1. Úvod

Statistika

- statistické údaje, informace o tzv. hromadných jevech
- praktická činnost, kterou získáváme údaje a zpracováváme je
- vědní disciplína

Základní pojmy

- statistický soubor
- statistická jednotka
- statistický znak
- statistické proměnné

Statistický soubor

- skupina prvků/jednotek, od které shromažďujeme data
- množina prvků/jednotek, které mají alespoň jednu společnou vlastnost (musí mít alespoň jednu společnou vlastnost, abychom byli schopni soubor nadefinovat)
- množina podniků, automobilů, zvířat
- základní soubor = obsahuje všechny existující statistické jednotky
    - populace
    - všechny společnosti v obchodním rejstříku
    - všichni obyvatelé Prahy
- výběrový soubor = obsahuje pouze část statistických jednotek, podmnožina základního souboru

Statistická jednotka

- konkrétní prvek statistického souboru, který má různé vlastnosti

Statistický znak

- pojmenování vlastnosti statistické jednotky
    - identifikační znak
        - společná vlastnost pro všechny jednotky souboru
    - statistická proměnná
        - nabývá různých obměn u jednotlivých jednotek v souboru

Statistické proměnné

- slovní (kvalitativní) = taková proměnná, která je reprezentována slovem
    - nominální (názvové)
        - jednotlivé varianty proměnné nelze seřadit podle pořadí
        - barvy, pohlaví, rodinný stav
    - ordinální (pořadové)
        - jednotlivé varianty proměnné lze seřadit od nejnižší po nejvyšší, ale není možné říct, o kolik se liší
        - rok narození, ukončené vzdělání, známky
- číselné (kvantitativní, numerické) = taková proměnná, která je reprezentována číslem (ve smyslu objem, počet, měřitelnost)
    - spojité
        - nabývají hodnot z konečného nebo nekonečného intervalu

        - výška, váha, věk, měsíční příjem
    - nespojité (diskrétní)
        - nabývají hodnot malého počtu jednoznačně izolovaných hodnot
        - počet let strávených na univerzitě, počet dětí, počet automobilů

2. Popisná statistika

- popisuje statistický soubor z různých hledisek a pomocí různých nástrojů

Tabulka rozdělení četností

- četnost = udává, kolikrát se něco vyskytne v daném souboru nebo kolik procent
    - absolutní četnost n_i = říká, kolikrát se v daném souboru vyskytuje příslušná varianta sledovaného znaku

      $$\sum_{i=1}^{k} n_i = n$$

        - k = počet variant sledovaného znaku v souboru
        - n_i = příslušná absolutní četnost
        - n = počet hodnot neboli rozsah souboru
    - relativní četnost p_i = říká, kolik procent z celku tvoří ta příslušná varianta sledovaného znaku

      $$p_i = \frac{n_i}{n}$$

        - p_i = příslušná relativní četnost
        - n_i = příslušná absolutní četnost
        - n = počet hodnot neboli rozsah souboru
    - kumulativní četnost = postupně načítaná četnost jednotlivých (vzestupně uspořádaných) hodnot statistického znaku ve statistickém souboru
        - kumulativní absolutní četnost N_i

          $$N_i = \sum_{j=1}^{i} n_j$$

        - kumulativní relativní četnost P_i

          $$P_i = \sum_{j=1}^{i} p_j$$

- tabulka rozdělení četností pro nespojitý znak
    - x_i = sledovaný znak = počet automobilů firmy

x_i	n_i	p_i	N_i	P_i
0	5	0,05	5	0,05
1	15	0,15	20	0,20
2	56	0,56	76	0,76
3	14	0,14	90	0,90
4	10	0,10	100	1,00
součet	**100**	**1,00**	-	-

- tabulka rozdělení četností pro spojitý znak
    - pravidlo pro stanovení počtu intervalů $k = \sqrt{n}$
    - pravidlo pro stanovení šíře intervalů
      $$\frac{x_{max} - x_{min}}{k}$$
        - intervaly musí být jasně stanovené – např. příjmy 20 000-25 000, 50 001-30 000; ne 20 000-25 000, 25 000 – 30 000, protože každá hodnota musí být jednoznačně přiřaditelná
        - šíře intervalů by měly být stejné

    - x_i = sledovaný znak – měsíční obrat firmy v milionech Kč

x_i	n_i	p_i	N_i	P_i
0-20	5	0,05	5	0,05
21 – 40	15	0,15	20	0,20
41 – 60	56	0,56	76	0,76
61 – 80	14	0,14	90	0,90
81 – 100	10	0,10	100	1,00
součet	**100**	**1,00**	-	-

Číselné charakteristiky

- charakteristiky polohy – popisují soubor z hlediska úrovně neboli velikosti hodnot
    - střední hodnoty
    - kvantily
- charakteristiky variability – popisují soubor z hlediska proměnlivosti hodnot souboru
    - absolutní variability
    - relativní variability
- použití charakteristik prostých a vážených – v závislosti na způsobu zadání vstupních údajů
- **prosté charakteristiky používáme v případě, že vstupní údaje nejsou uspořádané do tabulky četností, vážené pokud jsou vstupní údaje uspořádané do tabulky četností**
- charakteristiky polohy
    - střední hodnoty
        - aritmetický průměr
            - prostý aritmetický průměr = součet jednotlivých hodnot dělený jejich počtem
                - používáme, pokud nemáme k dispozici tabulku rozdělení četností
                $$\bar{x} = \sum_{i=1}^{n} x_i \,/n$$
            - vážený = používáme, pokud máme k dispozici tabulku rozdělení četností
                - pomocí absolutních četností
                $$\bar{x} = \sum_{i=1}^{k} x_i n_i / \sum_{i=1}^{k} n_i$$

- pomocí relativních četností

$$\bar{x} = \sum_{i=1}^{k} x_i p_i$$

- průměrná hodnota

$$\bar{x} = \sum_{i=1}^{k} x_i^{STŘ} n_i / \sum_{i=1}^{k} n_i$$

- vlastnosti AP
  - $\overline{x + k} = \bar{x} - k$
    - k = nějaká libovolná konstanta
    - jestliže ke každé hodnotě v souboru přičteme stejnou nenulovou konstantu, tak se o tuto konstantu zvýší také aritmetický průměr
  - $\bar{k} = k$
    - aritmetický průměr konstanty je konstanta
  - $\overline{x \cdot k} = \bar{x} \cdot k$
    - jestliže všechny hodnoty v souboru vynásobíme stejnou nenulovou konstantou k, tak se aritmetický průměr taktéž zvýší k-násobně
  - $$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$
    - součet odchylek hodnot od svého průměru je vždycky nulový
  - $\overline{x \pm y} = \bar{x} \pm \bar{y}$
    - aritmetický průměr součtu nebo rozdílu dvou znaků je roven součtu nebo rozdílu průměrů těchto znaků
- použití aritmetického průměru
  - nejčastěji používaná charakteristika polohy
  - používá se v případě, kdy má smysl hodnoty proměnné sčítat

- harmonický průměr
  - prostý harmonický průměr
    - používáme, pokud nemáme k dispozici tabulku rozdělení četností

$$\overline{x_H} = n / \sum_{i=1}^{n} \frac{1}{x_i}$$

  - vážený harmonický průměr
    - používáme, pokud máme k dispozici tabulku rozdělení četností

$$\overline{x_H} = \sum_{i=1}^{k} n_i \, / \sum_{i=1}^{k} \frac{n_i}{x_i}$$

$$\overline{x_H} = 1 / \sum_{i=1}^{k} \frac{p_i}{x_i}$$

  - použití harmonického průměru
    - harmonický průměr se používá málo, typicky pro výpočet průměrné rychlosti (tam jedeme 50 zpátky 70, průměr není 60, protože tam jedeme další čas – použijeme prostý

harmonický průměr, protože jedeme stejnou trasu – kdyby byla jinak dlouhá – vážený harmonický průměr)

- geometrický průměr
  - prostý geometrický průměr
    - používáme, pokud nemáme k dispozici tabulku rozdělení četností

      $\overline{x_G} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$
  - vážený geometrický průměr
    - mizivé použití – používáme, pokud máme k dispozici tabulku rozdělení četností

      $\overline{x_G} = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \cdot \ldots \cdot x_n^{n_k}}$
  - použití geometrického průměru
    - používá se při výpočtu průměrného koeficientu růstu – což je jiné označení pro průměrné tempo růstu (HDP, inflace, zisku, nezaměstnanosti…); koeficient růstu = podíl dvou hodnot
- modus $\hat{x}$
  - nejčastěji se vyskytující obměna sledovaného znaku v souboru vzhledem ke svému okolí
  - v případě intervalového rozdělení – tabulka četností – modální interval
  - nepočítá se, pouze se hledá

