# Statistická analýza dat

Alena Černíková

alena.cernikova@ujep.cz

14. října 2026

# Náplň předmětu

**Cílem tohoto předmětu je připravit Vás na statistické zpracování dat do Vašeho Bakalářského projektu.**

Probereme jednotlivé body správného postupu při statistickém výzkumu.
Zopakujeme základní analýzy a přidáme k nim pokročilejší metody lépe pasující na Vaše data.
Cvičení budou probíhat v SPSS.

# Podmínky zápočtu

Průběžné odevzdávání domácích úkolů (budou celkem 4)

1. data týkající se **Vaší bakalářské práce**
2. popisné statistiky včetně intervalu spolehlivosti
3. vztah dvou veličin
4. vztah více veličin – zobecněné lineární modely
5. kapitola do bakalářské práce

# Obsah kurzu

- Cíle statistického zpracování
- Konstrukce výběru a dotazníku
- Popisné statistiky
- Bodový vs intervalový odhad
- Testy pro vztah dvou náhodných veličin
- Modely pro vztah dvou náhodných veličin
- Modely s více nezávisle proměnnými

# Cíle statistického zpracování

Cílem statistického výzkumu není jen popsat naměřená data, ale především zobecnění výsledků na populaci.

- zajímá nás chování náhodné veličiny na vybrané populaci
- celou populaci změřit neumíme, uděláme z ní náhodný výběr
- pravidla pro tvorbu náhodného výběru jsou přesně daná – **výběr musí být reprezentativní**
- po provedení výběru je třeba realisticky ohodnotit, na jakou populaci se budou naměřené výsledky zobecňovat

**Příklad.** *Zajímá nás názor studentů na možnosti kulturního vyžití v Ústí nad Labem. Bude položena otázka: "Je pro Vás nabídka kulturních akcí v UL vyhovující?" Možné odpovědi:*

- *Rozhodně ano*
- *Spíše ano*
- *Spíše ne*
- *Rozhodně ne*

# Sběr dat

Jen při správně nastaveném výběru a správně sestaveném dotazníku dostanu výsledky, které budou mít výpovědní hodnotu.
Příprava sběru dat je mnohdy důležitější než statistická analýza.

**Na špatných datech ani sebelepší statistika nic nezmůže.**

# Konstrukce výběru

**Před provedením výzkumu** je třeba si rozmyslet

- Kdo je cílová populace a jak tedy realizovat náhodný výběr
- Jakým způsobem se budou data sbírat (pozorováním/ osobním dotazníkem/ on-line/ experimentem/ nalezením dat na internetu)
- Co chci dotazníkem zjistit, tedy jak navrhnout dotazník a jednotlivé otázky, abychom došli cíle
- Kolik dat chci získat (bude výstup kvalitativní nebo kvantitativní)

# Konstrukce výběru

Jak zajistit, aby byl výběr **reprezentativní**?

- **náhodný výběr** ze seznamu celé populace – dělá se někde?
- **systematický výběr** ze seznamu (beru každého desátého)
- **kvótní výběr** – vybrat několik základních znaků a u nich určit **kvóty**, jak často mají být zahrnuty ve výběru, aby výběr odpovídal populaci
  - Kvótní výběr pro dospělé obyvatele ČR – dodržují se kvóty pro pohlaví, věk (5 věkových skupin), vzdělání, kraj a příjem (5 příjmových skupin)
- **stratifikovaný výběr** – výběr i vyhodnocení realizuji ve stratech
- **vícestupňový výběr** – vybírám z několika velkých oblastí, v rámci každé vybrané pak z několika podoblastí, atd.

A co anketa?

# Konstrukce výběru

U pravděpodobnostních výběrů (náhodný, systematický) je možné dopředu určit minimální rozsah výběru, aby výsledné zjištění mělo požadovanou přesnost.

- Chci aby **interval spolehlivosti** pro populační průměr měl délku maximálně $2\Delta$. Délka poloviny intervalu spolehlivosti je $z(1-\alpha/2)\frac{\sigma}{\sqrt{n}}$, kde $\alpha$ je hladina významnosti, $n$ je počet pozorování a $\sigma$ je směrodatná odchylka proměnné. Pak

$$n \geq \left( z(1-\alpha/2)\frac{\sigma}{\Delta} \right)$$

- Chci aby **dvouvýběrový t-test** měl spolehlivost $\alpha$ a sílu testu $1-\beta$. Při sdružené směrodatné odchylce obou výběrů $s$ a rozdílu mězi průměry $\Delta$ (při tomto rozdílu má být požadovaná síla testu) by rozsah výběru měl být

$$n \geq ((z(1-\alpha)+z(1-\beta))/\Delta)^2 s^2$$

# Konstrukce dotazníku

Nejprve je třeba si stanovit výzkumné cíle

- Chci zjistit nějaká východiska – **explorační studie**
  – např. co vede lidi k návštěvě vybrané cukrárny
- Chci popsat nějakou situaci – **popisná studie**
  – např. jak se změnila spokojenost studentů FSE se studiem za posledních 10 let,
- Chci rozhodnout o pravdivosti nějakého tvrzení – explanační studie
  – např. Je návštěvnost cukrárny závislá na vybavenosti blízkého dětského hřiště?

# Konstrukce dotazníku

Dotazník by měl zjistit vše potřebné a přitom nebýt příliš dlouhý

- příliš **málo otázek** – na některé výzkumné cíle nenajdu odpověď
- příliš **mnoho otázek** – respondent bude unavený a už nebude odpovídat podle pravdy

**Otázky by měly být**

- jednoznačné,
- složitostí přizpůsobené respondentům,
- zodpověditelné, tj. měly by dotazovat to, co respondent ví,
- přehledné (pokud respondent vyplňuje dotazník sám)

Vlastnímu šetření by měla předcházet pilotáž.

# Konstrukce dotazníku

**Typy otázek**

- **Uzavřená otázka** – klasická otázka v dotazníku s několika položkami
- **Otevřená otázka** – těžké na zpracování, ale mnohdy přinese zajímavé podněty, dávají se jen výpisy
- **Polootevřená otázka** – kompromis
- **Podstatné (meritorní) otázky** – občas se doplňují kontrolními otázkami, abychom měli jistotu "správné" odpovědi
- **Filtrační otázky** – podle odpovědi pak dotazník pokračuje jednou z několika větví (Př. máte auto? Jaké značky?)
- **Identifikační** – měly by být v každém dotazníku, umožňují identifikovat skupiny (např. demografie)

Pozor na citlivé otázky.

# Konstrukce dotazníku

Každá otázka v dotazníku reprezentuje náhodnou veličinu/proměnnou. Při zpracování dat je důležité rozlišovat následující typy proměnných.

- **Číselné proměnné**, kardinální/intervalová škála – př. výška, váha, věk, atd.
- **Kategorické proměnné** – př. barva, kraj, povolání, nebo taky známka ve škole, číslo, které padne na kostce, atd.
- Kategorické proměnné se dále dělí na
  - **Nominální**, nominální škála – neuspořádané, př. barva, kraj
  - **Ordinální**, ordinální škála – uspořádané, př. známka, míra spokojenosti na stupnici

Otázky vedoucí na ordinální veličiny (**škálové otázky**) bývají v dotaznících nejoblíbenější.

# Konstrukce dotazníku

Příklady vybraných škál

- **Nominální škála** – seznam položek, př. barva
- **Ordinální škála** – seznam položek, př. míra souhlasu (rozhodně souhlasím, spíše souhlasím, spíše nesouhlasím, nesouhlasím) – u ordinální škály nelze zaručit, že všichni respondenti vnímají stejně intervaly mezi jednotlivými body škály
- **Kardinální škála** – snaha o zajištění stejné vzdálenosti mezi body, př.
  - vyberte hodnotu na stupnici od 1 do 10, kde 1 je velmi nespokojen a 10 velmi spokojen
  - zakreslete svou míru spokojenosti do úsečky, kde vlevo je velmi nespokojen a vpravo velmi spokojen
  - podle obrázků ohodnoťte svou míru spokojenosti (5 smajlíků)

# Konstrukce dotazníku

**Likertova metoda**. Jednu věc můžeme hodnotit i vícerozměrně. Několik charakteristik sledované jednotky hodnotíme na (stejné) škále a následně použijeme jedno souhrnné číslo, např. průměr, medián, součet, ...
Výsledná veličina je již spojitá (kromě mediánu).

**Příklad.** *Hodnotíme kvalitu výuky z několika pohledů: srozumitelnost vyjadřování, ochota odpovídat na dotazy, zajímavost vybraných příkladů, náročnost látky, náročnost u zkoušky, atd.*

Likertův koeficient diferenciace - málo diferencující otázky se vynechají (buďto přímo z šestření po pilotáži, nebo z vyhodnocení)

$$L_D = \frac{\text{součet čtvrtiny nejvyšších skóre} - \text{součet čtvrtiny nejnižších skóre}}{n/2}$$

kde $n$ je počet pozorování. Pro pětibodovou škálu nabývá hodnot od 0 do 2.

# Konstrukce dotazníku

Výběr ordinální škály

- neměla by být příliš **detailní** – př. 1) nikdy 2) vzdáleně 3) příležitostně 4) poměrně často 5) často 6) velmi často 7) téměř vždy 8) vždy
- zamyslet se nad možností/vhodností **sloučit položky** – př. 1) zřídka či nikdy 2) občas 3) často
- umožnit **střed** a v jaké podobě – př. 1) souhlasím 2) spíše souhlasím 3) nejsem si jistý 4) spíše nesouhlasím 5) nesouhlasím
- co s variantou **nevím** – často je možné zcela vynechat (ti, co neví, prostě neodpoví)
- škála by měla zahrnovat **"celý prostor"** –
  1) souhlasím 2)spíše souhlasím 3) spíše nesouhlasím 4) nesouhlasím
  versus
  1) silně souhlasím 2) souhlasím 3) nesouhlasím 4) silně nesouhlasím

Pozor na **objektivitu** odpovědí, odpověď respondenta vždy popisuje subjektivní postoj, ne nutně objektivní závěr.

# Čištění dat

Po sběru dat následuje jejich **čištění**

- vymazání nesmyslných odpovědí
- sloučení kategorií u otevřených otázek
- kontrola logických vztahů mezi otázkami
- kontrola opakujících se odpovědí (zda respondent u všech otázek odpověděl stejnou variantu)
- kontrola odlehlých pozorování (týká se číselných proměnných)
- kontrola chybějících hodnot (nemá smysl pracovat s dotazníkem, kde většina odpovědí chybí)
- nastavení správného typu proměnných

# Zpráva ze statistického výzkumu

Každá zpráva ze statistického výzkumu by měla obsahovat:

- Cíl výzkumu (výzkumná otázka/výzkumné otázky)
- Popis sběru dat
  - **počet pozorování**
  - kde a jak byla data sbírána
  - popis komplikací a jejich řešení
- Popisné statistiky a grafy pro důležité proměnné
- Výběr statistické metody a jeho odůvodnění včetně kontroly předpokladů
- Závěry ze statistických analýz
  - **nevkládají se přímo printscreeny ze softwaru**
  - výsledky se shrnují do textu / do přehledných tabulek / graficky
- Celkový závěr vzhledem ke splnění cílů výzkumu a vzhledem k budoucím výzkumům

# Statistický software

- **excel** – vhodný pro přípravu dat, popisné statistiky a grafy
  neumí všechny potřebné analýzy
- **SPSS** – software vyrobený na míru pro analýzy dotazníkových dat
  nainstalovaný v učebnách, jinak drahý
  v tomto softwaru budeme pracovat především
- **jamovi** – volně stažitelný na `www.jamovi.org`
  software na bázi R
- **JASP** – volně stažitelný na `www.jasp-stats.org`
  software na bázi R
- **Gretl** – software specializovaný na analýzu (ekonomických) časových řad
  volně stažitelný

# Jamovi

**Obsluha** softwaru

- otevírá přímo excelovské databáze
- ovládá se přes menu
- okamžitá odezva – něco naklikám a hned to vidím, když se mi to nelíbí, tak to zase vykliknu pryč

**Výhody** softwaru

- **jednoduchost** obsluhy
- obsahuje i složitější analýzy narozdíl od excelu
- kreslí pěkné grafy
- umožňuje export výstupů do html/pdf/jiné typy

**Nevýhody** softwaru

- umožňuje jen základní úpravu dat potřebnou pro analýzy (výpočet nové proměnné, transformace původní ...)
- negeneruje kód, analýzy se pokaždé musí znovu naklikat
- neumožňuje hromadné analýzy "přes všechny proměnné"

# JASP

Nabízí velmi podobné prostředí jako *jamovi*

- otevírá se přímo excelovská databáze
- jednoduchost ovládání (přes menu)
- okamžité výstupy
- nabízí přehledné výstupy a grafy
- výpočty zprostředkovává statistický software R
- umožňuje export výstupů do html/pdf
- neumožňuje uložení kódu

**Rozdíly** oproti *jamovi*

- trochu jiný design
- umožňuje lepší editaci dat
- **jiná základní nabídka analýz**
- jiné možnosti nastavení parametrů analýz

# Gretl

Jde o **specializovaný** software a proti *JASP* nebo *jamovi* nabízí méně intuitivní ovládání

- nenabízí tolik základních analýz
- výstupy nejsou tak intuitivní a přehledné
- očekává se, že uživatel má potřebné teoretické znalosti používaných analýz
- umožňuje zadání analýz kódem i přes menu
- nenabízí náhled na data

# SPSS

Veškerou práci s daty budeme realizovat ve **statistickém softwaru SPSS**.

