VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA STROJNÍHO INŽENÝRSTVÍ
ÚSTAV MATEMATIKY

FACULTY OF MECHANICAL ENGINEERING
INSTITUTE OF MATHEMATICS

STATISTICKÁ ANALÝZA DAT V EXCELU

STATISTICAL DATA ANALYSIS IN EXCEL

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — RADEK HUDEC
AUTHOR

VEDOUCÍ PRÁCE — doc. RNDr. ZDENĚK KARPÍŠEK, CSc.
SUPERVISOR

BRNO 2011

Vysoké učení technické v Brně, Fakulta strojního inženýrství

Ústav matematiky
Akademický rok: 2010/2011

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

student(ka): Radek Hudec

který/která studuje v **bakalářském studijním programu**

obor: **Matematické inženýrství (3901R021)**

Ředitel ústavu Vám v souladu se zákonem č.111/1998 o vysokých školách a se Studijním a zkušebním řádem VUT v Brně určuje následující téma bakalářské práce:

Statistická analýza dat v Excelu

v anglickém jazyce:

Statistical Data Analysis in Excel

Stručná charakteristika problematiky úkolu:

Popis vstupů a výstupů při řešení statistických úloh pomocí modulu Analýza dat v Excelu.

Cíle bakalářské práce:

Vypracování základní metodiky pro řešení statistických úloh s využitím modulu Analýza dat a statistických funkcí v software Excel. Popsat možnosti a omezení modulu a funkcí.

Seznam odborné literatury:

1. Brož, K.: Microsoft Excel. Základní příručka. Computer Press, 2001.
2. Šťastný, Z.: Matematické a statistické výpočty v Microsoft Excelu. Praha: Computer Press, 1999.
3. Firemní literatura firmy Microsoft a WWW stránky dle pokynů vedoucího práce.

Vedoucí bakalářské práce: doc. RNDr. Zdeněk Karpíšek, CSc.

Termín odevzdání bakalářské práce je stanoven časovým plánem akademického roku 2010/2011.

V Brně, dne 20.11.2009

L.S.

prof. RNDr. Josef Šlapal, CSc.	prof. RNDr. Miroslav Doupovec, CSc.
Ředitel ústavu	Děkan fakulty

Abstrakt

Tato bakalářská práce se zabývá statistikou v programu Excel. Cílem této práce je vypracování metodiky pro řešení statistických funkcí v software Excel. Popsat možnosti a omezení modulu a funkcí.

Klíčová slova

Statistika v Excelu, analýza dat, soubor, Excel.

Abstract

This bachelor thesis deals with statistics in Excel. The aim of this thesis is to develop methodology to deal with the statistical functions in Excel software – to describe capabilities and limitations of the module and function.

Keywords

Statistics in Excel, data analysis, sample, Excel.

HUDEC, R. Statistická analýza dat v Excelu. Brno: Vysoké učení technické v Brně, Fakulta strojního inženýrství, 2011. 27 s. Vedoucí bakalářské práce doc. RNDr. Zdeněk Karpíšek, CSc..

Prohlašuji, že jsem bakalářskou práci vypracoval samostatně pod vedením doc. RNDr. Zdeňka Karpíška, CSc., s použitím materiálů v seznamu literatury.

V Brně dne 27. 5. 2011 Radek Hudec

Rád bych poděkoval vedoucímu práce doc. RNDr. Zdeňkovi Karpíškovi, CSc. za vedení mé práce, trpělivost a rady, které pomohly ke zlepšení obsahové stránky této práce.

Radek Hudec

OBSAH

1. Úvod

Proč dělat statistiku v Excelu? Je výhodou, že na většině počítačů je tento program nainstalovaný, neboť je součástí MS OFFICE, a jednoduché statistické funkce jsou implementovány. Dále bych rád upozornil na to, že tento program obsahuje vlastní programovací jazyk, Visual Basic pomocí kterého můžeme sami funkce programovat.

2. Přehled statistických funkcí

Excel nabízí řadu základních funkcí, které je možné použít k rychlému statistickému zpracování dat. Omezení funkcí je 247 argumentů, což příjemné, neboť argument může být oblast buněk.

Musím upozornit, že funkce končící písmenkem A berou text a logickou hodnotu NEPRAVDA jako 0 a logickou hodnotu PRAVDA jako 1.

2.1. Popisná statistika

2.1.1. Základní informace o souboru dat

COUNTBLANK() – počet prázdných buněk.

ČETNOSTI() – vypočte počet výskytů hodnoty z oblasti hodnot.