- kvantily
  - p-procentní kvantil $\widetilde{x_p}$
    - číslo, které rozdělí uspořádaný soubor podle velikosti na dvě části, a to na p % hodnot, které jsou menší nebo rovny hodnotě daného kvantilu a 100-p % hodnot, které jsou větší nebo rovny hodnotě daného kvantilu
      - seřadíme hodnoty souboru podle velikosti
      - podle vzorce určíme pořadí jednotky

        $n\frac{p}{100} < z_p < n\frac{p}{100} + 1$
      - v případě, že mezi těmi dvěma mezemi (= pořadí jednotek) neleží žádné celé číslo, tak ty dvě čísla zprůměrujeme (ne ty dvě pořadí, ale čísla, která jim náleží)
  - pojmenované kvantily
    - kvartily (25%, 50%, 75% kvantily)
    - decily (10%, 20%,…, 90% kvantily)
    - percentily (1%, 2%,…, 99% kvantily)
    - názvy vznikají podle významných částí, které oddělují
    - vždycky rozdělují soubor na DVĚ části – čtvrtina a tři čtvrtiny, půl a půl atd.
  - medián = 50% kvantil = prostřední hodnota uspořádaného souboru hodnot podle velikosti

- charakteristiky variability
  - míry absolutní variability
    - rozpětí
      - variační rozpětí
        - nevýhodou je, že se určí pouze na základě dvou hodnot v souboru – nemusí být úplně objektivní

    - výhodou je triviální výpočet a vychází ve stejných měrných jednotkách

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

- kvartilové rozpětí
    - odečteme nejvyšší a nejnižší kvartil

$$R_Q = \widetilde{x_{75}} - \widetilde{x_{25}}$$

- decilové
    - odečteme nejvyšší a nejnižší decil

$$R_D = \widetilde{x_{90}} - \widetilde{x_{10}}$$

- percentilové
    - odečteme nejvyšší a nejnižší percentil

$$R_P = \widetilde{x_{99}} - \widetilde{x_1}$$

- průměrná absolutní odchylka
    - prostý tvar

$$d_x = \sum_{i=1}^{n} |x_i - \bar{x}| \, / n$$

    - vážený tvar

$$d_x = \sum_{i=1}^{k} |x_i - \bar{x}| \cdot n / \sum_{i=1}^{k} n_i$$

$$d_x = \sum_{i=1}^{k} |x_i - \bar{x}| \cdot p_i$$

- rozptyl = aritmetický průměr čtverců odchylek hodnot od svého průměru (průměrná čtvercová odchylka)
    - prostý tvar

$$s_x^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 \, / n$$

    - vážený tvar

$$s_x^2 = \sum_{i=1}^{k} (x_i - \bar{x})^2 \, n_i / \sum_{i=1}^{k} n_i$$

$$s_x^2 = \sum_{i=1}^{k} (x_i - \bar{x})^2 \, p_i$$

    - při počítání rozptylu
        - vypočítáme průměr $\bar{x}$
        - zjistíme odchylku $x_i - \bar{x}$
        - umocníme na druhou
        - vynásobíme četnostmi (kdyby to byl prostý rozptyl, tak tento krok vypadne)
        - sečteme $\sum$
        - vydělím celkovým počtem $/ \sum n_i$
    - vlastnosti rozptylu
        - $s_{x+k}^2 = s_x^2$
            - jestliže ke každé hodnotě v souboru přičteme stejnou nenulovou konstantu k, rozptyl se nezmění
        - $s_k^2 = 0$
            - rozptyl konstanty je nulový

- $s^2_{x\cdot k} = k^2 \cdot s^2_x$
  - jestliže každou hodnotu v souboru vynásobíme stejnou nenulovou konstantou k, rozptyl se zvýší k^2 násobně
- $s^2_{x\pm y} = s^2_x + s^2_y$
  - rozptyl součtu nebo rozdílu dvou nezávislých znaků je roven vždycky a pouze **součtu** rozptylů

- věta o rozkladu rozptylu
  - jestliže máme soubor rozdělen do k podskupin, u kterých známe průměr a rozptyl (a samozřejmě četnosti), anebo jsme schopni si je jednoduše dopočítat, pak rozptyl celého souboru spočítáme jako součet dvou složek – meziskupinové a vnitroskupinové variability

$$s_x^2 = \overline{s^2} + s_{\bar{x}}^2 = \sum_{i=1}^{k} s_i^2 n_i / \sum_{i=1}^{k} n_i + \sum_{i=1}^{k} (\overline{x_i} - \bar{x})^2 \, n_i / \sum_{i=1}^{k} n_i$$

$$s_x^2 = \sum_{i=1}^{k} s_i^2 p_i + \sum_{i=1}^{k} (\overline{x_i} - \bar{x})^2 \, p_i$$

  - výpočtový tvar rozptylu
    - jiný způsob výpočtu, ale vyjde stejný výsledek

$$s_x^2 = \overline{x^2} - \bar{x}^2 = \sum_{i=1}^{n} x_i^2 / n - \left(\frac{1}{n} \cdot \sum_{i=1}^{n} x_i \right)^2$$

$$s_x^2 = \overline{x^2} - \bar{x}^2 = \sum_{i=1}^{k} x_i^2 \, n_i / \sum_{i=1}^{k} n_i - \left(\sum_{i=1}^{k} x_i \, n_i / \sum_{i=1}^{k} n_i \right)^2$$

$$s_x^2 = \overline{x^2} - \bar{x}^2 = \sum_{i=1}^{k} x_i^2 \, p_i - \left(\sum_{i=1}^{k} x_i \, p_i \right)^2$$

- směrodatná odchylka (kladná hodnota) = druhá odmocnina z rozptylu

$$s_x = \sqrt{s_x^2}$$

  - používá se k výpočtu míry rizika v bankovním sektoru

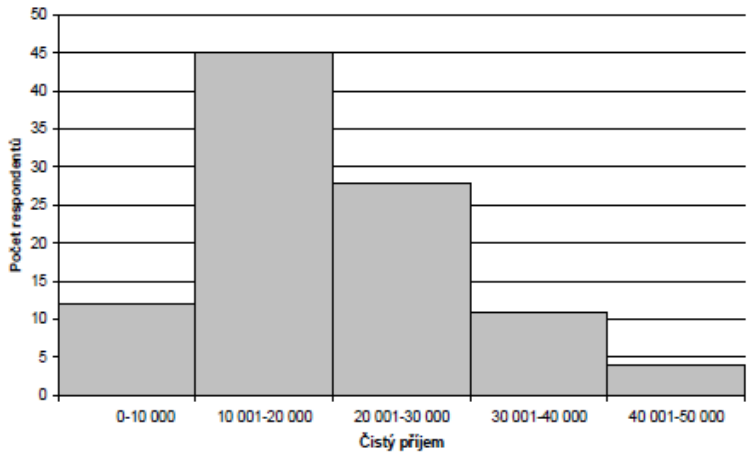
- míry relativní variability
  - variační koeficient

$$v_x = s_x / \bar{x}$$

    - podíl směrodatné odchylky a aritmetického průměru
    - po vynásobení 100 říká, jaká je variabilita v %
    - slouží ke srovnávání variability různých souborů

