1 Statistický Program NCSS 6.0.

Manuál k programu NCSS 6.0. začíná stručným seznámením s uživatelským prostředím programu NCSS 6.0. (kapitola 1.1.). Kapitola 1.2. zmiňuje typy proměnných používaných v programu. Následuje několik slov o testování statistických hypotéz (kapitola 1.3.), které je klíčem k správnému pochopení dalších statistických metod a pomůže uživateli snáze se vyvarovat vyvozování chybných závěrů a chybným formulacím výsledků. Za touto teoretickou částí následuje podrobnější seznámení s nejčastěji používanými statistickými metodami: testy o jednom, dvou nebo několika výběrech (kapitola 1.4.), korelace a regrese (kapitoly 1.5. a 1.6.), následuje diskriminační analýza (kapitola 1.7.), logistická regrese (kapitola 1.8.), kontingenční tabulky (kapitola 1.9.), shluková analýza (kapitola 1.10.), ordinační metody (1.11.) a konečně anglicko-český slovník základních statistických pojmů.

V manuálu je u některých metod vysvětlen i princip, na kterém pracují, zadávání dat do procedury; dále jsou zde vysvětleny nejdůležitější výsledky programu NCSS a jejich interpretace. I zde platí, že pochopení principu procedury zjednodušuje uživateli čtení výsledků a jejich interpretaci. Manuál k programu NCSS má jistě své "mouchy", proto pro zájemce o podrobnější studium doporučuji literaturu použitou při psaní tohoto textu a uvedenou za manuálem. Vesměs to jsou čtivé publikace psané i pro nematematicky vzdělaného člověka.

1.1 *Úvod*

Statistický program NCSS je kompatibilní s programy Excel 95, 97 a Word 95, 97. Mezi těmito programy mohu tudíž jednoduše a podle libosti kopírovat a vkládat data či výsledky statistických procedur. V programu NCSS jsou tři základní okna: *Spreadsheet* slouží pro zápis dat, *Procedure Template* k zadání konkrétní statistické procedury a *Output* znázorňuje výsledky procedury. Nechybí ani nápověda v anglickém jazyce.

Spreadsheet

Pro vytvoření nového souboru si můžeme vybrat jeden ze dvou formátů: **S0 formát** se doporučuje pro soubor s rozsahem menším než 50 sloupečků a 500 řádků, **ZDB formát** může obsahovat až 5 000 sloupečků a 32 000 řádků.

Data můžeme zadávat buď přímo do programu NCSS nebo je přenést z programu Excel. Zadáváme je tak, aby sloupečky byly generovány jednotlivými proměnnými a každý řádek jedním pozorováním (měřením). Stanovujeme-li např. hmotnost sacharidů u rostlin pěstovaných v kyselém a neutrálním prostředí, pak pro některé procedury (dvouvýběrové testy) můžeme zadat data dvojím způsobem. **Typ 1**: hodnoty hmotnosti sacharidů rostlin napíšeme pro každý typ prostředí zvlášť do nového sloupečku (dostaneme tak 2 proměnné). **Typ 2**: hodnoty hmotnosti sacharidů pro oba typy prostředí zadáme do jednoho sloupečku (do jedné proměnné) pod sebe a typ prostředí do druhého sloupečku (např. 0 - kyselé, 1 - neutrální). Je možné použít textové označení místo číselného kódování 0, 1. Při použití analýzy rozptylu je již nutné mít zadání podle typu 2.

Ve *Spreadsheetu* můžeme přepínat mezi dvěma okny: ***Sheet*** (v ZDB zvaný ***Data Window***) slouží pro zápis dat a ***Variable Info*** pro práci s proměnnými.

a) Funkce ve ***Variable Info***: označení proměnných (v kolonce *Name;* nedopuručuji nechávat v názvech proměnných mezery: občas to může činit potíže), dále transformace (kolonka *Transformation*), zaokrouhlování hodnot či jejich převádění na procenta (*Format*). Podrobněji bych na příkladu vysvětlil použití transformací. Mějme za úkol získat odmocninovou **transformaci** proměnné "Hmotnost_sacharidů". Pod prázdný sloupeček do kolonky *Transformation* zadáme: "sqrt(Hmotnost_sacharidů)" (sqrt je zkratka *square root*, čili druhá odmocnina). Poté musíme s kurzorem utéct z okénka a znovu se na daný řádek vrátit. Spustíme *Edit*; *Transform Current* a v novém sloupečku v *Sheetu* se objeví transformované hodnoty veličiny "Hmotnost_sacharidů". Někdy se může stát, že transformace neproběhne. Stává se to v případech, kdy máme data v souboru v několika *Sheetech*. Proto doporučuji soubory pouze s jedním *Sheetem*.

Pokud bychom, ve výše popsaném příkladu, chtěli mít ve výsledcích pro proměnnou „Typ_prostředí" místo hodnot 0, 1 slovní označení „kyselé", „neutrální", postupovali bychom takto: Do *Sheetu* (viz dále) zadáme do prázdného sloupečku pod sebe dvě čísla 0 a 1 a označíme ho jako "kod_Typ_prostředí" a do následujícího prázdného sloupečku do řádku, ve kterém je č. 0 napíšeme "kyselé" a do řádku s č. 1 "neutrální". Tento sloupeček označíme "text_ Typ_prostředí". Potom ve *Variable Info* do kolonky *Value Label* řádku „Typ prostředí" zadáme "kod_Typ_prostředí".

A konečně abychom dokončili požadovaný úkol, musíme v *Template* (viz dále) před spuštěním procedury aktivovat kolonku *Value Labels* na *Value Labels*.

b) V okénku ***Sheet*** (***Data Window***) je jednak celá řada funkcí adekvátních k programu Excel či Word a jednak speciální nabídka funkcí. Jednou z těch speciálních je *Filter* uložený v nabídce *Data*. Umožňuje odfiltrovat určité hodnoty definované proměnné. Př.: Chceme pracovat pouze s určitými, nikoliv všemi, hodnotami proměnné; např. odfiltrovat hodnoty nižší než číslo 100. Použití je vysvětleno v nápovědě pod heslem *Filter - using*. K tomuto heslu se v nápovědě dostaneme po spuštění procedury "Hledej" a zadáním hesla. V nabídce *Analysis* najdeme výběr celé řady statistických metod; o mnohých bude řeč dále. A konečně v nabídce *Graphics* si můžeme vybrat jeden z nabízených grafů. Přesto z vlastních zkušeností doporučuji sestrojovat grafy spíše v programu *Excel* či v jiném statistickém programu. Výběrem příslušné procedury či grafu se dostaneme do dalšího okna: *Procedure Template*.

Procedure Template slouží k aktivaci procedury: zadání proměnné, počátečních podmínek, volba výpočetního algoritmu, atd. Každá statistická procedura má vlastní *Template*. V jeden čas může být otevřeno maximálně jedno *Template* (mohu pracovat vždy jen v jedné proceduře).

U začátečníků bývá problém, zejména u metod se dvěma a více výběry, určit, do které kolonky zadat analyzovanou proměnnou. Je-li zadání typu 1 (viz výše), pak zkoumané proměnné zařadíme do *Response Variable*. Je-li zadání typu 2, pak do *Response Variable* zadáme pouze proměnnou určující obsah sacharidů a do řádku *Group Variable* (v některých procedurách *Factor Variable*) proměnnou udávající typ prostředí. Proměnné vybíráme pomocí okénka *Search Variables.* Pro zjednodušení práce uživatele můžeme námi nadefinovaný *Template* uschovat (*File; Save Template*) a kdykoli ho opět vyvolat (*File; Load Template*).

Máme-li zvolené proměnné, pak ke spuštění procedury slouží výběr *Run*; *Run Procedure*, přes který se dostaneme do posledního okénka *Output*.

Okénko ***Output*** zobrazuje výsledek poslední procedury, po každém spuštění procedury se přepisuje. Aktivní okno výsledku si můžeme uložit zvlášť do souboru (*File*; *Save As*), nebo do souboru společně s výsledky jiných procedur - ukládáme si je průběžně do okénka zvaného ***Log*** pomocí *Add Output to Log* a celý *Log* uložíme do jednoho souboru kdykoli uznáme za vhodné. Z okénka *Output* se můžeme dostat do okénka *Log* přes výběr *Window*. Takto uložené výsledky procedur programu NCSS se automaticky otevírají v programu *Word*.

1.2 . Typy proměnných

- INTERVALOVÁ PROMĚNNÁ (*interval variable*) - data mohou nabývat všech hodnot z určitého intervalu (Př.:výška, čas, teplota).
- ORDINÁLNÍ PROMĚNNÁ (*ordinal variable*) - měření utříděna podle čísel ve smyslu "více" nebo "méně". (Př.: 5-silně neodpovídá, 4-neodpovídá, 3-neutrální, 2-odpovídá, 1-silně odpovídá).
- POMĚROVÁ PROMĚNNÁ (*ratio variable*) - mezi přilehlými jednotkami je konstantní rozdíl (mezi 22 a 23 cm je stejný rozdíl jako mezi 4 a 5 cm). Pro tuto proměnnou má smysl mluvit o poměrech, například 9 cm je třikrát více než 3 cm.
- NOMINÁLNÍ PROMĚNNÁ (*nominal variable*) - čísla reprezentují prostý výčet slovních kategorií. Užívají se pouze pro identifikační účely. (Př.: pohlaví, typ absolvované střední školy, barva vlasů).
- SYMETRICKÁ-BINÁRNÍ PROMĚNNÁ (*symmetric-binary variable*) - má dvojí možný výsledek: 1-ano; 0-ne. (Př.: členství ve straně, pohlaví).
- ASYMETRICKÁ-BINÁRNÍ PROMĚNNÁ (a*symmetric-binary variable*) - má také dvojí možný výsledek: buď přítomnost nebo nepřítomnost (také i malá pravděpodobnost) určitého znaku, ale zde nám absence znaku nic nenapoví. (Př.: člověk s jizvou na tváři může být snadno identifikovatelný; ale jestliže jizvu nemá, tak ho stěží můžeme identifikovat).

1.3 . Statistické rozhodování- testování statistických hypotéz

Statistické rozhodování je formalizované vynášení soudů o základním souboru na základě zjištění učiněných na výběrovém souboru. **Základním souborem** (**populací** - *population*) rozumíme celou populaci (např. populace všech lidí narozených po roce 1980) a **výběrovým souborem** (zkráceně výběrem - *sample*) pouze reprezentativní část populace, kterou zkoumáme a na níž provádíme měření. Výběr nám má sloužit k odvození závěrů platných pro celou populaci.

Prvním krokem při statistickém rozhodování je formulace tvrzení o populaci zvaného **nulová hypotéza H_0** (*null hypothesis*). Většinou to bývá opak toho, co chci dokázat. Formulujeme ji jako:

data se neliší, průměry se shodují, atd. Doplněk k nulové hypotéze se označuje jako **alternativní hypotéza H_1** (*alternative hypothesis*). Jestliže tedy není splněna nulová hypotéza, pak je splněna alternativní hypotéza (a naopak).

Alternativní hypotéza může mít jednak formu **jednostranného testu** (*one-sided test*), kdy je populační průměr buď vyšší než daná hodnota resp. průměr jiné populace nebo je nižší, a jednak může mít formu **testu dvoustranného** (*two-sided*), kdy se průměr od dané hodnoty liší (nehovoří se zde o uspořádání menší, větší). Př.: U jednostanného testu bychom formulovali alternativní hypotézu takto: průměrná výška mužů je větší než 170 cm resp. je větší než průměrná výška žen, a u dvoustranného testu takto: průměrná výška mužů je různá od 170 cm resp. se liší od průměrné výšky žen.