- software je dostupný v počítačových učebnách FSE
- na internetu je volně dostupná 30-ti denní zkušební verze
  - doporučuji stáhnout až na vlastní zpracování BP
- obsahuje velké množství snadno dostupných statistických metod
- **výhoda** je kombinace volby metod přes menu a zdrojového kódu
  - pokud se přes menu dostanete k požadovanému výstupu, můžete si uložit zdrojový kód a příště ho využít pro jinou proměnnou
  - vygenerování zdrojového kódu přes tlačítko **Paste** ve spodní části každé šablony
  - zdrojový kód je možné doplnit komentářem

# Práce v SPSS

Načtení dat z excelu

- File -> Import Data -> Excel
- a vybrat databázi Prij.xlsx
- tlačítkem *OK* potvrdíte výběr, tlačítkem *Paste* vygenerujete kód

```
GET DATA
/TYPE=XLSX
/FILE='C:\Users\Cernikova\Documents\Statisticka_AD\Prij.xlsx'
/SHEET=name 'Prij'
/CELLRANGE=FULL
/READNAMES=ON
/DATATYPEMIN PERCENTAGE=95.0
/HIDDEN IGNORE=YES.
EXECUTE.
DATASET NAME DataSet2 WINDOW=FRONT.
```

# Práce v SPSS

Nastavení vlastností / typu proměnných

- Data -> Define variable properties
- zvolíte proměnnou a změníte typ, pojmenujete úrovně, nastavíte kreté hodnoty patří do chybějících atd.
- opět si můžete vygenerovat kód přes tlačítko *Paste* a pak už jen přidat další proměnné se stejnými vlastnostmi
- následující kód přidá popisky ke kategoriím promněnných *Obor* a *Pohlavi*, nastaví proměnnou *Mat* jako ordinální

```
DATASET ACTIVATE DataSet1.
* Define Variable Properties.
*Obor.
VARIABLE LABELS  Obor 'Obor studia'.
VALUE LABELS Obor
'MAN' 'Management'
'FIN' 'Finance'
'MNV' 'Mezinárodní vztahy'
'HOSP' 'Hospodářství'.

*Pohlavi.
VALUE LABELS Pohlavi
'z' 'Zena'
'm' 'Muz'.

*Mat.
VARIABLE LEVEL  Mat(ORDINAL).
VARIABLE LABELS  Mat 'Matematika'.
EXECUTE.
```

# Práce v SPSS

Tvorba nových proměnných

- Transform
  - *Compute Variable* – vytvoří novou proměnnou
  - *Count Values within Cases* – počítá, kokrát nastane určitá hodnota
  - *Recode into Same/ Different Variable* – překódování proměnné
  - *Create Dummy Variables* – vytvoří pomocné 0-1 proměnné
  - *Rank Cases* – vytvoří novou proměnnou s pořadím
- následující kód vytvoří nové proměnné s průměrnou známkou z VŠ předmětů, s počtem jedniček z VŠ předětů, překódovaným sportem na Ano x Ne a s pořadím pro celkový počet bodů u přijímaček

```
DATASET ACTIVATE DataSet1.
COMPUTE prum_zn=MEAN(eknm, aj, nj, mat, stat).
EXECUTE.

COUNT jedna=eknm aj nj mat stat(1).
VARIABLE LABELS  jedna 'Pocet jednicek'.
EXECUTE.

RECODE Sport ('denne'=1) ('casto'=1) ('pravidelne'=1) ('obcas'=1) ('ne'=0) INTO SportAN.
VARIABLE LABELS  SportAN 'Sport: Ano Ne'.
EXECUTE.

RANK VARIABLES=celprij (A)
/RANK
/PRINT=YES
/TIES=MEAN.
```

# Popisné statistiky

Máme-li k dospozici vyčištěná data, na jejichž základě můžeme odpovědět na výzkumnou otázku, je třeba naměřené výsledky **popsat**

- nezapomenout na počet měření (je rozdíl, když své závěry zakládáte na deseti odpovědích nebo na stu)
- je dobré popsat i kdo odpovídal (např. kolik odpovídalo mužů a kolik žen, případně s jakým věkovým složením)
- kromě číselných statistik použijte i grafy
- všechny výstupy je dobré okomentovat
- mezi popisné statistiky patří i interval spolehlivosti
- je lépe prezentovat výsledky v tabulce nebo v textu?

# Popisné statistiky

- **Číselné proměnné**
  - popisné statistiky polohy – průměr, medián, vybrané percentily (kvartily, extrémy)
  - popisné statistiky variability – rozptyl, směrodatná odchylka, mezikvartilové rozpětí, koeficient variace
  - popisné statistiky tvaru rozdělení – šikmost, špičatost
  - grafické charakteristiky – krabicový graf, histogram
- **Nominální proměnné**
  - číselné charakteristiky – absolutní a relativní četnosti
  - modus – nejčastěji uváděná hodnota
  - grafické charakteristiky – sloupcový a koláčový graf
- **Ordinální proměnné**
  - vhodné jsou absolutní i relativní četnosti, k obojímu též kumulativní četnosti
  - lze použít také průměr, medián atd.

# Opakování – popisné statistiky

Popisné statistiky pro číselnou proměnnou

- **průměr**: sečtu všechny hodnoty a vydělím počtem pozorování
  nejlepší odhad střední hodnoty
- **medián**: hodnota prostřední podle velikosti, nebo průměr prostředních dvou
- vybrané percentily, především **extrémy** a **kvartily**: hodnoty v jedné a ve třech čtvrtinách podle velikosti
- **Směrodatná odchylka**: základní charakteritika variability měřená ve stejných jednotkách jako původní veličina
  průměrná odchylka od průměru na druhou
- **Mezikvartilové rozpětí**: $IQR(X) = Q_3 - Q_1$, kde $Q_3$ je třetí kvartil a $Q_1$ je první kvartil

# Popisné statistiky – číselné proměnné

Popisné statistiky tvaru rozdělení pro číselnou proměnnou se počítají ze standardizovaných proměnných, tak zvaných **Z-skórů**

$$z_i = \frac{x_i - \overline{x}}{sd(x)}$$

- **Šikmost** – průměr ze třetích mocnin z-skórů

$$\text{Skew}(X) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{sd(x)}\right)^3 = \frac{\sum_{i=1}^{n} z_i^3}{n}$$

- **Špičatost** – průměr ze čtvrtých mocnin z-skórů mínus 3

$$\text{Kurt}(X) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{sd(x)}\right)^4 - 3 = \frac{\sum_{i=1}^{n} z_i^4}{n} - 3$$

Použitelné jen pro jednovrcholové rozdělení.

# Popisné statistiky – číselné proměnné

## Šikmost

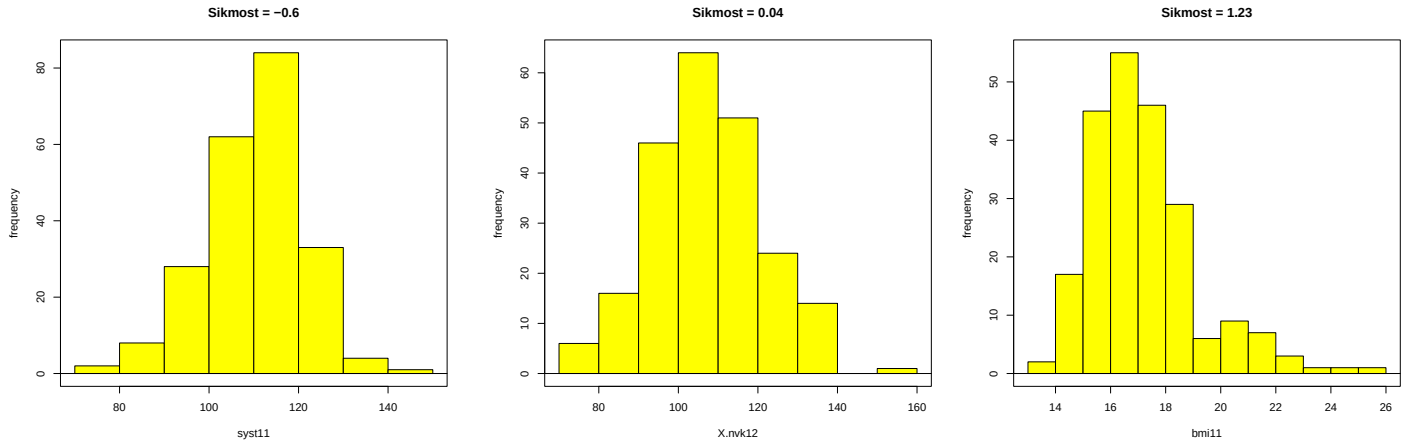
- nulová šikmost značí symetrická data
- záporná šikmost značí protažení histogramu doleva
- kladná šikmost značí protažení histogramu doprava
- sešikmení se šikmostí v intervalu od $-0.25$ do $0.25$ není na histogramu vidět

## Špičatost

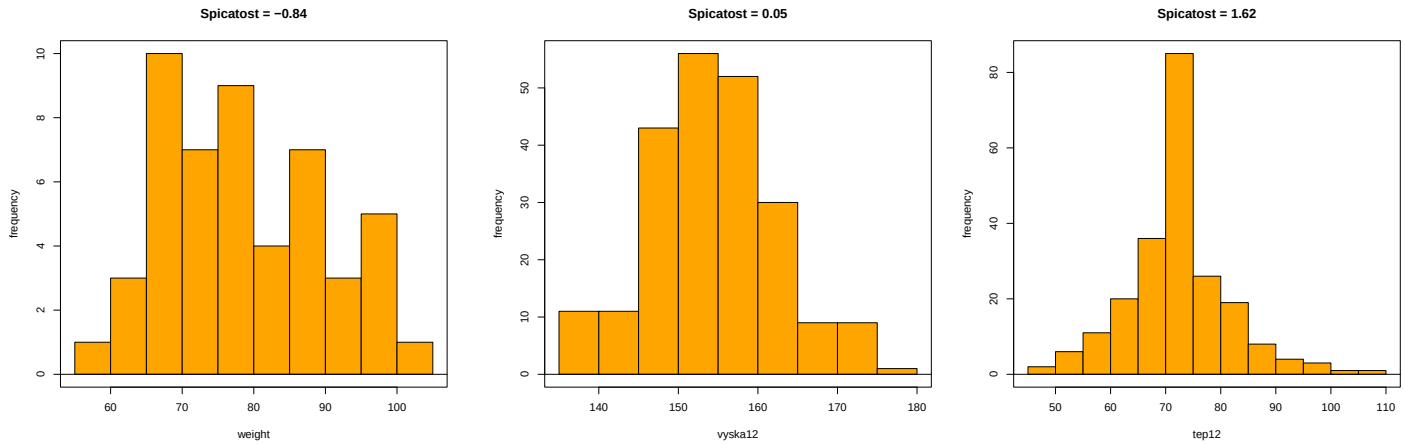
- nulová špičatost značí špičatost Gaussovy křivky
- záporná špičatost značí plochý histogram s podobně vysokými sloupci
- kladná špičatost značí histogram s jedním/ dvěma vysokými sloupci uprostřed a spoustou malých sloupců okolo
- špičatost v intervalu od $-0.5$ do $0.5$ se dá považovat za blízkou nule
- na sešikmených datech se špičatost špatně odhaduje