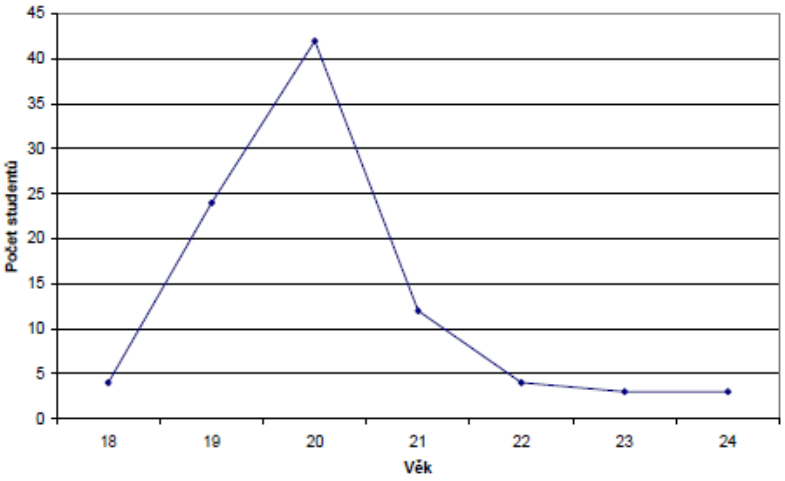
Grafy

- histogram četností



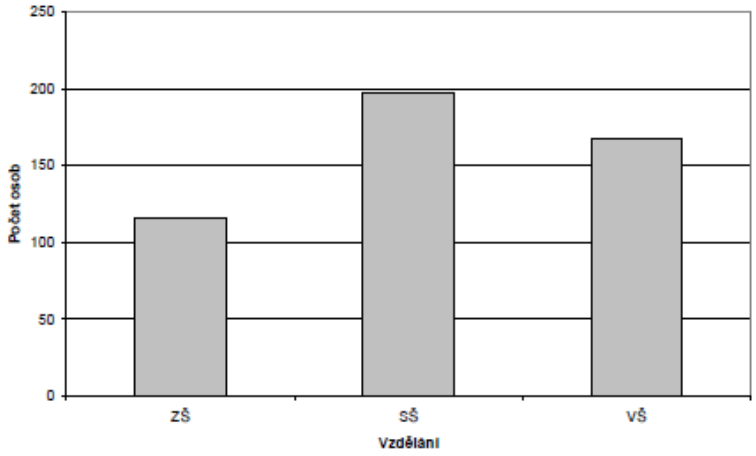
  - používá se pro kvantitativní spojité proměnné
  - šířka sloupce znamená šířku intervalu v tabulce četností

- polygon četností

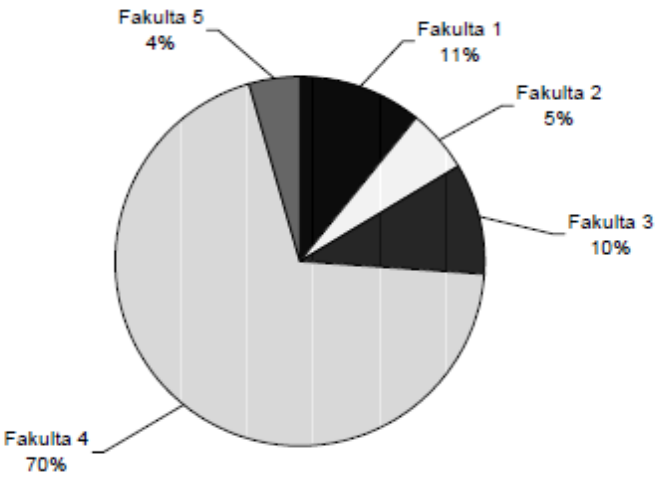


  - spojení jednotlivých středů základen v histogramu

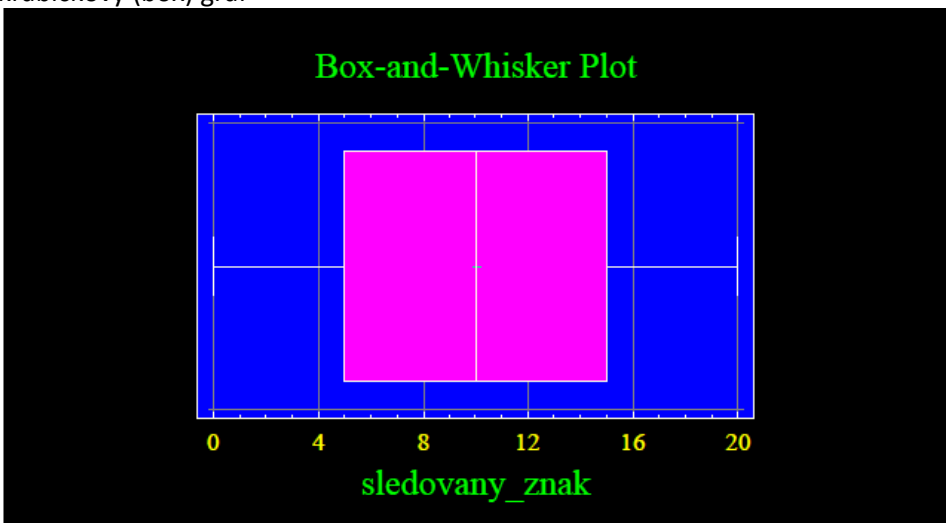
- sloupcový graf
  - není vhodný pro spojitou proměnnou



- koláčový graf



- krabičkový (box) graf



  - říká, v jakých mezích se pohybuje prostředních 50% respondentů
  - obsahuje:
    - minimální hodnotu souboru
      - vlevo nebo dole
    - maximální hodnotu souboru
      - vpravo nebo nahoře
    - spodní kvartil
      - levá nebo dolní hrana
    - horní kvartil
      - pravá nebo horní hrana
    - kvartilové rozpětí
      - růžová plocha
    - aritmetický průměr
      - +
    - extrémní a odlehlé hodnoty (pokud v souboru jsou)

Příklad číslo 1

Na základě hodnot v prvním sloupci vypočítejte zástupce všech skupin popisných charakteristik a výsledné hodnoty interpretujte. Zároveň porovnejte oba soubory hodnot.

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$x_i^{(uspoř)}$
1	-2	4	1
2	-1	1	2
4	1	1	2
2	-1	1	4
6	3	9	6
15	-	**16**	**15**

n	5	
$\bar{x}$	3	
s^2_x	3,20	
s_x	1,79	
V_x	0,60	
z_{50}	2,5	3,5
x_{50}	2	
R	5	

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$x_i^{(uspoř)}$
-5	-8	64	-5
2	-1	1	2
4	1	1	2
2	-1	1	4
12	9	81	12
15	-	**148**	**15**

n	5	
$\bar{x}$	3	
s^2_x	29,60	
s_x	5,44	
V_x	1,81	
z_{50}	2,5	3,5
x_{50}	2	
R	17	

- použijeme prosté charakteristiky
- vysoká variabilita snižuje vypovídací hodnotu aritmetického průměru
- u mezd – raději používat medián

Příklad číslo 2

Na základě hodnot zadaných v prvním a druhém sloupci vypočítejte zástupce charakteristik polohy a variability.

x_i	n_i	$x_i * n_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 n_i$	N_i
0	5	0	-2,09	4,37	21,84	5
1	15	15	-1,09	1,19	17,82	20
2	56	112	-0,09	0,01	0,45	76
3	14	42	0,91	0,83	11,59	90
4	10	40	1,91	3,65	36,48	100
Σ	**100**	**209**	-	-	**88,19**	-

n	100	
$\bar{x}$	2,09	
s^2_x	0,88	
s_x	0,94	
V_x	0,45	
z_{50}	50	51
x_{50}	2	
R	4	

- použijeme vážené charakteristiky

Příklad číslo 3

Řidič jel z města A do města B rychlostí 90 km/h a z města B do města C rychlostí 110 km/h. Vypočítejte průměrnou rychlost řidiče na celé trase, tj. z města A do města C, pokud víme, že vzdálenost měst A-B a B-C je stejná.

- použijeme harmonický průměr

Příklad číslo 4

Na základě tabulky, která obsahuje tempa růstu ceny výrobku A (koeficient růstu je definován jako podíl dvou sousedních cen, kde v čitateli je hodnota daného roku a ve jmenovateli hodnota předchozího roku), stanovte průměrné tempo růstu cen za celé sledované období.

- použijeme geometrický průměr

Příklad číslo 5

Hodnota HDP (měřeného v mld. Kč, v běžných cenách)vzrostla od roku 1999 do roku 2007 z 1466,5 na 3231,6. Určete, jaký byl průměrný relativní meziroční přírůstek hodnoty HDP.