Protože při testování hypotézy o populaci jde o úsudek prováděný z údajů získaných náhodným výběrem, můžeme se ve svých úsudcích dopustit chybných závěrů. Může se nám stát, že zamítneme testovanou hypotézu H_0, ačkoli ve skutečnosti platí. Pak se dopouštíme **chyby prvního druhu** (*type I error*), jejíž pravděpodobnost se označuje jako α. Druhá možnost chybného závěru spočívá v tom, že přijmeme nulovou hypotézu, ačkoli ve skutečnosti platí alternativní hypotéza. Vzniká tak **chyba druhého druhu**, jejíž pravděpodobnost se označuje jako **ß**. Pravděpodobnost chyby prvního druhu α si na začátku volíme. Volíme si tedy, jak nepravděpodobný výsledek za předpokladu platnosti nulové hypotézy musíme dostat, abychom se rozhodli pro závěr, že nulová hypotéza neplatí. Většinou volíme α rovno 5% nebo 1% (píšeme α= 0,05, resp. 0,01) a označujeme ho jako **hladina významnosti testu** (*significance level*). Pravděpodobnost chyby druhého druhu je minimalizována testovacím postupem – maximalizuje se **síla testu** (*power*), což je hodnota 1-ß udávající, s jakou pravděpodobností zamítneme nulovou hypotézu, platí-li alternativní hypotéza.

Množina všech výsledků svědčících ve prospěch H_0 se označuje jako **obor přijetí** a množina všech výsledků při kterých zamítneme H_0 jako tzv. **kritický obor** (*region of rejection*). Hranice oddělující kritický obor a obor přijetí nazýváme **kritické hodnoty** (*critical values*). Můžeme je nalézt ve statistických tabulkách.

Z výběrového souboru spočítáme na základě daného testu **hodnotu testového kritéria** (*F-value*, *t-value*, *chi-square*,...-to záleží na použitém testu). **HYPOTÉZU H_0 ZAMÍTNEME NA HLADINĚ VÝZNAMNOSTI α, JESTLIŽE HODNOTA TESTOVÉHO KRITÉRIA PADNE DO KRITICKÉHO OBORU.** Abychom se nemuseli zabývat vyhledáváním kritických hodnot, udává se v testech **dosažená hladina testu** (*Prob Level, P-value*), což je za platnosti H_0 počítaná pravděpodobnost, že nám v příštím pokusu vyjde výsledek stejně nebo ještě více odporující H_0. **HYPOTÉZU H_0 ZAMÍTNEME NA HLADINĚ VÝZNAMNOSTI α, JESTLIŽE DOSAŽENÁ HLADINA TESTU VYJDE NIŽŠÍ NEBO ROVNA NEŽ HLADINA VÝZNAMNOSTI TESTU. Př.: Zvolíme-li hladinu významnosti testu 5% a vyjde-li dosažená hladina testu nižší nebo rovna hodnotě 0,05, pak nulovou hypotézu zamítneme (obdobně vyjde-li hodnota vyšší, nemůžeme zamítnout nulovou hypotézu).**

1.3.1 Správná formulace výsledku:

a) Jestliže nulová hypotéza není přijímána (*H_0 reject*), pak formulujeme výsledek takto: **hypotéza H_0 byla na hladině významnosti α zamítnuta**.
b) Je-li nulová hypotéza přijímána (*H_0 accept*), pak jsou chybné formulace typu "nulová hypotéza je prokázána", neboť bychom se mohli dopustit chyby s velkou, nekontrolovatelnou pravděpodobností. Správná interpretace je typu**: na základě dat nemůžeme zamítnout nulovou hypotézu**.

1.4 Testy o jednom, dvou nebo několika výběrech

Tyto testy můžeme rozdělit podle několika hledisek:

- Podle předpokladů testů a rozdělení proměnných rozlišujeme testy parametrické a neparametrické. Pro obě skupiny je nutné, aby výběr byl náhodný, dále je nutné spojité rozdělení a (vyjma párového testu) nezávislost proměnných.

  Parametrické testy - předpokládají normální rozdělení proměnných, shodu rozptylů a absenci odlehlých pozorování (kapitola 1.4.1.)

  Neparametrické testy - výše uvedené předpoklady nevyžadují; nepracují s aritmetickými průměry, ale s mediány nebo na úplně jiném principu.

  Poznámka: Pokud jsou splněny předpoklady parametrického testu, je lepší použít tento test než-li test neparametrický.

  Poznámka: Pro malý počet pozorování je vliv narušení předpokladů parametrického testu větší.
- Podle počtu sledovaných výběrů (proměnných):

Jednovýběrové testy (*One-Sample Tests*) na základě výběrového průměru rozhoduji o nulové hypotéze, že populační průměr μ je roven hodnotě μ_0 (kapitola 1.4.2.). Př.: Byly určeny hodnoty defoliace koruny stromů. Testuji hypotézu, že průměrná hodnota defoliace koruny je 30%.

Dvouvýběrové testy (*Two-Sample Tests*) rozhodují o nulové hypotéze shodnosti dvou populačních průměrů (popř. rozptylů) (kapitola 1.4.3.). Př.: V Krušných horách a na Šumavě byly u náhodně vybraných stromů určeny hodnoty defoliace koruny. Testuji hypotézu, že průměrná hodnota defoliace koruny stromů v Krušných horách je shodná s průměrnou hodnotou defoliace koruny stromů na Šumavě. Jednou populací v tomto případě rozumíme soubor stromů v Krušných horách a druhou soubor stromů na Šumavě.

Analýza rozptylu (*Analysis of variance*) je zobecněním dvouvýběrových testů. Místo dvou populací máme *k* populací ($k \geq 2$) a testujeme nulovou hypotézu, že jsou populační průměry shodné (kapitola 1.4.4.).

1.4.1 Jak ověříme v programu NCSS předpoklady parametrických testů?

A) **Odlehlá pozorování** (= vybočující hodnoty = *outliers*). Velmi silně ovlivňují zejména aritmetický průměr a směrodatnou odchylku. Mohou být příčinou toho, že je zamítnuta hypotéza normálního rozdělení nebo shody rozptylů. Odlehlá pozorování můžeme detekovat pomocí bodového grafu (*Scatter Plot*), histogramu (*Histogram*) nebo pomocí procedury *Data Screening* ve výběru *Descriptive Statistic*.

B) **Shoda rozptylů** Kritérium na testování nulové hypotézy shody rozptylů dvou či více výběrů je zahrnuto v těchto procedurách *One-Way ANOVA, Two-Sample T-Test, Descriptive Statistics*. Jako kritérium se zde používá modifikovaný Leveneův test shody rozptylů (*Modified-Levene Equel-Variance test*). Graficky můžeme rozhodovat o shodě rozptylů pomocí krabicových diagramů všech výběrů (procedura *Graphics - Box plot*).

C) **Normální rozdělení proměnných**. Testuje se ve stejných procedurách jako B), dále také v proceduře *One-Sampled T-Test*. K ověření slouží tyto testy: *Skewness Normality* (koeficient šikmosti), *Kurtosis Normality* (koeficient špičatosti), *Omnibus Normality of Residuals* (= $Skewness^2 + Kurtosis^2$). Někde jsou tyto tři koeficienty nahrazeny modifikovanými: *D´Agostino Skewness*, *D´Agostino Kurtosis* a *D´Agostino Omnibus*. Každý z těchto koeficientů testuje nulovou hypotézu, že proměnná má normální rozdělení, ovšem každý z jiného hlediska. **JESTLIŽE JE NULOVÁ HYPOTÉZA ALESPOŇ U JEDNOHO Z TĚCHTO KOEFICIENTŮ ZAMÍTNUTA (*REJECT*), MĚLI BYCHOM ZAPOCHYBOVAT O NORMÁLNÍM ROZDĚLENÍ PROMĚNNÉ!!!**

Pozn.: Jsou-li výběry velké (počet pozorování v nejméně rozsáhlém z nich je alespoň 50), pak se prakticky můžeme opřít o centrální limitní větu, podle níž např. průměr má obvykle asymptoticky normální rozdělení a můžeme tedy předpokládat normální rozdělení dat.

Jestliže není splněn předpoklad normálního rozdělení či shody rozptylů mezi výběry, pak, dříve než-li použijeme neparametrický test, pokusme se proměnnou **TRANSFORMOVAT** a znovu ověřit předpoklady parametrických testů. Transformace bývá často účinná. Používá se zejména druhá odmocnina (*square root*), přirozený logaritmus (*ln*), ale i jiné funkce.

1.4.2 Jednovýběrový t-test

Testuje nulovou hypotézu o rovnosti populačního průměru μ hodnotě μ_0. Předpokládá normální rozdělení dat bez odlehlých pozorování. Testové kritérium: $T = \dfrac{\overline{X} - \mu_0}{s/\sqrt{n}}$, kde $\overline{X}$ je výběrový průměr, *s* směrodatná odchylka a *n* počet hodnocených objektů.

Oboustrannou hypotézu H_0 zamítneme, je-li $|T| \geq t_{n-1}(1-\alpha/2)$ (*t* symbol pro Studentovo rozdělení, *n-1* počet stupňů volnosti (*DF – Degree of freedom*, α hladina významnosti testu).

Jestliže nejsou splněny ani po transformaci předpoklady testu, použijeme Wilcoxonův test nebo znaménkový test (*tab.1*)

❄ Při zadání v *Template* do kolonky *Response Variable* doplníme název dané proměnné a kolonku *Paired Variable* necháme prázdnou (vyplňuje se v případě použití párového testu). Konkrétní hodnotu μ_0 zapíšeme do kolonky *H_0 value*.

1.4.3 Dvouvýběrový t-test (dvouvýběrový Studentův test)

Testuje nulovou hypotézu o rovnosti dvou populačních průměrů (H_0: $\mu_1=\mu_2$). Předpokládá normální rozdělení obou výběrů, shodu rozptylů a nepřítomnost odlehlých pozorování. Testové kritérium: $T = \frac{\overline{X}-\overline{Y}}{s}\sqrt{\frac{nm}{n+m}}$, kde $\overline{X}$ a $\overline{Y}$ jsou výběrové průměry, *s* směrodatná odchylka a *n* resp. *m* počet pozorování pro první resp. druhý výběr. Oboustrannou hypotézu H_0 zamítneme, je-li $|T| \geq t_{m+n-2}(1-\alpha/2)$.

Jestliže nejsou předpoklady testů splněny ani po transformaci, použijeme jiný z dalších čtyř dostupných testů pro dva výběry v programu NCSS (*tab.1*).

❄ V *Template* do kolonky *Response Variable* doplníme název sledovaných proměnných a v kolonce *H_0 mean difference* necháme č. 0, pokud chceme testovat shodu populačních průměrů.

1.4.4 Analýza rozptylu jednoduchého třídění

Testuje nulovou hypotézu o rovnosti populačních průměrů na základě *k* výběrů (H_0: $\mu_1=\mu_2=\mu_3=...=\mu_k$). Alternativní hypotéza: Alespoň jeden populační průměr je odlišný od jiného.

Předpoklady testu: Nezávislé, náhodné výběry, normální rozdělení, shoda rozptylů.

Výpočet: Mějme nezávislé výběry $Y_{11},.....,Y_{1n1} \sim N(\mu_1, \sigma^2)$
$Y_{21},.....,Y_{2n2} \sim N(\mu_2, \sigma^2)$
.....
$Y_{k1},.....,Y_{knk} \sim N(\mu_k, \sigma^2)$,

kde Y_{ij} *j*-té pozorování ($j=1,2,....,n_i$) pořízené z *i*-tého výběru ($i=1,2,...,k$), *k* počet výběrů, *ni* počet prvků *i*-tého výběru, μ_i populační průměr *i*-tého výběru, σ^2 populační rozptyl, N symbol pro normální rozdělení.

V *k* nezávislých populacích máme tedy celkem $n=n_1+...+n_k$ pozorování, která musí být nezávislá.

Myšlenka této metody spočívá v porovnání variability mezi průměry v jednotlivých výběrech s celkovou variabilitou jednotlivých pozorování uvnitř výběrů. Pokud se výběrové průměry mezi sebou liší více, než očekáváme podle variability jednotlivých pozorování uvnitř výběrů, svědčí to proti testované nulové hypotéze, a tedy ve prospěch alternativní hypotézy tvrdící, že populační průměry nejsou shodné.