POČET() – počet buněk obsahující čísla.

PERMUTACE() – počet permutací.

POČET() – počet buněk obsahující čísla.

POČET2() – počet buněk, které jsou neprázdné.

Percentil

PERCENTIL()

Minimum

MIN()

Maximum

MAX()

Kvartily

QUARTIL(pole;k) pro k = 0 minimum, pro k =1 dolní kvartil, pro k = 2 medián, pro k = 3 horní kvartil, pro k = 4 maximum.

SMALL(pole;k) vrátí k. nejmenší hodnotu výběru pole.

LARGE(pole;k) vrátí k. největší hodnotu výběru pole.

RANK(číslo;odkaz;pořadí) vrátí pořadí čísla z množiny odkaz. Parametr pořadí určí, zda se třídí vzestupně s parametrem 1 či sestupně (výchozí) s parametrem 0.

PERCENTRANK() – vrátí pořadí hodnoty vyjádřené procentuální částí množiny dat.

2.2. Charakteristiky polohy

2.2.1. Aritmetický průměr

PRŮMĚR(), AVERAGEA() – klasický aritmetický průměr

AVERGEAIF(),AVERGEAIFS() – aritmetický průměr omezený podmínkou, případně podmínkami.

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

TRIMMEAN() – průměrná hodnota vnitřní části množiny.

2.2.2. Geometrický průměr

GEOMEAN()

$$Geo = \sqrt[n]{\prod x_i}$$

2.2.3. Harmonický průměr

HARPMEAN()

$$Harp = \frac{1}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

2.2.4. Medián

MEDIAN() – střed souboru.

$\tilde{x} = x_{(\frac{n+1}{2})}$ pro n liché

$\tilde{x} = \frac{1}{2} \cdot \left(x_{(\frac{n}{2})} + x_{(\frac{n+1}{2})} \right)$ pro n sudé

2.2.5. Modus

MODE() – nejčetnější hodnota.

V případě, že máme více stejně četných hodnot, Excel zobrazí tu nejmenší hodnotu.

2.3. Charakteristiky variace

2.3.1. Rozptyl

VAR()

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

2.3.2. Výběrový rozptyl

VAR.VÝBĚR()

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

2.3.3. Směrodatná odchylka

SMODCH(), STDEVPA()

$s = \sqrt{s^2}$

2.3.4. Výběrová směrodatná odchylka

SMODCH.VÝBĚR (), STEDEVA()

$\hat{s} = \sqrt{\hat{s}^2}$

2.3.5. Výběrový koeficient (šikmosti) asymetrie

$A_3 = \frac{n}{(n-1)\cdot(n-2)} \cdot \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{s^3}$

SKEW()

2.3.6. Výběrový koeficient (špičatosti) excesu

$A_4 = \frac{n\cdot(n+1)}{(n-1)\cdot(n-2)\cdot(n-3)} \cdot \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{s^4} - \frac{3 \cdot (n-1)^2}{(n-2) \cdot (n-3)}$

KURT()

2.4. Kvantily

2.4.1. Spojitá rozdělení

BETADIST(x ; α ; β ; A ; B) – hodnota distribuční funkce beta rozdělení $F(x)$, kde $X \sim \beta(\alpha,\beta,A,B)$.

BETAINV(p;α;β;A;B) – kvantil $x_P = F^{-1}(P)$ beta rozdělení, kde $X \sim \beta(\alpha,\beta,A,B)$.

FDIST($x;k_1;k_2$) – hodnota distribuční funkce Fisherovo-Snedecorova rozdělení s k_1, k_2 stupni volnosti, kde $X \sim F(k_1, k_2)$.

FINV(1–p; $k_1;k_2$) – kvantil $F_P(k_1, k_2) = F^{-1}(P)$ Fisherovo-Snedecorova rozdělení, kde $X \sim F(k_1, k_2)$. $F_{1-\alpha}(\nu_1; \nu_2) = FINV(\alpha; \nu_1; \nu_2)$.

GAMMADIST($x;\alpha;\beta$;součet) –pro parametr součet = PRAVDA je vypočtena hodnota distribuční funkce gama rozdělení $F(x)$. A pro parametr součet = NEPRAVDA je vypočtena hodnota hustoty pravděpodobnosti gama rozdělení $f(x)$, kde $X \sim \Gamma(\alpha, \beta)$.

GAMMAINV(p; $\alpha;\beta$) – kvantil $x_P = F^{-1}(P)$ gama rozdělení, kde $X \sim \Gamma(\alpha, \beta)$.