- použijeme geometrický průměr
- Výsledek = 1,104 (osmá odmocnina z 3231,6/1466,5)

3. Teorie pravděpodobnosti

Základní pojmy

- náhoda = souhrn vlivů, které způsobují, že nejsme oprávněni s jistotou jednoznačně předpovědět výsledek náhodného pokusu (= vlivy, které úplně nebo částečně, které nejsme schopni kvantifikovat)
- náhodný pokus = takový pokus, jehož výsledek je ovlivněn náhodou (loterie, hod kostkou…)
- náhodné jevy = představují jednotlivé výsledky náhodného pokusu, značení: A, B, C…
    - elementární jev = nejjednodušší výsledek náhodného pokusu, který se dále nerozkládá
    - prostor elementárních jevů = množina všech elementárních jevů, značí se E
- Vennovy diagramy = diagramy, které zkoumají vztahy mezi náhodnými jevy
    - jev A je součástí jevu B: $A \subset B$
    - rovnocenné jevy A a B: $A = B$
    - průnik jevů A a B: $C = A \cap B$
        - jev C je definován jako průnik těch dvou jevů
    - neslučitelné jevy A a B: $A \cap B = \emptyset$
        - jevy A a B jsou neslučitelné tehdy, když je jejich průnikem prázdná množina
        - nemohou nastat zároveň
    - sjednocení jevů A a B: $D = A \cup B$
        - jev D představuje sjednocení jevů A a B
        - nastane jev A nebo jev B
    - jev opačný k jevu A
        - spočívá v nenastoupení jevů A
    - jevy A a B jsou nezávislé
        - jevy A a B jsou nezávislé v případě, že výskyt jednoho z jevů neovlivňuje výskyt toho druhého
- pravděpodobnost = číslo, které vyjadřuje míru možnosti (šanci) výskytu daného jevu v náhodném pokusu
    - klasická definice: je založena na předpokladu stejné možnosti výskytu jevu v každém z pokusů
        - uvažujeme n nezávislých pokusů, ve kterých se jev A vyskytl právě m-krát
        - pravděpodobnost jevu A se vypočítá podle vzorce

$$\mathrm{P(A)} = \frac{\mathrm{m}}{\mathrm{n}}$$

        - m = počet jevů příznivých
        - n = počet jevů možných (počet pokusů)
    - statistická definice: používá se v případě, že není splněn předpoklad klasické definice
        - se zvyšujícím se počtem opakování pokusů, ve kterých sledujeme výskyt jevu A, dochází ke stabilizaci relativní četnosti a tu nazveme jako pravděpodobnost jevu A
          $\mathrm{P(A)} \sim \frac{\mathrm{m}}{\mathrm{n}}$
    - subjektivní pravděpodobnost: představuje subjektivní míru ohodnocení šance daného jevu hodnotitelem
- pro pravděpodobnost náhodného jevu A platí:
    - obor možných hodnot
        - $0 \leq P(A) \leq 1$
        - jevy, které nabývají krajních pravděpodobností
            - jev nemožný (P = 0)
            - jev jistý (P = 1)
    - pravidla pro sčítání pravděpodobnosti
      $P(A \cup B) = P(A) + P(B)$ (pro neslučitelné jevy)
      $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ (pro slučitelné jevy)
        - kdy nastane jev A nebo jev B
    - pravidla pro násobení pravděpodobností
      $P(A \cap B) = P(A) \cdot P(B)$ (pro nezávislé jevy)
      $P(A \cap B) = P(A) \cdot P(B \setminus A) = P(B) \cdot P(A \setminus B)$ (pro závislé jevy)
        - kdy nastane jev A a současně jev B
    - pravděpodobnost jevu opačného
      $P(\bar{A}) = 1 - P(A)$
- úplná pravděpodobnost
    - nechť jevy B_i (i = 1, ..., s) tvoří tzv. úplný systém neslučitelných jevů
      $\sum_{i=1}^{s} P(B_i) = 1$
    - nechť jev A může nastat pouze ve spojení s některým z těchto jevů B_i
      $P(A) = \sum_{i=1}^{s} P(B_i) \cdot P(A \setminus B_i)$

Příklad číslo 1

V měšci je celkem 10 karet. 2 listy, 1 kára a 7 srdcí. Jaká je pravděpodobnost, že mezi třemi po sobě taženými kartami s vracením budou:

a) list, kára, srdce,
b) list, list, srdce,
c) srdce, srdce, srdce?

Příklad číslo 2

V měšci je celkem 10 karet. 2 listy, 1 kára a 7 srdcí. Jaká je pravděpodobnost, že mezi třemi po sobě taženými kartami bez vracení budou:

a) list, kára, srdce,
b) list, list, srdce,
c) srdce, srdce, srdce?

Příklad číslo 3

Ze 100 studentů složí zkoušku 95 studentů. Jaká je pravděpodobnost, že mezi třemi studenty:

a) všichni zkoušku složí,
b) právě jeden zkoušku nesloží,
c) ani jeden zkoušku nesloží,
d) alespoň jeden zkoušku složí?

Náhodná veličina

= veličina, jejíž hodnoty jsou jednoznačně určeny výsledky náhodných pokusů

- označení náhodné veličiny: X,Y,Z (pak indexy)
- označení hodnot náhodné veličiny: x,y,z
- druhy náhodné veličiny
    - nespojitá (diskrétní) : nabývá konečného nebo spočetně konečného počtu hodnot
    - spojitá: nabývá libovolných hodnot z konečného intervalu nebo celého oboru reálných čísel
        - příjem domácností, HDP ...
- zákon rozdělení náhodné veličiny: pravidlo, které každé hodnotě nebo množině hodnot z určitého intervalu přiřazuje pravděpodobnost, že náhodná veličina nabude právě této hodnoty nebo množiny hodnot

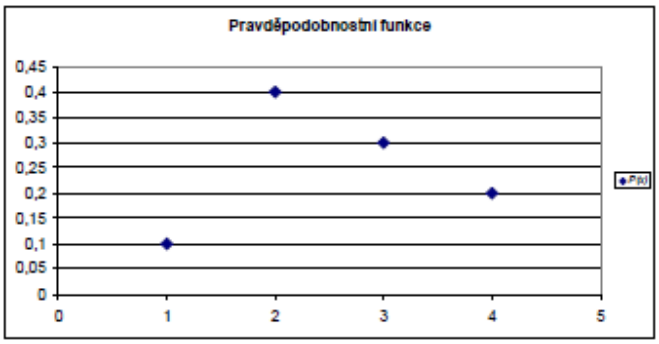
Nespojitá NV	Spojitá NV
Pravděpodobnostní funkce	*Distribuční funkce*
Distribuční funkce	*Hustota pravděpodobnosti*

- pravděpodobnostní funkce
    - definice: $P(x) = P(X = x)$
    - pravděpodobnost, že náhodná veličina X nabude hodnoty právě x
    - budeme se tázat na pravděpodobnost jedné konkrétní hodnoty (jaká je pravděpodobnost, že počet vybraných výrobků bude 5…)
    - vyjádření
        - matematickou formulí

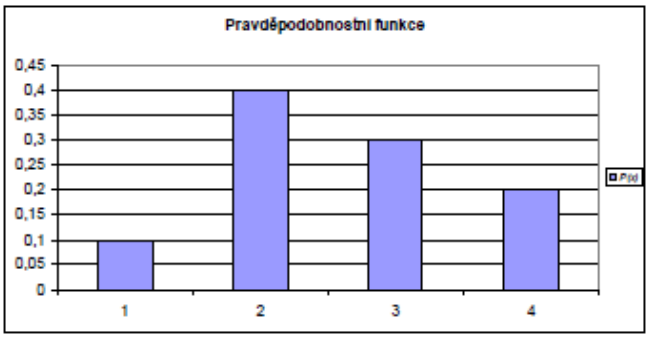
          $P(x) = \binom{n}{x} \pi^x (1-\pi)^{n-x}$
        - tabulkou rozdělení pravděpodobnosti

x	0	1	2	3	Σ
P(x)	0,10	0,20	0,25	0,45	1,00

        - graficky
            - bodovým grafem



- sloupcovým grafem



  - vlastnosti pravděpodobnostní funkce
    - $0 \leq P(x) \leq 1$
      - nabývá hodnot od 0 do 1
    - $\sum_x P(x) = 1$
      - jejich součet musí být 1
    - $P(x_1 \leq X \leq x_2) = \sum_{x=x_1}^{x_2} P(x)$
      - jestliže chceme znát pravděpodobnost, že náhodná veličina bude ležet v intervalu od x_1 do x_2 tak ty dílčí pravděpodobnosti výskytu posčítáme

- distribuční funkce
  - definice: $F(x) = P(X \leq x)$= pravděpodobnost, že náhodná veličina X nabude nejvýše hodnoty x
  - výpočet distribuční funkce:
    - nespojitá náhodná veličina

$$F(x) = \sum_{x \leq x_0} P(x)$$

    - spojitá náhodná veličiny

$$\int_{-\infty}^{x_0} f(x)dx$$

  - vlastnosti distribuční funkce
    - $0 \leq F(x) \leq 1$
      - nabývá hodnot od 0 do 1
    - $\forall_{x_2 > x_1} \rightarrow F(x_2) \geq F(x_1)$
    - $P(x_1 \leq X \leq x_2) = F(x_2) - F(x_1)$
      - je neklesající
      - je spojitá zprava a má nejvýš spočetně bodů nespojitosti