Celkový součet čtverců (*Total Sum of Squares - Adjusted* ve výsledku procedury v tabulce *Analysis of Variance Table Section*), roven celkové variabilitě dat $\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\overline{Y}\right)^2$ ($\overline{Y}$ je průměrná hodnota ze všech pozorování), rozložíme do dvou složek:

a) součet čtverců odchylek **mezi výběry** (ve výsledné tabulce NCSS horní řádek označen jako "*A*"):
$SS_A = \sum_{i=1}^{k} n_i\left(\overline{Y}_i-\overline{Y}\right)^2$, $\overline{Y}_i$ je průměrná hodnota zjištěná v *i*-tém výběru

b) součet čtverců odchylek **uvnitř výběrů** (v NCSS dolní řádek, označený "*S(A)*":
$SS_E = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\overline{Y}_i\right)^2$.

Označme dále tzv. **průměrné součty čtverců** odchylek (*Mean Sum of Squares*; v NCSS označeno "*Mean Squares*") mezi výběry a uvnitř výběrů jako $MS_A = \frac{SS_A}{k-1}$, $MS_E = \frac{SS_E}{n-k}$. Nulovou hypotézu budeme zamítat v případě, kdy je $F = \frac{MS_A}{MS_E} \geq F_{k-1,n-k}(\alpha)$, kde *k*-1 a *n-k* jsou počty stupňů volnosti (*DF*), *F* symbol pro Fisherovo rozdělení (v NCSS označen jako *F-Ratio*). Je-li tedy dosažená hladina testu (*Prob Level*) nižší než-li zvolená hladina významnosti α (většinou 0,05 nebo 0,01), pak nulovou hypotézu o shodě průměrů zamítneme.

Jestliže nulovou hypotézu zamítneme, mohlo by nás zajímat, které výběry jsou zodpovědné za toto zamítnutí (u kterých výběrů je nulová hypotéza o shodnosti populačních průměrů zamítnuta). K tomu nám slouží **mnohonásobné srovnávací testy** (*Multiple Comparison tests)*, z nichž se nejčastěji užívá test *Tukey-Kramer*. Pro výpočet je nutné v *Template* aktivovat okénko *Tukey-Kramer* na *Yes*.

Nejsou-li splněny předpoklady normality či shody rozptylů výběrů, použijeme výsledky **Kruskal-Wallisova testu** (druhý řádek "*Corrected for Ties*" - tzn. upravené pro opakující se hodnoty). Jako adekvátní mnohonásobný srovnávací test použijeme *Kruskal-Wallis Z test*. Opět i u něho musíme v *Template* aktivovat požadované okénko.
❄ Při zadávání dat do procedury píšeme analyzovanou proměnnou v *Template* do kolonky *Response Variable(s)* a proměnnou odlišující dané výběry do kolonky *Factor Variable*.

1.4.5 Testy o jednom, dvou a více výběrech v programu NCSS

Tab.1 udává přehled testů o jednom, dvou a více výběrech v programu NCSS.

Tab1: Přehled testů o jednom, dvou a více výběrech v programu NCSS. Sloupec normální rozdělení a shoda rozptylů nabývá + pokud je v daném testu nutný požadavek splnění tohoto předpokladu a - pokud daný požadavek nemusí být splněn.

	Použitý test (*název testu v NCSS*) (P- parametrický, N- neparametrický)	**Normální rozdělení**	**Shoda rozptylů**	**Poznámky**
Jednovýběrové testy (*One-Sample Tests*)	Jednovýběrový t-test *(One-Sample T-Test)* (P)	+		V proceduře *T-Tests*; *One-Sample T-Tests*
	Wilcoxonův test *(Wilcoxon Signed- Rank Test)* (N)	-		V proceduře *T-Tests*; *One-Sample T-Tests* Předpokládá symetrické rozložení dat okolo mediánu
	Znaménkový test (*Sign (Quantile) Test)* (N)	-		V proceduře *T-Tests*; *One-Sample T-Tests* Rozložení dat kolem mediánu nemusí být symetrické
Dvouvýběrové testy (*Two-Sample Tests*)	Dvouvýběrový t-test *(Equal-Variance T-Test)* (P)	+	+	V proceduře *T-Tests*; *Two-Sample T-Tests*
	Aspin-Welchův test neshodných rozptylů (*Aspin-Welch Unequal-Variance T-Test*) (P)	+	-	V proceduře *T-Tests*; *Two-Sample T-Tests*
	Mann-Whitneyův test (*Mann-Whitney U-Test*) (N)	-	-	V proceduře *T-Tests*; *Two-Sample T-Tests*. Ve výsledcích ekvivalentní s Wilcoxonovým testem pro dva výběry
	Wilcoxonův test pro 2 výběry *(Wilcoxon Rank-Sum Test)* (N)	-	-	V proceduře *T-Tests*; *Two-Sample T-Tests.* Ve výsledcích ekvivalentní s Mann-Whitneyovým testem
	Kolmogorov- Smirnovův test pro dva výběry (*Kolmogorov-Smirnov Test*) (N)	-	-	V proceduře *T-Tests*; *Two-Sample T-Tests*. Porovnává empirické distribuční funkce obou výběrů. Testuje jakoukoli neshodu mezi výběry (v průměru i rozptylu)
	Párový t-test (*Paired T-Test*) (P)	+		Pro zpracování **dvou závislých veličin**. Pro populace, kde je hypotéza o normálním rozdělení zamítnuta, se používá Wilcoxonův test (pro jeden výběr). V proceduře *T-Tests*; *Paired T-Tests* (*)
Analýza rozptylu- ANOVA (*Analysis of Variance)*	Analýza rozptylu jednoduchého třídění *(One-Way Analysis of Variance*) (P)	+	+	V proceduře *ANOVA; One-Way ANOVA*
	Kruskal- Wallisův test *(Kruskal-Wallis Test)* (N)	-	-	V proceduře *ANOVA; One-Way ANOVA*. Testuje se nulová hypotéza, že v každém z porovnávaných souborů je stejné rozdělení náhodné veličiny.
Mnohonásobné srovnávací testy (*Multiple-Comparison T.)*	Tukey- Kramerův test *(Tukey-Kramer Test)*	+	+	V proceduře *ANOVA; One-Way ANOVA.* Nutné v *Template* aktivovat kolonku *Tukey-Kramer Test* na *Yes*.
	Kruskal-Wallisův Z-test *(Kruskal-Wallis Z-Test)*	-	-	V proceduře *ANOVA; One-Way ANOVA.* Nutné v *Template* aktivovat kolonku *Kruskal-Wallis Z-Test* na *Yes*.

(*) **Poznámka k párovému testu:** Používá se při porovnání dvou populačních průměrů jestliže nejsou příslušné dvojice (páry) výběrových souborů nezávislé. Př.: Měříme sílu stisku levé (jedna proměnná) a pravé (druhá proměnná) ruky u 20 osob. Nulová hypotéza: síla stisku levé ruky se shoduje se sílou stisku pravé ruky. Přitom je pravděpodobné, že síla stisku levé ruky závisí na síle stisku pravé ruky (každý člověk je jinak silný).

1.5 . Korelace

Korelace určuje sílu lineární závislosti mezi dvěma proměnnými.
Poznámka: rozdíl mezi korelací a regresí. Korelace společně s regresí udávají vztah mezi dvěma proměnnými. Zatímco **korelace určuje sílu závislosti a obě proměnné v ní vystupují symetricky,** tak **regrese udává formu (způsob) závislosti** mezi dvěma veličinami (např. lineární závislost, exponenciální, polynomická, atd). U korelace nás tedy zajímá těsnost vzájemného vztahu a u regrese možnost ze známých hodnot jedné nebo několika náhodných veličin predikovat průměr jiné náhodné veličiny.

1.5.1 Korelační koeficient

Korelační koeficient udává **sílu lineární závislosti** mezi dvěma proměnnými. Dosahuje hodnot od -1 do 1. Hodnota nula udává, že mezi proměnnými není žádná lineární závislost, hodnoty blízké absolutní hodnotě 1 naopak predikují vysokou lineární závislost. Přičemž kladné hodnoty korelačního koeficientu udávají vztah přímé úměrnosti mezi proměnnými a záporné hodnoty vztah nepřímé úměrnosti. K hodnocení výsledků závislostí mezi více jak dvěma proměnnými se korelační koeficienty zapisují do **korelační matice**, jejíž diagonála je tvořena jedničkami.
Poznámka: Korelační koeficient, na rozdíl od regresního koeficientu, je bezrozměrné číslo, které vyjadřuje těsnost vztahu a jehož hodnota je nezávislá na použitých jednotkách.

Nejpoužívanější korelační koeficienty:
- PEARSONŮV KORELAČNÍ KOEFICIENT (*Pearson correlation coefficient*) - vyžaduje normální rozdělení dat, negativně ho ovlivňují odlehlá pozorování a neshoda rozptylů.
- SPEARMANŮV KORELAČNÍ KOEFICIENT (*Spearman correlation coefficients*) – Předpoklady Pearsonova korelačního koeficientu nepožaduje, ale je náchylný na mnoho nul v datech a velký počet chybějících údajů.

❄ V programu NCSS slouží k výpočtu korelačního koeficientu procedura *Correlation Matrix* (v *Regression/Correlation*). Je v ní zabudován výpočet Pearsonova a Spearmanova pořadového korelačního koeficientu (v kolonce *Correlation Type* si můžeme jeden z nich či oba vybrat).

Mohli bychom také potřebovat korelační koeficient udávající sílu vztahu mezi jednou binární a jednou spojitou veličinou (tzv. bodový biseriální korelační koeficient).
- BODOVÝ BISERIÁLNÍ (*POINT-BISERIAL*) KORELAČNÍ KOEFICIENT: jedna proměnná je binomická, druhá spojitá

$$r_p = \frac{|M_p - M_q|}{s} * \sqrt{(p * q)}$$

M_p resp. M_q aritmetický průměr skupiny, kde je binomická proměnná rovna č. 0 resp.1,
s směrodatná odchylka spojité proměnné, p,q proporce počtu pozorování M_p a M_q (např. je-li n_p=8 a n_q=12, pak p=0,4 a q=0,6)

1.5.2 Testovací kritéria korelačních koeficientů

Tab. 2, 3, a *4* udávají testovací kritéria korelačních koeficientů.

Tab.2: Testování nulové hypotézy o nezávislosti náhodných veličin (o rovnosti korelačního koeficientu číslu nule).
❄ V programu NCSS výběru *Other* použijeme *Probability Calculator*. Za *Probability Distribution* vybereme *Correlation* a v levé části se objeví pod sebou tři kolonky. Do první zadáme celkový počet měření, do prostřední číslo 0 a do spodní hodnotu korelačního koeficientu. Po vyplnění těchto tří kolonek zadáme *Calculate* a vpravo dole se nám objeví dvě vypočtené hodnoty. Menší z nich vynásobíme dvěma a vyjde nám dosažená hladina testu pro dvoustranný test. Nulovou hypotézu o nezávislosti náhodných veličin zamítneme na 5% hladině významnosti, jestliže je dosažená hladina testu nižší než hodnota 0.05.

Tab.3: K **testování nulové hypotézy o rovnosti korelačního koeficientu *r* konstantnímu číslu *c*** ($c \neq 0$)
❄ Postupujeme naprosto stejně jako v *tab.2*, pouze do prostřední kolonky zadáme konkrétní číslo *c*.

Tab.4: Testování nulové hypotézy, podle které je ve dvou populacích stejně těsná závislost mezi dvěma veličinami (korelační koeficienty se rovnají).

