definována.**u získáme pomocí příkazů

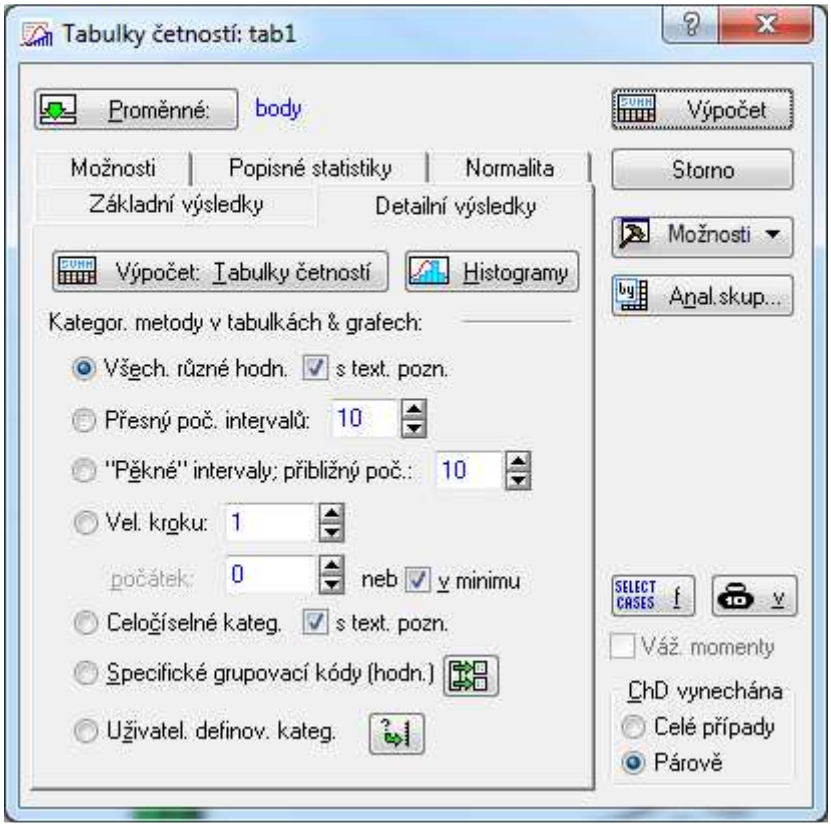
Potup: Statistika – Základní statistiky a tabulky – Tabulky četností – Histogram



Obr. 1 Histogram

Počet tříd můžeme měnit podle dialogového okna v záložce Detailní výsledky (obr. 2).

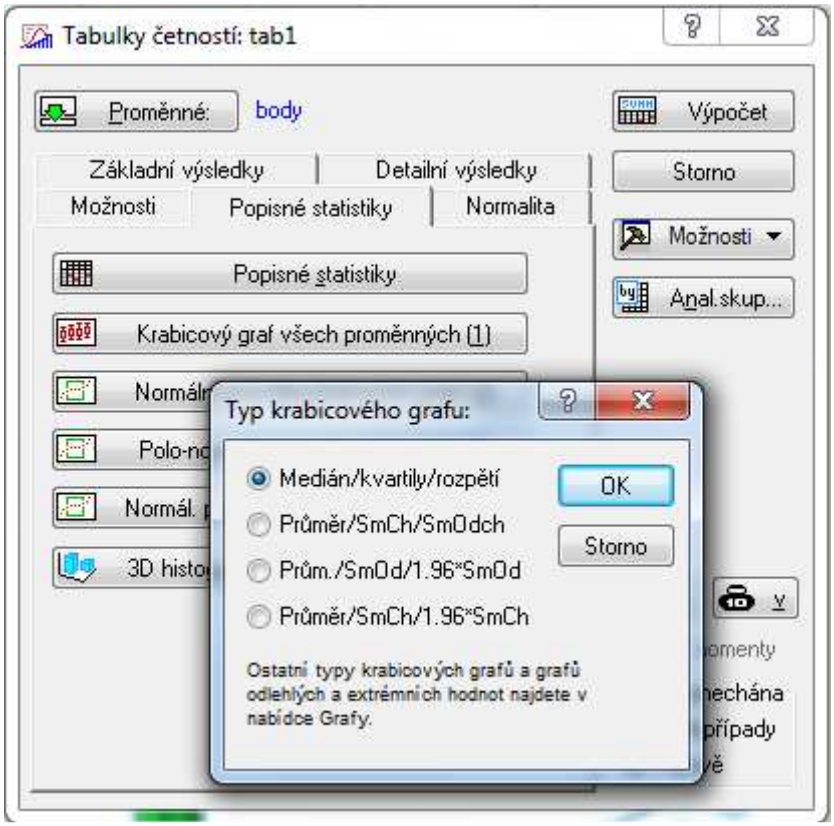
Potup: Statistika – Základní statistiky a tabulky – Tabulky četností – Detailní výsledky



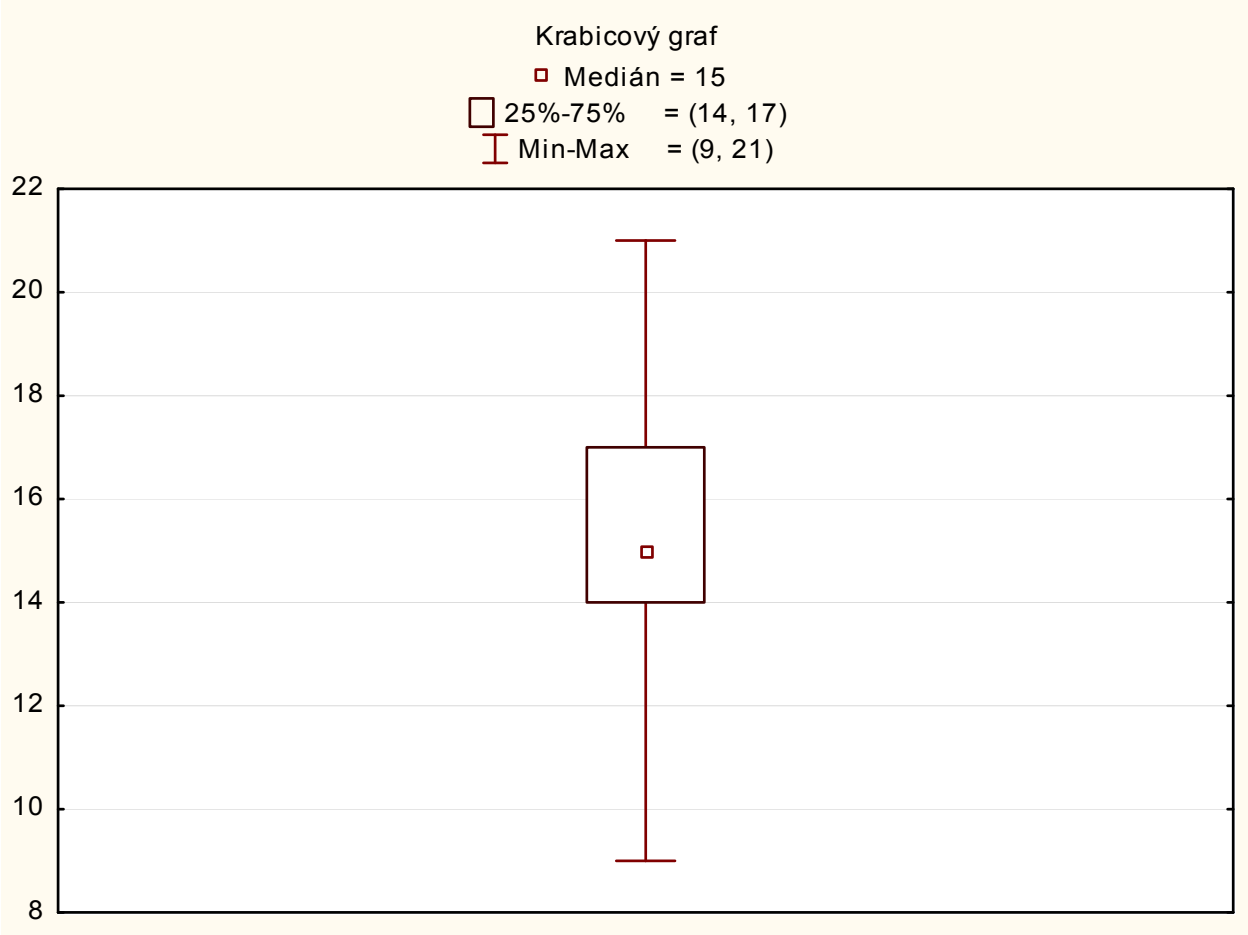
Obr. 2 Dialog pro změnu tříd u histogramů

Prostý krabicový graf (obr. 4) získáme příkazy (obr. 3)

Potup: Statistika – Základní statistiky a tabulky – Tabulky četností – Popisné statistiky – Krabicové grafy všech proměnných



Obr. 3 Postup vykreslení krabicového grafu



Obr. 4 Krabicový graf

Závěr příkladu

Histogramy a krabicové grafy ukázaly, že rozdělení výběru je zešikmené k vyšším hodnotám. Normální pravděpodobnostní graf**Chyba! Záložka není definována.**y signalizují, že odchylky výběru od normálního rozdělení jsou malé a podporují předpoklad, že analyzovaná data můžeme považovat za **výběr z normálního rozdělení**.

Grafické metody exploratorní analýzy dat**Chyba! Záložka není definována.Chyba! Záložka není definována.** jsou sice názorné, ale výsledky nejsou tak jednoznačné jako při numerickém zpracování. **Čtení grafů vyžaduje znalosti a zkušenosti.**

Při řešení příkladů většinou předpokládáme, že analyzovaná data mají normální rozdělení. Naskýtá se otázka, jak postupovat v případě **nesplnění předpokladu normality**Chyba! Záložka není definována. **dat**. Jde o případy, kdy rozdělení dat je jiné než normální, nebo jsou v datech vybočující měření. V případě symetrických rozdělení používáme pro odhady parametrů polohy a rozptýlení **robustní techniky**, nebo vyzkoušíme **mocninnou transformaci dat**. U sešikmených rozdělení je vždy výhodné začít mocninnou transformací. Není-li mocninná transformace úspěšná**, hledáme vhodné aproximující rozdělení**. Tyto postupy však převyšují obsah našeho základního kurzu statistiky. Podrobný výklad postupu při nesplnění předpokladů nezávislosti, homogenity**Chyba! Záložka není definována.Chyba! Záložka není definována.** a normality**Chyba! Záložka není definována.** dat najde zájemce např. v knize Meloun & Militký, 1998).

Kontrolní otázky a úkoly

1. Jaké informace obsahuje krabicový graf?
2. Jaký je rozdíl mezi procentuálním vyjádřením a vyjádřením relativních četností?
3. Co definuje Sturgesovo pravidlo?
4. Jak může vzniknout odlehlá / extrémní hodnota?
5. Co je na ose X a na ose Y u histogramu?

3 Základní statistické charakteristiky

3.1 Popisná statistika

Procedury popisné statistiky použijeme k prvotnímu posouzení předložených dat. Nejčastěji používané statistické charakteristiky jsou

- aritmetický průměr $$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (3)$$

Definice následujících charakteristik předpokládají uspořádaný výběr, tj. $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$

- minimální hodnota $x_{min} = x_{(1)}$
- maximální hodnota $x_{max} = x_{(n)}$
- medián $\tilde{x}_{0,50}$ pro n sudé $\tilde{x}_{0,50} = \frac{x_{(n/2)} + x_{(n/2+1)}}{2}$, (4)
  pro n liché $\tilde{x}_{0,50} = x_{(n+1)/2}$
- dolní kvartil $\tilde{x}_{0,25} = x_{(k)}$, kde pro pořadový index *k* platí
  $n \cdot 0{,}25 < k < n \cdot 0{,}25 + 1$
- horní kvartil $\tilde{x}_{0,75} = x_{(k)}$, kde pro pořadový index *k* platí
  $n \cdot 0{,}75 < k < n \cdot 0{,}75 + 1$

Charakteristiky variability

- variační rozpětí $R = x_{max} - x_{min}$ (5)
- kvartilové rozpětí $R_Q = \tilde{x}_{0,75} - \tilde{x}_{0,25}$ (6)
- výběrový rozptyl $s^2_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ (7)
- výběrová směrodatná odchylka $s_{n-1} = \sqrt{s^2_{n-1}}$ (8)
- variační koeficient $v = \frac{s_n}{|\bar{x}|}$ nebo $v = \frac{s_{n-1}}{|\bar{x}|}$ (9)

Charakteristiky kategoriální proměnné

- *Modus* - hodnota nejčetnější kategorie
- *Četnost* - počet pozorování spadajících do příslušné kategorie
- Stanovení četností – absolutní a relativní

3.2 Příklad

Máme k dispozici údaje o 569 respondentech a to konkrétně proměnnou BMI (Body Mass Index) a poměru pas/boky – viz příloha 1. Vypočítáme základní statistické charakteristiky (tab. 3).