CHIDIST(x;k) – hodnota distribuční funkce Pearsnova rozdělení s k stupni volnosti, kde $X \sim \chi^2(k)$.

CHIINV($1 - p$;k) – kvantil $\chi_P^2 = F^{-1}(P)$ Pearsnova rozdělení s k stupni volnosti. Inverzní funkce k CHIDIST(x;k). Tedy $\chi_\alpha^2(\nu) = CHINV(1 - \alpha; \nu)$.

NORMDIST($x;\mu;\sigma$; součet) – pro parametr součet = PRAVDA je vypočtena hodnota distribuční funkce normálního rozdělení $F(x)$. A pro parametr součet = NEPRAVDA je vypočtena hodnota hustoty pravděpodobnosti normálního rozdělení $f(x)$, kde $X \sim N(\mu, \sigma^2)$.

NORMINV(p ; μ ; σ) – kvantil normálního rozdělení $x_P = F^{-1}(P)$, kde $X \sim N(, \sigma^2)$. Inverzní funkce k NORMDIST($x;\mu;\sigma$; PRAVDA).

NORMSDIST(z) – hodnota distribuční funkce normovaného normálního rozdělení $F(x)$, kde $X \sim N(0,1)$. Stejné výsledky získáme NORMDIST(x;0;1; PRAVDA).

NORMSINV(p) – kvantil $x_P = F^{-1}(P)$ normovaného normální rozdělení, kde $X \sim N(0,1)$. Inverzní funkce k NORMSDIST(z).

TDIST($x;k$;strany) – hodnota distribuční funkce Studentova t-rozdělení s k stupni volnosti, kde $X \sim S(k)$. Parametr strany nabývá pouze hodnot 1 či 2 podle toho, zda se jedná o jednostranné nebo dvoustranné rozdělení.

TINV(2·(1-p);k) – kvantil $t_P = F^{-1}(P)$ Studentova t-rozdělení s k stupni volnosti, kde $X \sim S(k)$. Inverzní funkce k TDIST($x;k$;2). Tedy kvantil $t_{1-\frac{\alpha}{2}}(\nu) = TINV(\alpha; \nu)$.

WEIBULL($x;\alpha;\beta;typ$) – pro parametr typ = PRAVDA získáme hodnotu distribuční funkce Weibullova rozdělení, nebo pro parametr typ = NEPRAVDA je vypočtena hodnota hustoty pravděpodobnosti Weibullova rozdělení, kde $X \sim W(\lambda)$.

2.4.2. Diskrétní rozdělení

BINOMDIST($x;n;p$;počet) – pro parametr počet = PRAVDA získáme hodnotu distribuční funkce binomického rozdělení $F(x)$, nebo pro parametr počet = NEPRAVDA je vypočtena hodnota hustoty pravděpodobnosti binomického rozdělení $f(x)$, kde $X \sim Bi(n,p)$.

CRITBINOM(n; p; α) – je (1- α) kvantil binomického rozdělení $X \sim Bi(n,p)$.

EXPONDIST($x;\lambda$; součet) – pro parametr počet = PRAVDA získáme hodnotu distribuční funkce exponenciálního rozdělení $F(x)$, nebo pro parametr počet = NEPRAVDA je vypočtena hodnota hustoty pravděpodobnosti exponenciálního rozdělení $f(x)$, kde $X \sim Exp(\lambda)$.

HYPGEOMDIST($x;n;M;N$) – hodnota distribuční funkce hypergeometrického rozdělení $F(x)$, kde $X \sim Hg(N,M$,n).

LOGNORMDIST($x;\mu;\sigma$) – hodnota distribuční funkce logaritmicko-normálního rozdělení $F(x)$, kde $X \sim LN(\mu;\sigma^2)$.

LOGINV(p; $\mu;\sigma$) – inverzní funkce LOGNORMDIST().

NEGBINOMDIST($x;s;p$) – hodnota distribuční funkce negativně binomického rozdělení $F(x)$, kde $X \sim NegBi(s,p)$.

POISSON(x; λ;součet) – pro parametr součet = PRAVDA získáme hodnotu distribuční funkce Poissonova rozdělení, nebo pro parametr součet = NEPRAVDA je vypočtena hodnota pravděpodobnostní funkce Poissonova rozdělení, kde $X \sim Po(\lambda)$.

2.5. Regrese

2.5.1. Lineární regrese

$y = a + b \cdot x$

INTERCEPT(data_y; data_x) – vypočte průsečík regresní přímky s osou y. Parametr a.