Jsou-li r_1 a r_2 korelační koeficienty spočítané z n_1 a n_2 dvojic pozorování získaných nezávisle na sobě ze dvou populací, pak o nulové hypotéze rovnosti těchto koeficientů budeme rozhodovat pomocí:

$$c_1 = \frac{1}{2}\log\frac{1+r_1}{1-r_1}, \quad c_2 = \frac{1}{2}\log\frac{1+r_2}{1-r_2}, \quad u = \frac{c_1 - c_2}{\sqrt{\frac{1}{n_1-3}+\frac{1}{n_2-3}}},$$

přičemž nulovou hypotézu zamítáme ve prospěch oboustranné alternativní hypotézy v případě, že $|u| \geq z(\alpha/2)$, kde z je symbol pro standardizované normální rozdělení (z=N(0,1)). V programu NCSS toto testovací kritérium není zahrnuto.

- spíše než-li jednostranné závislosti jsou lepší interpretace typu "vzájemné vztahy" mezi veličinami
- pozor při interpretaci proměnných, které jsou vyjádřeny v procentech a jejichž součet dá 100 % (jinak řečeno které jsou příčinně závislé).
- vysoké hodnoty korelačního koeficientu neznamenají nezbytně příčinnou závislost mezi proměnnými. Důvodem statistické závislosti může být nějaký třetí faktor. Uveďme si příklad: Vysocí bratři mají sklon mít vysoké sestry. Pozitivní vysoká korelace v tomto případě je způsobena třetím faktorem, který je genetického původu.
- velmi opatrně je třeba interpretovat výsledky s malým počtem pozorování
- nezapomeňme, že korelační koeficient udává sílu **lineární** závislosti
- pro výběry velkého rozsahu ($n \geq 50$) se na základě praxe přikládá koeficientu $|r|$ různá významnost. Tak např. považujeme stupeň závislosti za: nízký pro$\forall$ $|r| < 0,3$; mírný pro$\forall$ $|r| \in$ <0,3;0,5); význačný pro$\forall$ $|r| \in$ <0,5;0,7); vysoký pro$\forall$ $|r| \in$ <0,7;0,9); velmi vysoký pro$\forall$ $|r| \in$ <0,9;1>. Mnozí statistici ovšem s takovým škatulkováním korelačního koeficientu nesouhlasí.

1.5.3.1 Koeficient determinace

Koeficient determinace *(coefficient of determination* popř. *explanation)* udává míru, jaké množství variability dat jedné proměnné je vysvětlitelné druhou proměnnou. Koeficient determinace se vypočítá jako druhá mocnina korelačního koeficientu. Vyjde-li např. r^2=0,63, pak hodnoty jedné proměnné vysvětlují 63% variability dat druhé proměnné.

1.6 . Regrese

Regrese udává formu závislosti mezi dvěma nebo více proměnnými (např. lineární, polynomická, exponenciální, logaritmická, atd.). Umožňuje nám ze známých hodnot jedné či několika veličin (tzv. **nezávisle proměnné**) předpovídat průměr jiné spojité veličiny (**závisle proměnná**).

Typy regresí můžeme roztřídit např. podle těchto kritérií:

- **Předpoklady**. Rozlišujeme **klasické metody**, které požadují normální rozdělení reziduí (kapitola 1.6.1.), konstantní rozptyl a negativně je ovlivňují vybočující hodnoty a **metody robustní**, jež dané předpoklady nevyžadují. Dáváme přednost, je-li to možné, klasickým metodám. Jestliže jejich podmínky nejsou splněny (kapitola 1.6.1.2.), můžeme se pokusit o transformaci dat, ta bývá často účinná (kaitola 1.6.3.).
- **Funkční vztah mezi nezávisle a závisle proměnnými**. Pokud jsou funkce lineární z hlediska parametrů hovoříme o **lineární regresi**, pokud ne, jedná se o **nelineární regresi** (kapitola 1.6.3.).
- **Počet nezávisle proměnných**. Je-li pouze jedna, hovoříme o **jednoduché regresi** a je-li jich více jedná se **mnohonásobnou regresi** (kapitola 1.6.4.).

1.6.1 Hlavní myšlenka regresní analýzy, příklad jednoduché lineární regrese

Rozlišujeme mezi **teoretickou** (**hypotetickou, populační**) **regresní funkcí**, která je nepozorovatelná (neměřitelná) a mezi **empirickou** (**výběrovou**) **regresní funkcí**, která je vypočítaná

na základě empirických údajů. Empirickou regresní funkci můžeme považovat za odhad teoretické regresní funkce. Považujeme-li teoretickou regresní funkci za model (idealizaci) průběhu proměnné *y* při systematických změnách vysvětlující proměnné *x*, pak empirickou regresní funkci pokládáme za odhad modelu na základě získaných pozorování z výběru. Označme teoretickou regresní funkci jako *Y*, pak pro každé konkrétní pozorování bude platit rovnice

$y_i = Y_i + \varepsilon_i$, ve které y_i je hodnota závisle proměnné *y* pro *i*-té pozorování, Y_i je hodnota teoretické regresní funkce pro *i*-té pozorování a ε_i je odchylka y_i od Y_i.

Označme dále parametry regresní funkce jako α, β, takže

$Y_i = \alpha + \beta x_i$, kde α je absolutní člen (*itercept*) souřadnice průsečíku přímky s osou *y*; β směrnice přímky, sklon (*slope*) a x_i je *i*-té pozorování vysvětlující proměnné *x*. Našim úkolem je najít konkrétní formu této funkce a odhadnout její parametryα a β tak, aby přímka byla co nejtěsněji svázána s naměřenými body $[x_i, Y_i]$ (aby hodnoty závisle proměnné *Y* byly co nejlépe vysvětleny hodnotami nezávisle proměnné *x*). Označme jako *a*, *b* odhady parametrů α, β. Dále označme $\hat{y}_i$ jako odhad hodnoty y_i, takže zápis $\hat{y}_i$ vyjadřuje, že hodnota empirické regresní funkce y_i pro *i*-té pozorování je zároveň odhadem teoretické hodnoty Y_i odpovídající hodnotě nezávisle proměnné x_i.

Metod, jak odhadnout koeficienty *a*, *b*, je více. Nejznámější používanou metodou je ta, která vede k minimalizaci výrazu

$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ a označuje se jako **metoda nejmenších čtverců**. Hodnotu výrazu označujeme jako reziduální součet čtverců. **Reziduem** rozumíme hodnotu $y_i - \hat{y}_i$.

❄ Odhady parametrů α, β jsou jedním z výsledků lineární regrese v programu NCSS, ve výsledcích procedury *Multiple Regression* v sekci *Regression Equation Section*.

1.6.1.1 Jaká je interpretace parametrů α a β?

- **parametr** β. Využijme poznatku, že tento parametr je směrnicí přímky. Pak při jednotkové změně nezávisle proměnné *x* se změní průměr závisle proměnné *Y* právě o β. Je-li parametr β kladný, pak hovoříme o přímé úměře mezi proměnnými (s rostoucí hodnotou nezávisle proměnné *x* roste i průměr závisle proměnné *y*). Občas se můžeme setkat s touto špatnou úvahou: Čím vyšší je hodnota parametru β, tím silnější je vztah mezi závisle a nezávisle proměnnou. To samozřejmě není pravda. Jestliže změníme měřítko u jedné z proměnných, tak se změní i hodnota parametru β.
- **parametr** α. Málokdy se interpretuje. Udává průsečík regresní přímky s vertikální osou *y*.

Dále bychom se mohli ptát, zda je populační průměr závisle proměnné ovlivněn hodnotami nezávisle proměnné.

- Kdyby populační průměr závisle proměnné nebyl ovlivněn hodnotami nezávisle proměnné, byla by hodnota parametru β rovna nule (a regresní přímka rovnoběžná s osou *x*). Hypotéza H_o pro β=0 (resp. α=0) se testuje t-testem a je také výsledkem regresní analýzy v NCSS (sekce *Regression Equation Section*). Je-li tedy hypotéza H_o pro β=0 akceptována, je na dané hladině významnosti prokázáno, že hodnoty nezávisle proměnné neovlivňují populační průměr závisle proměnné.
- Pokud bychom chtěli otestovat to, že regresní přímka prochází počátkem (bodem [0,0]), testovali bychom hypotézu, že hodnota parametru α je rovna nule.

1.6.1.2 Ověření platnosti předpokladů při použití metody nejmenších čtverců

Předpoklady: nezávislá pozorování, konstantní rozptyl a normální rozdělení reziduí.

Předpoklad o **normálním rozdělení reziduí** lze ověřit vizuálně buď na histogramu nebo na normálním diagramu reziduí (*Normal Probability Plot of Residuals*). V něm by při normálním rozdělení reziduí měly všechny hodnoty ležet v intervalu spolehlivosti určeném dvěma křivkami. Oba grafy jsou obsaženy ve výsledné proceduře *Multiple regression*.

Předpoklad o **konstantním rozptylu** se nejčastěji ověřuje pomocí bodového diagramu *Residual versus Predicted*, v němž by při konstantním rozptylu měly být body rozmístěny nahodile, ale rovnoměrně kolem vodorovné osy na nulové úrovni (pro *y*=0). Tento graf je opět jedním z výsledných grafů procedury *Multiple regression*.

Poznámka: Jestliže po uvážení zamítneme některý z těchto předpokladů, tak se pokusme data ztransformovat. Pokud ani transformace nepomůžou, pak místo procedur založených na metodě nejmenších čtverců můžeme použít metody robustní regrese - v NCSS procedura *Robust Regression*.

1.6.2 Jak zvolit vhodný typ regresní funkce?

Ne vždy je vztah mezi nezávisle a závisle proměnnou lineární. Jak zvolíme vhodný typ regresní funkce?

- Víme-li z literatury v jakém vztahu je závisle a nezávisle proměnná, zvolíme regresní funkci podle tohoto poznatku.
- Jestliže nám není známo v jakém vztahu proměnné jsou, pak u jednoduché regrese volíme regresní funkci po prohlédnutí bodového grafu (procedura *Scatter Plots* ve výběru *Graphics*) .
- U mnohonásobné regrese je dobré mít před výběrem vhodné regresní funkce nějakou představu o ní. Jestliže tato představa chybí, můžeme jako vhodnou funkci použít součet funkcí stanovených na základě bodového grafu či jednoduchou regresí zvlášť s každou nezávisle proměnnou.

1.6.3 Jak postupovat při nelineární závislosti?

- Linearizujeme vztah transformací. Př.: $Y = \beta_0 x^{\beta_1}$ můžeme linearizovat touto transformací ln *Y* = ln β_0 + β_1 ln *x*

Transformaci nezávisle proměnné tak, aby výsledná závislost byla lineární, můžeme provádět dle libosti. Naproti tomu, transformace závisle proměnné mění jak tvar, tak typ rozdělení dat a stálost rozptylu. Logaritmická transformace závisle proměnné může zlepšit stálost rozptylu, byla-li směrodatná odchylka lineárně závislá na průměru; pokud ovšem byly předpoklady lineární regrese splněny u netransformovaných dat, nebudou splněny u transformovaných.

- Pokud jsou předpoklady pro použití metody nejmenších čtverců splněny u netransformovaných dat přičemž závislost není lineární a transformace nezávisle proměnné nepomáhá, je možné použít polynomiální nebo nelineární regresi (přehled v *tab.5*).

1.6.4 Mnohonásobná regrese

Na obdobném principu jako jednoduchá regrese pracuje i mnohonásobná regrese. Vezměme si jako příklad tuto lineární závislost: $Y = \alpha + \beta_1 x_1 + \beta_2 x_2$, kde x_1 a x_2 jsou dvě nezávisle proměnné, β_1, β_2 parciální regresní koeficienty. Koeficient β_1 lze interpretovat jako změnu populačního průměru *Y* při jednotkové změně x_1 a nezměněné hodnotě x_2. Podobně β_2 se interpretuje jako změna populačního průměru *Y* při jednotkové změně x_2 a nezměněné hodnotě x_1. Hypotéza H_o o nulovosti parciálního regresního koeficientu β_1, znamená, že proměnná x_1 nepřináší pro *Y* žádnou další informaci nad tu, která je již obsažena v informaci o x_2.

Poznámka: U složitějších lineárních modelů mnohonásobné regrese je interpretace obdobná.