Postup: Statistika – Základní statistiky a tabulky – Popisné statistiky – Detailní výsledky

Tab. 3 Výsledky popisné statistiky

	Poměr Pas/boky	**BMI**	Popis
N platných	569	569	počet hodnot
Aritmetický průměr	0,88	25,01	• statistická veličina, která v jistém smyslu vyjadřuje typickou hodnotu popisující soubor mnoha hodnot • nejčastější chybou je aplikace aritmetického průměru tam, kde je na místě využít jinou statistiku. Např. aritmetickým průměrem souboru { 1, 1, 1, 1, 16 } je 4, přestože čtyři z pěti hodnot tohoto souboru je menších. V obdobných případech je mnohem vhodnější použít pro vyjádření typické hodnoty medián (který je u této množiny roven 1, což je mnohem lepší popis střední hodnoty)
Minimum	0,60	16,77	• nejmenší hodnota
Maximum	1,03	40,67	• nejvyšší hodnota
 **Medián**	0,88	24,39	• medián (označován Me nebo $\tilde{x}$) je hodnota, jež dělí řadu podle velikosti seřazených výsledků na dvě stejně početné poloviny. • není ovlivněný extrémními hodnotami. • medián lze definovat na každém souboru uspořádaném relací „menší nebo rovno“, i když se nejedná o soubor čísel. Například medián souboru {absolvent ZŠ, vyučen, vyučen s maturitou, vysokoškolák} je roven hodnotě „vyučen“, pokud kategorie vzdělání považujeme za seřazené podle náročnosti školy.
Spodní kvartil	0,84	22,16	• kvartily oddělují ze statistického souboru čtvrtiny. Rozlišuje se spodní kvartil $Q_{0,25}$ a horní kvartil $Q_{0,75}$. Data předpokládají uspořádaný výběr.
Horní kvartil	0,92	27,36	
Rozpětí	0,43	23,9	• rozdíl mezi maximem a minimem
Kvartilové rozpětí	0,08	5,2	• pomocí horního a spodního kvartilu lze zavést mezikvartilové rozpětí, které definujeme jako hodnotu $Q_{0,75} - Q_{0,25}$.
Rozptyl	0,00284	15,40	• rozptyl - jedná se o charakteristiku variability rozdělení pravděpodobnosti náhodné veličiny, která

			vyjadřuje variabilitu rozdělení souboru kolem střední hodnoty.
Směrodatná odchylka	0,0533	3,92	• jedná se o kvadratický průměr odchylek hodnot znaku od jejich aritmetického průměru. Vypovídá o tom, jak moc se od sebe navzájem liší typické případy v souboru zkoumaných čísel. Je-li malá, jsou si prvky souboru většinou navzájem podobné, a naopak velká směrodatná odchylka signalizuje velké vzájemné odlišnosti.
Variační koeficient	6,05	15,69	• variační koeficient je použitelný i při porovnávání variability proměnných, které jsou v různých jednotkách

Kontrolní otázky a úkoly

1. Uveďte rozdíl mezi aritmetickým průměrem a mediánem.
2. Uveďte nevýhody aritmetického průměru
3. Uveďte výhody aritmetického průměru
4. Jaký je rozdíl mezi rozptylem a směrodatnou odchylkou?
5. K čemu slouží variační koeficient?

4 Testování hypotéz

V této kapitole se budeme zabývat odhady parametrů normálního rozdělení**Chyba! Záložka není definována.** $N(\mu,\sigma^2)$ a testováním hypotéz o těchto parametrech. Rovněž budeme testovat hypotézu o shodě normálního a empirického rozdělení.

4.1 Odhady parametrů

Bodovými odhady Chyba! Záložka není definována. parametrů μ a σ^2 jsou výběrový průměr $\bar{x}$ a výběrový rozptyl s_{n-1}^2. Ze statistického hlediska nemají bodové odhady velký význam, protože hodnoty výběrových statistik kolísají kolem neznámého parametru. Více informací poskytují **intervalové odhady**Chyba! Záložka není definována. **(konfidenční intervaly**Chyba! Záložka není definována.**)**, které určují interval, v němž se zadanou pravděpodobností (1-α) se nachází skutečná hodnota daného parametru. **Koeficient spolehlivosti**Chyba! Záložka není definována. **(1-α)** se obvykle volí 0,95 nebo 0,90 případně 0,99. Parametr α se nazývá **hladina statistické významnosti**Chyba! Záložka není definována.**.**

100(1-α)% intervaly spolehlivosti pro střední hodnotu Chyba! Záložka není definována. μ

$$< \bar{x} - t_{1-\alpha/2}(n-1)\frac{s_{n-1}}{\sqrt{n}}; \bar{x} + t_{1-\alpha/2}(n-1)\frac{s_{n-1}}{\sqrt{n}} > \text{ ... oboustranný interval}$$

Chyba! Záložka není definována. (10)

$$< -\infty; \bar{x} + t_{1-\alpha}(n-1)\frac{s_{n-1}}{\sqrt{n}} > \text{ ... horní odhad}$$

Chyba! Záložka není definována. (11)

$$< \bar{x} - t_{1-\alpha}(n-1)\frac{s_{n-1}}{\sqrt{n}}; +\infty > \text{ ... dolní odhad}$$

Chyba! Záložka není definována. (12)

100(1-α)% intervaly spolehlivosti pro rozptyl Chyba! Záložka není definována. σ^2

$$< \frac{(n-1)s_{n-1}^2}{\chi_{1-\alpha/2}^2(n-1)}; \frac{(n-1)s_{n-1}^2}{\chi_{\alpha/2}^2(n-1)} > \text{ ... oboustranný interval}$$

Chyba! Záložka není definována. (13)

$$< 0; \frac{(n-1)s_{n-1}^2}{\chi_{\alpha}^2(n-1)} > \text{ ... horní odhad}$$

Chyba! Záložka není definována. (14)

$$< \frac{(n-1)s_{n-1}^2}{\chi_{1-\alpha}^2(n-1)}; \infty > \text{ ... dolní odhad}$$

Chyba! Záložka není definována. (15)

4.2 Testování hypotéz

Testování hypotéz jsou klasické statistické úsudky založené na nějakém apriorním předpokladu. Vyslovíme-li předpoklad o hodnotě neznámého parametru nebo o zákonu rozdělení sledované náhodné veličiny, vyslovíme tak **statistickou hypotézu**Chyba! Záložka není definována.. Ověřování, zda hypotéza platí či nikoliv, je předmětem testování, které provádíme na základě nějakého výběru (měření, pozorování).

Test statistické hypotézy Chyba! Záložka není definována. **H proti alternativní hypotéze**Chyba! Záložka není definována. **A** je pravidlo, podle něhož na základě náhodného výběru rozhodneme mezi dvěma tvrzeními - sledovanou hypotézou H a alternativní hypotézou A. Výsledkem našeho rozhodování je buď zamítnutí hypotézy H ve prospěch alternativy A či její nezamítnutí. Skutečnost, že hypotézu nezamítáme, neznamená, že naměřená data tuto hypotézu potvrzují, ale pouze to, že ji nevyvracejí.

Ve většině programů je testovaná hypotéza označovaná jako **nulová hypotéza**Chyba! Záložka není definována. **H_0 a alternativní hypotéza H_1 nebo H_A.**

Rozhodovací pravidlo je určeno **testovou statistikou (testovým kritériem**Chyba! Záložka není definována.**)** T(**X**) a intervalem $\mathbf{W_\alpha}$, kterému říkáme **kritický obor**Chyba! Záložka není definována.**.** Kritický obor je ohraničený tzv. **kritickými hodnotami**Chyba! Záložka není definována., což jsou **kvantily**Chyba! Záložka není definována. rozdělení příslušného testového kritéria. Jestliže hodnota testové statistiky $T(\mathbf{X}) \in W_\alpha$, potom hypotézu H zamítáme.

Při testování hypotéz se můžeme dopustit chyby dvěma způsoby: Buď zamítneme hypotézu, která platí - to je **chyba prvního druhu**Chyba! Záložka není definována. α - nebo naopak tuto hypotézu nezamítneme i když je nesprávná - v tomto případě se jedná o **chybu druhého druhu**Chyba! Záložka není definována. β. Při konstrukci testu požadujeme, aby pravděpodobnost chyby 1. druhu byla menší nebo rovna danému číslu α, kterému říkáme **hladina významnosti**Chyba! Záložka není definována. **testu.** Obvykle volíme $\alpha = 0{,}05$ nebo 0,1.

Testování probíhá tak, že vypočítáme hodnotu testové statistiky, porovnáme ji s kritickými hodnotami**Chyba! Záložka není definována.**, odpovídajícími hladině významnosti α, a rozhodneme o zamítnutí či nezamítnutí hypotézy H.

Při **testování pomocí statistických programů** se používá jiný postup: Spočte se hodnota testové statistiky a k ní nejmenší kritický obor**Chyba! Záložka není definována.**, při kterém bychom ještě mohli na základě této hodnoty zamítnout hypotézu H_0 proti dané alternativě. Hladina významnosti, odpovídající tomuto kritickému oboru, se nazývá **minimální hladina významnosti**Chyba! Záložka není definována.Chyba! Záložka není definována. (**p-hodnota**Chyba! Záložka není definována.**)**.

Pokud je $p > \alpha$, pak hypotézu H_0 nezamítáme. V opačném případu, kdy $p \leq \alpha$, pak hypotézu H_0 zamítáme.

Věcná významnost

- používání nestatistického hodnocení velikosti rozdílu či vztahu ve výzkumných výsledcích, tzv. „size of effect“, zvláště pomocí tzv. koeficientu ω^2 jakožto podílu, resp. procenta vysvětleného rozptylu
- Např. ke kvantifikování velikosti účinku, tj. k hodnocení věcné významnosti je možné použít *Cohenův koeficient účinku d.* Jednou z hlavních výhod koeficientu je jeho nezávislost na rozsahu výběru. Platí pro něj konvenční hodnoty, jež usnadňují rozhodnutí, kdy lze hovořit o velkém efektu. Pokud je *d* větší než 0,8, je efekt velký; pro *d* z intervalu 0,5 – 0,8 je efekt střední; efekt pod hodnotou 0,2 lze považovat za malý.

Postup při práci s hypotézami by měl vypadat následovně: 1. nejprve zhodnotit věcnou významnost jak absolutně (v jednotkách měření), tak i relativně k podílu vlivu ostatních faktorů (pomocí ω^2), a jen jde-li o randomizovaný výzkum pak 2. použít statistickou významnost α jakožto riziko zobecnění.