- hustota pravděpodobnosti
  - definice: $f(x) = F'(x)$
    - je to derivace distribuční funkce, popisuje průběh pravděpodobnostního rozdělení
  - vlastnosti hustoty pravděpodobnosti
    - $f(x) \geq 0$
      - nabývá hodnot větších nebo rovno 0
    - $\int_{-\infty}^{\infty} f(x)dx = 1$
      - jestliže zintegrujeme hustotu pravděpodobnosti, tak získáme distribuční funkci
      - jestliže budeme integrovat přes všechny meze, dostaneme 1

        - integrál od $-\infty$ do ∞ bude vždy roven 1
    - $P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(x)dx = F(x_2) - F(x_1)$
        - jestliže nás zajímá hodnota v intervalu od x_1 do x_2, bude to integrál od x_1 do x_2 → plocha pod křivkou
- charakteristiky polohy náhodné veličiny
    - střední hodnota = hodnota, kolem které kolísají hodnoty náhodné veličiny X
        - označení: E(X); očekávaná hodnota
        - pro nespojité náhodné veličiny

$$E(X) = \sum_x xP(x)$$

        - pro spojité náhodné veličiny

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

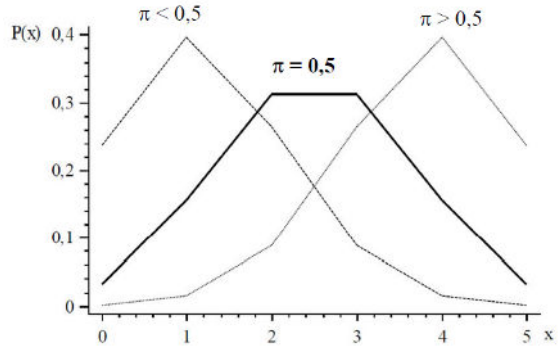
        - vlastnosti
            - $E(k) = k$
                - střední hodnota konstanty je konstanta (k = nenulová konstanta)
            - $E(kX) = kE(X)$
                - jestliže každou hodnotu náhodné veličiny vynásobíme stejnou nenulovou konstantou k, pak se střední hodnota zvýší k-násobně
            - $E(X + k) = E(X) + k$
                - jestliže ke každé hodnotě náhodné veličiny připočteme stejnou nenulovou konstantu k, pak se o tuto konstantu zvýší i střední hodnota
            - $E(X \pm Y) = E(X) \pm E(Y)$
                - střední hodnota součtu nebo rozdílu dvou náhodných veličin se rovná součtu nebo rozdílu středních hodnot těchto náhodných veličin
            - $E(XY) = E(X)E(Y)$
                - střední hodnota součinu dvou náhodných veličin se rovná součinu středních hodnot těchto náhodných veličin
            - $E\{[X - E(X)]\} = 0$
                - střední hodnota rozdílu jednotlivých hodnot od své střední hodnoty je rovna 0
    - kvantily $F(x_p) = P$ = p% kvantil v tomto případě je bod, ve kterém hodnota distribuční funkce kvantilu nabývá hodnoty P
    - modus = bod, hodnota náhodné veličiny, ve které pravděpodobnostní funkce nebo hustota pravděpodobnosti nabývá svého maxima
- charakteristiky variability náhodné veličiny
    - rozptyl = kolísání kolem střední hodnoty náhodné veličiny X
        - označení: D(X); $D(X) = E[X - E(X)]^2$
        - pro nespojité náhodné veličiny

$$D(X) = \sum_x x^2 P(x) - \left[\sum_x xP(x)\right]^2$$

        - pro spojité náhodné veličiny

$$D(X) = \int_{-\infty}^{\infty} x^2 f(x)dx - \left[\int_{-\infty}^{\infty} xf(x)dx\right]^2$$

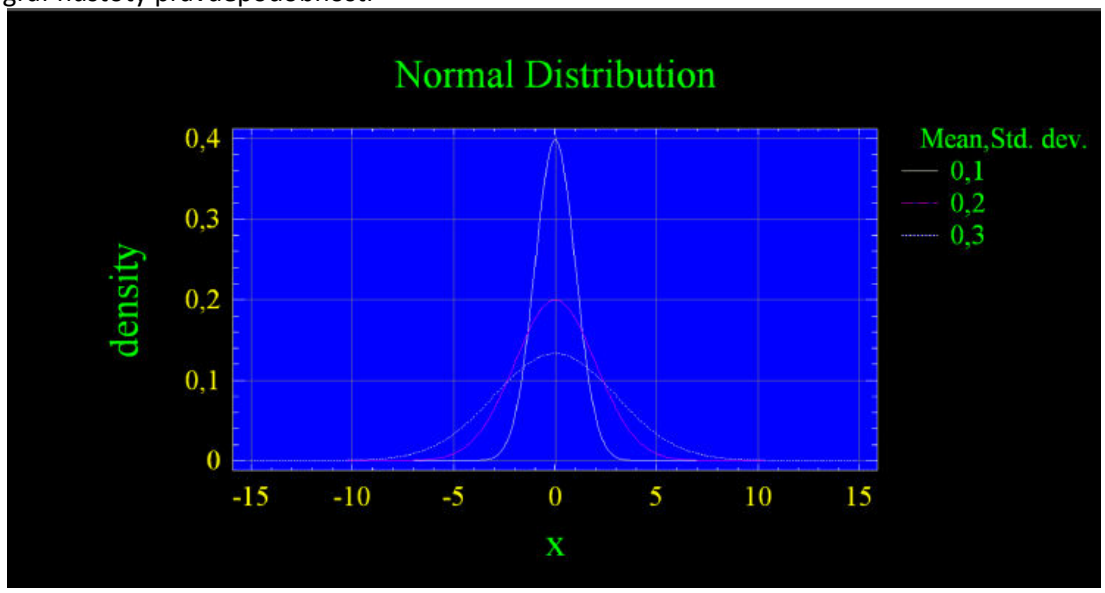
        - vlastnosti rozptylu
            - $D(k) = 0$
                - rozptyl konstanty je nula
            - $D(kX) = k^2 D(x)$
                - jestliže každou hodnotu náhodné veličiny vynásobíme stejnou nenulovou konstantou k, rozptyl se zvedá k^2 násobně
            - $D(X + k) = D(X)$
                - jestliže ke každé hodnotě přičteme stejnou nenulovou konstantu k, rozptyl se nezmění
            - $D(X \pm Y) = D(X) + D(Y)$
                - rozdíl nebo součet rozptylů je roven součtu rozptylů (platí pouze pro nezávislé náhodné veličiny)
    - směrodatná odchylka
        - označení: $\sigma_x = \sqrt{D(X)}$
- rozdělení nespojitých náhodných veličin
    - alternativní rozdělení
        - značení: $X \sim A(\pi)$
        - použití: v případě jednoho pokusu, během kterého je sledován výskyt náhodného jevu, který může nastat s pravděpodobností π (nebo nenastane)
        - X → počet výskytů jevu během jednoho pokusu
        - pravděpodobnostní funkce
          $P(x) = \pi^x (1 - \pi)^{1-x} \quad x = 0, 1; 0 < \pi < 1$
        - střední hodnota
          $E(X) = \pi$
        - rozptyl
          $D(X) = \pi(1 - \pi)$
    - binomické rozdělení
        - označení: $X \sim Bi(n; \pi)$ (= tvrzení, že náhodná veličina X se řídí binomickým rozdělením se dvěma parametry)
        - použití: u náhodné veličiny, která představuje počet výskytů sledovaného jevu během n vzájemně nezávislých pokusů a parametr π představuje konstantní pravděpodobnost výskytu sledovaného jevu v každém z pokusů
        - pravděpodobnostní funkce binomického rozdělení
        - X → počet výskytů jevu během n nezávislých pokusů
        - pravděpodobnostní funkce
          $P(x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x} \quad x = 0,1,2 \ldots, n; n > 0, 0 < \pi < 1$
        - střední hodnota
          $E(X) = n\pi$
        - rozptyl
          $D(X) = n\pi(1 - \pi)$
        - pravděpodobnostní funkce binomického rozdělení