1.6.4.1 Jak vybrat v množině nezávisle proměnných podmnožinu těch nejlepší?

Můžeme stát před problémem, jak vybrat z velkého množství nezávisle proměnných ty, které tvoří nejlepší regresní model. K částečnému řešení problému nám může pomoci stupňovitá regrese (v programu NCSS procedura *Stepwise Regression*). Ta mechanicky vybere skupinu nejlepších nezávisle proměnných. Ovšem interpretace této skupiny může být nesmyslná. Naproti tomu skupina s velmi zajímavou a pro praxi důležitou interpretací, může mít jen o maličko horší model a my se o něm pomocí této procedury nedovíme. Proto je dobré použít metodu stupňovité regrese pouze k hrubému výběru proměnných- odstranit ty nejméně významné) a zbylé- i ty jejichž vliv v modelu není průkazný, zkoumat „ručně“. Je možné použít i proceduru *All possible regression*, jejímž výsledkem je více nejlepších množin nezávisle proměnných.

V proceduře stupňovitá regrese si můžeme v *Template* v kolonce *Selection Method* vybrat jednu ze čtyř metod:

a) ***Forwards (step-up)****;* metoda postupných přidávání jednotlivých proměnných do modelu, pokud na zvolené hladině významnosti nová veličina ovlivňuje závisle proměnnou. Nedostatkem tohoto modelu je, že výběr je ovlivněn pořadím, jakým proměnné do regresní funkce vstoupily. Nedostatek se odstraní užitím procedury *Stepwise Selection* v NCSS.

b) ***Stepwise selection***; po každém kroku, kdy se přidá proměnná do modelu se přezkoumají všechny do modelu již zahrnuté proměnné, jestli se jejich vliv na závislou proměnnou nesnížil. Jestliže se snížil na statisticky nevýznamnou hladinu, je daná proměnná odstraněna. **Tato metoda je nejpoužívanější**.

c) ***Backward (step down)****;* do modelu jsou nejprve zahrnuty všechny proměnné, nejméně významné se odstraní.

d) ***Min MSE****;*

Podmínkou pro nalezení nejlepší podmnožiny nezávislých proměnných je, abychom měli alespoň pětkrát více měření než proměnných.

1.6.5 Jak lze hodnotit kvalitu nahrazení dat regresní funkcí?

V NCSS k tomu máme několik možností:

- **Koeficient determinace (= *R-squared*, R^2, *coefficient of determination, explanation*)**;

$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{Y})^2}$, kde $\overline{Y}$ průměrná hodnota závisle proměnné, y_i, $\hat{y}_i$ viz kapitola 1.6.1.

$R^2 \in \langle 0,1 \rangle$; čím vyšší, tím lepší model.

Koeficient determinace udává míru, jaké množství variability závisle proměnné je vysvětlitelné nezávisle proměnnou. Vyjde-li např. R^2=0,63, pak nezávisle proměnné vysvětlují 63% variability závisle proměnné.

- **F-statistika** (*F-Ratio*) testuje celkovou významnost regrese (nulovou hypotézu o rovnosti všech parciálních regresních koeficientů - připomeňme že mezi ně nepatří absolutní člen - číslu nula). Zamítneme-li nulovou hypotézu, říkáme, že regrese je průkazná. Ve výsledcích NCSS v proceduře *Analysis of Variance Section*.
- **Root MSE=** $\sqrt{MSE}$ je odhad rozptylu; čím menší, tím lepší model
- ***C*$_p$**; optimální model má hodnotu C_p blízkou číslu *p+1* (*p* je počet nezávisle proměnných); Pokud je $C_p > p+1$, pak některé proměnné v modelu jsou zbytečné- může být problém s multikolinearitou (závislost mezi nezávisle proměnnými); pokud je $C_p < p+1$, pak je stále proměnných v modelu málo.

1.6.5.1 Co lze vyčíst z koeficientu determinace a parciálních regresních koeficientů?

- Vyjde-li některý z parciálních regresních koeficientů průkazně odlišný od nuly, zatímco celkový test (F-statistika) vyjde neprůkazně, potom jsou některé proměnné v modelu zbytečné. Pokud F-statistika pro celý regresní model nevyjde průkazná, nedoporučuje se již příliš důvěřovat průkaznosti jednotlivých parciálních regresních koeficientů.
- Vyjde-li celková regrese průkazná, zatímco žádný z parciálních regresních koeficientů průkazně odlišný od nuly není, pak to většinou znamená, že nezávisle proměnné jsou korelované (jsou mezi sebou závislé), což by být neměly.

1.6.6 Regrese a program NCSS

Tab.5 znázorňuje vybrané regresní procedury statistického programu NCSS

Tab.5: Vybrané procedury programu NCSS zaměřené na regresi. Sloupeček "Hledá nejlepší model" znamená, jestli procedura z množství nezávisle proměnných vybírá nejlepší model.

Typ regrese	Procedura v NCSS (CF)...pod procedurou *Curve Fitting* (R/C)... *Regression /Correlation*	Hledá nejlepší model?	Regresní funkce	Poznámky
Grafy	*Scatter Plot Matrix* (CF)	-		Matice bodových grafů pro dvě proměnné a jejich 6 transformací (celkem 7 x 7 bodových grafů)
	Function Plots (CF)	-		Kreslí grafy speciálních funkcí, které si zadám (nepotřebuje data)
Jednoduchá nelineární regrese	*Nonlinear regression* (CF; R/C)	Ne	Modely nelineární v parametrech (př. *Y=a+b exp(cx); Y=a+b/(c+x))*	
	Growth and other models (CF)	Ne	Testuje 22 různých naprogramovaných regresních funkcí	růstové křivky
	Ratio of polynomials fits (CF)	Ne	Podíl dvou polynomů, z nichž každý je řádu menšího než pět	
	Ratio of polynomials search (CF)	Ano	Hledá nejlepší modely typu podílu dvou polynomů maximálně pátého řádu	Nezávisle proměnné v polynomu jsou tvořeny různými transformacemi jedné proměnné
	Piecewise polynomial models (CF)	Ne	Polynomický model	Vícefázové modely sestrojované po úsecích (př. pro $x<a$ polynom, pro $x>a$ lineární funkce)
	Sum of functions (CF)	Ano	Modely podílu dvou lineárních výrazů,ve kterých se vyskytuje maximálně pět funkcí	funkce jsou transformované, standardní typu: sin (x); ln(x+1); sqrt(x/2)
Mnohonásobná lineární regrese	*All possible regression* (R/C)	Ano	Studuje lineární závislosti	Vybere nejvhodnější nezávisle proměnné pro regresi (do 15 nezávisle proměnných)
	Multiple regression (R/C)	Ne	Studuje lineární závislosti	Závisle proměnná je spojitého typu
	Stepwise regression (R/C)	Ano	Studuje lineární závislosti	Vybere nejvhodnější nezávisle proměnné pro regresi (i více než-li 15 nezávisle proměnných)
	Robust regression (R/C)	Ne	Studuje lineární závislosti	Není náchylná na nekonstantní rozptyl, odlehlá pozorování a nenormalitu reziduí.
	Multivariate variable selection (R/C)	Ano	Studuje lineární závislosti	Mnohorozměrná regrese (pro více než 1 závisle proměnnou)
Mnohonásobná nelineární regrese	*Multivariate ratio search* (CF)	Ano	Hledá nejlepší polynomy či podíly polynomů	
	Multivariate ratio fit (CF)	Ne	Studuje vhodnost podílu dvou polynomických funkcí řádu < 5	

1.6.7 Porovnání dvou regresních přímek

V některých případech potřebujeme porovnat regresní přímky. Zajímá nás, jestli jsou rovnoběžné, shodné či mají odlišný průběh. Př.: Zkoumáme vztah mezi **veličinami počet ročníků jehlic** a **procento defoliace koruny stromu**. Mohlo by nás zajímat, zda je závislost shodná pro stromy s hřebenitým typem větvení koruny a pro stromy se svazčitým typem větvení koruny.

Program NCSS nám umožňuje znázornit vztahy u obou typů větvení koruny odlišnými barvami nebo odlišnými symboly **do** jednoho **grafu**. Potřebujeme k tomu sestrojit novou proměnnou nabývající hodnoty 0 pro stromy s hřebenitým a hodnoty 1 pro stromy se svazčitým typem větvení koruny. Označíme ji **typ větvení koruny**. V proceduře *Graphics* zvolíme volbu *Scatter Plot*, za *Horizontal Variable* dosadíme procento defoliace koruny stromu, za *Vertical Variable* počet ročníků jehlic a podle toho, chceme-li odlišit typy větvení koruny barvou nebo symboly, dosadíme tento sloupeček do příslušné kolonky *Grouping Variables*. Takto získáme přehled, jak dané závislosti vypadají.

Statisticky porovnáme regresní přímky takto:

- pokud se obě přímky liší ve směrnici (parametr β), pak se jedná o dvě různoběžné přímky
- pokud se obě přímky shodují ve směrnici, pak se může jednat buď o totožné přímky nebo o rovnoběžky- testujeme zda jsou přímky oproti sobě posunuty, tj. liší-li se v parametru α.

V programu NCSS ovšem nemáme možnost otestovat, kdy lze hodnoty směrnic obou přímek ještě považovat za stejné a kdy již ne. Postup ovšem můžeme obejít tímto: Sestrojíme novou proměnnou: **procento defoliace koruny stromu * typ větvení koruny**. Spustíme proceduru *Multiple regression*. Za *Dependent Variable* dosadíme **počet ročníků jehlic** a za *Independent Variables* **typ větvení koruny**, **procento defoliace koruny stromu** a **procento defoliace koruny stromu * typ větvení koruny**. Jestliže je regresní koeficient u proměnné **procento defoliace koruny stromu * typ větvení koruny** statisticky průkazně odlišný od čísla 0, pak můžeme přímky považovat za různoběžné. Pokud je nulová hypotéza o rovnosti regresního koeficientu číslu nula přijata, pak se přímky považují buď za rovnoběžné nebo shodné. To zjistíme přesně opět pomocí procedury *Multiple regression*. Za *Dependent Variable* necháme proměnnou **počet ročníků jehlic** a za *Independent Variables* dosadíme proměnné **typ větvení koruny** a **procento defoliace koruny stromu**. Jestliže je regresní koeficient u proměnné **typ větvení koruny** statisticky průkazně odlišný od čísla 0, pak lze přímky považovat za rovnoběžné, pokud není, pak za shodné.

1.7 . Diskriminační analýza

Uvažujme několik populací (nominální veličina; označme ji jako závisle proměnnou). Pro každý objekt každé populace mějme změřené údaje několika veličin (nezávisle proměnné). Každý objekt je tedy charakterizován vektorem složeným z hodnot naměřených nezávisle proměnných. Cílem diskriminační analýzy je nalézt klasifikační pravidlo, podle kterého bychom mohli určit, jestli nějaký nový objekt patří do té nebo do jiné populace. Přičemž klasifikační pravidlo je vyjádřené jako **lineární kombinace nezávisle proměnných**). K řešení tohoto problému slouží **diskriminační analýza** (*Discriminant Analysis*).

Př.: Mějme příbuzné, morfologicky velmi si podobné druhy, které se liší počtem chromozómů. Protože je obtížné počítat při každém určování chromozómy, ptáme se, zda je možné nalézt pravidlo, kde by pomocí kombinace morfometrických údajů bylo možné druh spolehlivě určit.