## Popisné statistiky tvaru rozdělení

Ukázka záporné, nulové a kladné šikmosti




Ukázka záporné, nulové a kladné špičatosti



# Odhady střední hodnoty

U číselné proměnné nejčastěji odhadujeme populační průměr

- **nejlepším bodovým odhadem** střední hodnoty je výběrový průměr $\overline{X} = \sum_{i=1}^{n} X_i/n$
- nestranný odhad
- platí Centrální limitní věta – rozdělení výběrového průměru konverguje k normálnímu pro $n \to \infty$
- platí Zákon velkých čísel – výběrový průměr konverguje ke střední hodnotě pro $n \to \infty$
- střední chyba průměru je $\text{SEM} = \text{sd(X)}/\sqrt{\text{n}}$
- **intervalový odhad** pro průměr je

$$\overline{X} \pm t_{n-1}(1-\alpha/2)\frac{\text{sd(X)}}{\sqrt{n}}$$

# Odhady pro pravděpodobnost

Podíl jedné konkrétní odpovědi na otázku v dotazníku

- **nejlepším bodovým odhadem** pravděpodobnosti je relativní četnost $p_i = n_i/n$
- nestranný odhad
- náhodná veličina $p = (p_i - \pi_i)/\sqrt{\pi_i(1-\pi_i)/n}$ konverguje k normálnímu rozdělení $N(0,1)$ pro $n \to \infty$
- **intervalový odhad** pro pravděpodobnost (Waldův se spojitostní korekcí) je

$$p_i \pm \frac{1}{2n} \pm z(1-\alpha/2)\sqrt{\frac{p_i(1-p_i)}{n}}$$

- pro použití tohoto intervalu musíme mít dostatečně velké $n$ a $p_i$, má platit $np_i(1-p_i) > 9$

**S většinou škálovaných proměnných z dotazníku pracujeme jako s kategorickými a počítáme pro ně podíly jednotlivých odpovědí!**

# Práce v SPSS

Popisné statistiky číselné proměnné

- Analyze -> Descriptive statistics -> Explore
- zvolit číselnou proměnnou Celprij
- mohu nastavit i třídění podle kategorické proměnné, kterou bychom dali do pole *Factor*
- vpravo přes tlačítko *Statistics* je možné nastavit spolehlivost intervalu spolehlivosti a výpočet vybraných percentilů
- vpravo přes tlačítko *Graphs* zaškrtněte Histogram namísto Stem-and-leaves a můžete přidat grafické a číselné testy normality

```
DATASET ACTIVATE DataSet1.
EXAMINE VARIABLES=celprij
/PLOT BOXPLOT HISTOGRAM
/COMPARE GROUPS
/PERCENTILES(5,10,25,50,75,90,95) HAVERAGE
/STATISTICS DESCRIPTIVES
/CINTERVAL 95
/MISSING LISTWISE
/NOTOTAL.
```

# Práce v SPSS

Popisné statistiky nominální proměnné

- Analyze -> Descriptive statistics -> Frequencies
- zvolit kategorickou proměnnou Obor
- vpravo přes tlačítko *Bootstrap* je možné přidat bootstrapové intervaly spolehlivosti pro jednotlivé kategorie
- vpravo přes tlačítko *Graphs* zvotle sloupcový nebo koláčový graf případně ve vygenerovaném syntaxu obojí
- vygenerovaný graf je možné upravit (třeba přidat popisky): dvojím kliknutím na graf se aktivuje jeho editace

```
BOOTSTRAP
/SAMPLING METHOD=SIMPLE
/VARIABLES INPUT=Obor
/CRITERIA CILEVEL=95 CITYPE=PERCENTILE  NSAMPLES=1000
/MISSING USERMISSING=EXCLUDE.
FREQUENCIES VARIABLES=Obor
/BARCHART FREQ
/PIECHART FREQ
/ORDER=ANALYSIS.
```

Potřebujete-li interval spolehlivosti pro procento

- Analyze -> Descriptive statistics -> Proportion Confidence Interval
- nastavíte počet úspěchů a počet možností

```
PROPOR NUM=23 DENOM=60 /LEVEL ALPHA=.05.
```

# Bootstrap

**Bootstrap** je statistická metoda umožňující počítat intervaly spolehlivosti a testy bez znalosti pravděpodobnostního rozdělení dat

- metoda je založená na generovaných bootrapových výběrech
- aktuální naměřené hodnoty se berou jako vstupní populace
- ze vstupní populace se generuje bootstrapový výběr náhodným výběrem s opakováním
- bootstrapových výběrů se vygeneruje velký počet, např. $B = 1000$
- pro každý výběr se spočítá odhadovaná statistika, např. průměr
- 95%-ní interval spolehlivosti se určí tak, že vezmu 2.5% a 97.5%-ní kvantil

# Práce v SPSS

Grafy

- některé grafy jsou dosažitelné přes základní popisné statistiky - viz. výše
- další grafy jsou k dispozici přes menu Graphs
- Graphs -> Graphboard Template Chooser
- nebo Graphs -> a typ grafu
- zvolit proměnnou a typ grafu
- editace grafů možná přes dvojí kliknutí na vykreslený graf (můžete měnit barvy, tloušťku čar, styl popisků atd.)

```
GGRAPH
/GRAPHDATASET NAME="graphdataset"
VARIABLES=celprij[LEVEL=scale]
MISSING=LISTWISE REPORTMISSING=NO
/GRAPHSPEC SOURCE=VIZTEMPLATE(NAME="Histogram"[LOCATION=LOCAL]
MAPPING( "x"="celprij"[DATASET="graphdataset"] "Summary"="count" "Title"='Histogram pro '+
'celkový počet bodů u přijímaček'))
VIZSTYLESHEET="Traditional"[LOCATION=LOCAL]
LABEL='HISTOGRAM: celprij'
DEFAULTTEMPLATE=NO.

GRAPH
/BAR(SIMPLE)=COUNT BY Obor
/TITLE='Sloupcový graf'.
```

# Interpretace intervalu spolehlivosti

Interval spolehlivosti nestačí správně spočítat, ale je nutné znát i jeho správnou **interpretaci**.

- Interval spolehlivosti pro průměr: Se spolehlivostí *xx*% se skutečná střední hodnoty nachází v intervalu od ... do ....
- Interval spolehlivosti pro pravděpodobnost: Se spolehlivostí *xx*% se skutečná pravděpodobnost nachází v intervalu od ... do ....

# Základy testování hypotéz

Při statistickém rozhodování testujeme proti sobě 2 hypotézy

- Nulovou hypotézu – značíme $H_0$
  – je v ní vždy pouze jedna varianta
- Alternativní hypotézu – značíme $H_A$
  – obsahuje více možností (např. interval)

Nejčastěji požívané testované hypotézy

- $H_0$ : mezi skupinami není rozdíl
  $H_A$ : mezi skupinami je rozdíl
- $H_0$ : proměnné spolu nesouvisí
  $H_A$ : proměnné spolu souvisí

# Základy testování hypotéz

Na základě testu uděláme jedno ze dvou rozhodnutí

- Zamítneme nulovou hypotézu – platí alternativa
- Nezamítneme nulovou hypotézu

Při rozhodování můžeme udělat chybu

- chyba prvního druhu – zamítneme $H_0$, přestože platí
  – značí se $\alpha$, a jmenuje se **hladina významnosti**
  – závažnější z obou chyb
- chyba druhého druhu – nezamítneme $H_0$, přestože neplatí
  – značí se $\beta$ a hodnota $1 - \beta$ se nazývá **síla testu**
  – za dané hladiny významnosti chceme test co nejsilnější

# Základy testování hypotéz

| | Zamítáme $H_0$ | Nezamítáme $H_0$ |
|---|---|---|
| Skutečně platí $H_0$ | Chyba I. druhu<br>$\leq \alpha$ | OK |
| Skutečně platí $H_1$ | OK<br>síla testu | Chyba II. druhu<br>$\beta$ |

# Základy testování hypotéz

Testovat mohu buď přes porovnání *testové statistiky* a *kritické hodnoty* (kvantil vybraného teoretického rozdělení), nebo přes porovnání *p-hodnoty* a *hladiny významnosti*.

Statistické softwary a tedy i my na tomto předmětu budeme využívat ***p*-hodnotu**

- aktuální dosažená hladina testu
- pravděpodobnost, že za platnosti $H_0$ nastal výsledek, jaký nastal, nebo jakýkoliv jiný, který ještě více odpovídá alternativě
- *p*-hodnota $\leq \alpha$ potom ZAMÍTÁME $H_0$
- *p*-hodnota $> \alpha$ potom NEZAMÍTÁME $H_0$

# Vztah dvou proměnných

Použité metody se liší podle typu vstupních proměnných

- **dvě číselné proměnné:** bodový graf
  – korelační koeficient, regresní modely (lineární, nelineární, poissonova regrese, atd.)
- **číselná a kategorická proměnná:** krabicové grafy
  – dvouvýběrové testy, ANOVA
- **dvě kategorické proměnné:** sloupcový graf
  – $\chi^2$-test, Fisherův test, poměr šancí, Cramérovo V
  – ukazatele závislosti pro dvě ordinální veličiny

# Kontrola normality

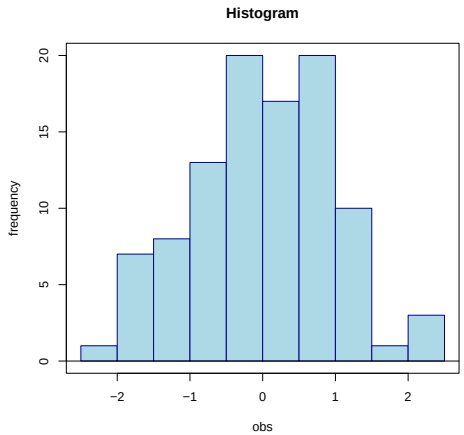
U korelačního koeificientu, jednovýběrových, párových a dvouvýběrových testů a u ANOVY je třeba rozlišovat jejich **parametrickou** a **neparametrickou** verzi.