4.2.1 Jednovýběrové testyChyba! Záložka není definována.

*Testy o střední hodnotě**Chyba! Záložka není definována.***

Jestliže testujeme hypotézu H: $\mu = \mu_0$ proti alternativní hypotéze**Chyba! Záložka není definována.** A, pak testovým kritériem**Chyba! Záložka není definována.** je statistika

$$t = \frac{\overline{x} - \mu_0}{s_{n-1}} \sqrt{n}\,, \tag{16}$$

která má při platnosti hypotézy H: $\mu = \mu_0$ Studentovo rozdělení**Chyba! Záložka není definována.** t(n-1). Kritické obory jsou dány vztahy (17) - (19).

alternativa	kritický obor**Chyba! Záložka není definována.**	
$A_1: \mu > \mu_0$	$W_1 = \{t; t \geq t_{1-\alpha}(n-1)\}$	(17)
$A_2: \mu < \mu_0$	$W_2 = \{t; t \leq -t_{1-\alpha}(n-1)\}$	(18)
$A_3: \mu \neq \mu_0$	$W_3 = \{t; \lvert t\rvert \geq t_{1-\alpha/2}(n-1)\}$	(19)

4.2.2 Dvouvýběrové testyChyba! Záložka není definována.

a) *Test homogenity dvou rozptylů**Chyba! Záložka není definována.***

Testujeme hypotézu H: $\sigma_1^2 = \sigma_2^2$ proti alternativě $A_3: \sigma_1^2 \neq \sigma_2^2$. Testovým kritériem je statistika

$$F = \frac{s_1^2}{s_2^2}\,, \tag{20}$$

která má za předpokladu správnosti hypotézy $H: \sigma_1^2 = \sigma_2^2$ rozdělení F**Chyba! Záložka není definována.**$(n_1 - 1, n_2 - 1)$.

Hypotézu H zamítneme na hladině významnosti α ve prospěch alternativy $A: \sigma_1^2 \geq \sigma_2^2$, jestliže

$$F \geq F_{1-\alpha}(n_1 - 1, n_2 - 1). \tag{21}$$

b) *Testy o středních hodnotách dvou výběrů**Chyba! Záložka není definována.***
s homogenními rozptyly

Jestliže nezamítneme hypotézu o homogenitě rozptylů obou výběrů, pak testujeme hypotézu $H: \mu_1 = \mu_2$ proti alternativní hypotéze**Chyba! Záložka není definována.** A pomocí testového kritéria

$$t = \frac{\bar{x} - \bar{y}}{S} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}, \tag{22}$$

kde

$$S = \left[\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \right]^{1/2}. \tag{23}$$

Testové kritérium (22) má při platnosti hypotézy $H: m_1 = m_2$ rozdělení $t(n_1 + n_2 - 2)$. Kritické obory jsou dány vztahy (24) - (26).

alternativa	kritický obor**Chyba! Záložka není definována.**	
$A_1: \mu_1 > \mu_2$	$W_1 = \{t; t \geq t_{1-\alpha}(n_1 + n_2 - 2)\}$	(24)
$A_2: \mu_1 < \mu_2$	$W_2 = \{t; t \leq -t_{1-\alpha}(n_1 + n_2 - 2)\}$	(25)
$A_1: \mu_1 \neq \mu_2$	$W_3 = \{t; \lvert t \rvert \geq t_{1-\alpha/2}(n_1 + n_2 - 2)\}$	(26)

Velmi důležitým předpokladem použití testu je **nezávislost obou výběrů**.

c) *Testy o středních hodnotách dvou výběrů**Chyba! Záložka není definována.** s nehomogenními rozptyly*

Jestliže zamítneme hypotézu o homogenitě rozptylů obou výběrů, pak testujeme hypotézu $H: \mu_1 = \mu_2$ proti alternativní hypotéze**Chyba! Záložka není definována.** A pomocí testového kritéria

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \tag{27}$$

které má při platnosti hypotézy $H: m_1 = m_2$ přibližně rozdělení $t(\mu)$. Pro počet stupňů volnosti n platí vztah

$$\nu \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1} \right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2} \right)^2}. \tag{28}$$

Kritické obory jsou dány vztahy (29) - (31).

alternativa kritický obor**Chyba! Záložka není definována.**

$A_1: \mu_1 > \mu_2 \qquad W_1 = \{t;\ t \geq t_{1-\alpha}(\nu)\}$ (29)

$A_2: \mu_1 < \mu_2 \qquad W_2 = \{t;\ t \leq -t_{1-\alpha}(\nu)\}$ (30)

$A_1: \mu_1 \neq \mu_2 \qquad W_3 = \{t;\ |t| \geq t_{1-\alpha/2}(\nu)\}$ (31)

Rovněž u tohoto testu se předpokládá **nezávislost obou výběrů.**

4.2.3 Párový testChyba! Záložka není definována.

Párový t-test se používá pro test hypotézy o rovnosti středních hodnot ve dvou výběrech za předpokladu normality rozdělení, pokud jsou **oba výběry závislé**. Většinou se jedná o dvojice dat $(x_1, y_1), \ldots, (x_n, y_n)$, kde veličina X má střední hodnotu μ_1 a veličina Y má střední hodnotu μ_2.

Testujeme hypotézu H_0: $\mu_1 - \mu_2 = 0$, to znamená, že testujeme shodu středních hodnot. Test hypotézy o rozdílu středních hodnot náhodných veličin X a Y převádíme na test hypotézy o střední hodnotě veličiny D = X - Y.

Jako testové kritérium volíme statistiku

$$t = \frac{\bar{d}}{s_d}\sqrt{n}, \tag{32}$$

která má při platnosti hypotézy H_0 rozdělení t(n-1).

Alternativním hypotézám odpovídají kritické obory (33), (34) a (35).

alternativa kritický oborChyba! Záložka není definována.

$A_1: \mu_1 - \mu_2 > 0 \qquad W_1 = \{t;\ t \geq t_{1-\alpha}(n-1)\}$ (33)

$A_2: \mu_1 - \mu_2 < 0 \qquad W_2 = \{t;\ t \leq -t_{1-\alpha}(n-1)\}$ (34)

$A_3: \mu_1 - \mu_2 \neq 0 \qquad W_3 = \{t;\ |t| \geq t_{1-\alpha/2}(n-1)\}$ (35)

4.2.4 Testy dobré shodyChyba! Záložka není definována.

Zatímco při parametrických testech předpokládáme, že typ rozdělení souboru je znám, testy dobré shody prověřují hypotézy právě o typu rozdělení, z něhož byl výběr pořízen. Jedná se tedy o hypotézy **o shodě teoretického a empirického rozdělení**.

a) *Kolmogorovův - Smirnovův jednovýběrový test**Chyba! Záložka není definována.***

Test vychází z netříděných hodnot a porovnává hodnoty teoretické (hypotetické) distribuční funkce $F_0(x)$ a empirické distribuční funkce $F_n(x)$, která je definována z výběrových hodnot uspořádaných podle velikosti předpisem.

$$F_n(x) = \begin{cases} 0 & \text{pro} \quad x < x_{(1)}, \\ \frac{i}{n} & \text{pro} \quad x_{(i-1)} \leq x < x_{(i)}, i = 1, \ldots, n, \\ 1 & \text{pro} \quad x \geq x_{(n)}. \end{cases} \tag{36}$$

Testovým kritériem je statistika

$$D = \max\left\{\max\left|F_0(x_{(i)}) - \frac{i}{n}\right|, \left|F_0(x_{(i)} - \frac{i-1}{n}\right|\right\}a. \tag{37}$$

Hypotéza H: $F_n(x) = F_0(x)$ bude zamítnuta na hladině významnosti α ve prospěch alternativy A: $F_n(x) \neq F_0(x)$, jestliže

$$D \geq D_{1-\alpha/2}(n), \tag{38}$$

kde $D_{1-\alpha/2}(n)$ je kvantil pro Kolmogorov-Smirnovův test.

b) χ^2 - *test dobré shody*

Testujeme hypotézu H_0: základní soubor má rozdělení určitého typu proti
A: základní soubor nemá rozdělení určitého typu.

Test vychází z tříděných dat a předpokládá výběr velkého rozsahu. Empirické četnosti n_j se porovnávají s teoretickými četnostmi $n\pi_j$, to znamená s četnostmi, které očekáváme v případě, že platí hypotéza H_0.

Testovací kritérium je statistika

$$\chi^2 = \sum_{j=1}^{k} \frac{(n_j - n\pi_j)^2}{n\pi_j}, \tag{39}$$

která má za předpokladu správnosti hypotézy H_0 asymptoticky Chí-kvadrát rozdělení**Chyba! Záložka není definována.Chyba! Záložka není definována.Chyba! Záložka není definována.Chyba! Záložka není definována.**
$\nu = k-c-1$ stupni volnosti, kde k je počet tříd a c je počet neznámých parametrů ověřovaného rozdělení.

Hypotézu H_0 zamítneme na hladině významnosti α ve prospěch alternativy A, jestliže

$$\chi^2 \geq \chi^2_{1-\alpha}(k-c-1). \tag{40}$$

4.3 Příklad

Určete s 95% spolehlivostí horní odhad**Chyba! Záložka není definována.**y srdeční frekvence (SF) ve dvou skupinách sportovců a na 5% hladině významnosti tvrzení, že průměr SF je u obou skupin stejný.

Tab. 4 Naměřené hodnoty SF ve dvou skupinách

Skupina 1	68	133	144	106	154	175	141	148	75	50	130	151
	199	134	183	137	127	101	157	119	112	115	88	168
	142	103	135	115	195	133	105	82	78	143	85	85
	179	124	113	97	80	84	135	99	116	133	118	200
	145	165	123	155	131	98	148	44	125	82	110	111

Skupina 2	148	127	174	132	125	139	158	140	108	146	125	154
	132	128	127	111	134	111	118	105	150	109	112	114
	115	81	132	112	148	162	124	159	198	134	134	157

	158	73	137	154	138	168	151	136	117	104	141	171
	145	151	140	113	147	146	105	141	128	167	152	131
	167	133	108	148								

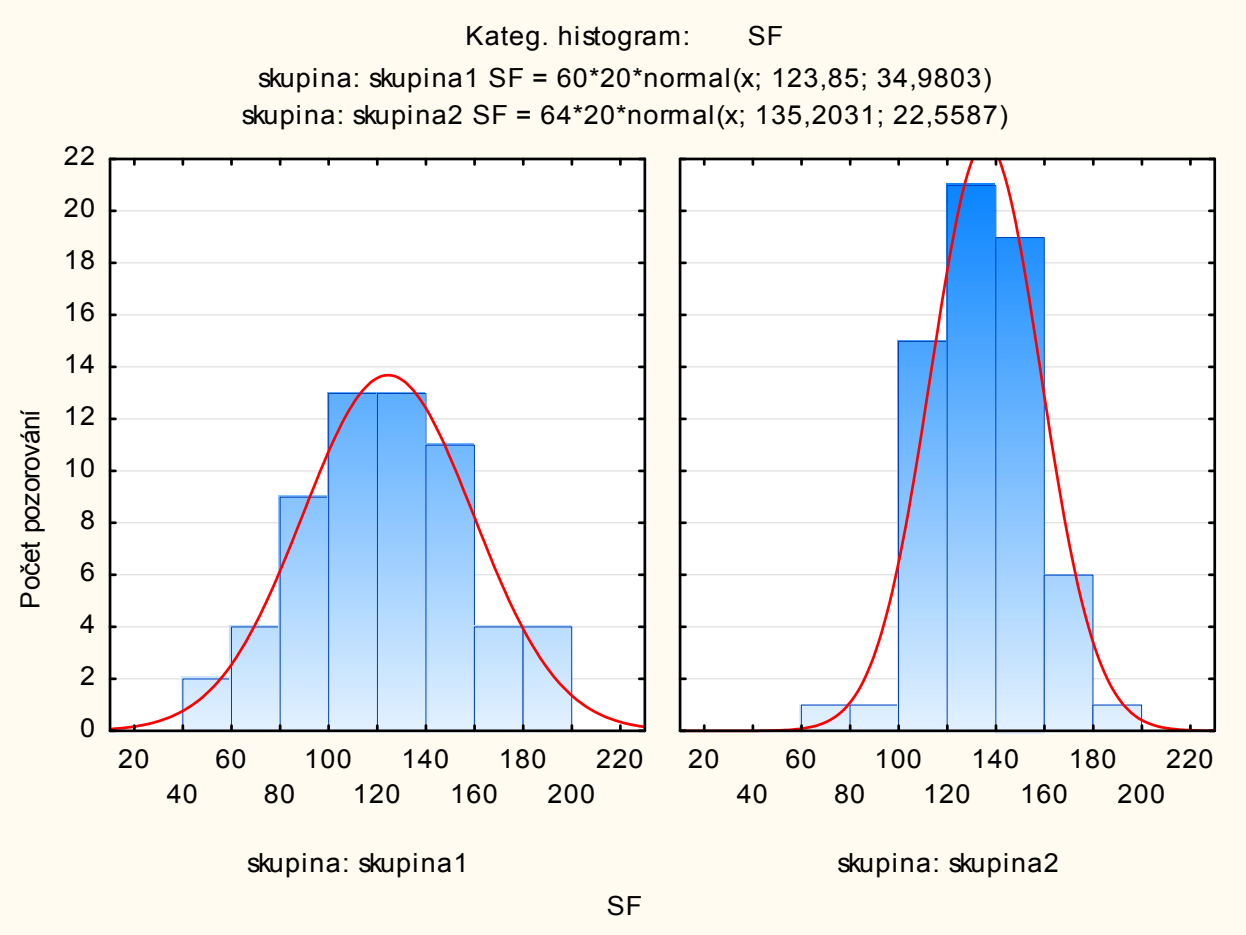
POZOR na zápis vstupních dat a jeho alternativní formu, kterou vyžadují některé metody. V tomto případě musíme data zapsat ve formě grupovací proměnné!