1.7.1 Předpoklady diskriminační analýzy:

- pro každou populaci je nutné mít **větší počet pozorování než počet nezávisle proměnných**
- **normální rozdělení** proměnných
- **homogenita kovariančních matic** - testuje se pomocí Boxova testu (*Box´s M test* v proceduře *Equality of Covariances v Multivariate Analysis*), který je velmi náchylný na nenormální rozdělení dat; hypotézu o homogenitě kovariančních matic zamítnu, je-li dosažená hladina testu v řádku *Box's M* nižší než předem zvolená hladina významnosti testu; nehomogenitu matic mohu někdy odstranit transformací proměnných
- **není přípustná multikolinearita** (= závislost mezi nezávisle proměnnými)
- **nepoužívat proměnné s častými chybějícími hodnotami** (*missing values*); pokud u daného objektu chybí hodnota třebas jen pro jednu proměnnou, pak diskriminační analýza s objektem nepracuje
- **nepřítomnost odlehlých pozorování**

1.7.2 Výběr nejlepší množiny nezávisle proměnných:

Podobně jako ve stupňovité mnohonásobné regresi i zde je možnost výběru nejvhodnějšího modelu nezávisle proměnných, který nejlépe vysvětluje závisle proměnnou. Opět i zde je nutná velká opatrnost při použití automatického výběru- platí zde to, co bylo již řečeno u stupňovité regrese. V *Template* v kolonce *Variable Selection* můžeme přepínat mezi těmito možnostmi

a) **bez výběru** (*No Selection*) - analýza se provede se všemi zadanými proměnnými.
b) **s výběrem** (*Automatic Selection*) - *Stepwise Selection*. Prinncip podobný jako u stupňovité regrese.

1.7.3 Jak přečíst výsledky diskriminační analýzy

Algoritmus diskriminační analýzy spočítá zvlášť **lineární diskriminační funkci** pro každou populaci. V programu NCSS jsou koeficienty těchto funkcí prezentovány vertikálně v sekci *Linear Discriminant Functions Report*.

Na základě lineárních diskriminačních funkcí je možné predikovat populaci, do které daný objekt patří. Popišme si, jak k tomu dochází. Hodnotu lineární diskriminační funkce pro každý objekt udává **diskriminační skór** (v *Outputu* označeno *Score*):

Diskriminační skór = konstanta + $\sum$ (hodnota proměnné * koeficient lineární diskriminační funkce)

Diskriminační skór se tedy počítá pro každý objekt a každou populaci zvlášť. Objekt je klasifikován k té populaci u které je hodnota diskriminačního skóru nejvyšší. Diskriminační skóry přepočítané na procenta udává v NCSS sekce *Percent Chance of Each Group*.

Výsledek, počet správně a chybně klasifikovaných řádků je shrnutý v **klasifikační tabulce** (*Classification Count Table*) ve formě matice. Řádky matice jsou určeny empirickými populacemi (tedy těmi, do kterých objekt skutečně patří) a sloupce populacemi predikovanými algoritmem. Diagonála matice udává počty správně klasifikovaných objektů, zatímco počty chybně klasifikovaných objektů jsou uvedeny mimo diagonálu matice. Touto tabulkou ovšem nezjistíme, jsou-li populační průměry vypočítané ze všech nezávisle proměnných pro každou populaci shodné. To testuje mnohorozměrná analýza rozptylu (MANOVA), která je také součástí statistického vybavení programu NCSS. Navíc je třeba dodat, že tato tabulka je ve většině případů optimistická, neboť zpětná predikce se provádí na těch samých datech, ze kterých byly vypočteny diskriminační skóry.

Některé z dalších důležitých výstupů diskriminační analýzy:

Wilks´ Lambda- zobecnění indexu determinace (R^2). Testuje významnost predikce populace pomocí nezávisle proměnných. (u *Automatic Selection* testuje významnost predikce na základě proměnných uvedených na tomto řádku a výše). Nabývá hodnot od nuly do jedné. Hodnoty blízké nule indukují vysokou schopnost predikce a hodnoty blízké jedné naopak nízkou (opačně než u indexu determinace).

Overall Wilks´ Lambda- je hodnota Wilksovy Lambdy pro všechny proměnné zahrnuté do modelu.

Removed Lambda- hodnota Wilksovy Lambdy udávající na vliv odstranění dané proměnné. Čím nižší je tato hodnota, tím horší bude výsledek po odstranění dané proměnné.

Alone Lambda- udává hodnotu Wilksovy Lambdy, která testuje významnost pouze té dané proměnné zahrnuté do celého modelu.

F- value- testuje významnost hodnoty Wilksovy Lambdy. Čím vyšší hodnota, tím lépe se dají danou proměnnou klasifikovat objekty do populací.

1.8 . Logistická regrese

Logistická regrese (*Logistic Regression*) vyšetřuje závislost binární (nula-jedničkové) veličiny na jiných veličinách. Nula-jedničkovou veličinou můžeme rozumět nevýskyt či výskyt sledovaného znaku. Snažíme se vyjádřit závislost pravděpodobnosti výskytu nějakého jevu na sledovaných doprovodných veličinách (nezávisle proměnné). Při tom chceme umět rozhodovat, zda tato pravděpodobnost je vůbec doprovodnou veličinou ovlivňována. K řešení obdobných problémů v minulosti sloužila i **mnohonásobná regrese** (dnes je již překonaná) a **diskriminační analýza** v případě, že je závisle proměnná binární. Logistická regrese na rozdíl od diskriminační analýzy nepožaduje homogenitu kovariančních matic ani normální rozdělení proměnných.

Lineární logistická funkce má tento tvar: $P(Y = y_1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + ... + \beta_k X_k)}}$, kde *Y* je binární závisle proměnná, která nabývá dvou hodnot y_1 a y_2, dále $X_1,...,X_k$ jsou nezávisle proměnné, *k* jejich počet a $\beta_0, \beta_1,...,\beta_k$ logistické regresní koeficienty.

Neznámé parametry $\beta_0, \beta_1...,\beta_k$ se odhadnou podobně, jako v klasické lineární regresi, byť samotný výpočet je podstatně komplikovanější. Vypočítané odhady parametrů jsou ve výsledcích logistické regrese programu NCSS zobrazeny v sekci *Parameter Estimation Section.* Statistika *Chi-Square* testuje nulovou hypotézu rovnosti logistických regresních koeficientů nule. Je-li příslušná dosažená hladina testu (*Prob Level*) nižší než hladina významnosti testu (α), pak je proměnná v modelu významná. Naopak je-li vyšší, pak proměnná nepřispívá k predikci binární veličiny o nic více než proměnné do modelu již zahrnuté. Celkový model můžeme posuzovat v sekci *Model Summary Section* (podobná interpretace jako v *Parameter Estimation Section*, ale zde se hodnotí vliv všech proměnných dohromady na nula-jedničkovou proměnnou). Podobně jako v diskriminační analýze, i zde je výsledkem procedury klasifikační tabulka (v sekci *Classification Table*); platí o ní co již bylo řečeno v diskriminační analýze.

Podobně jako v lineární regresi i zde se můžeme ptát na očekávanou změnu ve výskytu či nevýskytu znaku při změně nezávisle proměnné o jednotku. Označme poměr mezi pravděpodobností výskytu jevu a pravděpodobností jeho nevýskytu jako tzv. **šance** - *odds*). Výraz $\mathbf{e^{\beta i}}$ udává podíl šance pro hodnoty nezávislé proměnné X_i+1 a X_i; **vyjadřuje** tedy **očekávanou změnu šancí při změně nezávisle proměnné** X_i **o jednotku, jestliže se ostatní proměnné nezmění**.

Podobně jako v diskriminační analýze i zde je možné automatickým výběrem zvolit ty nezávisle proměnné, které nejlépe predikují pravděpodobnost výskytu znaku binární závisle proměnné. Připomeňme jenom, že automatický mechanický výběr může být dosti nebezpečný (viz stupňovitá regrese – kapitola 1.6.4.1.). V *Template* v kolonce *Variable Selection* k tomu máme tyto možnosti:

- *All Variables-No Selection*
- *Forward Selection*
- *Backward Selection*

(podrobnější vysvětlení ve stupňovité regresi - kapitola 1.6.4.1. a diskriminační analýze kapitola 1.7.2.)

1.9 . Kontingenční tabulky

Při vyšetřování závislosti dvou nominálních veličin se používají **kontingenční tabulky** (*Contingency Tables*). Př.: Testujeme nulovou hypotézu, jestli je pravděpodobnost výskytu lišejníků (zde binární veličina) na větvích koruny stromu závislá na druhu stromu.

V programu NCSS jsou tyto testy zahrnuty v proceduře *Descriptive Statistics*; *Cross Tabulation*. Jestliže obě z proměnných nabývají pouze dvou hodnot, používá se **Fisherův test** (kapitola 1.9.2.), pokud alespoň jedna z proměnných nabývá více hodnot, použijeme **chí-kvadrát test** (kapitola 1.9.1.); při podezření, že dvojice proměnných jsou závislé **McNemarův test** (kapitola 1.9.3.), a je-li jedna z proměnných ordinálního typu, použijeme **Armitagův test** (pouze tabulka 2 x *n*; kapitola 1.9.4.).

1.9.1 **Chí-kvadrát test *(Chi-Square Test)***

Nechť nominální veličina *X* nabývá hodnot *1,2,...,r* a veličina *Y 1,2,...,c*. Zabývejme se jejich možnou závislostí. Pokud jsou obě náhodné veličiny nezávislé, pak by relativní četnosti (procentní podíly) možných hodnot jedné veličiny byly podobné při všech hodnotách druhé veličiny. V našem případě by pravděpodobnost výskytu lišejníků byla stejná pro všechny druhy stromů. Na tomto základě se počítá statistika χ^2 používaná k rozhodování o hypotéze nezávislosti:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(N_{ij} - \frac{N_{i+}N_{+j}}{n}\right)^2}{\frac{N_{i+}N_{+j}}{n}}$$, kde N_{ij} je počet objektů v řádku *i* a sloupci *j*, N_{i+} resp. N_{+j} řádkové resp. sloupcové marginální četnosti ($N_{i+} = \sum_{j=1}^{c} N_{ij}$; $N_{+j} = \sum_{i=1}^{c} N_{ij}$) a *n* počet všech objektů. Podíl $\frac{N_{i+}N_{+j}}{n}$ znamená tedy teoretickou, očekávanou četnost v řádku *i* a sloupci *j* za platnosti nulové hypotézy.

Nulovou hypotézu (tvrzení, že *X* a *Y* jsou nezávislé) zamítáme v případě, je-li $\chi^2 > \chi^2_{(r-1)(c-1)}(\alpha)$ (kde χ^2 symbol pro dané zozdělení, (r-*1)(c-1)* počet stupňů volnosti a α hladina významnosti testu), neboli když je dosažená hladina testu-*Prob Level* nižší než předem zvolená hladina významnosti. Podmínkou pro zdárný průběh testu je, aby všechny teoretické četnosti dosahovaly hodnoty alespoň 5.

Tímto postupem zjistíme pouze to, jsou-li veličiny závislé nebo nezávislé. Mohlo by nás ovšem zajímat jak silná je tato závislost. K tomu nám slouží statistika **Cramerovo V** (*Cramer's V*), která je obdobou (i co se týká interpretace) korelačního koeficientu.

❄ Pro výpočet chí-kvadrát testu a Cramerova V je nutné v Template aktivovat kolonku *Chi-Square Stats*. na *All Statistics*. Pokud chceme zobrazit teoretické četnosti, musíme v *Template* aktivovat kolonku *Show Expected Values*. Nominální i ordinální proměnné zadáváme do kolonek *Discrete Variable*.

1.9.2 Fisherův test *(Fisher´s Test)*

Pro *r=2* a *c=2* můžeme použít tzv. **čtyřpolní tabulky**. Patří k nim Fisherův test (*Fisher's Exact Test*), který testuje hypotézu o nezávislosti nominálních proměnných. K určení síly závislosti použijeme korelační koeficient *Correlation Coefficient*. Interpretace je shodná jako u Cramerova V.

1.9.3 McNemarův test *(McNemar Test)*

McNemarův test se užívá v případě veličin, u kterých máme podezření, že dvojice proměnných jsou závislé. **Př.:** Ve dvou sezónách posuzujeme zdravotní stav stejných náhodně vybraných stromů na základě výskytu lišejníků. Zajímá nás, zda došlo mezi sezónami ke změně.
Pro objektivní průběh testu by mělo platit, že součet prvků mimo diagonálu je alespoň 8.