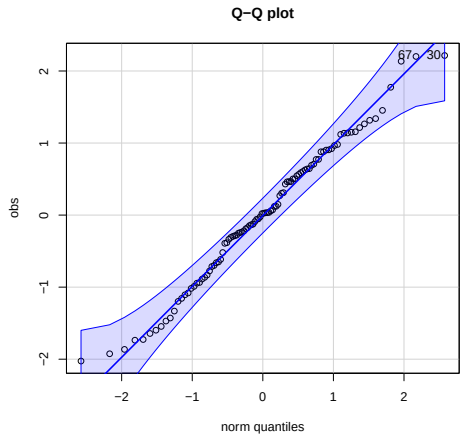
- pro data s přibližně normálním rozdělením se používají parametrické testy: t-testy, klasická ANOVA, Pearsonův korelační koeficient
- pro data, která nemají normální rozdělení se používají neparametrické testy: Wilcoxonův test, Mann-Whitneyův test, Kruskal-Wallisův test, Spearmanův korelační koeficient

# Kontrola normality

Nejprve je tedy třeba otestovat normalitu

- **Grafické testy** – histogram a pravděpodobnostní graf




- **Číselné testy** – např. Shapiro-Wilkův, Kolmogorovův-Smirnovův
- Testované hypotézy
  - $H_0$ : data mají normální rozdělení
  - $H_1$ : data nemají normální rozdělení

# Práce v SPSS

Kontrola normality

- Analyze -> Descriptive statistics -> Explore
- zvolit proměnnou
- vpravo přes tlačítko *Graphs* zaškrtnout Normality plots with tests a vyšrtnout nepotřebné grafy

```
EXAMINE VARIABLES=celprij
/PLOT HISTOGRAM NPPLOT
/STATISTICS DESCRIPTIVES
/CINTERVAL 95
/MISSING LISTWISE
/NOTOTAL.
```

- pokud máte v datech shodné hodnoty, je potřeba nakreslit jiný Q-Q plot
- Analyze -> Descriptive statistics -> Q-Q Plots
- zvolit proměnnou
- nastavit vpravo dole *Rank Assigned to Ties* *Break ties arbitrarily*

```
PPLOT
/VARIABLES=celprij
/NOLOG
/NOSTANDARDIZE
/TYPE=Q-Q
/FRACTION=BLOM
/TIES=BREAK
/DIST=NORMAL.
```

- normalitu můžete testovat i ve výběrech zvlášť

# Neparametrické testy

Hlavní důvody porušení normality:

- odlehlé hodnoty,
- sešikmená data

**Neparametrické testy jsou založeny na pořadích**

**Příklad.** *Uvažujme naměřené věky otců 30, 28, 36, 38, 28, 26, 29, 37, 25, 50. Data věků rodičů bývají sešikmena a často obsahují odlehlé hodnoty. Přiřadíme-li hodnotám pořadí podle velikosti, získáme řadu 6, 3.5, 7, 9, 3.5, 2, 5, 8, 1, 10. Takto získaná řada není sešikmená a nemá odlehlé hodnoty.*

# Jednovýběrový test

Nejjednodušším testem je **jednovýběrový test o střední hodnotě**.

- $H_0$ : střední hodnota $= \mu_0$
- $H_1$ : střední hodnota $\neq \mu_0$, nebo $< \mu_0$, nebo $> \mu_0$

Nejčastěji používané jednovýběrové testy:

- **jednovýběrový t-test** s testovou statistikou

$$T = \frac{\overline{X} - \mu_0}{\mathrm{sd}(X)}\sqrt{n}$$

  za platnosti $H_0$ má tato statistika $t$-rozdělení o $n-1$ stupních volnosti.

- **jednovýběrový Wilcoxonův test,** který má pro větší počet pozorování testovou statistiku

$$U = \frac{S^+ - \frac{1}{4}n(n+1)}{\sqrt{\frac{1}{24}n(n+1)(2n+1)}}$$

  kde $S^+$ je součet pořadí kladných rozdílů $X_i - \mu_0$,
  za platnosti $H_0$ má tato statistika $N(0,1)$ rozdělení.

# Práce v SPSS

Jednovýběrový t-test

- Analyze -> Compare Means and Proportions -> One-Sample T-test
- zvolit proměnnou a testovanou hodnotu
- přes tlačítko *Options* možno nastavit hladinu významnosti pro interval spolehlivosti

```
T-TEST
/TESTVAL=145
/MISSING=ANALYSIS
/VARIABLES=celprij
/ES DISPLAY(TRUE)
/CRITERIA=CI(.95).
```

Jednovýběrový Wilcoxonův test

- Analyze -> Nonparametric Tests -> One-Sample
- v záložce *Fields* zvolit proměnnou
- v záložce *Statistics* nastavit Customized test Wilcoxon test a testovanou hodnotu

```
NPTESTS
/ONESAMPLE TEST (celprij) WILCOXON(TESTVALUE=145)
/MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
/CRITERIA ALPHA=0.05 CILEVEL=95  SEED=RANDOM.
```

# Test o pravděpodobnosti

Testované hypotézy

- $H_0$ : pravděpodobnost daného jevu $= \pi_0$
- $H_1$ : pravděpodobnost daného jevu $\neq \pi_0$, nebo $< \pi_0$, nebo $> \pi_0$

**Testová statistika** nejčastěji používaného testu o pravděpodobnosti je

$$Z = \frac{p_i - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$$

za platnosti nulové hypotézy má tato statistika $N(0,1)$-rozdělení.

- v SPSS budeme používat alternativu odpovídající Vám známému intervalu spolehlivosti

# Práce v SPSS

Jednovýběrový test o pravděpodobnosti

- Analyze -> Compare Means and Proportions -> One-Sample Proportions
- zvolit proměnnou a kód pro testovanou hodnotu pravděpodobnosti
- přes tlačítko *Confidence intervals* a *Tests* nastavte variantu *Exact* a *Wald (Continuity correction)* a můžete změnit i požadovanou spolehlivost a testovanou hodnotu

```
PROPORTIONS
/ONESAMPLE eknm TESTVAL=0.5 TESTTYPES=EXACT WALDCC  CITYPES=CLOPPER_PEARSON WALDCC
/SUCCESS VALUE=LEVEL(1 )
/CRITERIA CILEVEL=95
/MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE.
```

# Párový test

Porovnání dvou závislých výběrů
Testované hypotézy

- $H_0$ : střední hodnota rozdílu párů $= \mu_0$
- $H_1$ : střední hodnota rozdílu $\neq \mu_0$, nebo $< \mu_0$, nebo $> \mu_0$

U t-testu i u Wilcoxonova testu je postup stejný. Nejprve se spočítají rozdíly

$$R_i = X_i - Y_i$$

kde $X_i$ a $Y_i$ jsou párová měření.
Tyto rozdíly se pak testují příslušným **jednovýběrovým testem**.

# Práce v SPSS

Párový t-test

- Analyze -> Compare Means and Proportions -> Paired Samples T-test
- zvolit proměnné
- na výstupu kromě párového testu najdu i hodnotu korelačního koeficientu a výstupy týkající se věcné významnosti (*Effect size*)

```
T-TEST PAIRS=ss2 WITH ss3 (PAIRED)
/ES DISPLAY(TRUE) STANDARDIZER(SD)
/CRITERIA=CI(.9500)
/MISSING=ANALYSIS.
```

Párový Wilcoxonův test

- Analyze -> Nonparametric Tests -> Related Samples
- v záložce *Fields* zvolit proměnné
- v záložce *Statistics* nastavit Wilcoxon matched-pair signed-rank (2 samples)

```
NPTESTS
/RELATED TEST(ss2 ss3) WILCOXON
/MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
/CRITERIA ALPHA=0.05  CILEVEL=95.
```

# Věcná významnost

Porovnání dvou skupin

- Cohenovo $d$

$$d = \frac{\overline{X} - \overline{Y}}{\sqrt{S^2}}, \qquad S^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2}$$

  do 0.5 malý efekt, 0.5-0.8 střední efekt, nad 0.8 velký efekt

- Hedgesovo $g$

$$g = \frac{\overline{X} - \overline{Y}}{\sqrt{MSe}},$$

  $MSe$ je zbytkový průměrný součet čtverců z tabulky analýzy rozptylu
  do 0.5 malý efekt, 0.5-0.8 střední efekt, nad 0.8 velký efekt

# Dvouvýběrový t-test

Porovnání dvou nezávislých výběrů.
Testované hypotézy

- $H_0$ : rozdíl středních hodnot $= \mu_0$
- $H_1$ : rozdíl středních hodnot $\neq \mu_0$, nebo $< \mu_0$, nebo $> \mu_0$

**Testová statistika** dvouvýběrového t-testu pro shodné rozptyly je

$$T = \frac{\overline{X} - \overline{Y} - \mu_0}{S} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

kde

$$S = \frac{1}{n_1 + n_2 - 2} \left( \sum_{i=1}^{n_1} (X_i - \overline{X})^2 + \sum_{i=1}^{n_2} (Y_i - \overline{Y})^2 \right)$$

a $n_1, n_2$ je rozsah výběru $X$, respektive $Y$. Za platnosti nulové hypotézy má tato statistika $t$-rozdělení o $n_1 + n_2 - 2$ stupních volnosti.

# Dvouvýběrový t-test

V případě, že výběry shodné rozdělení nemají, používá se **Welchova varianta dvouvýběrového t-testu**. Její **testová statistika** má tvar

$$T = \frac{\overline{X} - \overline{Y} - \mu_0}{\sqrt{\frac{\mathrm{Var(X)}}{n_1} + \frac{\mathrm{Var(Y)}}{n_2}}}$$

Tato statistika má za platnosti nulové hypotézy $t$-rozdělení o $\nu$ stupních volnosti, kde

$$\nu = \frac{(\mathrm{Var(X)}/\mathrm{n_1} + \mathrm{Var(Y)}/\mathrm{n_2})^2}{\frac{(\mathrm{Var(X)}/\mathrm{n_1})^2}{n_1 - 1} + \frac{(\mathrm{Var(Y)}/\mathrm{n_2})^2}{n_2 - 1}}.$$

kritické hodnoty je možno odvodit, přestože $\nu$ není celé číslo.

# Porovnání dvou rozptylů

Pokud tedy máte přibližně normální data a hodláte použít dvouvýběrový test, je nejprve nutné otestovat shodu rozptylů.
Testované hypotézy

- $H_0$ : rozptyly jsou stejné
- $H_1$ : rozptyly se liší

Na výběr je více testů

- F-test pro porovnání dvou rozptylů
- Bartlettův test
- Levenův test

# Wilcoxonův dvouvýběrový test

Testová statistika dvouvýběrového Wilcoxonova testu je založena na pořadích dat ve sdruženém výběru. Označme $T_1$, $T_2$ součet pořadí v prvním, respektive druhém výběru. Dále vypočteme

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - T_1, U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - T_2,$$

kde $n_1, n_2$ jsou rozsahy jednotlivých výběrů. Přesný test porovnává hodnotu $min(U_1, U_2)$ s kritickou hodnotou. Asymptoticky platí, že

$$U_0 = \frac{U_1 - \frac{1}{2} n_1 n_2}{\sqrt{\frac{n_1 n_2}{12}(n_1 + n_2 + 1)}}$$

má za platnosti $H_0$ $N(0,1)$ rozdělení.