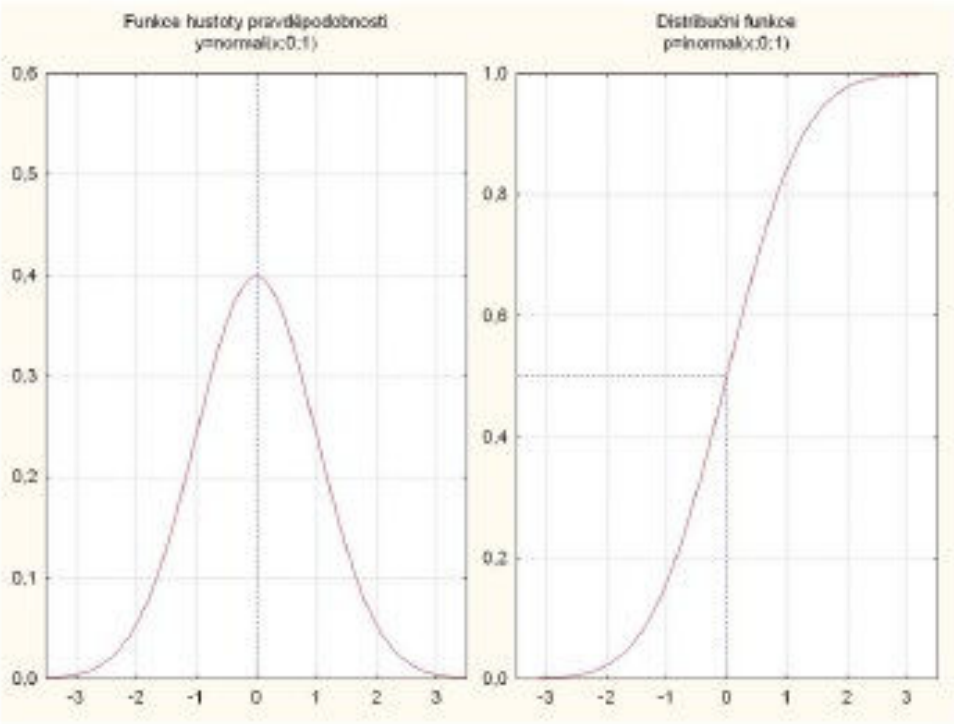
- Poissonovo rozdělení
  - označení: $X \sim Po(\lambda)$
  - použití: sleduje se výskyt jevů v časovém intervalu dané délky (za určitých podmínek), u nezávislých pokusů pro tzv. řídké jevy s malou pravděpodobností výskytu $(n > 30 \wedge \pi \leq 0{,}1)$
  - X → počet výskytů jevů během n pokusů (resp. v intervalu)
  - pravděpodobnostní funkce (λ = střední hodnota počtu výskytů náhodné veličiny – nπ)
    $P(x) = \frac{\lambda^x}{x!} e^{-\lambda} \quad x = 0{,}1, \ldots;\ \lambda > 0$
  - střední hodnota
    $E(X) = \lambda$
  - rozptyl
    $D(X) = \lambda$
- hypergeometrické rozdělení
  - označení: $X \sim Hy(N, M, n)$ (= tvrzení, že náhodná veličina X se řídí přibližně hypergeometrickým rozdělením se třemi parametry M, N, n (N = celkový počet prvků, M = počet prvků se sledovanou vlastností, n = rozsah výběru))
  - použití: v případě závislých pokusů (například pokusů bez vracení)
  - X → počet prvků se sledovanou vlastností během závislých pokusů
  - pravděpodobnostní funkce
    $P(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$
  - střední hodnota
    $E(X) = n\frac{M}{N}$
  - rozptyl
    $D(X) = n\frac{M}{N}\left(1 - \frac{M}{N}\right)\frac{N-n}{N-1}$

- rozdělení spojitých náhodných veličin
  - normální rozdělení (Gaussovo normální rozdělení)
    - označení: $X \sim N(\mu; \sigma^2)$
    - použití: u náhodné veličiny, jejíž kolísání je způsobeno součtem drobných vzájemně nezávislých jevů
    - Gaussova křivka
      $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad -\infty < x < \infty; -\infty < \mu < \infty; \sigma^2 > 0$
    - střední hodnota
      $E(X) = \mu$
    - rozptyl (udává tvar rozdělení)
      $D(X) = \sigma^2$
    - kvantily
      $x_p = \mu + \sigma u_p$
    - $P(x_1 \leq X \leq x_2) = P\left(\frac{x_1-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{x_2-\mu}{\sigma}\right) = P(u_1 \leq U \leq u_2) = \Phi(u_2) - \Phi(u_1)$

- graf hustoty pravděpodobnosti



- normované normální rozdělení
  - označení: $U \sim N(0;1)$; (= normovaná náhodná veličina U má normované normální rozdělení s nulovou střední hodnotou a jednotkovým rozptylem)
    $U = \frac{X-\mu}{\sigma} \quad X \sim N(\mu;\sigma^2)$
  - střední hodnota
    $E(U) = 0$
  - rozptyl
    $D(U) = 1$
  - pravděpodobnostní funkce
  - rozdělení je symetrické kolem nulové střední hodnoty (μ=0), z toho vyplývá
    - $\Phi(u) = 1 - \Phi(-u)$
    - $\mathrm{u_p} = -u_{1-P}$



- tabulky kvantilů normálního normovaného rozdělení

Tabulka IV. **Kvantily normovaného normálního rozdělení (u_P)**

P	u_P	P	u_P	P	u_P	P	u_P
0,50	0,000	0,75	0,674	0,950	1,645	0,975	1,960
0,51	0,025	0,76	0,706	0,951	1,655	0,976	1,970
0,52	0,050	0,77	0,739	0,952	1,665	0,977	1,995
0,53	0,075	0,78	0,772	0,953	1,675	0,978	2,014
0,54	0,100	0,79	0,806	0,954	1,685	0,979	2,034
0,55	0,126	0,80	0,842	0,955	1,695	0,980	2,054
0,56	0,151	0,81	0,878	0,956	1,706	0,981	2,075
0,57	0,176	0,82	0,915	0,957	1,717	0,982	2,097
0,58	0,202	0,83	0,954	0,958	1,728	0,983	2,120
0,59	0,228	0,84	0,994	0,959	1,739	0,984	2,144

- tabulky distribuční funkce normovaného normálního rozdělení

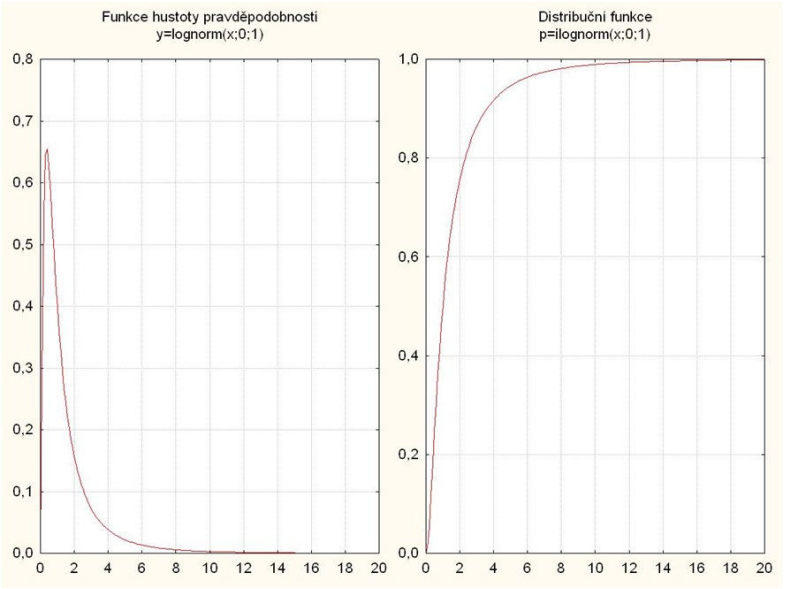
Tabulka III /pokračování/

Distribuční funkce normálního rozdělení

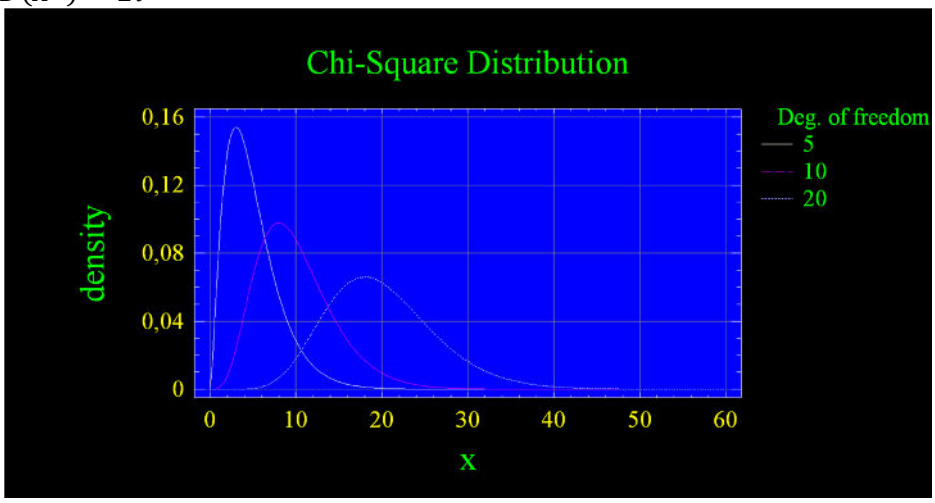
$$\Phi(u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy$$