Řešení pomocí STATISTICA

a) *Grafické ověření normality dat*

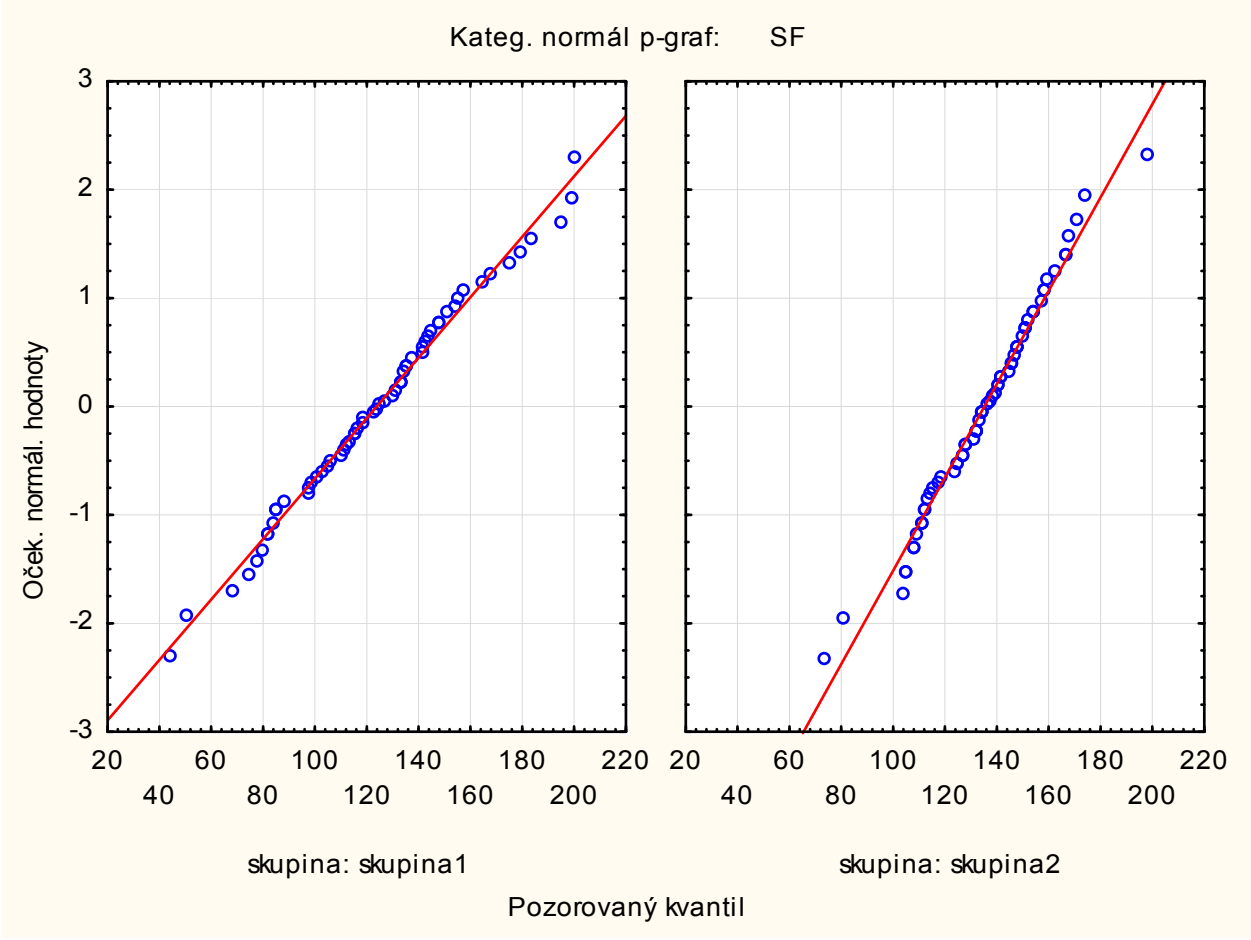
Protože se při odvozování intervalů spolehlivosti i testů hypotéz předpokládá normální rozdělení výběrového souboru, měli bychom vždy tento předpoklad ověřit. Nesplnění předpokladu normality dat, vede k přibližným a někdy i chybným řešením.

Potup: Statistika – Základní statistiky a tabulky – t-test (nezávislé, dle skupin) – detailní výsledky – kategorizované histogramy



Obr. 5 Ověření symetrie a homogenity dat SF u obou skupin

Potup: Statistika – Základní statistiky a tabulky – t-test (nezávislé, dle skupin) – detailní výsledky – kategorizované pravděpodobnostní normální grafy



Obr. 6 Ověření normality dat SF u obou skupin

b) *Testy dobré shody***Chyba! Záložka není definována.**

Posloupností příkazů

Potup: Statistika – Základní statistiky a tabulky – tabulky četností – normalita – test normality

Získáme textový výstup Kolmogorov-Smirnovova testu (tab. 14), který obsahuje hodnotu statistiky (65) = maximální rozdíl, K-S tabulkovou statistiku a vypočítanou oboustrannou pravděpodobnost (p-hodnotu). K-S test potvrdil dobrou shodu výběrových rozdělení s rozdělením normálním.

Tab. 5 Testy normality (SF)

	N	max D	K-S - p	Lilliefors - p	W	p
SF-skupina 1	60	0,049968	p > .20	p > .20	0,988634	0,850352
SF-skupina 2	64	0,052919	p > .20	p > .20	0,986257	0,698076

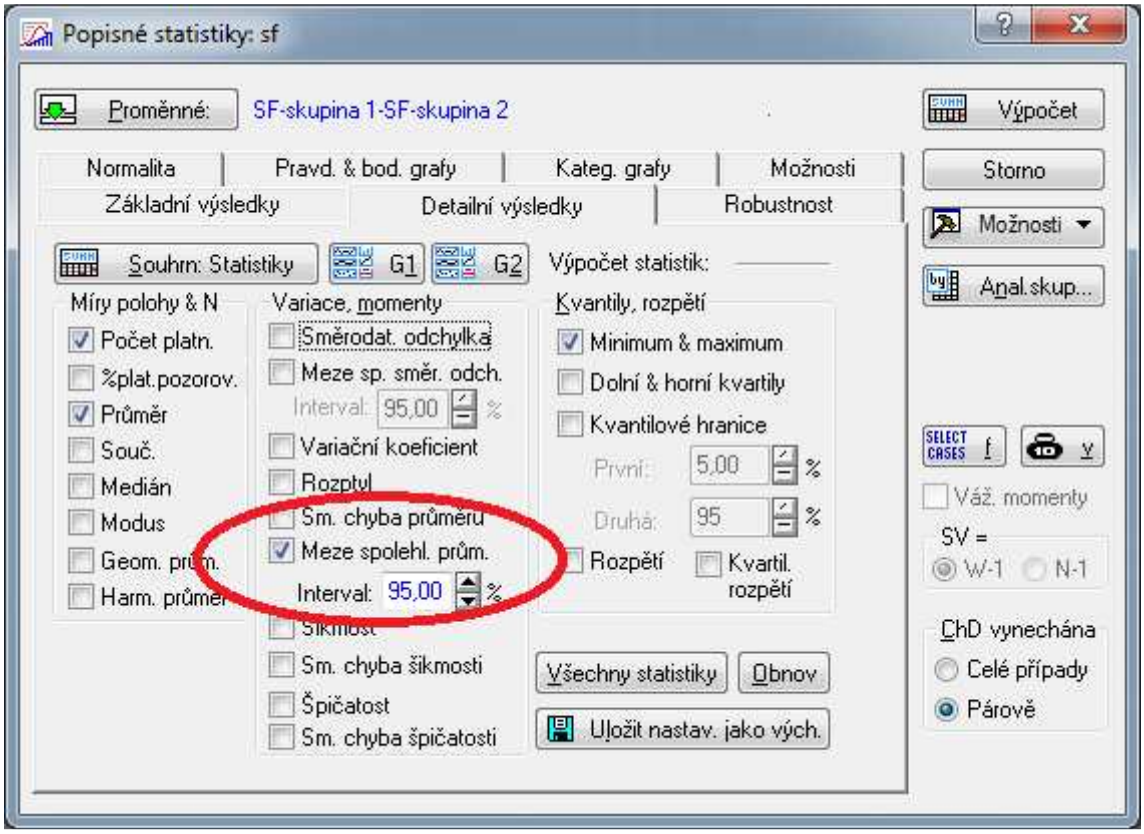
Na základě údajů v tabulce 5 nezamítáme hypotézu o normalitě obou výběrů.

c) *Intervaly spolehlivosti*

Zadání příkladu požaduje 95% jednostranné intervaly spolehlivosti - horní odhad**Chyba! Záložka není definována.**y očekávaných výdělků v obou městech. Posloupností příkazů

Postup: Statistika – Základní statistiky a tabulky – Popisné statistiky

a vypsáním v okénku **Interval** vstupního panelu **0.90**, získáme 90% oboustranné intervaly spolehlivosti uvedené v tab. 6.



Obr. 7 Dialogové okno pro výpočet oboustranného intervalu spolehlivosti pro průměr

Tab. 6 Výstupu procedury Meze spolehlivosti průměru

	N platných	**Průměr**	**Int. spolehl. -95,000%**	**Int. spolehl. 95,000%**
SF-skupina 1	60	123,85	114,81	132,89
SF-skupina 2	64	135,20	129,57	140,84

d) *Test rovnosti rozptylů*

Proceduru, která provádí výpočet F-testu**Chyba! Záložka není definována.** (48) vyvoláme příkazy

Postup: Statistika – Základní statistiky a tabulky – t-test nezávislé dle proměnných – Výpočet t-testy

Výsledek testu je tab. 7. V horní části tabulky najdeme hodnoty základních výběrových statistik: průměr, rozsah výběru, směrodatnou odchylku. Následuje hodnota testového F-kritéria = 2,40 a vypočtená pravděpodobnost (p = 0,00, která představuje minimální hladinu jednostranného testu, pro kterou ještě H_0 zamítáme.

F-test zamítl hypotézu o rovnosti rozptylů (p = 0,00 < 0,05 = α).