# Wilcoxonův dvouvýběrový test

**Příklad.** *Chceme porovnat výsledky testů studentů v Ústí nad Labem a v Liberci. Studenti v Ústí dostali bodová ohodnocení 45, 79, 81, 56, 53, 77. Studenti v Liberci získali ohodnocení 76, 62, 84, 80, 41, 79, 66. Testujeme*

- $H_0$ : *Studenti v Ústí a v Liberci jsou stejní*
- $H_1$ : *Studenti v Ústí a v Liberci se liší.*
- *V prvním kroku srovnám všechny hodnoty do řady*
  *41, 45, 53, 56, 62, 66, 76, 77, 79, 79, 80, 81, 84*
- *následně jim přiřadím pořadí*
  *1, 2, 3, 4, 5, 6, 7, 8, 9.5, 9.5, 11, 12, 13*
- *pak vypočtu* $T_1 = 38.5$, $T_2 = 52.5$, $U_1 = 24.5$, $U_2 = 17.5$, $U_0 = 0.5$, $p = 0.6678$

*P-hodnota* $> \alpha$ *a tedy nezamítám nulovou hypotézu, neprokázal se rozdíl mezi studenty v Ústí a v Liberci.*

# Práce v SPSS

Dvouvýběrový t-test

- Analyze -> Compare Means and Proportions -> Independent Samples T-test
- zvolit proměnné a nastavit porovnávaný rozdíl skupin
- SPSS automaticky porovná rozptyly a nabídne oba t-testy

```
T-TEST GROUPS=Pohlavi('m' 'z')
/MISSING=ANALYSIS
/VARIABLES=celprij
/ES DISPLAY(TRUE)
/CRITERIA=CI(.95).
```

Dvouvýběrový Wilcoxonův test

- Analyze -> Nonparametric Tests -> Independent Samples
- v záložce *Fields* zvolit proměnné
- v záložce *Statistics* nastavit Mann-Whitney U (2 samples)

```
NPTESTS
/INDEPENDENT TEST (matprij) GROUP (Pohlavi) MANN_WHITNEY
/MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
/CRITERIA ALPHA=0.05  CILEVEL=95.
```

# Testy srovnávající pravděpodobnosti

Stejně jako můžete porovnávat průměry ve dvou výběrech, můžete porovnávat i procenta

- **Párový test** pro dva závislé výběry je opět založen na rozdílu
- **Dvouvýběrový test** porovnává pravděpodobnost ve dvou nezávislých skupinách
- Testová statistika dvouvýběrového testu má tvar

$$Z = \frac{p_1 - p_2 - \pi_0}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \text{ kde } p = \frac{x_1 + x_2}{n_1 + n_2}$$

kde

- $n_1, n_2$ jsou počty pozorování v prvním a druhém výběru
- $x_1, x_2$ jsou počty výskytů sledovaného jevu v prvním a druhém výběru
- $p_1, p_2$ jsou relativní četnosti jevu v prvním a druhém výběru

# Práce v SPSS

Párový test pro dvě pravděpodobnosti

- Analyze -> Compare Means and Proportions -> Paired Samples Proportions
- zvolit proměnné a nastavit hodnotu, jejíž pravděpodobnost chceme testovat
- v záložce *Tests* a *Confidence Intervals* nastavit *Exact Binomial*/ *Bonett-Price* a *Wald (Continuity Corrected)*

```
PROPORTIONS
/PAIREDSAMPLES aj WITH nj (PAIRED) TESTTYPES=EXACT WALDCC  CITYPES=BONETT_PRICE WALDCC
/SUCCESS VALUE=LAST
/CRITERIA CILEVEL=95
/MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE.
```

Dvouvýběrový test pro dvě pravděpodobnosti

- Analyze -> Compare Means and Proportions -> Independent Samples Proportions
- zvolit proměnné a nastavit hodnotu, jejíž pravděpodobnost chceme testovat
- v záložce *Tests* a *Confidence Intervals* nastavit *Agresti-Caffo* a *Wald (Continuity Corrected)*

```
PROPORTIONS
/INDEPENDENTSAMPLES stat BY Pohlavi SELECT=LEVEL("m" , "z" ) CITYPES=AGRESTI_CAFFO WALDCC
TESTTYPES=WALDH0CC
/SUCCESS VALUE=LAST
/CRITERIA CILEVEL=95
/MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE.
```

# Analýza rozptylu (ANOVA)

Rozlišujeme **klasickou ANOVU** pro normálně rozložená data ve všech výběrech a **Kruskal-Wallisovu ANOVU** pro data, která nemají normální rozdělení

- Klasická ANOVA porovnává variabilitu mezi výběry s variabilitou v rámci výběrů
- Výstupem z analýzy rozptylu je tzv. **tabulka analýzy rozptylu**

| | Součty čtverců | Stupně volnosti | Průměrné čtverce | Testová statistika | $p$-hodnota |
|---|---|---|---|---|---|
| Faktor $A$ | $SSA$ | $dfA = k - 1$ | $MSA$ | $F = MSA/MSe$ | $p$ |
| Chyba $e$ | $SSe$ | $dfe = n - k$ | $MSe$ | | |
| Celkem | $SST$ | $dft = n - 1$ | | | |

- používá se pro výběry se shodnými rozptyly

# Welchova analýza rozptylu

Nemají-li všechny porovnávané skupiny stejné rozptyly, používá se tzv. **Welchova ANOVA**.

- založena na myšlence vážení skupinových průměrů vahou odpovídající jejich variabilitě
- místo celkového průměru se používá průměr vážený

$$\overline{X}_w = \frac{\sum_{i=1}^{k} w_i \overline{X}_{i.}}{\sum_{i=1}^{k} w_i}, \text{ kde } w_i = \frac{n_i}{\text{Var(X)}_\text{i}}$$

- variabilita mezi výběry

$$SSA_w = \sum_{i=1}^{k} w_i (\overline{X}_{i.} - \overline{X}_w)^2, MSA_w = \frac{SSA_w}{k-1}$$

- testová statistika

$$F_w = \frac{SSA_w/(k-1)}{1 + \frac{2\Lambda(r-2)}{3}}, \text{ kde } \Lambda = \frac{3\sum_{i=1}^{k} \frac{\left(1 - \frac{w_i}{\sum_{i=1}^{k} w_i}\right)^2}{n_i - 1}}{k^2 - 1}$$

# Párové srovnání

Vyjde-li mezi skupinami významný rozdíl, hledáme, které dvojice se od sebe významně liší.

- párové srovnání provádíme pomocí **Tukeyho testu**

Testuje se

- $H_0$ : střední hodnoty $\mu_i$ a $\mu_j$ jsou stejné
- $H_1$ : střední hodnoty $\mu_i$ a $\mu_j$ se liší

pro všechny dvojice $i$ a $j$.
Testová statistika má tvar

$$Q = \frac{|\overline{X}_{i.} - \overline{X}_{j.}|}{s^*}, \text{ kde } s^* = \sqrt{\frac{SSe}{n(n-k)}}$$

Rozdělení těchto statistik se jmenuje studentizované rozpětí a má své vlastní tabelované kritické hodnoty.

# Věcná významnost

Porovnání více než dvou skupin (ANOVA)

- Fisherovo $\eta^2$

$$\eta^2 = \frac{SSA}{SST}$$

kde *SSA* a *SST* jsou součty čtverců z tabulky analýzy rozptylu
procento vysvětlené variability

- Haysova $\omega^2$

$$\omega^2 = \frac{SSA - (k-1)MSe}{SST + MSe}$$

kde *SSA*, *SST* a *MSe* jsou součty čtverců/průměrné čtverce z tabulky analýzy rozptylu
procento vysvětlené variability

# Práce v SPSS

Klasická analýza rozptylu

- Analyze -> Compare Means and Proportions -> One-Way ANOVA
- zvolit proměnné (z nějakého důvodu chce SPSS kategorickou proměnnou kódovanou číselně)
- v nabídce *Options* zašktněte *Descriptive, Homogeneity of variance test, Welch test, Mean plot*
- v nabídce *Post Hoc ...* zaškrtněte *Tukey, Dunett's T3*

```
ONEWAY celprij BY eknm
/ES=OVERALL
/STATISTICS DESCRIPTIVES HOMOGENEITY WELCH
/PLOT MEANS
/MISSING ANALYSIS
/CRITERIA=CILEVEL(0.95)
/POSTHOC=TUKEY T3 ALPHA(0.05).
```

Kruskal-Wallisova ANOVA

- Analyze -> Nonparametric Tests -> Independent Samples
- v záložce *Fields* zvolit proměnné
- v záložce *Statistics* nastavit *Kruskal-Wallis 1-way ANOVA (k samples)* a *Median test (k samples)* pro testovani mezi výběry

```
NPTESTS
/INDEPENDENT TEST (celprij) GROUP (Obor) KRUSKAL_WALLIS(COMPARE=PAIRWISE)
/MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
/CRITERIA ALPHA=0.05  CILEVEL=95.
```

# Korelační koeficient

Závislost dvou číselných proměnných se popisuje pomocí

- bodového grafu
- **korelačního koeficientu**

Vlastnosti korelačního koeficientu

- standardizovaná míra lineární závislosti
- nabývá hodnot mezi -1 a 1
  - absolutní nepřímá závislost má $\mathrm{Cor}(X, Y) = -1$
  - lineární nezávislost/ nekorelovanost má $\mathrm{Cor}(X, Y) = 0$
  - absolutní přímá závislost má $\mathrm{Cor}(X, Y) = 1$

# Korelační koeficient

- pro proměnné s normálním rozdělením se používá **Pearsonův korelační koeficient**

$$Cor(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

- pro proměnné, které nemají normální rozdělení se používá **Spearmanův korelační koeficient**
  - pro každou proměnnou nahradím originální hodnoty jejich pořadími
  - tato pořadí dosadím do vzorce pro výpočet Pearsonova korelačního koeficientu

# Pearsonův korelační koeficient

Test nezávislosti pomocí korelačního koeficientu

- $H_0$ : korelační koeficient $= 0$
- $H_1$ : korelační koeficient $\neq 0, > 0, < 0$

Za platnosti nulové hypotézy platí, že testová statistika

$$T = \frac{\mathrm{Cor}(X, Y)}{\sqrt{1 - \mathrm{Cor}(X, Y)^2}} \sqrt{(n-2)}$$

má $t$-rozdělení o $n - 2$ stupních volnosti.

# Práce v SPSS

Korelační matice

- Analyze -> Correlations -> Bivariate
- zvolit proměnné a vhodný typ korelačního koeficientu

```
CORRELATIONS
/VARIABLES=celprij jazprij matprij
/PRINT=TWOTAIL NOSIG FULL
/MISSING=PAIRWISE.
```

Pokud chcete odhad hodnoty korelačního koeficientu spolu s intervalem spolehlivosti

- Analyze -> Correlations -> Bivariate with Confidence Interval
- nastavit proměnné a spolehlivost
- počítá pouze pro Pearsonův korelační koeficient

```
STATS CORRELATIONS VARIABLES=celprij jazprij matprij
/OPTIONS CONFLEVEL=95 METHOD=FISHER
/MISSING EXCLUDE=YES PAIRWISE=YES.
```

Nakreslení bodového grafu

- Graphs -> Scatter/Dot
- zvolit *Simple Scatter* a pak nastavit proměnné

```
GRAPH
/SCATTERPLOT(BIVAR)=jazprij WITH celprij
/MISSING=LISTWISE.
```

# Chí-kvadrát test

Test nezávislosti dvou kategorických proměnných

- založen na rozdílu pozorovaných a očekávaných četností
- Testová statistika má tvar

$$X^2 = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(pozorovane_{i,j} - ocekavane_{i,j})^2}{ocekavane_{i,j}} = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(n_{i,j} - n_{i.}n_{.j}/n)^2}{n_{i.}n_{.j}/n}$$

- předpokladem je, že všechny očekávané četnosti jsou větší než 5

Není-li předpoklad splněn, používá se Exaktní **Fisherův test**

- počítá přímo přesnou p-hodnotu
- pro čtyřpolní tabulku se p-hodnota spočítá jako

$$p = \frac{n_{1.}!n_{2.}!n_{.1}!n_{.2}!}{n!n_{11}!n_{12}!n_{21}!n_{22}!}$$

# Míry asociace dvou nominálních proměnných

Obdoby korelačního koeficientu, hodnoty mezi 0 a 1, kde

- 0 znamerá nezávislost
- 1 znamená závislost

Symetrické míry nehodnotící směr závislosti

- koeficient $\phi = X^2/n$ – pro tabulky $2 \times 2$
- kontingenční koeficient $C = \sqrt{\frac{X^2}{n+X^2}}$
- Cramerovo $V = \sqrt{\frac{X^2}{n \times min\{k-1, r-1\}}}$ – pro libovolně velké tabulky

# Míry asociace dvou nominálních proměnných

Nesymetrické míry umožňující hodnotit i směr závislosti

- do jaké míry znalost jedné proměnné redukuje chybu při předpovědi hodnot druhé proměnné
- koeficient $\lambda = \frac{\sum_{i=1}^{r} r_i - \max_j n_{.j}}{n - \max_j n_{.j}}$, kde $r_i = \max_j n_{ij}$
  kolik variability sloucové proměnné se vysvětlí proměnnou řádkovou
- Goodman and Kruskal tau $\tau = \frac{\sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij}^2 / p_{i.} - \sum_{j=1}^{c} p_{.j}^2}{1 - \sum_{j=1}^{c} p_{.j}^2}$.
  koeficient je obdobný jako $\lambda$, ale počítá s relativními četnostmi
- koeficient nejistoty $U = -\frac{\sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij} \ln \frac{p_{ij}}{p_{i.} p_{.j}}}{\sum_{j=1}^{c} p_{.j} \ln p_{.j}}$,