❄ Ke spuštění této (i jiných) statistik je nutné v *Template* aktivovat kolonku *Chi-Square Stats* na *All Statistics*.

1.9.4 Armitagův Test *(Armitage test)*

Používá se pro jednu ordinální a jednu nominální veličinu. Testuje lineární trend mezi veličinami.

❄ Ke spuštění této statistiky je nutné v *Template* aktivovat kolonku *Armitage Proportion Trend Test*.

1.10 Shluková analýza

Cílem shlukové analýzy (*cluster analysis*) je nalézt v celém souboru dat takové skupiny objektů, které jsou si navzájem blízké či podobné, ale které se liší od objektů ostatních skupin. Jde v ní tedy především o sloučení objektů (individuí) do skupin (do shluků) na základě jejich vlastností. Každá skupina pak obsahuje objekty s velmi podobnými vlastnostmi**. Shluková analýza je především metodou prvního stupně analýzy dat, která má navrhnout určité hypotézy**. Neměla by být konečným cílem žádné práce, ale spíše prvním vodítkem k použití dalších statistických metod (např. diskriminační analýzy). Jelikož ve shlukové analýze nedochází k testování hypotéz, tak ji někteří autoři nepovažují za statistickou metodu. V každém případě je shluková analýza vhodná pro exploratorní práci, kdy se snažíme v datech nějak orientovat.

Příklad použití shlukové analýzy: Mějme soubor stromů a pro každý z nich řadu naměřených parametrů. Shluková analýza nám vytvoří takové shluky (*clusters*) stromů, uvnitř kterých jsou stromy s podobnými parametry. A také obráceně: stromy zahrnuté do různých shluků se v daných parametrech liší více, než stromy obsažené v jednom shluku.

1.10.1 Princip shlukové analýzy:

- Seřazení dat do tabulky; sloupce jsou tvořeny jednotlivými proměnnými a řádky objekty (v našem případě stromy). **Pozn.:** Proměnné do *Template* se zadávají do příslušných kolonek podle typu proměnné (kapitola 1.2.).
- Transformace dat. V souboru mohu mít proměnné s různými stupnicemi (cm, %, poměry-bezjednotková proměnná, atd.). Proto se data transformují na standartizovanou stupnici (kapitola 1.10.1.1.).
- Výpočet matice podobnosti či nepodobnosti mezi objekty (pomocí vzdálenosti mezi objekty) (*tab.6*)
- Aplikace třídící strategie: vezmou se objekty, které mají v matici nepodobnosti nejnižší koeficient (tudíž jsou si nejbližší), sloučí se do stejné skupiny (do stejného shluku), pak se spočítá opět matice nepodobnosti mezi skupinami a opět se spojí nejbližší skupiny, atd. Byla vyvinuta celá řada třídících strategií (kapitola 1.10.1.2.).
- Výsledkem shlukové analýzy může být např. dendrogram (*tab.7*)

Počet shluků může být předem zadán (v *Template* kolonka *Number of Clusters*), nebo je součástí procedury podle nějakého kritéria určit optimální počet shluků (kapitola 1.10.1.5.).

1.10.1.1 Typy transformací

V *Template* ve výběru *Scaling Types* jsou 4 typy transformací, přičemž pro všechny platí: $Z_{ij} = \frac{X_{ij} - A_i}{B_i}$, kde X_{ij} je původní hodnota proměnné i a řádku j.

Jednotlivé transformace se liší ve členech A_i a B_i. (*N*...počet objektů)

Typy transformací;	**Hodnota A_i**	**;**	**Hodnota B_i**
Average Absolute Deviation;	$\frac{\sum_{j=1}^{N} X_{ij}}{N}$	;	$\frac{\sum_{j=1}^{N} \lvert X_{ij} - A_i \rvert}{N}$
Standard deviation;	$\frac{\sum_{j=1}^{N} X_{ij}}{N}$	;	$\sqrt{\frac{\sum_{j=1}^{N} (X_{ij} - A_i)^2}{N-1}}$
Range	$\underset{overj}{Min}(X_{ij})$	;	$\underset{overj}{Max}(X_{ij}) - \underset{overj}{Min}(X_{ij})$
None	0	;	1

1.10.1.2 Třídící strategie v NCSS

(v *Template* ve výběru *Linkage Type*)

1) Single linkage (*nearest neighbour clustering*, metoda nejbližšího souseda). Vzdálenost mezi dvěma shluky se definuje jako vzdálenost mezi nejbližšími členy shluků.
2) Complete linkage (*furthest neighbour*, metoda nejvzdálenějšího souseda). Vzdálenost mezi dvěma shluky se definuje jako vzdálenost nejodlehlejších objektů.
3) Simple average (*weighted pair-group method*, metoda průměrné vzdálenosti). Vzdálenost mezi shluky se definuje jako průměrná vzdálenost mezi všemi dvojicemi obou shluků, přičemž obě skupiny mají stejný vliv na výsledek (i když v každé je jiný počet členů).
4) Centroid (*unweighted pair-group centroid method*, centroidní metoda). Vzdálenost mezi dvěma shluky se definuje jako vzdálenost mezi těžišti.
5) Median (*weighted pair-group centroid method*, mediánová metoda). Vzdálenost mezi dvěma shluky se definuje jako medián vzdáleností.
6) Group average (*unweighted pair-group method*). Vzdálenost mezi shluky se definuje jako průměrná vzdálenost mezi všemi členy obou shluků. **Nejpoužívanější metoda**.
7) Wards minimum variance. Minimalizace součtu čtverců objektů uvnitř shluku.

8) Flexible strategy. Umožňuje nastavit libovolnou třídící strategii.
Poznámka: Strategie 4, 5, 7 a 8 se používají pouze ve spojení s Euklidovskou vzdáleností (*tab.6*).

Tab.6: Počítání matice podobnosti či nepodobnosti mezi objekty

Matici nepodobnosti můžeme získat jednak z matice vzdáleností a jednak z matice korelací.

V NCSS si v *Template* ve výběru *Distance Type* můžeme zvolit jednu ze dvou **vzdáleností**

*a)***Euklidovská (*Euclidean distance*)**

$d_{jk} = \sqrt{\frac{\sum_{i=1}^{P} \delta_{ijk}^2}{P}}$ ...nejkratší vzdálenost mezi objekty (délka spojnice). **Používá se častěji**.

P... počet proměnných

$\delta_{ijk}=Z_{ij}-Z_{ik}$ pro intervalové, ordinální a poměrové proměnné.

$\delta_{ijk}= 1$ pro $X_{ij}\neq X_{ik} \wedge \delta_{ijk}= 0$ pro $X_{ij}=X_{ik}$ pro binární a nominální proměnné.

b) **Manhatanovská (*Manhattan distance*)**

$d_{jk} = \frac{\sum_{i=1}^{P} |\delta_{ijk}|}{P}$ ...součet délek na sebe kolmých úseček mezi objekty

Korelace; v programu NCSS jsou tři způsoby, jak převést korelační koeficienty na vzdálenosti.

a) ***Correlations 1***

$d_{ij} = \frac{1 - r_{ij}}{2}$, kde r_{ij} je korelační koeficient mezi i-tou a j-tou proměnnou

b) *Correlations 2* $d_{ij} = 1 - |r_{ij}|$

c) *Correlations 3* $d_{ij} = 1 - r_{ij}^2$

❄ **Poznámka**: Chceme-li spustit shlukovou analýzu na základě korelačního koeficientu, musí být data zadána maticí korelačních koeficientů mezi proměnnými a v *Template* musíme nastavit v *Input Format* podle výběru *Correlations 1-3*. Proměnné, které obsahují korelační matici dosadíme do výběru *Interval Variables.*

a použije se ta, která má nejvyšší hodnotu tzv. "C*ophenetic Correlation Coefficient*". Je to korelace mezi původními vzdálenostmi objektů a těmi, které způsobí zařazení do shluku. Hodnoty nad 0,75 značí, že byla nalezena dobrá struktura. Druhou možností je použít hodnotu *Delta*. Zde jsou zase žádoucí hodnoty blízké 0.Volba vhodného počtu shluků podle dendrogramu viz tab.5.

2) ***Fuzzy clustering***. Je to metoda shlukové analýzy, která dovolí, aby proměnná byla zařazena do více jak jednoho shluku. Zatímco u jiných shlukových analýz je hodnota odpovídající zařazení objektu do shluku rovna buď 1 (objekt do shluku patří) nebo 0 (objekt do shluku nepatří), pak zde tato hodnota reprezentuje pravděpodobnost, že objekt patří do daného shluku. Může tedy nabývat hodnot od 0 do 1. Jsou-li u objektu všechny hodnoty příslušející všem shlukům stejné, pak objekt leží mezi shluky. Nejvhodnější počet shluků zjistíme pomocí hodnot $F_c(U)$, $D_c(U)$ a $S_c(U)$ (kapitola 1.10.1.5.).

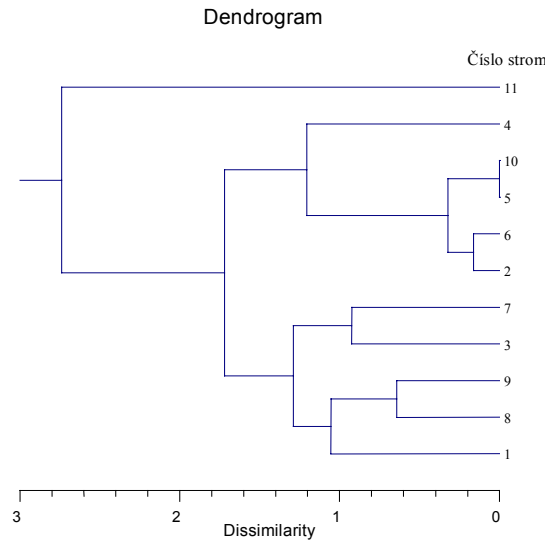
3) ***K-means cluster analysis***. Vhodné pro tvorbu malého počtu shluků z velkého množství dat. Požadavky: spojité proměnné bez odlehlých hodnot; je možné zahrnout i diskrétní proměnné, ty ovšem mohou působit potíže. Algoritmus je založen na minimalizaci sumy čtverců vzdáleností uvnitř shluků (*within cluster sum of squares*) = WSS_f. *Percent of variation* je vnitřní suma čtverců převedená na % sumy čtverců objektů, které nejsou shlukovány. Nejvhodnější počet shluků zvolíme tak, kde se toto číslo výrazně snižuje.

4) ***Medoid clustering***. Medoid je objekt uvnitř shluku, pro který je průměrná hodnota koeficientu nepodobnosti ke všem jiným objektům v tom samém shluku minimální. Algoritmy minimalizují tuto hodnotu.

5) ***Regression clustering***. Jde o shlukování v mnohonásobné regresi, ve které máme závisle proměnnou a jednu či více nezávisle proměnných. Algoritmus rozdělí data do dvou a více shluků a v každém z nich provede mnohonásobnou regresi. Algoritmus je založen na tom (shluky jsou zvoleny tak), aby regrese uvnitř shluků byla nejlepší (maximalizuje se index determinace = *R-squared*). Počet shluků určíme tak, aby byl maximalizován index determinace a přitom minimalizován počet shluků.

Tab.7: Dendrogram

Jak číst z dendrogramu? Dendrogram má na horizontální ose vynesený koeficient nepodobnosti (*Dissimilarity*) a na ose vertikální jsou vyneseny objekty (v našem případě čísla stromů). Čím delší jsou ve stromovém diagramu horizontální úsečky, tím větší je rozdíl mezi objekty. Nejvíce podobné byly stromy č. 5 a 10, které mají nulovou délku horizontální úsečky (koeficient nepodobnosti je roven č. 0. To znamená, že se ve všech proměnných, vzhledem ke kterým proběhl shlukový algoritmus, tyto dva stromy shodují. Naopak nejvíce odlišným stromem od všech dalších byl strom č.11, jehož koeficient nepodobnosti je téměř 3.