SLOPE(data_y; data_x) – vypočte směrnici b regresní přímky.

LINREGRESE(pole_y;pole_x;b;stat) – jedná se o maticovou funkci, která má výstup větší než jedna buňka. Před použitím vybereme dostatečně velkou oblast buněk a do té zadáme funkci a pak zmáčkneme kombinaci kláves Ctrl+Shift+Enter, tím se vyplní vybrané buňky.

Kdybychom vynechali argument pole_x, Excel doplní řadu 1, 2, 3 atd. Argumenty b a $stat$ jsou volitelné a pouze typu logická hodnota. Výchozí nastavení je $b = PRAVDA$ a stat = NEPRAVDA. Parametr b určuje, zda chceme parametr b počítat normálním způsobem, když je NEPRAVDA, předpokládáme $b = 0$.

Parametr $stat$ = NEPRAVDA spočítá pouze parametry a a b, nebo když $stat$ = PRAVDA vypočte další regresní statistiky.

LINTREND(pole_y; pole_x; nove_x; b) – vrátí hodnoty y k novým x proložením regresním přímkou. Parametr b není nutno zadávat, pokud nechceme, aby regrese procházela počátkem.

FORECAST(x; pole_y;pole_x) – vrátí pouze jednu hodnotu x.

STEYX(pole_y;pole_x) vrátí standardní chybu při výpočtu lineární regrese.

2.5.2. Exponenciální regrese

$y = b \cdot m_1^{x1} \cdot m_2^{x2} \cdot \ldots \cdot m_n^{xn}$ případně $y = b \cdot m^x$

LOGLINREGRESE(pole_y;pole_x;b;$stat$) – vypočte parametry m_i i další parametry.

LOGLINTREND(pole_y;pole_x;nove_x;b) – proloží regresní parabolu.

2.5.3. Vícenásobná regrese

$y = m_1 \cdot x_1 + m_2 \cdot x_2 + \cdots x_n \cdot x_n + b$

2.6. Koeficienty kovariance a korelace

COVAR() – vrátí hodnotu kovariance dvou proměnných.

CORREL() – korelační koeficient dvou proměnných.

PEARSON() – Pearsonův koeficient korelace r.

RKQ() – druhá mocnina Pearsonova korelačního koeficientu pro lineární regresi.

2.7. Testy

Podrobnosti později u Analýzy dat.

FTEST() – F-test, kde je třeba zadat první soubor s větším rozptylem.

CHITEST() – test χ^2, kdy zjišťujeme nezávislost dvou výběrů.

TTEST() – pravděpodobnost Studentova t-testu.

ZTEST() – P-hodnota z-testu.

2.8. Další funkce

CONFIDENCE() – interval spolehlivosti pro střední hodnotu.

DEVSQ() – součet druhých mocnin odchylek od střední hodnoty výběru.

FISHER() – hodnota Fisherovy transformace.

GAMMALN() – přirozený logaritmus gama funkce.

PROB() – hodnoty oblasti bude mezi dvěma limitami.

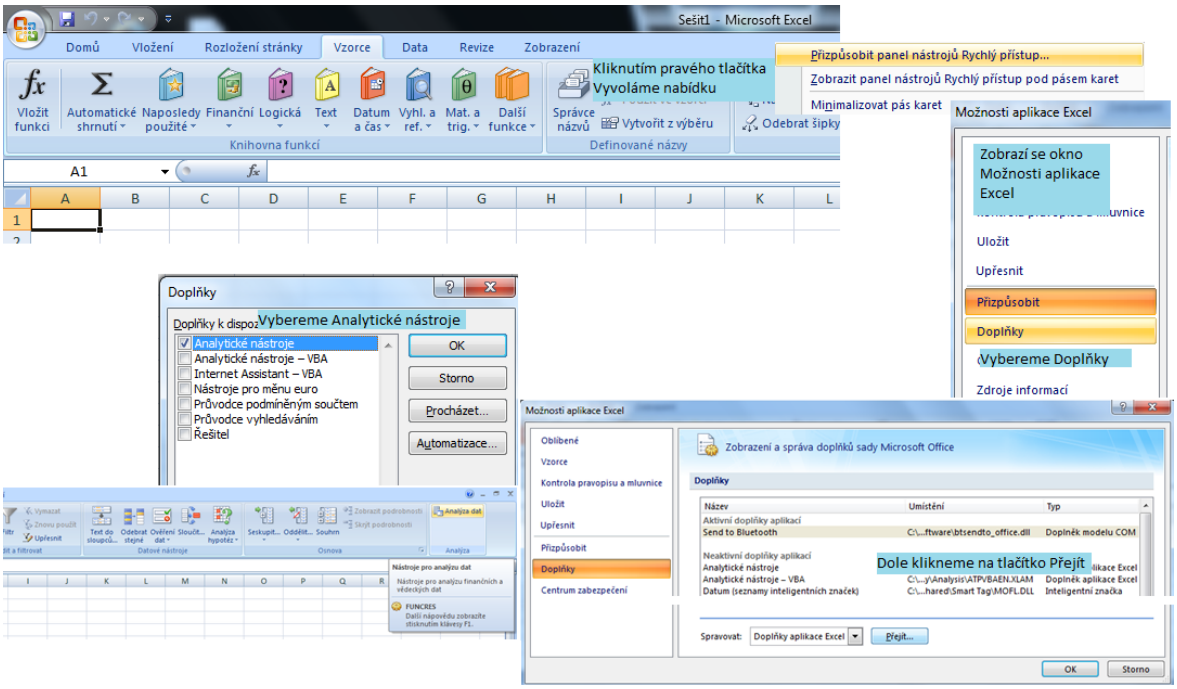
PRŮMODCHYLKA() průměrná hodnota absolutních odchylek od střední hodnoty výběru.