# Míry asociace dvou ordinálních proměnných

Označme

- **konkordantní** páry: $X_i > X_j \wedge Y_i > Y_j$ nebo $X_i < X_j \wedge Y_i < Y_j$
  $P$ počet konkordantních párů
- **diskondartní** páry: $X_i > X_j \wedge Y_i < Y_j$ nebo $X_i < X_j \wedge Y_i > Y_j$
  $Q$ počet diskondartních párů

**Míry asociace**

- nejjednodušší, ale nemá přirozenou interpretaci $\gamma = \frac{P-Q}{P+Q}$
- Kendallovo $\tau_b = \frac{P-Q}{\sqrt{n^2-\sum_{i=1}^{r} N_{i\cdot}^2}\sqrt{n^2-\sum_{j=1}^{c} N_{\cdot j}^2}}$
  korelační koeficient znaménkových veličin
- Kendallovo (Stuartovo) $\tau_c = \frac{P-Q}{n^2}\frac{m}{m-1}$, kde $m = \min\{r, c\}$
- Sommersovo $d_{Y|x} = \frac{P-Q}{n^2-\sum_{j=1}^{c} n_{\cdot j}^2}$
  založeno na předpovídání hodnot jedné veličiny ze znalosti hodnot veličiny druhé

# Práce v SPSS

Vztah dvou nominálních proměnných

- Analyze -> Descriptive Statistics -> Crosstabs
- zvolit proměnné
- v nabídce *Statistics* zvolte míry závislosti, které Vás zajímají
- v nabídce *Cells* zvolte, co chcete vidět v kontingenční tabulce

Nominální proměnné

```
CROSSTABS
/TABLES=Obor BY aj
/FORMAT=AVALUE TABLES
/STATISTICS=CHISQ CC PHI LAMBDA UC
/CELLS=COUNT EXPECTED ROW
/COUNT ROUND CELL
/METHOD=EXACT TIMER(5).
```

Ordinální proměnné

```
CROSSTABS
/TABLES=nj BY aj
/FORMAT=AVALUE TABLES
/STATISTICS=CHISQ GAMMA D BTAU CTAU
/CELLS=COUNT EXPECTED COL
/COUNT ROUND CELL
/METHOD=EXACT TIMER(5).
```

# Vztah párových měření

Měříme dvakrát to samé a sledujeme, jestli nastala změna

- opakovaná měření po nějaké době
- měření toho samého pro různé objekty
- od každé jednotky mám dvě měření

Sledujeme míru shody

- koeficient shody **Kappa** $\kappa = \frac{P_{obs} - P_{chance}}{1 - P_{chance}}$
  porovnává zjištěnou shodu s očekávanou shodou při absolutní náhodnosti
- **McNemarův test** – test symetrie $\chi^2 = \sum_{i<j} \frac{(N_{ij} - N_{ji})^2}{N_{ij} + N_{ji}}$
- odhady rizika – jen ve čtyřpolní tabulce
  **poměr šancí** $OR = \frac{N_{11}/N_{21}}{N_{12}/N_{22}}$
  **relativní risk** $RR = \frac{N_{11}/N_{\cdot 1}}{N_{12}/N_{\cdot 2}}$

# Poměr šancí

Vyhodnocuje se ve čtyřpolní tabulce

- Je větší šance dostat chřipku ve městě nebo na venkově?

| | Chřipku má | Chřipku nemá | |
|---|---|---|---|
| Město | $n_{11}$ | $n_{12}$ | $n_{1.}$ |
| Venkov | $n_{21}$ | $n_{22}$ | $n_{2.}$ |
| | $n_{.1}$ | $n_{.2}$ | $n$ |

**Šance** mít chřipku je definovaná jako

$$Odds = \frac{P(\text{má chřipku})}{P(\text{nemá chřipku})}$$

# Poměr šancí

**Poměr šancí** je šance ve městě dělená šancí na vesnici

$$OR = \frac{n_{11} n_{22}}{n_{12} n_{21}}$$

**Interpretace**: kolikrát je větší šance na chřipku ve městě než na venkově.

Porovnání města a venkova

- testované hypotézy
- $H_0 : OR = 1$ vs. $H_1 : OR \neq 1$
- testová statistika
- $Z = \frac{\ln(OR)}{\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}}$
- je možné testovat i Fisherovým exaktním testem

# Práce v SPSS

Vztah párových nominálních proměnných

- Analyze -> Descriptive Statistics -> Crosstabs
- zvolit proměnné
- v nabídce *Statistics* zvolte míry závislosti, které Vás zajímají (Kappa, McNemar, Risk)
- v nabídce *Cells* zvolte, co chcete vidět v kontingenční tabulce

Měření toho samého

```
CROSSTABS
/TABLES=aj BY nj
/FORMAT=AVALUE TABLES
/STATISTICS=CHISQ KAPPA MCNEMAR
/CELLS=COUNT
/COUNT ROUND CELL.
```

Měření rizika v $2 \times 2$ tabulkách

```
CROSSTABS
/TABLES=Ekn1 BY Matur.drive
/FORMAT=AVALUE TABLES
/STATISTICS=CHISQ RISK
/CELLS=COUNT
/COUNT ROUND CELL.
```

# Lineární regrese

Jednoduchá **lineární regrese**, popisuje, jakým způsobem závisí jedna číselná proměnná na jiné.

Předpokládáme lineární model ve tvaru

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

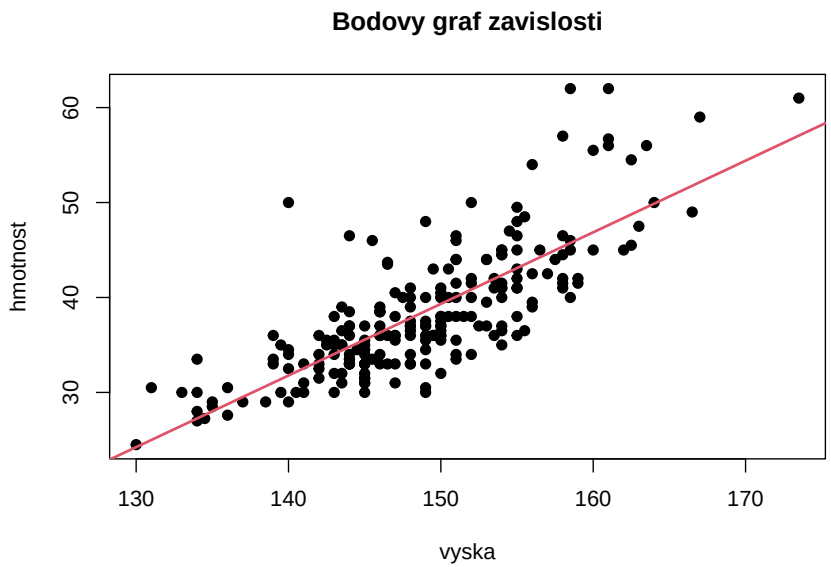
kde

- $Y_i$ jsou hodnoty závisle proměnné
- $X_i$ jsou hodnoty nezávisle proměnné
- $\beta_0$ je absolutní člen
- $\beta_1$ je lineární člen
- $e_i$ jsou náhodné chyby

# Lineární regrese

Graficky popisujeme pomocí bodového grafu, ale není jedno, která proměnná je na které ose

- na $x$-ovou osu se kreslí nezávisle proměnná
- na $y$-ovou osu se kreslí závisle proměnná



# Lineární regrese

Odhad probíhá **metodou nejmenších čtverců**, která minimalizuje součet druhých mocnin residuí

$$\min \sum_{i=1}^{n} R_i^2 = \min \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \min_{b_0, b_1} \sum_{i=1}^{n} (Y_i - (b_0 + b_1 X_i))^2$$

- $\hat{Y}_i$ jsou odhady/ predikce
- $b_0, b_1$ jsou pak odhady regresních koeficientů
- očekávaná hodnota pro $x_0$ je

$$\hat{Y}_0 = b_0 + b_1 x_0$$

**Koeficient determinace**

Procento variability vysvětlené modelem

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = \text{cor}(\text{X}, \text{Y})^2$$

**Test nezávislosti**. Testujeme

- $H_0$ : Proměnná $Y$ na proměnné $X$ lineárně nezávisí, $\beta_1 = 0$
- $H_1$ : Proměnná $Y$ na proměnné $X$ lineárně závisí, $\beta_1 \neq 0$

Test je založen na faktu, že $b_1/\text{se}(\text{b}_1) \sim \text{N}(0, 1)$, kde $b_1$ je odhad lineárního členu $\beta_1$ a $\text{se}(\text{b}_1)$ je jeho střední chyba.

# Lineární regrese

**Předpoklady** lineární regrese

- Mezi proměnnými je skutečně lineární vztah
  (je vidět z grafu)
- Residua jsou nezávislá
  (Durbin-Watsonova statistika by měla být mezi 1.5 a 2.5 statisti)
- Residua mají normální rozdělení
  (k dispozici jsou grafy normality)
- Stabilita rozptylu residuí
- V datech nejsou vlivná pozorování
  (Cookova vzdálenost by neměla přesáhnout 0.5)

# Práce v SPSS

Bodový graf - Scatterplot

- Graphs -> Scatter/Dot... -> Simple Scatter
- zvolit proměnné, případně název

```
GRAPH
/SCATTERPLOT(BIVAR)=matprij WITH celprij
/MISSING=LISTWISE.
```

- kliknout na graf a v editoru pridat *Add fit line at Total*

Model jednoduché lineární regrese s kontrolou předpokladů

- Analyze -> Regression -> Linear
- nastavit proměnné
- v nabídce *Statistics* zaškrtnout *Estimates, Confidence intervals, Model fit, Durbin-Watson*
- v nabídce *Plots* zvolte *Histogram, Normal probability plot* a přidat závislost **ZRESID* na **ZPRED*
- v nabídce *Save* zaškrtnout *Cook's*

```
REGRESSION
/MISSING LISTWISE
/STATISTICS COEFF OUTS CI(95) R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT celprij
/METHOD=ENTER matprij
/RESIDUALS DURBIN HISTOGRAM(ZRESID) NORMPROB(ZRESID)
/SAVE COOK.
```

# Lineární regrese

Závislost na více proměnných řeší **mnohonásobná lineární regrese**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + \epsilon$$

**Příklad.** *Zkoumáme, o kolik stoupne/klesne voda v řece v závislosti na srážkách, na teplotě, na typu půdy, na nasycenosti půdy, na nadmořské výšce, atd.*
Některé proměnné mají na závisle proměnnou větší vliv, jiné menší.

# Lineární regrese

Optimální model závislosti hledáme pomocí **krokové regrese**

- **Dopředná (forward)**, která začíná s modelem bez nezávisle proměnných a v každém kroku přidá jednu s největším, statisticky významným vlivem
- **Zpětná (backward)**, která začíná s úplným modelem a v každém kroku vynechá jednu proměnnou s nejmenším, statisticky nevýznamným vlivem
- **Kombinace obou předchozích (both sided)**, která začíná s prázdným modelem bez nezávisle proměnných a v každém kroku přidá jednu proměnnou s největším, statisticky významným vlivem a poté zkontroluje, zda nelze jinou proměnnou vynechat.

Cílem je získat model, kde budou nezávisle proměnné pouze se statisticky významným vlivem.

# Lineární regrese

Model s nezávislou kategorickou proměnnou.