Tab. 7 Test rovnosti rozptylů

	Průměr skupina1	Průměr skupina2	Poč.plat - skupina1	Poč.plat. skupina2	Sm.odch. skupina1	Sm.odch. skupina2	F-poměr Rozptyly	p Rozptyly
SF	123,85	135,20	60	64	34,98	22,56	2,40	0,00

e) *Test rovnosti středních hodnot*

Na základě výsledků F-testu**Chyba! Záložka není definována.**, který zamítl hypotézu o rovnosti rozptylů, bychom měli správně zvolit zvolit test (55) o středních hodnotách s nehomogenními rozptyly posloupností příkazů. Sw Statistica však neumožňuje tuto volbu a tak vybereme jedinou nabídku pro t-test.

Postup: Statistika – Základní statistiky a tabulky – t-test nezávislé dle proměnných – Výpočet t-testy

Tab. 8 Test rovnosti středních hodnot

	Průměr skupina1	Průměr skupina2	t separ. odh.prom	sv	p oboustr.
SF	123,85	135,20	-2,13	99,77	0,04

V tab. 8 najdeme hodnotu testového kritéria (50) = t-statistika = -2,13, počet stupňů volnosti = 99,77 a minimální oboustranná pravděpodobnost p = 0,04 při které ještě zamítáme hypotézu o rovnosti středních hodnot.

Protože vypočítaná dvoustranná pravděpodobnost p = 0.04 < 0,05 = α, zamítáme hypotézu $H_0: \mu_1 - \mu_2 = 0$ o rovnosti středních hodnot a s pravděpodobností 0,95 tvrdíme, že průměrné hodnoty srdeční frekvence (SF) uvedených skupin jsou různé.

Závěr příkladu

Grafy pro ověření normality a testy dobré shody ukázaly, že oba výběry mají normální rozdělení. Hypotézu o rovnosti středních hodnot naměřených srdečních hodnot v obou skupinách jsme s 95% spolehlivostí zamítli.

Při testování se neomezíme pouze na jeden test. Je-li to možné, **hypotézu ověříme více testy**.

Kontrolní otázky a úkoly

1. Popište průběh testování hypotéz pomocí tabulek
2. Popište průběh testování hypotéz pomocí software
3. Co je hladina statistické významnosti?

4. Hypotézu potvrzujeme nebo nezamítáme? Proč?
5. Může být hodnota 0,12 statisticky nulová?
6. Jak je rozdíl mezi F-testem a t-testem?
7. Uveďte dvojici proměnných, kterou lze považovat za závislá pozorování.
8. Uveďte dvojici proměnných, kterou lze považovat za nezávislá pozorování.
9. Jak ověříte normalitu vstupních dat?
10. Jak je definována věcná významnosti?

5 Neparametrické testy

Teoretický aparát byl převzat z učebnice Cyhelský, Kahounová & Hindls (2001).

Vybrané neparametrické testy

Často vycházíme z toho, že testování statistických hypotéz probíhá při apriorní znalostí toho, jaké je pravděpodobnostní rozdělení základního souboru, z něhož byl výběrový soubor pořízen. Mlčky předpokládáme pravdivost a funkčnost této znalosti. Existují metody, které slouží k ověřování předpokladů o typu rozdělení. Jinak řečeno, chceme testovat hypotézu o tom, že existuje shoda mezi teoreticky předpokládaným rozdělením a rozdělením empirickým (tj. rozdělením hodnot pořízených náhodným výběrem). Této skupině testů se často říká – jak už z povahy problému vyplývá – *testy dobré shody*.

Testy dobré shody řadíme mezi velmi rozsáhlou skupinu dalších testů ve statistice, pro které používáme pojmenování *neparametrické testy*.

Půjde zejména o testy shody úrovně (např. Wilcoxonův test, Friedmanův test, Mannův-Whitneyův test). Společnou předností pro tyto testy je, že nevyžadují praktickou žádnou znalost pravděpodobnostního rozdělení zkoumané veličiny. Vedle této okolnosti k dalším výhodám uvedené skupiny testů patří jejich účinné použití i při poměrně malém rozsahu výběru, dále možnost aplikovat je i pro ordinální a nominální proměnné a rovněž větší *robustnost* (to je obecně důležitá statistická vlastnost, kterou bychom mohli zjednodušeně charakterizovat asi takto: zvolený postup, např. statistický test, je tím robustnější, čím je kvalita jeho výsledků méně závislá na povaze konkrétních dat a na případném „narušení jejich kvality“ v důsledku výrazných odchylek od ideálních předpokladů). Na druhé straně použití neparametrického testu obvykle vede za jinak nezměněných podmínek k rozšíření oboru přijetí na úkor oboru kritického, což v konečných důsledcích může při použití neparametrického testu mít za následek zvýšení (v porovnání s analogickým parametrickým testem) pravděpodobnost chyby druhého druhu, tj. může dojít k chybnému nezamítnutí nepravdivé testované (nulové) hypotézy. Jinak řečeno, důsledkem aplikace neparametrického testu je nižší síla testu.

5.1 χ^2 test dobré shody

Základem tohoto často užívaného testu je možnost roztřídit výsledky náhodného výběru jednoznačným a vyčerpávajícím způsobem do určitého počtu navzájem se nepřekrývajících tříd. Nulová hypotéza pak vyjadřuje teoretické pravděpodobnosti obsazení těchto tříd a porovnává se se skutečnými výběrovými výsledky (čili s empirickými četnostmi). Odtud také název testy dobré shody.

Problematika tohoto testu může mít dvě podoby. Buď nulová hypotéza H_0 (tedy teoretické pravděpodobnosti) je udána některým standardním rozdělením náhodné veličiny, které známe z počtu pravděpodobnosti (např. Poissonovo, normální, exponenciální či řada jiných), anebo obecněji hypotézu H_0 tvoří jakékoliv teoretické rozdělení pravděpodobností, které může být formulováno intuitivně, např. jako zobecněná zkušenost apod.

Nulová hypotéza, udávající pravděpodobnost obsazení j-té třídy, (označíme ji π_j) má tvar

$$H_0: \pi_j = \pi_{0j} \text{ pro } j = 1, 2, \ldots, k$$

alternativní je H_1: non H_0.

Testovým kritériem je veličina

$$G = \sum_{j=1}^{k} \frac{(n_j - \Gamma_j)^2}{\Gamma_j}, \qquad (41)$$

kde $\Gamma_j = n \,.\, \pi_{0j}$ udává teoretické (očekávané) obsazení j-té třídy při rozsahu výběru n, zatímco n_j je empirická četnost v téže j-té třídě. Statistika (41) má rozdělení $\chi^2(k\text{-}1)$ a pro hypotézu H_1 se vyslovíme, pokud G překročí kvantil $\chi^2_{1-\alpha}(k-1)$.

Test klade celkem vysoké požadavky na rozsah výběru, protože by měl být ve všech třídách splněn požadavek $\Gamma_j = n \,.\, \pi_{0,j} > 1$ a alespoň v 80 % třídách $\Gamma_j = n \,.\, \pi_{0,j} > 5$. Pokud tomu v průběhu průzkumu tak není, je možné toho dosáhnout případným dodatečným navýšením rozsahu výběru (což ovšem někdy nemusí být již uskutečnitelné) nebo vhodným slučováním sousedním, popř. věcně příbuzných tříd). Původní počet k tříd se potom ale musí patřičným způsobem snížit.

V předešlých odstavcích jsme se věnovali případu, kdy testované rozdělení bylo jednoznačně určeno (např. Possionovo, $\lambda = 2$). Říkáme, že jde o tzv. úplně specifikovaný model rozdělení. Častá je však situace, kdy některý parametr nebo dokonce parametry všechny v testovaném rozdělení známy nejsou. (tzv. neúplně specifikovaný model rozdělení). Potom nezbývá, než chybějící počet c parametrů odhadnout z údajů výběru (mělo by se tak dít tzv. modifikovanou metodou chí-kvadrát minima, ale v praxi se velmi často tento požadavek nesplňuje a k odhadu se s vědomím jisté nepřesnosti používá spíše jiných metod, např. maximálně věrohodných odhadů). Další postup je potom už stejný jako v případě úplně specifikovaného modelu rozdělení, jen kritický obor je vymezen těmi hodnotami G, pro které platí $G \geq \chi^2_{1-\alpha}(k-c-1)$, kde c je počet parametrů odhadnutých z výpočtu.

Další neparametrické testy

V této části se nejprve budeme věnovat testům o shodě úrovně. Úvodem upozorněme, že při výkladu následujících neparametrických testů budeme opět rozlišovat mezi *závislými* a *nezávislými* výběry. Tedy v případě nezávislých výběrů platí $n_1 = n_2 = \ldots = n_k = n$, zatímco v případě nezávislých výběrů budeme mít $n_1 \neq n_2 \neq \ldots \neq n_k$.

5.2 Wilcoxonův test pro dva závislé výběry

(neboli též pořadový znaménkový test)

Předpokládejme, že pro každou vybranou statistickou jednotku, máme k dispozici dvě pozorování, tj. máme celkem $2n$ pozorování. Ověřujeme, zda úroveň hodnot je v obou výběrech stejná. Nulová (testovaná) hypotéza má tvar

$$H_0: \mu_1 = \mu_2$$

proti alternativě H_1: non H_0 (dvoustranná alternativní hypotéza), resp. proti alternativě typu H_1: $\mu_1 \neq \mu_2$ (jednostranná alternativní hypotéza).

Určení hodnoty příslušného testového kritéria je postaveno na zjištění rozdílů mezi dvojicemi pozorování s tím, že budeme zohledňovat i faktickou těchto rozdílů. Pro každou dvojici pozorování totiž vypočteme rozdíly d_i pro $i = 1, 2, \ldots, n$. Nenulovým rozdílům

přiřadíme pořadová čísla s tím, že postupujeme od nejnižší absolutní hodnoty tohoto rozdílu k nejvyšší. Pořadová čísla poté rozdělíme do dvou skupin podle znamének diferencí a zjistíme součty těchto pořadových čísel. Menší ze součtů je hodnotou testového kritéria T_w, kterou porovnáme s kvantily speciálně zkonstruovaného rozdělení T_w. Kritický obor je vymezen jako množina hodnot T_w menších nebo rovných než (100 α/2) % kvantil rozdělení T_w v případě dvoustranné alternativní hypotézy ($T_w \leq T_{w;\alpha/2}$), resp. menších nebo rovných než 100 α % kvantil tohoto speciálního rozdělení v případě jednostranné alternativní hypotézy ($T_w \leq T_{w;\alpha}$)

Z technického hlediska ještě uvedeme, že pokud se během výpočtu testového kritéria objeví dvě nebo více stejných hodnot diferencí (což obecně nelze vyloučit), potom jim přiřazujeme tzv. průměrné pořadové číslo. Pokud se např. objeví stejné diference o velikosti dejme tomu 13, vyskytují se celkem čtyřikrát a těmto čtyřem hodnotám přísluší pořadová čísla např. 4, 5, 6, 7, pak průměrné pořadové číslo přiřazené každé z diferencí rovných 13 bude 5,5.

5.3 Mannův-Whitneyův test pro dva nezávislé výběry

(někdy též Wilcoxonův test pro dva nezávislé výběry)

Máme k dispozici dva nezávislé výběry, tj. máme-li celkem $n_1 + n_2 = n$ pozorování, kde $n_1 < n_2$, lze i v této situaci testovat hypotézu o shodě úrovně v těchto dvou nezávislých výběrech, tzn.

$$H_0: \mu_1 = \mu_2$$

proti alternativě H_1: non H_0 (dvoustranná alternativní hypotéza), resp. proti alternativě typu H_1: $\mu_1 \neq \mu_2$ (jednostranná alternativní hypotéza).