STANDARDIZE() – normalizovaná hodnota s normálním rozdělením.

3. Analýza dat

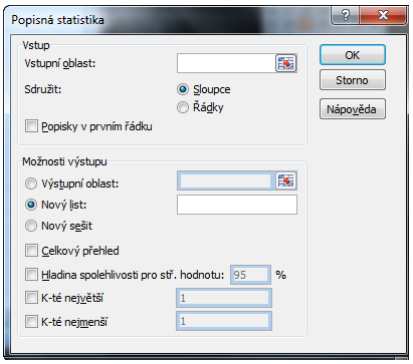
3.1. První spuštění Analýzy dat:

Nabídka  pak kliknutím na Možnosti aplikace Excel, kde na kartě Doplňky vybereme Spravovat: Doplňky aplikace Excel. Kliknutím na tlačítko Přejít, se nám zobrazí okno, kde vybereme Analytické nástroje a potvrdíme OK. Pak na panelu Data nalezneme kartu Analýzu a v ní Analýzu dat. Na obrázku je rychlejší cesta.



OBRÁZEK 3. 1

3.2. Popisná statistika

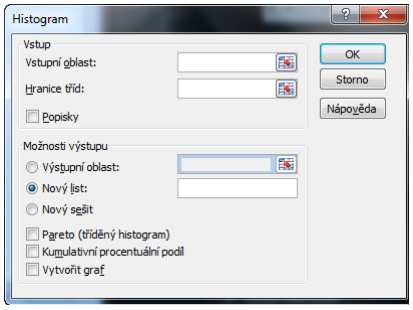


OBRÁZEK 3. 2

Výběrem popisné statistiky z analýzy dat získáme základní informace o souboru jako střední hodnota, medián, modus, směrodatná odchylka, rozptyl, špičatost, šikmost, minimální a maximální hodnotu. Oblast dat je brána jako výběr, tudíž všechny parametry jsou výběrové.

3.3. Histogram

Možnost histogram nám vytvoří četnostní tabulku, ale také umožní zobrazení histogramu – grafu.



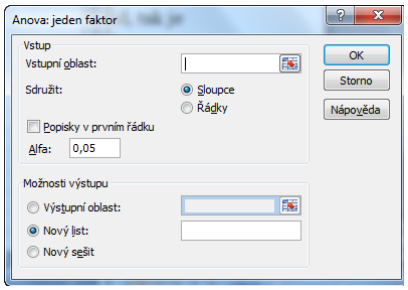
OBRÁZEK 3. 3

Pokud máme spojitou veličinu, je vhodné odstranit mezery mezi sloupci. Když nezadáme hranice tříd, pak si je Excel sám navrhne, ale u spojitých veličin to nejsou celá čísla. Zásadní chybou je, že v histogramu se zobrazují hranice tříd místo jejich středů. Dále bych rád upozornil, že v grafu u kumulativního procentuálního podílu, jsou na druhé ose y hodnoty větší než 100%.

3.4. Anova

ANOVA nebo taky analýza rozptylu. Výběr druhu záleží na počtu výběrů, které chceme testovat.