Má-li kategorická proměnná $k$ kategorií, do modelu dáme $k-1$ pomocných *dummy* proměnných. Jedná se o 0-1 proměnné a platí

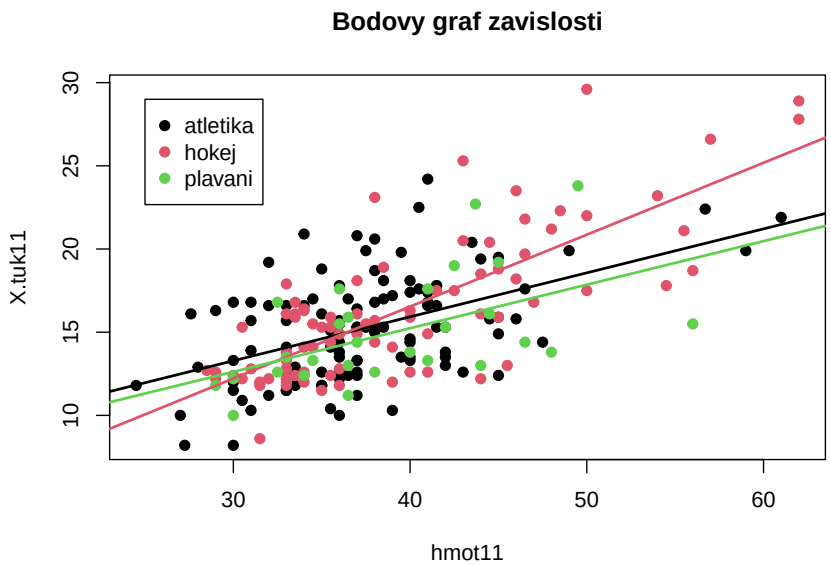
$$\begin{aligned} X_i &= 1 \ldots \text{nastala i-ta kategorie} \\ &= 0 \ldots \text{jinak} \end{aligned}$$

- poslední $k$-tá kategorie nastane, pokud všechny $X_1 = \ldots = X_{k-1} = 0$
- každá z těchto *dummy* proměnných má svou p-hodnotu
- významnost vlivu kategorické proměnné jako celku se řeší přes **tabulku analýzy rozptylu**

# Lineární regrese

Model s **interakcemi**.

- způsob, jímž se dvě nezávisle proměnné ovlivňují při jejich současném vlivu na proměnnou závislou
- Příklad. Do výběru bylo zařazeno 222 jedenáctiletých dětí a bylo u nich zjišťováno, jak závisí procento tuku v těle na jejich váze a na sportu, kterému se věnují.



- **Interakce** jsou vidět z grafu, do kterého se vykreslí závislost číselných proměnných zvlášť pro každou kategorii proměnné kategorické
- interakce se zobrazí jako různoběžné přímky
- rovnoběžné přímky znamenají vliv obou proměnných bez interakcí
- pokud rozdíl mezi skupinami není, přímky splývají

**Příklad.** *V modelu interakce existují - závislost procenta tuku na hmotnosti je jiná pro lední hokejisty a pro ostatní.*

# Zobecněná lineární regrese

Co dělat, když nejsou splněny předpoklady na rozdělení náhodné chyby modelu?

- Závisle proměnná je **spojitá** – normalizující transformace závisle proměnné (přirozený logaritmus, Box-Coxova transformace)
- Závisle proměnná je **dvouhodnotová** (0-1) – logistická regrese.
- Závisle proměnnou tvoří **počty** – Poissonova regrese.
- Závisle proměnná je **ordinální** – ordinální regrese.

# Logistická regrese

- závisle proměnná je dvouhodnotová
- modeluje se šance hodnoty 1 Odd $= \frac{\mathrm{P(X=1)}}{\mathrm{P(X=0)}}$
- odhaduje se model ve tvaru

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + \epsilon$$

- funkce závisle proměnné se jmenuje **logit link**
- interpretují se koeficienty $\exp \beta$: klikrát vzroste šance

# Logistická regrese

Pokud z odhadnutého modelu pak chceme zpětně získat vztah pro pravděpodobnost $\pi$, použijeme

$$\pi = \frac{\exp\{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k\}}{1 + \exp\{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k\}}$$

Koeficient $\beta_1$ se interpretuje následovně: "Šance vlastnost mít mi při nárůstu $X_1$ o 1 vzroste průměrně $\exp \beta_1$ krát při stejných hodnotách ostatních nezávisle proměnných."

# Logistická regrese

**Příklad.** *Uvažujme 150 cestujících na Titaniku. Ke každému cestujícímu máme uvedeno pohlaví, věk, třídu, ve které cestoval a informaci, zda se zachránil nebo ne. Následujícíc tabulky ukazují, jací cestující se zachránili, a jací se utopili.*

| | *Muž* | *Žena* |
|---|---|---|
| *Přežil* | *23* | *26* |
| *Nepřežil* | *89* | *12* |

| | *Dospělý* | *Dítě* |
|---|---|---|
| *Přežil* | *46* | *3* |
| *Nepřežil* | *97* | *4* |

| | *1. třída* | *2. třída* | *3. třída* | *Posádka* |
|---|---|---|---|---|
| *Přežil* | *11* | *10* | *11* | *17* |
| *Nepřežil* | *2* | *12* | *39* | *48* |

# Logistická regrese

**Příklad.** *Označme $\pi$ pravděpodobnost, že dotyčný přežil. Odhad modelu logistické regrese vyšel*

$$\ln\left(\frac{\pi}{1-\pi}\right) = 1.45 - 1.58(2.tr) - 3.04(3.tr) - 1.72(posadka) + \\ +2.32(zena) - 0.9(dospely)$$

*Jako vyznamné vyšly proměnné pohlaví ($p = 2.55 \times 10^{-6}$) a třída (p=0.0014) v níž dotyčný cestoval, konkrétně se významně liší třetí třída od první třídy a na hladině významnosti 0.1 i posádka od první třídy.*

*Ženy mají* $\exp(2.32) = 10.17$ *krát větší šanci na přežití než muži při ostatních parametrech neměnných.*
*Cestující v první třídě mají* $1/\exp(-3.04) = 20.9$ *krát větší šanci přežít než cestující ve třetí třídě při ostatních parametrech neměnných. Cestující v první třídě mají* $1/\exp(-1.72) = 5.6$ *krát větší šanci přežít než posádka při ostatních parametrech neměnných.*

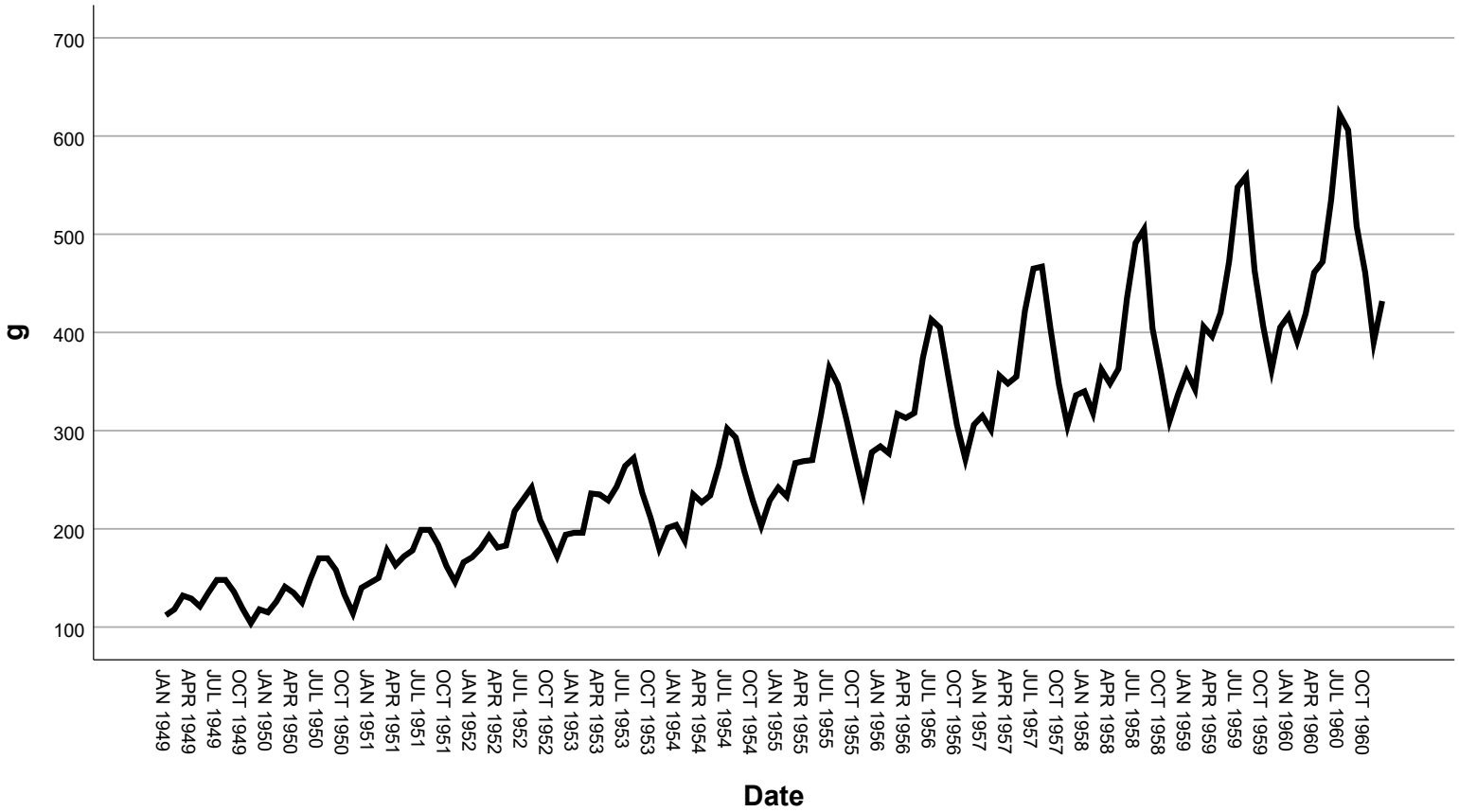
# Práce v SPSS

Regresní modely

- **Lineární regrese s více číselnými nezávisle proměnnými**
  Analyze -> Regression -> Linear
  umí krokovou regresi, ale dává až finální model
  umí testovat předpoklady
- **Lineární regrese s kategorickými nezávisle proměnnými**
  Analyze -> Generalized Linear Models -> Generalized Linear Models
  nastavit Scale response *Linear*
  nedělá krokovou regresi (musí se dělat ručně), netestuje předpoklady
  nebo je možné vytvořit dummy proměnné pro kategorickou veličinu ručně (Transform -> Create Dummz Variables, zvolit proměnnou a Root names - označení pro novou proměnnou) a použít klasickou lineární regresi výše
- **Poissonova regrese** se závisle proměnnou **počet**
  Analyze -> Generalized Linear Models -> Generalized Linear Models
  nastavit Scale response *Poisson*
- **Logistická regrese** s binární závisle proměnnou
  Analyze -> Regression -> Binary logistic
  umí krokovou regresi a výstupy dává po krocích
- **Ordinální regrese** s ordinální závisle proměnnou
  Analyze -> Regression -> Ordinal
  neumí krokovou regresi

# Časová řada je souhrn pozorování měřených v čase

- pozorování nebývají nezávislá
- znázorňuje se spojnicovým grafem v závislosti na čase



# Dekompozice

Časová řada se může skládat až **ze čtyř složek**

- $Tr_t$ časový trend
- $C_t$ cyklická složka
- $S_t$ sezónní složka
- $e_t$ náhodná složka

Dohromady se tyto složky skládají

- označme $Y_t$ časovou řadu v čase $t$
- **aditivně**

$$Y_t = Tr_t + C_t + S_t + e_t$$

  časová řada vykazuje stále stejně velké výchylky
- **multiplikativně**

$$Y_t = Tr_t * C_t * S_t * e_t$$

  výchylky časové řady v čase rostou

# Práce v SPSS

Nastavení časového rámce

- Data -> Define date and time ...
- zvolit vhodnou položku, např *Years, Quarters*