Při výpočtu hodnoty testového kritéria údaje získané z obou výběrů seřadíme vzestupně a jednotlivým údajům přiřadíme pořadová čísla. V každém ze souborů tato pořadová čísla sečteme. Součet pořadových čísel v souboru o rozsahu n_1 ($n_1 < n_2$) je statistikou B_1, kterou použijeme při výpočtu testového kritéria

$$T_M = B_1 - \frac{n_1\,(n+1)}{2}$$

která má při platnosti H_0 rozdělení T_M. Kritický obor je v případě dvoustranné hypotézy vymezen jako

$$W = \{T_M; T_M \geq T_{M;1-\alpha/2} \quad \text{a} \quad T_M \leq T_{M;\alpha/2}\}$$

Je-li alespoň $n_1 > 8$ a zároveň $n_2 > 14$, můžeme aproximativně použít testového kritéria

$$U_M = \frac{T_M\,\sqrt{12}}{\sqrt{n_1\,n_2\,(n+1)}}$$

které má při platnosti H_0 normované normální rozdělení. Kritický obor je v případě dvoustranné alternativní hypotézy vymezen množinou hodnot

$$W = \{U_M; U_M \geq u_{1-\alpha/2} \quad \text{a} \quad U_M \leq u_{\alpha/2}\}$$

5.4 Příklad I

V případě, kdy data nepocházejí z normálního rozdělení, použijeme neparametrické testy. Na datech z tab. 4 proveďte neparametrický t-test. Data jsou jasně nezávislé výběry, pocházejí pokaždé od jiné skupiny respondentů, proto použijeme Mann-Whitneyův t-test

Postup: Statistiky – Neparametrické statistiky – Porovnání dvou nezávislých vzorků (skupiny)

Tab. 9 Výsledek Mann-Whitneyova t-testu

Mann-Whitneyův U test (sf) Dle proměn. skupina Označené testy jsou významné na hladině p <,05000

	Sčt poř. skupina1	Sčt poř. skupina2	U	Z	p-hodn.	Z upravené	P hodn.	N skupina1	N platn. skupina2	2*1str. přesné p
SF	3315,00	4435,00	1485,00	-2,17	0,03	-2,17	0,03	60	64	0,03

V tabulce najdeme opět hodnoty základních výběrových statistik: tentokráte součet pořadí. Následuje hodnota testového kritéria: U = 1485, a minimální oboustranná pravděpodobnost p = 0,03, při které ještě zamítáme hypotézu o rovnosti středních hodnot.

I tento postup zamítl hypotézu o rovnosti středních hodnot naměřené srdeční frekvence obou skupin

5.5 Příklad II

Test nezávislosti χ^2 (chí-kvadrát), měření síly závislosti

V roce 1950 zkoumali Yule a Kendall barvu očí a vlasů u 6800 mužů.

Tab. 10 Data pro chí-2 test

Barva očí	Barva vlasů			
	světlá	Kaštanová	Černá	rezavá
modrá	1768	807	180	47
šedá nebo zelená	946	1387	746	53
hnědá	115	438	288	16

Na asymptotické hladině významnosti 0,05 testujte hypotézu o nezávislosti barvy očí a barvy vlasů. Simultánní četnosti znázorněte graficky.

Návod: Vytvořte nový datový soubor o 12 případech a třech proměnných (OCI, VLASY, CETNOST). Do proměnné OCI napište varianty barvy očí $x_{[1]} = 1$ (modrá), $x_{[2]} = 2$ (šedá nebo zelená), $x_{[3]} = 3$ (hnědá), přičemž každá varianta se objeví čtyřikrát pod sebou. Do proměnné VLASY napište třikrát pod sebe všechny varianty $y_{[1]} = 1$ (světlá), $y_{[2]} = 2$ (kaštanová), $y_{[3]} = 3$ (černá), $y_{[4]} = 4$ (rezavá).

Tab. 11 Data pro chí-2 test ve formátu Statisticy

	oci	vlasy	cetnost
1	modrá	světlá	1768
2	modrá	kaštanová	807
3	modrá	černá	180
4	modrá	rezavá	47
5	šedozelená	světlá	946
6	šedozelená	kaštanová	1387
7	šedozelená	černá	746
8	šedozelená	rezavá	53
9	hnědá	světlá	115
10	hnědá	kaštanová	438
11	hnědá	černá	288
12	hnědá	rezavá	20

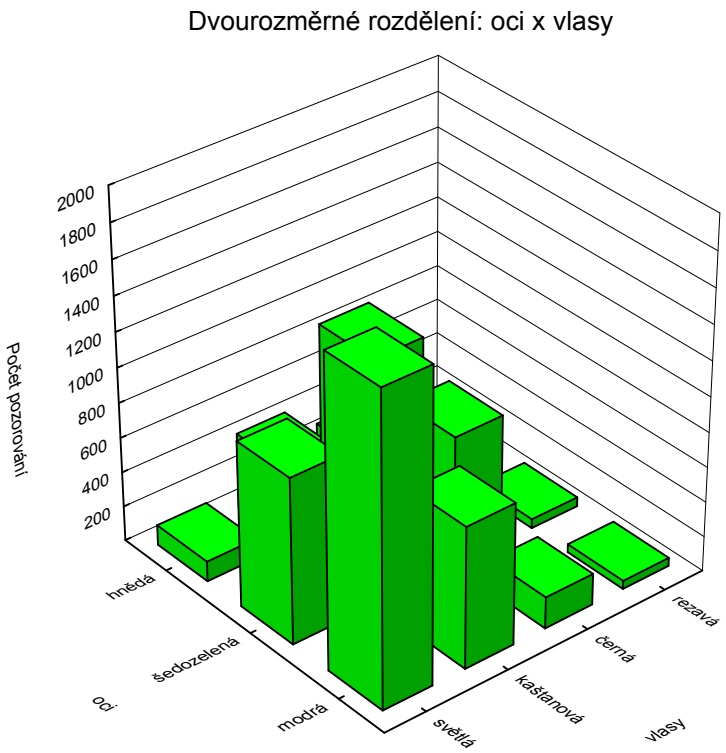
Postup: Statistiky – Základní statistiky a tabulky – Kontingenční tabulky – Specifikace tabulky (List1-oci, List2-vlasy) OK – V (váhy) proměnná vah: četnost – OK – Možnosti: **zaškrtněte Pearson & M-L Chi -square, Phi & Cramer's V –** Detailní výsledky – Detailní 2-rozm. Tabulky

Tab. 12 Výsledek chí-2 test

Statist. : oci(3) x vlasy(4) (oci)			
	Chí-kvadr.	**sv**	**p**
Pearsonův chí-kv.	1088,934	df=6	p=0,0000
M-V chí-kvadr.	1157,153	df=6	p=0,0000
Fí	,4003188		
Kontingenční koeficient	,3716458		
Cramér. V	,2830681		

Ve výstupní tabulce najdeme mj. hodnotu testové statistiky (Chi-square = 1088,149) s počtem stupňů volnosti (df = 6) a odpovídající **p-hodnotou** (p = 0,0000) i **Cramérův koeficient (V = 0,283)**. Pro grafické znázornění četností se vraťte do Detailní výsledky – 3D histogram. Po vytvoření grafu (obr. 8) je nutné manuálně zvětšit rozsah zobrazovaných hodnot na osách x a y.

Pomocí STATISTIKY je možno lehce ověřit splnění podmínek dobré aproximace (tzn., že teoretické četností mají být aspoň v 80 % případů větší než 5 a ve zbylých 20% případů nemají klesnout pod 2). Teoretické četnosti se vypočítají tak, že v Možnosti zaškrtneme Očekávané četnosti. V našem případě jsou podmínky dobré aproximace splněny.



Obr. 8 3D histogram

Závěr: Na základě tab. 12 zamítáme hypotézu o nezávislosti. Můžeme tak předpokládat, že existuje závislost mezi barvou očí a barvou vlasů. Závislost můžeme i docela přesně popsat. Světlovlasí lidé mají modré oči, tmavovlasí zas mají tmavé oči.

Kontrolní otázky a úkoly

1. Kdy se používají neparametrické testy?
2. Proč se neparametrickým metodám říká pořadové?
3. Definujte robustnost metody.
4. Uveďte neparametrické t-testy.
5. K čemu slouží metoda chí-2 (chí-kvadrát)?
6. Proč se vlastně říká neparametrické?

6 Korelační koeficient

Korelace znamená vzájemný vztah mezi dvěma procesy nebo veličinami. Pokud se mezi dvěma procesy ukáže korelace, je pravděpodobné, že na sobě závisejí, nelze z toho však ještě usoudit, že by jeden z nich musel být příčinou a druhý následkem. To samotná korelace nedovoluje rozhodnout.

Výpočet Pearsonova korelačního koeficientu je naznačena v následujícím vztahu

$$r_{x,y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \qquad (42)$$

Matematickými úpravami lze převést na tzv. výpočtový tvar:

$$r_{x,y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{[n\sum x_i^2 - (\sum x_i)^2] \cdot [n\sum y_i^2 - (\sum y_i)^2]}} \qquad (43)$$

V určitějším slova smyslu se pojem korelace užívá ve statistice, kde znamená vzájemný lineární vztah mezi znaky či veličinami x a y. Tento vztah může být kladný, pokud (přibližně) platí $y = kx$, nebo záporný ($y = -kx$). Míru korelace pak vyjadřuje korelační koeficient, který může nabývat hodnot od −1 až po +1.

Hodnota korelačního koeficientu −1 značí zcela nepřímou závislost, tedy čím více se zvětší hodnoty v první skupině znaků, tím více se zmenší hodnoty v druhé skupině znaků, např. vztah mezi uplynulým a zbývajícím časem. Hodnota korelačního koeficientu +1 značí zcela přímou závislost, např. vztah mezi rychlostí běhu a běžeckou frekvencí kroků sprintera. Pokud je korelační koeficient roven 0, pak mezi znaky není žádná statisticky zjistitelná lineární závislost. Je dobré si uvědomit, že i při nulovém korelačním koeficientu na sobě veličiny mohou záviset, pouze tento vztah nelze vyjádřit lineární funkcí, a to ani přibližně. Může jít např. o nelineární závislost (kvadratickou, …).

Hendl (1997) uvádí nevýhody korelačního koeficientu, který je citlivý k náhodné chybě. Proto se používá ve srovnávacím experimentu. Naneštěstí je citlivý také k rozmezí měření. Často zvětšením rozsahu měření, dosáhneme značného přiblížení korelačního koeficientu k 1. Snad největší chyba spočívá v tom, že přisuzujeme důležitost tomu, že korelační koeficient je významné různý od nuly. Ve srovnávacích experimentech není tento typ uvažování na místě, přesto se údaje o této významnosti pravidelně objevují v hodnotících zprávách. Závažná je skutečnost, že korelační koeficient neodhaluje ani přítomnost proporcionální chyby ani chyby konstantní. Odpůrci korelačního koeficientu tvrdí, že tato statistika by se neměla nikdy používat při hodnocení dat srovnávacích experimentů.

Doporučuje se nahradit/doplnit posouzení korelačního koeficientu, který je pouze mírou lineární závislosti výsledků, jinými postupy, např. Bland-Altmanovým rozdílovým grafem (Bland & Altman, 1986).

Koeficient determinace r^2 určuje, jaká část rozptylu výkonnosti v jednom testu je dána proměnlivostí výkonů v druhém testu. Čím více se bude r blížit hodnotě 1, tím považujeme danou závislost za silnější, čím více se bude r blížit hodnotě 0, tím považujeme danou závislost za slabší.

Poznámky ke korelacím:

Matematicko-statistické předpoklady výpočtu korelačního koeficientu: linearita, normalita, dostatečný rozsah souboru

Spearmanův koeficient pořadové korelace

Spearmanův koeficient pořadové korelace se používá pro výpočet těsnosti závislosti ordinálních znaků u souborů o nevelkém rozsahu při poruše normality rozložení četností. Vzorec pro výpočet koeficientu pořadové korelace je následující

$$r_{x,y} = 1 - \frac{6 \cdot \sum (i_x - i_y)^2}{n(n^2 - 1)}$$

kde i_x, resp. i_y je index pořadí hodnot x resp. y. (44)

Příklad

Zjistěte míru závislosti proměnné BMI a poměr pas/boky) z dat v příloze 1.