u	$\Phi(u)$	u	$\Phi(u)$	u	$\Phi(u)$	u	$\Phi(u)$
1,60	0,94520	2,00	0,97725	2,40	0,99180	3,10	0,99903
1,61	94630	2,01	97778	2,41	99202	3,12	99910
1,62	94738	2,02	97831	2,42	99224	3,14	99916
1,63	94845	2,03	97882	2,43	99245	3,16	99921
1,64	94950	2,04	97932	2,44	99266	3,18	99926
1,65	95053	2,05	97982	2,45	99286	3,20	99931
1,66	95154	2,06	98030	2,46	99305	3,22	99936
1,67	95254	2,07	98077	2,47	99324	3,24	99940
1,68	95352	2,08	98124	2,48	99343	3,26	99944
1,69	95449	2,09	98169	2,49	99361	3,28	99948

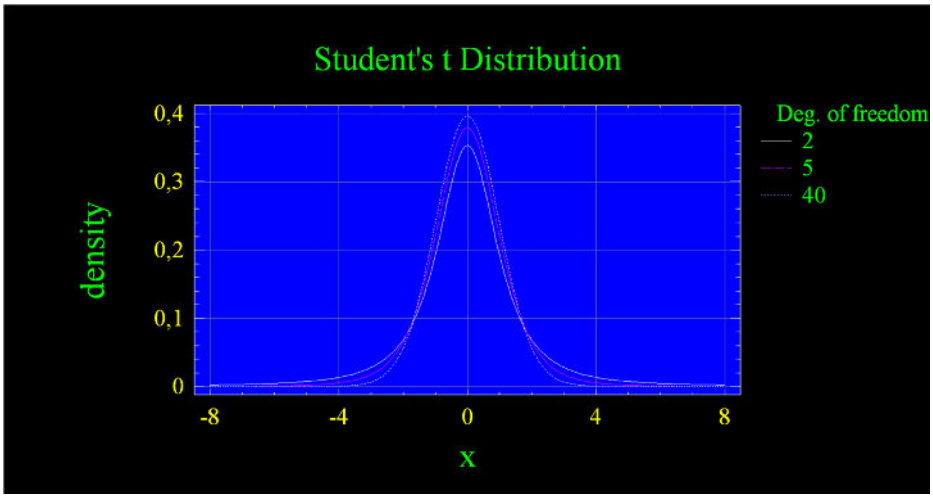
- logaritmicko-normální rozdělení
  - uvažujeme: $Y \sim N(\mu; \sigma^2)$ $X = e^Y$
  - označení: $X \sim LN(\mu; \sigma^2)$ (= nezáporná náhodná veličina X, jejíž logaritmy jsou normálně rozděleny, se řídí logaritmicko-normálním rozdělením - „lognormálním“)
  - použití: např. při zkoumání mzdových a příjmových rozdělení
  - střední hodnota
    $E(X) = e^{\mu + \frac{\sigma^2}{2}}$
  - rozptyl
    $D(X) = e^{2\mu+\sigma^2}\left(e^{\sigma^2} - 1\right)$
  - $U = \frac{lnX - \mu}{\sigma}$ $U \sim N(0; 1)$



- další významná rozdělení (zejména pro mat. stat. úlohy)
  - Chí-kvadrát rozdělení
    - označení: $X^2 \sim X^2(v)$
    - $U_i \sim N(0;1)$
    - pak NV:
      $$X^2 = \sum_{i=1}^{v} U_i^2 \sim X^2(v)$$
    - střední hodnota
      $E(X^2) = v$
    - rozptyl
      $D(X^2) = 2v$

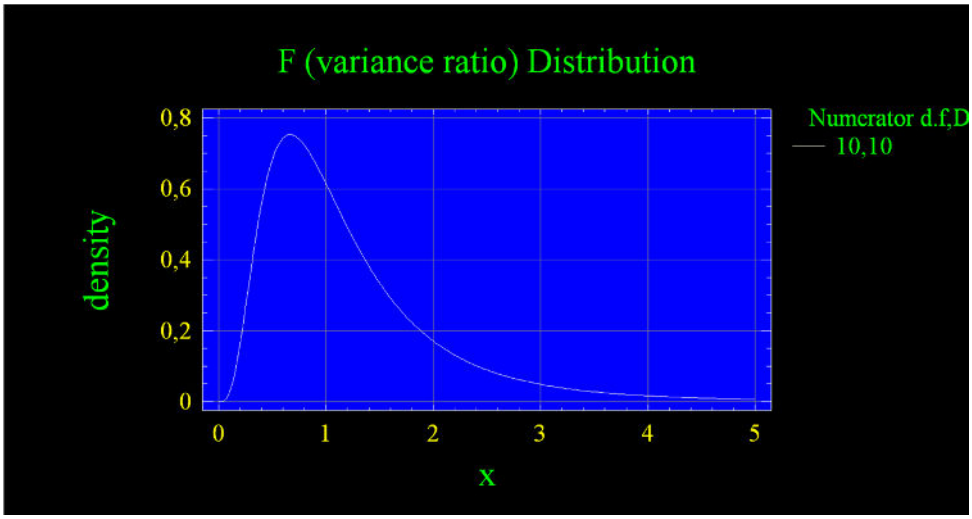


  - rozdělení t (Studentovo)
    - označení: $t \sim t(v)$
    - $U \sim N(0;1) \quad X^2 \sim X^2(v)$
    - pak NV:
      $$t = \frac{U}{\sqrt{\frac{X^2}{v}}} \sim t(v)$$
    - $t_p = -t_{1-P}$



- rozdělení F (Fisherovo-Snedecorovo)
  - označení: $F \sim F(v_1; v_2)$
  - $X_1^2 \sim X^2(v_1) \quad X_2^2 \sim X^2(v_2)$
  - pak NV:

    $$F = \frac{\frac{X_1^2}{v_1}}{\frac{X_2^2}{v_2}} \sim F(v_1, v_2)$$
  - pro $p < 0{,}5$: $F_p(v_1, v_2) = \frac{1}{F_{1-P}(v_2, v_1)}$



Limitní věty

- tvrzení o pravděpodobnostních rozděleních v případě rostoucího počtu náhodných pokusů
- zákon velkých čísel
- centrální limitní věty

Zákon velkých čísel

- definice = opakujeme-li nezávisle náhodný pokus, můžeme z napozorovaných hodnot sestavit rozdělení relativních četností a informací o tomto rozdělení shrnout do charakteristik (=empirická rozdělení četností)

- zvětšuje-li se počet pokusů, lze za jistých podmínek očekávat, že empirické rozdělení četností (i jeho charakteristiky) se bude blížit rozdělení teoretickému

Centrální limitní věty

- normální rozdělení je rozdělením limitním, k němuž se jiná rozdělení za určitých okolností blíží
- náhodná veličina X, která vznikla jako součet velkého počtu vzájemně nezávislých náhodných veličin X_1, X_2, X_3, X_n má za obecných podmínek přibližně normální rozdělení.
- náhodná veličina X, jejímž limitním zákonem rozdělení je rozdělení normální má asymptoticky normální rozdělení
- Moivre-Laplaceova věta
    - náhodná veličina X je součtem n vzájemně nezávislých náhodných veličin, z nichž každá má rozdělení A(π)
    - víme, že tato veličina má rozdělení Bi(n, π), se střední hodnotou E(X)=n π (1- π)
    - potom pro normovanou náhodnou veličinu

$$U = \frac{X - n\pi}{\sqrt{n\pi(1-\pi)}}$$

    - platí limitní vztah

$$\lim_{n\to\infty} = P(U < n) = Ø(\mathrm{n}) \qquad -\infty < n < \infty$$

    - obdobně i aritmetický průměr náhodných veličin X_i, tj. náhodná veličina

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

    - má asymptoticky normální rozdělení s parametry

$$E(\bar{x}) = \pi \qquad D(\bar{x}) = \frac{\pi(1-\pi)}{n}$$

    - při dostatečně velkém počtu nezávislých pokusů konverguje (blíží se) binomické rozdělení k rozdělení normálnímu, tj. Bi(n, π) → N(n π, n π (1- π))
    - nastane právě, když platí Empirické pravidlo: n π (1- π) > 9