3.4.1. Anova: jeden faktor



OBRÁZEK 3. 4

Testuje rovnost středních hodnot, pokud výběry pocházejí ze stejného rozdělení a mají stejný rozptyl. Pro dva výběry se jedná o t-test. Data uspořádáme do sloupců vedle sebe. Hodnotu alfa musíme zadat číselně, nemůžeme se odkázat na žádnou buňku.

Výstup obsahuje dvě tabulky. V první je počet hodnot, součet, průměr a rozptyl. V druhé tabulce je rozdíl, ten má význam stupňů volnosti, SS, součet čtverců odchylek. Sloupec MS obsahuje sumy vydělené rozdílem, z toho je vypočítáno F-kritérium.

Hypotézu, že jednotlivé sloupce jsou ze stejného statistického výběru, nezamítáme, když F je menší než F krit.

3.4.2. Anova: dva faktory bez opakování

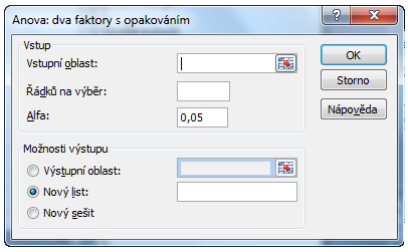
Když sledujeme vliv dvou faktorů měření, sestavíme data podle tabulky. Výstupem je o jednu tabulku více než v předchozí analýze s jedním faktorem. Nezamítáme, když F je menší než F krit. Jako vstupní oblast vybereme oblast podobnou níže uvedené.

	Skupina A	Skupina B
Data 1		
Data 2		

TABULKA 3.1

3.4.3. Anova: dva faktory s opakováním

Když sledujeme vliv dvou faktorů na měření, sestavíme data podle obrázku.

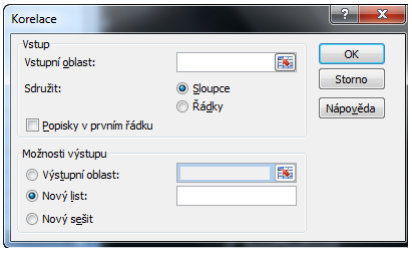


OBRÁZEK 3. 5

Použijeme, když máme data ve dvou různých dimenzích. Například: měříme výšku rostliny v závislosti na druhu hnojení a úrovni teploty. Musíme mít stejný počet měření pro všechny varianty. Do vstupní oblasti musíme zadat alespoň dvě sousední oblasti. Můžeme testovat, zda u různých druhů hnojení pochází výška ze stejného základního souboru, tedy ignorujeme úroveň teploty. Případně testujeme, zda u různých úrovní teplot pochází výška ze stejného základního souboru, tedy ignorujeme druh hnojení.

3.5. Korelace

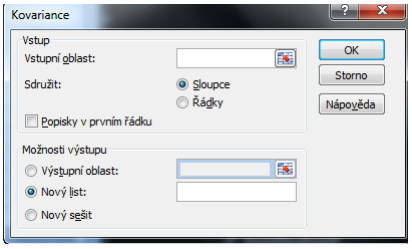
Spočte korelaci výběrů, kdy do vstupní oblasti zadáme sloupce s výběry.



OBRÁZEK 3. 6

3.6. Kovariance

Spočte kovarianci výběrů, kdy do vstupní oblasti
zadáme sloupce s výběry.



OBRÁZEK 3. 7

3.7. Exponenciální vyrovnání

Nástroj exponenciální vyrovnání nahradí hodnotu budoucí hodnotou lineární kombinací současné hodnoty a hodnoty vyhlazené v předchozím kroku.

$$y_{t+1} = ky_t + (1-k) \cdot x_t = y_t + (1-k) \cdot (x_t - y_t)$$

Vstupní oblastí jsou data sloupec nebo řádek alespoň 4 hodnot x_i, kde $i = 1, 2, \dots, t$. Pokud nezvolíme koeficient útlumu, pak $k = 0{,}3$.

3.8. Klouzavý průměr

Nástroj Klouzavý průměr je nejjednodušší ze všech dolních propustí. Nové hodnoty jsou nahrazeny aritmetickým průměrem předchozích n hodnot. $y_t = \frac{\sum x_i}{n}$, pro $i = t - n + 1, \dots, t$. Parametr interval je roven hodnotě n. Standardní chyby jsou vypočteny následovně: $s_t = \sqrt{\frac{\sum (y_i - x_i)^2}{n}}$.