Postup: Základní statistiky a tabulky – Korelační matice – 1 seznam proměnných – Možnosti – Zobrazit r, p, N: Souhrn

Tab. 13 Výsledek výpočtu Pearsonova korelačního koeficientu
Korelace. Označ. korelace jsou významné na hlad. p < ,05000 N=569 (Celé případy vynechány u ChD)

	poměr pas/boky	**BMI**
poměr pas/boky	1,0000	,8922
	p= ---	p=0,00
BMI	,8922	1,0000
	p=0,00	p= ---

Postup: Základní statistiky – Neparametrické statistiky – Korelace

Tab. 14 Výsledek výpočtu Spearmonova korelačního koeficientu

	poměr pas/boky	**BMI**
poměr pas/boky	1,000	0,919
BMI	0,919	1,000

Vyhodnocení příkladu: Velikost korelačního koeficientu mezi proměnnými „BMI“ a „poměr pas/boky“ je 0,8922 (Pearsonův), resp. 0,919 (Spearmonův). Znaménko plus značí přímou úměru. Korelační koeficient mezi proměnnými je poměrně vysoký a je roven hodnotě cca 0,9,

což značí hodnotu indexu determinace 0,81. Neboli daným modelem jsme schopni vysvětlit cca 81 % celkové variability.

Kontrolní otázky a úkoly

1. Co vyjadřuje korelační koeficient?
2. Co znamená znaménko plus (minus) u korelačního koeficientu?
3. Co vyjadřuje index determinace u korelačního koeficientu? Jak se počítá?
4. Jaký e rozdíl mezi Pearsonovým a Spearmonovým korelačním koeficientem?
5. Lze z existující závislosti určit příčinnost? Proč?
6. Může hodnota korelačního koeficientu rovna 0 mezi dvěma proměnnými znamenat závislost?

7 Regresní přímka

7.1 Lineární regrese

Regrese umožňuje postihnout povahu závislosti. Hledáme matematickou funkci, která by co nejlépe vyjadřovala charakter závislosti. Tato matematická funkce se nazývá regresní funkce a je vyjádřena regresní rovnicí. Regresní funkce může nabývat mnoha typů:

- lineární regrese
- kvadratická regrese
- kubická regrese
- polynomická regrese
- hyperbolická regrese
- logaritmická regrese

Hlavní úkoly regresní analýzy jsou:

❖ **Volba regresní funkce**

Zvolíme vhodný tvar regresní funkce, která respektuje teoretický model závislosti. Není-li teoretický model znám, provádíme analýzu bodového diagramu, grafu podmíněných průměrů a při volbě regresní funkce vycházíme ze zkušeností. Podle tvaru regresní funkce rozlišujeme **lineární a nelineární regresní modely.**

❖ **Odhad parametrů regresní funkce**

K určení parametrů regresní funkce byla navržena řada metod. Je-li **regresní funkce lineární vzhledem k parametrům**, pak nejpoužívanější metoda pro odhad parametrů je **metoda nejmenších čtverců.** Tato metoda vychází z požadavku, aby součet čtverců odchylek pozorovaných hodnot y_i od hodnot modelových $\hat{Y}_i$, ležících na regresní křivce, byl minimální.

❖ **Posouzení kvality zvolené regresní funkce**

Vystihneme-li průběh korelační závislosti regresní funkcí, zajímají nás velikosti odchylek experimentálních hodnot od vyrovnaných hodnot (hodnot ležících na výběrové regresní křivce). Přichází-li v úvahu více typů regresní funkce, můžeme při výběru využít následující kritéria:

➢ **Reziduální rozptyl s_R^2**

$$s_R^2 = \frac{\sum_{i=1}^{n}\left(y_i - \hat{Y}_i\right)^2}{n-(p+1)}, \tag{45}$$

kde součet

$$\sum_{i=1}^{n}\left(y_i - \hat{Y}_i\right)^2 = S_R \tag{46}$$

se nazývá **reziduální součet čtverců**, *n* je počet pozorování a *(p + 1) = c* je počet parametrů regresní funkce.

Za vhodnější se považuje ta regresní funkce, u níž má reziduální rozptyl (45) menší hodnotu

- **Index determinace i_{yx}^2**

$$i_{yx}^2 = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}, \tag{47}$$

kde součet

$$\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{y}\right)^2 = S_T \tag{48}$$

se nazývá **teoretický součet čtverců** a součet

$$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = S_y \tag{49}$$

se nazývá **celkový součet čtverců.**

Výběrovou regresní funkci považujeme za tím výstižnější, čím je index determinace bližší jedné.

Nejjednodušší z nich je *lineární regresní funkce*, která má ve své empirické podobě tvar

$$Y = a + b \,.\, x \tag{50}$$

Pro konkrétní závislost (např. tělesné výšky a hmotnosti) je třeba určit tzv. regresní koeficienty a, b, přičemž vycházíme z empirických údajů (znaků) sledované závislosti.
Pro výpočet regresních koeficientů a, b je výhodné použít následující vzorce

$$b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} \tag{51}$$

$$a = \frac{\sum y_i - b\sum x_i}{n} \tag{52}$$

Příklad

U dvaceti prodaných ojetých automobilů určité značky byla zjištěna cena a stáří auta. Závislost ceny na stáří automobilu popište regresní přímkou. Zjištěné hodnoty jsou uvedený v tab. 17. Odhadněte parametry regresní přímky a charakterizujte těsnost závislosti pomocí koeficientu determinace.

Tab. 15 Data k přímkové regresi

i	x_i [roky]	y_i [10 tis. Kč]	i	x_i [roky]	y_i [10 tis. Kč]
1	0,6	55,0	11	5,0	34,0
2	1,0	54,6	12	5,1	31,0
3	1,1	50,6	13	5,2	29,0
4	2,0	51,1	14	5,6	31,6
5	2,3	47,0	15	5,9	34,0
6	2,5	50,0	16	6,0	25,6
7	3,0	43,6	17	6,1	28,0
8	4,1	41,3	18	6,3	24,6
9	4,4	43,0	19	6,8	27,0
10	4,8	39,9	20	7,5	17,6

Přímková regrese ve STATIAITCE

Postup: Základní statistiky – Vícenásobná regrese

Závislou proměnnou stanovíme cen*u*, nezávislou *roky*.
Získáme vstupní panel pro přímkovou regresi a v Kroku 1 vybereme proměnné, viz obr. 17.

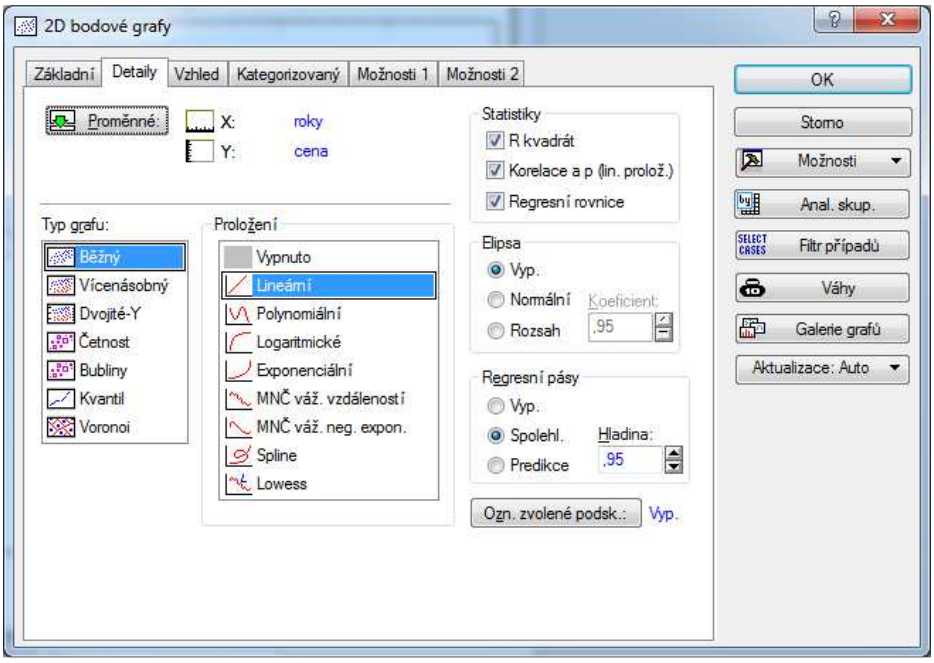
Výsledky regrese se závislou proměnnou : cena (auta)
R= ,96260583 R2= ,92660999 Upravené R2= ,92253277
F(1,18)=227,26 p<,00000 Směrod. chyba odhadu : 3,1047

N=20	b*	Sm.chyba z b*	b	Sm.chyba z b	t(18)	p-hodn.
Abs.člen			59,95	1,62	37,06	0,00
roky	-0,96	0,06	-5,16	0,34	-15,08	0,00

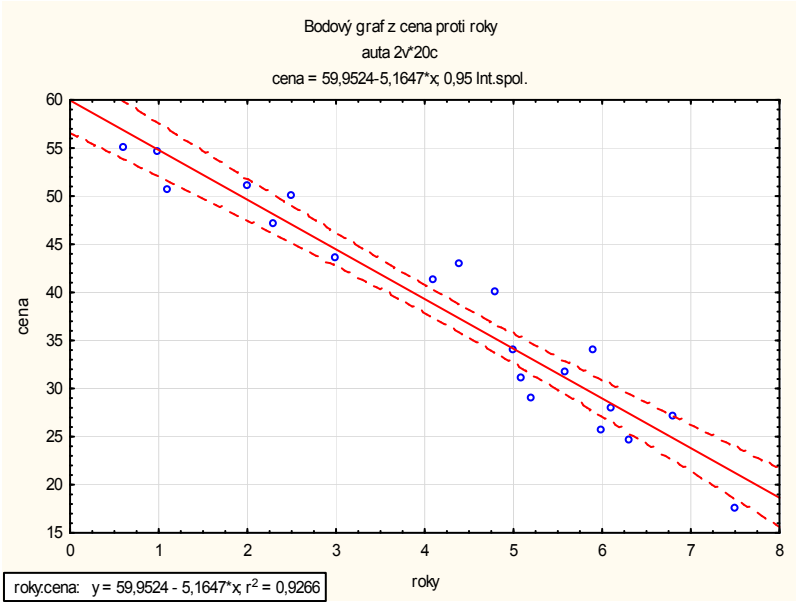
Obr. 9 Výsledky přímkové regrese

Stejný výsledek integrovaný do jednoho kroku lze vypočítat společně s grafickým znázorněním.

Postup: Grafy – Bodové grafy – další nastavenié podle obr.



Obr. 10 Nastavení parametrů přímkové regrese



Obr. 11 Alternativní výsledek přímkové regrese

Na obr. 9 a 11 jsou uvedeny odhady regresních koeficientů a jejich testování. Všechny regresní koeficienty jsou statisticky různé od nuly a výsledný model je ve tvaru:

$$cena = 59{,}9524 - 5{,}1647 * roky$$

Vypočítaný index determinace $r^2 = 0{,}9266$ naznačuj velmi vysokou přesnost tohoto modelu.

Kontrolní otázky a úkoly

1. V čem spočívá princip metody nejmenších čtverců?
2. Co vyjadřuje index determinace?
3. Uveďte dvojici proměnných, mezi nimiž existuje exponenciální vztah
4. Jaký je vztah korelačního koeficientu v lineární regresi?
5. Zkuste vymyslet nelineární závislost mezi dvěma proměnnými.

Seznam použitých zdrojů

- Bland, J. M., & Altman, D.. G. Statistical methods for assessing agreement between two methods of clinical measurement. In *Lancet*, (1) 1986, s. 307–10.
- Cyhelský, L., Kahounová, J., & Hindls, R. (2001). *Elementární statistická analýza*. (2. dopl. vyd., 318 s.) Praha: Management Press.
- Hendl, J. Statistické přístupy k porovnání biomedicínských metod měření. In. *Česká kinantopologie*. 1997, roč. 1, č. 2, s. 87-96.
- Hendl, J. (2006). *Přehled statistických metod zpracování dat: analýza a metaanalýza dat*. (Vyd. 2., opr., 583 s.) Praha: Portál.
- Meloun, M., & Militký, J. (1998). *Statistické zpracování experimentálních dat*. (2. vyd., xxi, 839 s.) Praha: East Publishing.
- Řezanková, H. (2005). *Analýza kategoriálních dat*. (Vyd. 1., 99 s.) Praha: Oeconomica.