Zobrazení časové řady

- Analyze -> Forecasting -> Sequence Charts
- zvolit časovou řadu (je možné nechat si zobrazit i logaritmované hodnoty)

```
TSPLOT VARIABLES=Income
/NOLOG
/FORMAT NOFILL NOREFERENCE.
```

Dekompozice

- Analyze -> Forecasting -> Seasonal Decomposition
- zvolit časovou řadu a nastavit tzp *Additive* nebo *Multiplicative*

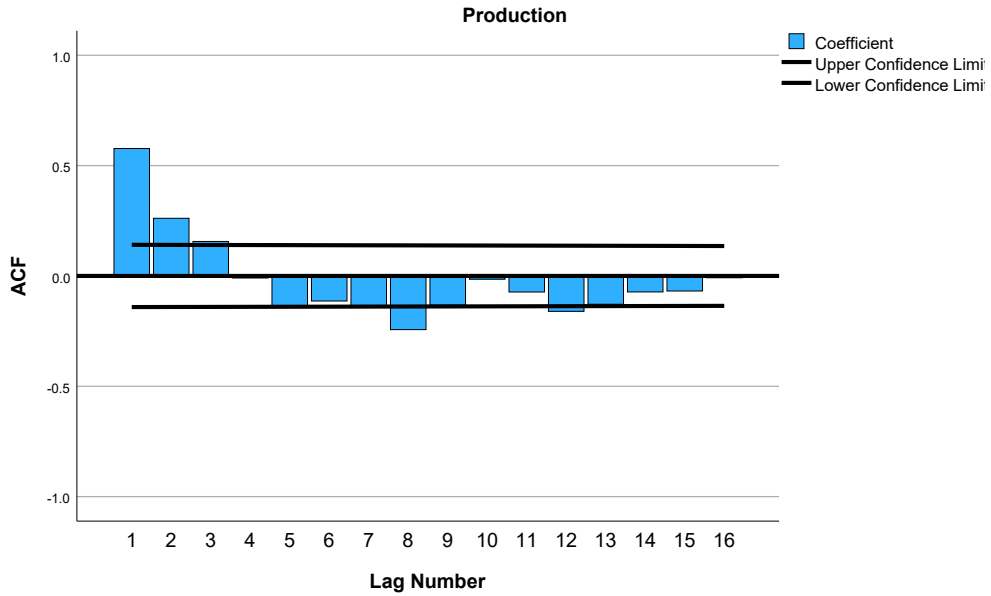
```
TSET PRINT=BRIEF NEWVAR=ALL MXNEWVAR=4.
SEASON
/VARIABLES=Income
/MODEL=ADDITIVE
/MA=EQUAL.
```

- v databázi se vytvoří 4 nové proměnné:
- ERR_1: náhodná chyba
- SAS_1: řada s odstraněným sezónním trendem
- SAF_1: sezónnost
- STC_1: časový trend a sezónnost

# Autokorelační funkce

Ukazuje míru závislosti jednotlivých pozorování mezi sebou

$$acf(k) = \text{cor}(Y_t, Y_{t+k})$$




- existuje i parciální autokorelační funkce, která říká, jak spolu souvisí $Y_t$ a $Y_{t-k}$ při odstínění vlivu pozorování mezi $(Y_{t-1}, \ldots, Y_{t-k+1})$
- dohromady pomohou odhadnout proces generující časovou řadu

# ARIMA model pro časovou řadu

- hledá proces generující časovou řadu v závislosti na minulých hodnotácj a minulých chybách
- základem je tzv. ARMA model
- ARMA model je určen pro **stacionární** časové řady: stabilní řada bez sezónních výkyvů
- např. ARMA(1,1)

$$Y_t = \phi_1 Y_{t-1} + \theta_1 e_{t-1} + e_t$$

- existuje i rozšíření pro řady s trendem a sezónností
- nejprve se řada tzv. stacionarizuje pomocí diferencí

# Práce v SPSS

Autokorelační funkce

- Analyze -> Forecasting -> Autocorrelations ...
- zvolit časovou řadu

```
ACF VARIABLES=Production
/NOLOG
/MXAUTO 16
/SERROR=IND
/PACF.
```

ARIMA model

- Analyze -> Forecasting -> Create Traditional Models
- zvolit časovou řadu do pole *Dependent*, v možnosti *Method* musí být *Expert Modeler* a přes *Criteria* nastavte *ARIMA models only*
- v záložce *Statistics* nastavte *Parameter estimates* a *Residual autocorrelation function*

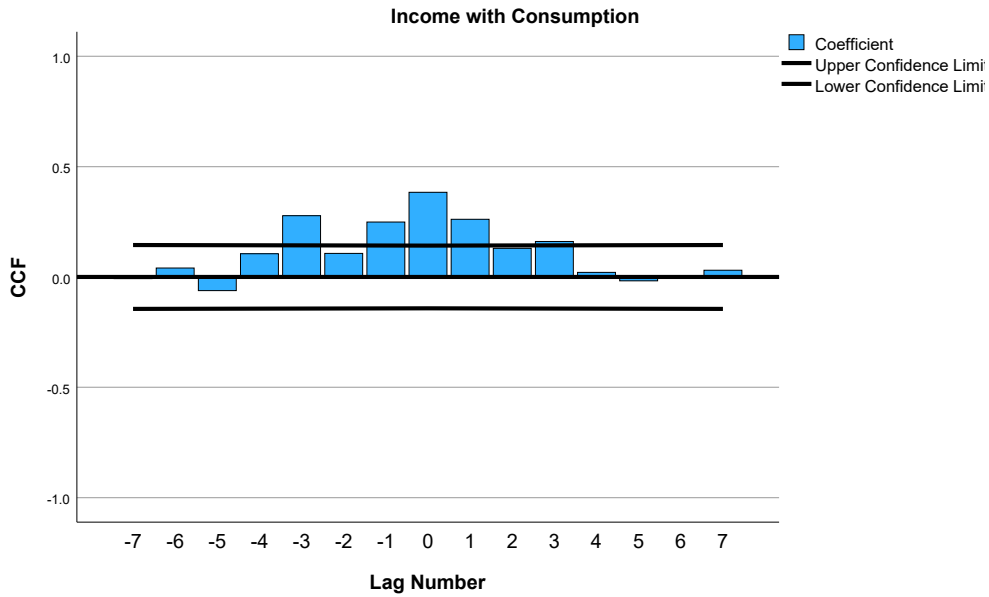
```
TSMODEL
/MODELSUMMARY  PRINT=[MODELFIT]
/MODELSTATISTICS  DISPLAY=YES MODELFIT=[ SRSQUARE]
/MODELDETAILS  PRINT=[ PARAMETERS RESIDACF]
/SERIESPLOT OBSERVED FORECAST
/OUTPUTFILTER DISPLAY=ALLMODELS
/AUXILIARY  CILEVEL=95 MAXACFLAGS=24
/MISSING USERMISSING=EXCLUDE
/MODEL DEPENDENT=Income
PREFIX='Model'
/EXPERTMODELER TYPE=[ARIMA] TRYSEASONAL=YES
/AUTOOUTLIER  DETECT=OFF.
```

# Závislost mezi časovými řadami – kroskorelace

Vztah mezi řadami v různých časových okamžicích se měří pomocí kroskorelační funkce

$$ccf(k) = \text{cor}(Y_t, Z_{t+k})$$

- kde $Y$ a $Z$ jsou dvě časové řady
- $k$ je zpoždění, s jakým se závislost projeví




- hledám zpoždění, kde je korelace maximální, a s tímto zpožděním pak modeluji závislost

# Práce v SPSS

Kroskorelační funkce

- Analyze -> Forecasting -> Cross-Correlation
- zvolit dvě časové řady

```
CCF
/VARIABLES=Income Consumption
/NOLOG  /MXCROSS 7.
```

# Závislost mezi časovými řadami – ARIMAX model

Hledáme model závislosti

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \varepsilon_t$$

kde náhodná chyba $\varepsilon_t$ se řídí ARIMA modelem.

- výsledkem je zjištění podstatných závislostí a nalezení predikčního modelu
- je možné modelovat i závislost posunutou v čase

# Práce v SPSS

ARIMAX model

- Analyze -> Forecasting -> Create Traditional Models
- zvolit časovou řadu do pole *Dependent*, do pole *Independent* nastavte nezávisle proměnné
- do pole *Method* nastavte *ARIMA*, přes tlačítko *Criteria* nastavte členy do ARIMA modelu a charakteristiky nezávisle proměnných
- v záložce *Transfer Function* dejte ke každé nezávisle proměnné 3 nuly a do pole *Lags* zpoždění závislosti v čase
- v záložce *Statistics* nastavte *Parameter estimates* a *Residual autocorrelation function*

```
TSMODEL
/MODELSUMMARY  PRINT=[MODELFIT]
/MODELSTATISTICS  DISPLAY=YES MODELFIT=[ SRSQUARE]
/MODELDETAILS  PRINT=[ PARAMETERS RESIDACF]
/SERIESPLOT OBSERVED FORECAST
/OUTPUTFILTER DISPLAY=ALLMODELS
/AUXILIARY  CILEVEL=95 MAXACFLAGS=24
/MISSING USERMISSING=EXCLUDE
/MODEL DEPENDENT=Income INDEPENDENT=Consumption Production Unemployment
PREFIX='Model'
/ARIMA AR=[1]  DIFF=0  MA=[1]  ARSEASONAL=[0]  DIFFSEASONAL=0  MASEASONAL=[0]
TRANSFORM=NONE  CONSTANT=YES
/TRANSFERFUNCTION VARIABLES=Production  NUM=[0] DENOM=[0] DIFF=0 NUMSEASONAL=[0]
DENOMSEASONAL=[0] DIFFSEASONAL=0 DELAY=0   TRANSFORM=NONE
/TRANSFERFUNCTION VARIABLES=Unemployment  NUM=[0] DENOM=[0] DIFF=0 NUMSEASONAL=[0]
DENOMSEASONAL=[0] DIFFSEASONAL=0 DELAY=0   TRANSFORM=NONE
/TRANSFERFUNCTION VARIABLES=Consumption  NUM=[0] DENOM=[0] DIFF=0 NUMSEASONAL=[0]
DENOMSEASONAL=[0] DIFFSEASONAL=0 DELAY=0   TRANSFORM=NONE
/AUTOOUTLIER DETECT=OFF.
```

# Panelová data

Data, kde máme několik jednotek měřených v čase

- je potřeba předpokládat, že pozorování v rámci jedné jednotky jsou závislé
- modelují se přes **smíšené modely** (Mixed models)
- jednotky vstupují do modelu jako **náhodné efekty**: ovlivňují variabilitu, každá jednotka může mít svůj vlastní absolutní člen
- ostatní proměnné vstupují do modelu jako **pevné efekty** a na nich se modeluje závislost

# Práce v SPSS

ARIMAX model

- Analyze -> Mixed Models -> Linear
- nastavit jednotky jako *Subjects* a časový údaj jako *Repeated*
- dále se nastavuje závislá a nezávislé proměnné
- pod tlačítkem *Random* nastavte jednotku jako *Combinations* a zaškrtněte *Include Intercepts*
- pod tlačítke *Fixed* nastavte všechny nezávisle proměnné do pole *Model*
- pod tlačítkem *Statistics* zaškrtněte *Parameter estimates for fixed effects*

```
MIXED INDOUTPT BY IND WITH EMP CAP
/CRITERIA=DFMETHOD(SATTERTHWAITE) CIN(95) MXITER(100) MXSTEP(10) SCORING(1)
SINGULAR(0.000000000001) HCONVERGE(0.00000001, RELATIVE) LCONVERGE(0, ABSOLUTE) PCONVERGE(0,
ABSOLUTE)
/FIXED=IND EMP CAP | SSTYPE(3)
/METHOD=REML
/PRINT=SOLUTION TESTCOV
/RANDOM=INTERCEPT | SUBJECT(unit) COVTYPE(VC)
/REPEATED=YEAR | SUBJECT(unit) COVTYPE(DIAG).
```