3.9. Fourierova analýza

Nástroj Fourierova analýza využívá metodu rychlé Fourierovy transformace. Počet hodnot vstupní oblasti musí být sudá mocnina čísla 2 a před zápornou hodnotu musíme zadat apostrof (‘), ale nejvíc 4096 hodnot.

3.10. Generátor pseudonáhodných čísel

Čísla se generují podle vzorců, které mají periodický charakter. To znamená, že se po určitém počtu opakují, proto se nazývají pseudonáhodná čísla. Počet proměnný je počet sloupců a počet náhodných čísel je počet řádků. Základ generátoru ovlivní generovaná čísla, proto se při stejném základě generují stejná čísla.

Velmi dobře generuje diskrétní hodnoty například falešná kostka.

3.11. Pořadové statistiky a percentily

Vstupní data jsou sdružena do sloupců, ale mohou být sdružena i do řádků.

3.12. Regrese

Nástroj Regrese provádí lineární regresi. Pomocí MNČ proloží body přímku. Vstupní data jsou seřazena do sloupců a omezení je 16 proměnných.

Testovací statistika $F = \frac{R^2}{1-R^2} \cdot \frac{n-p}{p-1}$

3.13. Vzorkování

Nástroj Vzorkování vytvoří vzorek ze souboru tak, že považuje vstupní oblast za soubor. Je-li soubor příliš rozsáhlý a nelze ho celý zpracovat, pak použijeme vzorek. Vstupní oblast obsahuje hodnoty souboru, které chceme vzorkovat. Vzorky jsou vybírány nejprve z prvního sloupce postupně další.

3.14. Dvouvýběrový F-test pro rozptyl

Test rovnosti rozptylů, ale nejdříve musíme ověřit, zda jsou výběry nezávislé pomocí testu korelace.

3.15. Dvouvýběrový párový t-test na střední hodnotu

Předpokládáme spárované výběry, neboť provádíme dvojice měření za jiných podmínek.

Testujeme $H_0: \mu_1 = \mu_2 + \Delta\mu$ proti $H_1: \mu_1 \neq \mu_2 + \Delta\mu$.

$t\ Stat = \frac{(\overline{x_1} - \overline{x_2})}{s_\Delta}$, kde $s_\Delta = \sqrt{\frac{\sum(R_i - \bar{R})^2}{N \cdot (N-1)}}$, kde $R_i = x_{1i} - x_{2i}$ a $R = \frac{\sum Z_i}{n}$

H_0 zamítáme na hladině významnosti α, když $|t\ Stat| > t\ krit\ (2)$, kde $t\ krit\ (2) = t_\alpha(n_1 + n_2 - 2)$.

3.16. Dvouvýběrový t-test s rovností rozptylů

Testujeme $H_0: \mu_1 = \mu_2 + \Delta\mu$ proti $H_1: \mu_1 \neq \mu_2 + \Delta\mu$.

$t\ Stat = \frac{(\overline{x_1} - \overline{x_2})}{\bar{s}_{12}}$, kde $\bar{s}_{12} = \sqrt{\frac{(n_1-1)\cdot s_1^2 + (n_2-1)\cdot s_2^2}{n_1+n_2-2} \cdot \frac{n_1+n_2}{n_1 \cdot n_2}}$

H_0 zamítáme na hladině významnosti α, když $|t\ Stat| > t\ krit\ (2)$, kde $t\ krit\ (2) = t_\alpha(n_1 + n_2 - 2)$.

3.17. Dvouvýběrový t-test s nerovností rozptylů

Testujeme $H_0: \mu_1 = \mu_2 + \Delta\mu$ proti $H_1: \mu_1 \neq \mu_2 + \Delta\mu$.

$t\ Stat = \frac{(\overline{x_1} - \overline{x_2})}{s_{12}} \sim t_\alpha(\nu)$, kde $s_{12} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ a $\nu \doteq \frac{{s_{12}}^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$ (zaokrouhlíme na celé číslo)

H_0 zamítáme na hladině významnosti α, když $|t\ Stat| > t\ krit\ (2)$, kde $t\ krit\ (2) = t_\alpha(\nu)$.

3.18. Dvouvýběrový z-test na střední hodnotu

Nutno mít dopředu zjištěné rozptyly, neboť je nepočítá a musí se při výpočtu zadat.

$z = \frac{(x_{prům1} - x_{prům2})}{s_{12}}$, kde $s_{12} = \sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}$ a s_1, s_2 jsou výběrové směrodatné odchylky

Když $|z| > z_{krit}$ tak jsou střední hodnoty různé.