Rejstřík

Příloha 1 - data

poměr pas/boky	BMI
0,60	18,68
0,73	16,77
0,75	19,60
0,76	19,22
0,76	18,95
0,77	18,08
0,77	18,07
0,77	18,82
0,77	17,39
0,77	20,19
0,78	18,98
0,78	17,40
0,78	19,08
0,78	18,96
0,78	17,26
0,78	19,39
0,78	17,57
0,78	17,93
0,79	25,13
0,79	19,82
0,79	22,01
0,79	21,16
0,79	20,25
0,79	19,54
0,79	20,15
0,79	18,93
0,80	17,54
0,80	20,73
0,80	20,55
0,80	20,69
0,80	18,75
0,80	19,69
0,80	21,31
0,80	21,43
0,80	19,98
0,80	20,27
0,80	19,30
0,80	20,15
0,80	19,55
0,81	19,40

poměr pas/boky	BMI
0,81	18,86
0,81	21,74
0,81	23,61
0,81	20,90
0,81	21,56
0,81	20,64
0,81	22,11
0,81	18,44
0,81	21,76
0,81	18,45
0,81	20,69
0,81	19,16
0,81	20,69
0,81	19,65
0,81	21,08
0,81	20,40
0,81	20,58
0,82	19,56
0,82	20,57
0,82	21,12
0,82	20,76
0,82	20,80
0,82	21,60
0,82	21,48
0,82	19,96
0,82	21,08
0,82	23,46
0,82	21,21
0,82	19,64
0,82	19,54
0,82	21,38
0,82	20,92
0,82	20,13
0,82	23,12
0,82	21,10
0,82	20,45
0,82	21,42
0,82	22,14
0,82	19,88
0,82	21,59

poměr pas/boky	BMI
0,82	22,31
0,82	22,31
0,82	20,09
0,82	20,87
0,82	19,34
0,83	22,11
0,83	21,59
0,83	21,35
0,83	23,05
0,83	22,44
0,83	21,76
0,83	21,15
0,83	22,28
0,83	20,17
0,83	22,07
0,83	20,44
0,83	21,25
0,83	22,03
0,83	22,49
0,83	20,01
0,83	21,15
0,83	22,39
0,83	21,64
0,83	21,62
0,83	20,68
0,83	20,47
0,83	20,90
0,83	21,30
0,83	21,45
0,84	22,50
0,84	20,82
0,84	21,44
0,84	21,92
0,84	22,68
0,84	21,91
0,84	21,45
0,84	21,43
0,84	22,66
0,84	23,17
0,84	22,68

poměr pas/boky	BMI
0,84	20,84
0,84	22,90
0,84	24,51
0,84	22,91
0,84	21,67
0,84	22,73
0,84	24,16
0,84	20,35
0,84	24,12
0,84	21,32
0,84	23,84
0,84	21,63
0,84	22,76
0,84	19,97
0,84	24,17
0,84	20,72
0,84	23,11
0,84	22,12
0,84	21,57
0,84	21,89
0,84	19,65
0,84	21,62
0,84	22,13
0,84	21,67
0,84	22,82
0,84	23,32
0,84	21,59
0,84	20,65
0,84	21,62
0,84	21,98
0,84	23,21
0,85	22,06
0,85	24,42
0,85	23,97
0,85	22,92
0,85	23,45
0,85	22,16
0,85	23,13
0,85	23,66
0,85	22,40

poměr pas/boky	BMI
0,85	23,64
0,85	22,91
0,85	23,96
0,85	20,65
0,85	21,80
0,85	22,26
0,85	20,90
0,85	24,12
0,85	22,90
0,85	20,87
0,85	22,01
0,85	22,68
0,85	22,66
0,85	22,56
0,85	21,50
0,85	21,68
0,85	23,67
0,85	22,97
0,85	23,19
0,85	21,23
0,85	21,07
0,85	21,46
0,86	23,14
0,86	25,47
0,86	22,62
0,86	22,53
0,86	21,68
0,86	22,77
0,86	24,60
0,86	24,26
0,86	23,86
0,86	23,69
0,86	23,81
0,86	24,49
0,86	24,63
0,86	23,55
0,86	23,26
0,86	24,78
0,86	22,59
0,86	24,35
0,86	23,15
0,86	21,41

poměr pas/boky	BMI
0,86	21,51
0,86	23,47
0,86	25,21
0,86	24,12
0,86	23,03
0,86	23,21
0,86	23,28
0,86	22,42
0,86	23,21
0,86	24,46
0,86	23,21
0,86	26,40
0,86	22,85
0,86	22,98
0,86	24,07
0,86	24,98
0,87	22,47
0,87	22,33
0,87	23,18
0,87	24,72
0,87	23,70
0,87	24,64
0,87	23,10
0,87	24,66
0,87	23,59
0,87	22,12
0,87	25,97
0,87	24,04
0,87	23,62
0,87	22,25
0,87	24,08
0,87	24,23
0,87	26,95
0,87	22,90
0,87	24,39
0,87	24,11
0,87	23,95
0,87	23,35
0,87	22,54
0,87	24,90
0,87	22,41
0,87	25,52

poměr pas/boky	BMI
0,87	25,48
0,87	22,56
0,87	24,45
0,87	25,39
0,87	25,61
0,87	23,91
0,87	23,43
0,87	23,44
0,87	22,67
0,87	25,66
0,87	23,36
0,87	23,63
0,87	22,97
0,87	22,15
0,87	24,71
0,87	22,08
0,87	23,37
0,87	24,49
0,87	25,42
0,87	25,30
0,87	24,47
0,87	22,64
0,87	25,25
0,87	23,49
0,88	22,20
0,88	24,54
0,88	25,83
0,88	27,20
0,88	24,25
0,88	24,97
0,88	24,09
0,88	26,28
0,88	25,28
0,88	24,30
0,88	24,89
0,88	25,29
0,88	24,13
0,88	24,35
0,88	25,13
0,88	25,74
0,88	24,36
0,88	23,33

poměr pas/boky	BMI
0,88	22,11
0,88	22,97
0,88	25,48
0,88	23,68
0,88	21,77
0,88	22,88
0,88	23,95
0,88	24,20
0,88	23,32
0,88	24,06
0,88	25,55
0,88	23,61
0,88	23,68
0,88	22,55
0,88	23,83
0,88	23,90
0,88	25,17
0,88	24,12
0,88	26,47
0,88	26,35
0,88	22,58
0,88	24,20
0,88	22,99
0,88	26,76
0,88	24,24
0,88	24,11
0,89	23,39
0,89	26,38
0,89	25,73
0,89	25,77
0,89	25,83
0,89	27,26
0,89	25,02
0,89	24,45
0,89	29,05
0,89	27,00
0,89	26,08
0,89	26,89
0,89	27,59
0,89	26,14
0,89	26,54
0,89	24,35

poměr pas/boky	BMI
0,89	24,77
0,89	30,20
0,89	22,09
0,89	25,84
0,89	24,23
0,89	25,44
0,89	26,89
0,89	23,61
0,89	23,38
0,89	26,34
0,89	25,88
0,89	26,05
0,89	25,17
0,89	26,07
0,89	27,51
0,89	23,86
0,89	23,50
0,89	24,99
0,89	25,66
0,89	23,00
0,89	25,59
0,89	24,08
0,89	25,15
0,89	24,94
0,89	24,74
0,89	24,80
0,89	23,06
0,90	26,52
0,90	25,49
0,90	27,18
0,90	27,22
0,90	24,70
0,90	29,38
0,90	24,22
0,90	22,24
0,90	24,51
0,90	26,91
0,90	24,06
0,90	22,63
0,90	26,23
0,90	27,64
0,90	26,26

poměr pas/boky	BMI
0,90	28,24
0,90	24,86
0,90	27,62
0,90	26,60
0,90	28,00
0,90	22,90
0,90	24,76
0,90	25,28
0,90	25,01
0,90	26,14
0,90	23,54
0,90	27,04
0,90	25,11
0,90	24,51
0,90	27,12
0,90	24,86
0,90	25,34
0,90	26,31
0,91	25,51
0,91	26,27
0,91	29,64
0,91	28,67
0,91	30,06
0,91	32,34
0,91	25,05
0,91	26,22
0,91	27,80
0,91	25,08
0,91	24,61
0,91	27,83
0,91	24,18
0,91	26,86
0,91	27,43
0,91	25,75
0,91	28,09
0,91	30,99
0,91	28,12
0,91	26,07
0,92	24,52
0,92	26,01
0,92	26,96
0,92	28,58

poměr pas/boky	BMI
0,92	27,40
0,92	28,34
0,92	30,65
0,92	31,67
0,92	28,30
0,92	28,18
0,92	27,41
0,92	28,56
0,92	27,27
0,92	27,39
0,92	27,50
0,92	27,96
0,92	30,13
0,92	29,65
0,92	26,62
0,92	28,74
0,92	28,22
0,92	31,28
0,92	29,30
0,92	24,78
0,92	27,21
0,92	26,08
0,92	27,01
0,92	24,16
0,92	26,75
0,92	27,41
0,92	27,18
0,92	25,63
0,92	26,02
0,92	26,54
0,92	29,55
0,92	28,85
0,92	25,63
0,92	27,11
0,93	25,42
0,93	27,99
0,93	27,65
0,93	31,33
0,93	26,26
0,93	26,34
0,93	25,26
0,93	27,37

poměr pas/boky	BMI
0,93	27,44
0,93	37,46
0,93	27,84
0,93	28,78
0,93	31,46
0,93	28,57
0,93	27,70
0,93	27,17
0,93	32,31
0,93	25,86
0,93	32,21
0,93	28,48
0,93	30,22
0,93	25,37
0,93	28,76
0,93	29,72
0,93	28,47
0,93	29,27
0,93	29,25
0,93	27,29
0,93	29,16
0,93	26,59
0,93	26,84
0,93	27,65
0,93	26,75
0,94	28,90
0,94	29,06
0,94	26,62
0,94	24,98
0,94	29,61
0,94	28,56
0,94	32,03
0,94	29,06
0,94	27,70
0,94	25,37
0,94	28,98
0,94	28,76
0,94	28,22
0,94	29,94
0,94	26,64
0,94	29,98
0,94	27,22

poměr pas/boky	BMI
0,94	29,66
0,94	28,85
0,94	28,02
0,94	26,46
0,94	27,56
0,94	32,80
0,94	27,36
0,95	28,79
0,95	31,89
0,95	31,20
0,95	28,62
0,95	29,37
0,95	29,21
0,95	28,50
0,95	30,56
0,95	29,31
0,95	32,85
0,95	25,84
0,95	29,34
0,95	30,66

poměr pas/boky	BMI
0,95	26,50
0,95	31,10
0,95	29,93
0,95	28,04
0,95	28,06
0,96	30,40
0,96	28,33
0,96	28,41
0,96	33,40
0,96	27,61
0,96	34,64
0,96	32,28
0,96	31,05
0,96	34,25
0,96	32,65
0,96	32,79
0,96	28,24
0,96	33,24
0,97	32,22
0,97	35,58

poměr pas/boky	BMI
0,97	32,65
0,97	31,85
0,97	31,74
0,97	29,36
0,97	30,59
0,97	35,29
0,97	30,55
0,97	31,18
0,97	30,39
0,97	33,98
0,98	29,69
0,98	31,56
0,98	30,40
0,98	29,22
0,98	31,86
0,98	33,55
0,98	36,61
0,98	27,99
0,98	31,83
0,98	32,35

poměr pas/boky	BMI
0,98	25,56
0,98	30,02
0,99	35,37
0,99	35,57
0,99	34,31
0,99	32,50
0,99	33,40
1,00	37,49
1,00	35,54
1,01	40,15
1,01	35,69
1,01	40,67
1,03	36,84