Jaký počet shluků zvolit? To je samozřejmě subjektivní otázka, obecně se postupuje takto: počet shluků zjistíme tak, že si pevně zvolíme vhodný koeficient nepodobnosti a v něm kolmo k horizontální ose vedeme přímku přes celý stromový diagram. Kolikrát nám přímka protne horizontální úsečky dendrogramu, takový počet shluků zvolíme. V našem případě bychom si mohli zvolit koeficient nepodobnosti 1,5 a dostaneme tak tři shluky [(1,8,9,3,7), (2,6,5,10,4), (11)], nebo 2 a dostaneme 2 shluky [(11), (zbylé stromy)].

1.10.1.4 Zvolení vhodného počtu shluků a určení úspěšnosti shlukového algoritmu

V programu NCSS k řešení tohoto problému slouží tři možnosti.

1) *Dun´s partition coefficient.* Označení *F(U);* koeficienty nabývají hodnot z intervalu $\langle 1/K ; 1 \rangle$, kde K je počet shluků. Užívá se i normovaná hodnota koeficientu *F(U)* a značí se $F_c(U)$; $F_c(U)$ tedy nabývá hodnot z intervalu $\langle 0;1 \rangle$. Čím vyšší jsou tyto koeficienty, tím lépe je objekt klasifikován. Je-li *F(U)=1/K* resp. $F_c(U)= 0$, pak objekt leží mezi všemi shluky a je tudíž špatně klasifikován.

2) *Kaufman partition coefficient D(U)*; $D(U) \in \langle 0;1-1/K \rangle$. Užívá se i normovaná hodnota *D(U)* a značí se $D_c(U)$; $D_c(U) \in \langle 0;1 \rangle$ Čím nižší jsou tyto koeficienty, tím lépe je objekt klasifikován. Je-li *D(U)=1-1/K* resp. $D_c(U)= 1$, pak objekt leží mezi všemi shluky.

3) *Silhouettes*- označení *s*; $s \in \langle -1;1 \rangle$; *s* blízké č.1...objekt je dobře klasifikovaný; *s* blízké č.0...objekt leží mezi shluky; *s* blízké č.-1..špatné, objekt by měl patřit do jiného shluku. Opět i zde se používá hodnota $S_c(U)$, což je maximální průměrná hodnota s přes všechny shluky. Interpretace S_c;
$S_c \in \langle 0{,}71; 1 \rangle$...nalezena silná struktura (shlukování je úspěšné);
$S_c \in \langle 0{,}51; 0{,}71 \rangle$...nalezena měřitelná struktura; $S_c \in \langle 0{,}26; 0{,}51 \rangle$...struktura je slabá, zkus jinou metodu v databázi; $S_c \in \langle -1; 0{,}26 \rangle$.....nebyla nalezena struktura

Počet shluků se zvolí tak, aby bylo $S_c(U)$ a $D_c(U)$ maximální a naopak $F_c(U)$ minimální.

1.11 Faktorová analýza a jiné příbuzné metody (tzv. ordinační metody)

Podobně jako shluková analýza i faktorová analýza (*factor analysis*) patří k technikám explorační analýzy dat, umožňuje především orientaci v datech a měla by dát spíše podnět pro další statistické zpracování dat. Má zpravidla následující cíl: nahradit základní proměnné, které jsou mezi

sebou často korelované, menším počtem hypotetických proměnných (tzv. **faktorů**) a tím původní proměnné roztřídit do skupin. Uvnitř skupin by měly být proměnné pokud možno co nejvíce korelované a proměnné patřící různým faktorům by měly být pokud možno co nejméně závislé. Přitom faktorů má být co nejméně a mají být navzájem nekorelované. Takto vzniklé faktory potom usnadní interpretaci dat.

Příklad: Mějme několik lesních stanovišť a na každém z nich údaje pro tyto proměnné: průměrná defoliace koruny stromů, výskyt chloróz, typ větvení koruny, nadmořská výška a vzdálenost stanoviště od silnice. Chceme zjistit, která stanoviště jsou si podobná a která se od sebe liší. Ordinační metody nám pomůžou zjistit, že tyto veličiny lze nahradit třemi hypotetickými faktory: první lze interpretovat jako zdravotní stav stromů na stanovišti, druhý jako genetická složka a třetí jako vnější faktory.

Podobně jako například v diskriminační analýze, i zde vznikají nové charakteristiky (faktory) jako lineární kombinace původních veličin. Při interpretaci koeficientů těchto lineárních kombinací je třeba vzít v úvahu, že metoda je určuje až na násobek nenulovým číslem. To speciálně znamená, že u jednoho faktoru můžeme změnit znaménko všech koeficientů. Koeficienty se dají interpretovat jako korelační koeficienty.

Metod, které se užívají, je celá řada. V programu NCSS jsou to tyto: **analýza hlavních komponent** (kapitola 1.11.1.), **faktorová analýza** (kapitola 1.11.2.), **mnohorozměrné škálování** (kapitola 1.11.3.), **korespondenční analýza** (kapitola 1.11.4.). Všechny jsou v programu NCSS součástí *Multivariate Analysis*.

1.11.1 Analýza hlavních komponent (*PCA* - Principal Component Analysis *)*

Zadání dat v programu NCSS: Obvyklé, řádky jsou tvořeny v našem případě konkrétními stanovišti a sloupce proměnnými (v *Template* ve výběru *Data Input Format* ponecháme výběr *Regular Data*). Je možné zadání i přes matici korelací.

Jak zvolit vhodný počet faktorů (*factors*)? Necháme proběhnout proceduru a ve výsledcích v sekci *Eigenvalues* vycházíme ze sloupečku *Cumulative Percent* udávajícího jaké množství z celkové variability proměnných je vysvětleno daným faktorem a faktory nad ním. Jestliže např. první dva faktory vysvětlují 85% celkové variability proměnných a první tři faktory 90% variability proměnných, tak zvolíme počet faktorů 2. (Třetí faktor vysvětluje již jen malé procento variability proměnných oproti prvním dvěma). Graficky je tato hodnota zanesena ve sloupečku *Scree Plot*. Sloupeček *Individual Percent* říká, jaké množství variability proměnných vysvětluje daný faktor.

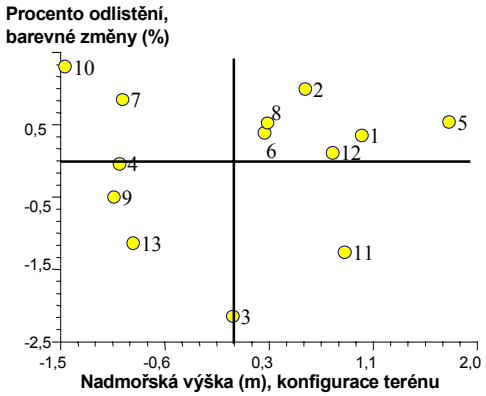
Z jakých proměnných je složen daný faktor? Korelaci mezi faktory a původními proměnnými nám udává ve výsledcích sekce *Factor Loadings*.

Sekce *Communality* udává, jaké faktory vysvětlují variabilitu dané proměnné nejvíce (nejvyšší hodnota), jaké méně a jaké variabilitu proměnné nevysvětlují téměř vůbec (téměř nulová hodnota).

To, které proměnné nejvíce přísluší k daným faktorům, nám udává slovně sekce *Factor Structure Summary Section*.

Jestliže chceme z našeho příkladu zjistit, která stanoviště jsou si blízká a která ne, tak v *Template* změníme kolonku *Show Factor Scores* na *Yes*. Abychom tento výsledek mohli graficky znázornit, změníme také výběr *Show Factor Score Plots* na *Yes- Points and Rows*. Ve výsledcích procedury se nám objeví jednak sekce *Factor Scores Section*, která v sobě obsahuje hodnoty faktorových skórů pro objekty a příslušné faktory. Hodnoty faktorových skórů jsou vyneseny do grafu *Factor Score Plots*.

Co můžeme z grafu vyčíst? Každý bod v grafu znamená konkrétní stanoviště, konkrétní pixel. Čím jsou si stanoviště v grafu blíž, tím jsou si podobnější. Na vertikální ose jsou naneseny skóry prvního faktoru, který vysvětluje zejména průměrné procento defoliace koruny a barevné změny, tedy proměnné týkající se zdravotního stavu stromů na stanovišti. Na horizontální ose je nanesena skupina proměnných, na něž mají největší vliv nadmořská výška a konfigurace terénu. Platí zde závislost: čím výše je stanoviště umístěné, tím je zdravější a čím více je vlevo, tím vyšší je jeho nadmořská výška a tím větší je jeho sklon. Do grafu jsou pro názornost naneseny osy, ve kterých jsou skóry obou faktorů nulové (tzn. že na stanoviště č. 4, 12, 1, 6, 9 nemá faktor týkající se zdravotního stavu stromů téměř žádný vliv; neovlivňuje



Graf 1: **Výsledný graf procedury analýzy hlavních komponent.** Čísla u symbolů pro jednotlivá stanoviště jsou převzatá z *tab.1* závěrečné zprávy. Podrobněji v textu.

je ani kladně, ani záporně- nejsou ani zdravá ani poškozená s porovnáním s ostatními stanovišti). Z grafu je navíc patrné, že většina nížinných stanovišť je zdravá a většina horských naopak poškozená. To nás vedlo k myšlence testovat nulovou hypotézu, že v nižších a vyšších polohách jsou stanoviště stejně poškozená proti alternativní hypotéze, že v nižších nadmořských výškách a na rovinatých plochách jsou stanoviště méně poškozená než-li v horských strmějších stráních. Nulovou hypotézu můžeme testovat procedurou mnohonásobné regrese.

1.11.2 Faktorová analýza (FA- Factor Analysis)

Podobná PCA, pracuje na jiném algoritmu.

1.11.3 Mnohorozměrné škálování (MDS- Multidimensional Scaling)

Příklad užití: Mějme několik lesních stanovišť a zabývejme se jejich podobností (např. co se týče druhového složení). Každé dvojici stanovišť přiřadíme hodnotu od 0 (stanoviště mají naprosto stejné druhové složení) do 10 (druhové složení je naprosto odlišné). Metoda MDS znázorní do dvourozměrného grafu všechna stanoviště, přičemž čím jsou si stanoviště blíž ší, tím jsou si podobnější.

1.11.4 Korespondenční analýza (CA- Correspondence Analysis = RA- Reciprocal Averaging)

Dnes hojně používaná metoda zejména pro floristická data. Je podobná metodě PCA, ovšem na rozdíl od ní nepoužívá spojitá data, nýbrž tabulky o *r* sloupcích a *k* řádcích. Příklad použití: Mějme opět několik stanovišť a na každém z nich zaznamenaný počet výskytu několika rostlinných druhů. Metoda nám umožní graficky porovnat stanoviště na základě druhového zastoupení rostlin a také naopak porovnat rostlinné druhy podle výskytu na stanovištích.

1.12 Seznam použité literatury:

- Havránek T. Statistika pro biologické a lékařské vědy. Academia, Praha 1993.
- Kassab J.Y. Applied Statistics, University College of North Wales - Centre for Applied Statistics. September 1989.
- Kassab J.Y. Experimental Design and Statistical Analysis, University College of North Wales - Centre for Applied Statistics September 1989.
- Kent M., Coker P. Vegetation Description and Analysis - A Practical Approach. John Wiley & Sons, Chichester 1997.
- Lepš J. Biostatistika. Jihočeská univerzita České Budějovice 1996.
- Seger J., Hindls R., Hronová S. Statistika v hospodářství. ETC Publishing, Praha 1998. Škrášek J., Tichý Z. Základy aplikované matematiky III. SNTL, Praha 1990.
- Zvára K., Biostatistika; Karolinum, Praha 1998.
- Zvára, K. Statistika v antropologii. Ve: Stloukal, M. a kol.:Antropologie. Příručka pro studium kostry. Národní muzeum Praha 1999.
- Nápověda statistického programu NCSS60.