4. Vstupy a výstupy v Excelu

4.1. Import

Data buď můžeme zadat ručně nebo je můžeme importovat z různých souborů, do kterých je uložili jiné program.

Import dat z textového souboru: Na kartě Data vybereme z oblasti Načíst externí data možnost z textu. Zobrazí se okno, ve kterém vybereme soubor. Máme dvě možnosti jak rozdělit data do sloupců, buď pomocí oddělovače, tím může být čárka nebo tabulátor, nebo víme, že se jedná o data, která mají stejnou délku, třeba telefonní nebo rodné čísla. Když vybereme pevnou šířku, zobrazí se nám soubor, ve kterém jsou šipky, s kterými můžeme pohybovat a rozdělit data do sloupců.

Dále určíme formát dat. Máme na výběr text, datum a obecný, ale tento typ je nepraktický, či můžeme daný sloupec přeskočit. Nakonec vybereme místo, kde se data uloží, proto by od vybrané buňky dolů a doprava nic nemělo být, jinak by se naše data mohly smazat.

4.2. Export

Výsledky můžeme různě uložit a tím zjednodušit další použití. Můžeme vytvořit různé textové soubory, kde jsou data od sebe odděleny středníky, tabulátory nebo jen mezerami. Dále můžeme exportovat tabulku nebo graf.

5. Zajímavé stránky

Rozšíření Excelu o Add-In seznam dostupných rozšíření:

http://www.dmoz.org/Computers/Software/Spreadsheets/EXCEL/Add-Ins/

Rozšíření o statistiXL lze najít na http://www.statistixl.com/ je to soubor statistických procedur.

Po nainstalování statistiXL se zobrazí na kartě doplňky zobrazí Příkazy nabídky, které obsahují: Lumenaut Decision, Lumenaut Monte, Lumenaut Statistics a statistiXL.

6. Závěr

Cílem bakalářské práce bylo vypracování základní metodiky pro řešení statistických úloh s využitím modulu Analýza dat a statistických funkcí v Excelu. Popsat možnosti a omezení modulu funkcí z hlediska uživatele při statistických výpočtech v aplikacích. Popis je orientován na verze Excelu 2007 a Excelu 2010. Výhodou Excelu je to, že každou funkci po změně dat v oblasti znova přepočítá, ale nefunguje to u nástrojů Analýzy dat.

- jsou popsána omezení a možnosti funkcí a modulu
- nevyužití potence
- špatná terminologie

Na práci je možné navázat a vytvořit vlastní makro, které by mohlo otestovat dva výběry tak, že by nejdříve zjistilo, jestli mají stejné rozdělení. Pak dále testovat, jestli můžeme předpokládat, že mají stejné rozptyly. Dále bychom otestovali, jestli mají stejné i střední hodnoty.

Excel neumí práci s operátory, tudíž nedokáže vybrat každou k. hodnotu. Nebo podle kritéria vůči jednomu sloupci vybrat data z jiného sloupce.

LITERATURA

[1] ANDĚL, Jiří. Matematická statistika. 1. vyd. Praha : SNTL/ALFA, 1978. 346 s.

[2] BARILLA, Jiří, SIMR, Pavel, SÝKOROVÁ, Květuše. Microsoft Excel 2010 : podrobná uživatelská příručka. 1. vyd. Brno : Computer Press, 2010. 416 s. ISBN 978-80-251-3031-5.

[3] ČERNÝ, Jaroslav. Excel 2000 - 2007: záznam, úprava a programování maker. 1. vyd. Praha : Grada, 2008. 184 s. ISBN 978-80-247-2305-1.

[4] DODGE, Mark, STINSON, Craig. Mistrovství v Microsoft Office Excel 2007. 1. vyd. Brno : Computer Press, 2008. 936 s. ISBN 978-880-251-1980-8.

[5] MELOUN, Milan, MILITKÝ, Jiří. Statistická analýza experimentálních dat. 2. vyd. Praha : Academia, 2004. 954 s. ISBN 80.200.1254-0.

[6] ŠŤASTNÝ, Zdeněk. Matematické a statistické výpočty v Microsoft Excelu. 1. vyd. Brno : Computer Press, 1999. 254 s. ISBN 80-7226-141-X.

[7] *Office.com* [online]. 2003 [cit. 2011-05-26]. About statistical analysis tools. Dostupné z WWW: <http://office.microsoft.com/en-us/excel-help/about-statistical-analysis-tools-HP005203873.aspx>.