- Lindeberg-Lévyho věta
    - součet n vzájemně nezávislých náhodných veličin, které jsou stejně rozdělené (mají libovolné stejné rozdělení) s končenou střední hodnotou E(X) a konečným rozptylem D(X), konverguje pro velká n k přibližně normálnímu rozdělení se střední hodnotou E(X)=n·μ a rozptylem $D(X)=n\sigma^2$
    - pro normovanou veličinu

$$U = \frac{X - n\mu}{\sqrt{n\sigma^2}} \quad \mathrm{U} \sim \mathrm{N}(0;1)$$

- platí limitní vztah

$$\lim_{n\to\infty} = P(U < n) = \text{Ø(n)} \qquad -\infty < n < \infty$$

- obdobě i aritmetický průměr náhodných veličin Xi, tj. náhodná veličina

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

- má asymptoticky normální rozdělení s parametry

$$E(\bar{x}) = \mu \qquad D(\bar{x}) = \frac{\sigma^2}{n}$$

- důležité: při dostatečně velkém n lze rozdělení náhodné veličiny $\sum x_i$ a $\overline{X}$ aproximovat (nahradit) normálním rozdělením

$$\sum x_i \sim \text{N}(\text{n}\mu; \text{n}\sigma^2)$$

$$\bar{x} \sim \text{N}\left(\mu; \text{n}\frac{\sigma^2}{\text{n}}\right)$$

- empirické pravidlo: n≥30

4. Matematická statistika

- zabývá se problematikou konstrukce zásad úsudků o základním souboru na základě poznání z části souboru pořízené náhodným výběrem
- statistické odhady
  - metody odhadování neznámých parametrů základního souboru na základě informací o charakteristikách náhodného výběru
- testování statistických hypotéz
  - induktivní postupy, které vedou k zamítnutí nebo potvrzení určitých tvrzení (hypotéz) o základním souboru
  - postup, který umožňuje na základě naměřených dat určit, zda náhodná veličina vykazuje určitou vlastnost
- značení charakteristik
  - základní soubor σ, μ, π
  - výběrový soubor $S'_x, \bar{x}, p$
- statistické šetření
  - příprava
  - provedené šetření
  - zpracování údajů
  - vyhodnocení výsledků
  - publikace výsledků

5. Teorie odhadu

Charakteristiky základního souboru (populace)

- základní střední hodnota

$$\mu = \sum_{i=1}^{N} x_i \, / N$$

- základní rozptyl

$$\sigma^2 = \sum_{i=1}^{N} (x_i - \mu)^2 \, / N$$

- základní relativní četnost (podíl)

$$\pi = \frac{M}{N}$$

Charakteristiky výběrového souboru (statistiky)

- výběrový průměr

$$\bar{x} = \sum_{i=1}^{n} x_i \, / n$$

- výběrový rozptyl

$$x_x^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 \, / (n-1)$$

- výběrová relativní četnost (podíl)

$$p = \frac{m}{n}$$

Bodový odhad

- odhad parametrů ZS pomocí jednoho jediného čísla
- neznámou hodnotu parametru G základního souboru odhadneme pomocí vypočítané hodnoty vhodné výběrové charakteristiky (statistiky) g (g ~ G …. statistika g je bodovým odhadem charakteristiky základního souboru G)
- zápis: $estT = t$; $\hat{T} = t$
- vlastnosti bodového odhadu
    - nezkreslenost (nestrannost)
        - statistika g je nezkreslený (nestranný, nevychýlený) odhad parametru G, pokud E(g)=G
            - tj. zvolená statistika systematicky nenad(pod)hodnocuje odhadovanou charakteristiku
            - rozdíl E(g)–G je zkreslení (vychýlení, výběrová chyba)

- konzistence
  - odhad je konzistentní tehdy, jestliže s rostoucím rozsahem výběru roste PP, že hodnota g se bude nalézat v blízkosti G

    $$\lim_{n\to\infty} P(|t - T| < \varepsilon) = 1$$

- vydatnost
  - z nezkreslených odhadů je nejvýhodnější takový odhad, který má nejmenší rozptyl – vydatný odhad charakteristiky základního souboru

- odhadované charakteristiky základního souboru
  - základní střední hodnota = odhad základní střední hodnoty je roven výběrovému průměru

    $\hat{\mu} = \bar{x}$

  - základní rozptyl = odhad populačního rozptylu je roven výběrovému rozptylu

    $\hat{\sigma}^2 = s'^2_x$

  - základní relativní četnost

    $\hat{\pi} = p$

Intervalový odhad

- spočívá v konstrukci intervalu spolehlivosti
- interval spolehlivosti = interval, od kterého očekáváme s předem zvolenou, hodně vysokou pravděpodobností, že obsáhne neznámou odhadovanou charakteristiku
- 3 druhy
  - oboustranný

    $P(T_d < T < T_h) = 1 - \alpha$

  - levostranný (když máme dolní mez)

    $P(T_d < T) = 1 - \alpha$

  - pravostranný (když máme horní mez)

    $P(T < T_h) = 1 - \alpha$

- $1 - \alpha =$ spolehlivost odhadu (nejčastěji 0,95, případně 0,99)
- s rostoucí spolehlivostí klesá přesnost intervalového odhadu
- přesnost intervalového odhadu roste (při stejné spolehlivosti) s rostoucím n (rozsahem výběrového vzorku)
- přesnost intervalového odhadu je dána jeho šířkou
- odhadové charakteristiky
  - základní střední hodnota = zkoumáme, jestli známe rozptyl a rozsah výběrového souboru (zkoumáme, jestli bude platit centrální limitní věta)

- výběr pochází z normálního rozdělení a známe σ^2

$$P\left(\bar{x} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

- výběr pochází z normálního rozdělení, neznáme $\sigma^2, n > 30$

$$P\left(\bar{x} - u_{1-\alpha/2}\frac{s'_x}{\sqrt{n}} < \mu < \bar{x} + u_{1-\alpha/2}\frac{s'_x}{\sqrt{n}}\right) = 1 - \alpha$$

- výběr pochází z normálního rozdělení, neznáme $\sigma^2, n < 30$

$$P\left(\bar{x} - t_{1-\alpha/2}\frac{s'_x}{\sqrt{n}} < \mu < \bar{x} + t_{1-\alpha/2}\frac{s'_x}{\sqrt{n}}\right) = 1 - \alpha$$

- výběr pochází z libovolného rozdělení, $n > 30$

$$P\left(\bar{x} - u_{1-\alpha/2}\frac{s'_x}{\sqrt{n}} < \mu < \bar{x} + u_{1-\alpha/2}\frac{s'_x}{\sqrt{n}}\right) = 1 - \alpha$$

  - přípustná chyba odhadu = součin toho kvantilu a tzv. směrodatné chyby odhadu = chyba, kterou s danou spolehlivostí odhadu připouštíme

$$\Delta = u_{1-\alpha/2}\frac{s'_x}{\sqrt{n}}$$

  - směrodatná chyba odhadu

$$s_{\bar{x}} = \frac{s'_x}{\sqrt{n}}$$

- základní relativní četnost (podíl lidé se sledovanou vlastností)

$$P\left(p - u_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} < \pi < p + u_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) = 1 - \alpha$$

  - jestliže máme jednostranný interval spolehlivosti, tak α nedělíme 2

Příklad číslo 1

Byl sledován obsah určité látky (v gramech) ve vytvářené směsi. Na základě 100 měření byl zjištěn výběrový průměr 4 a výběrový rozptyl 12,25. Stanovte 95% interval spolehlivosti pro neznámou základní střední hodnotu.

Příklad číslo 2

Byl proveden průzkum, ve kterém byl sledován zájem o nový film, který má být uveden do kin. Ze 100 respondentů odpovědělo 72 z nich, že by daný film měli zájem shlédnout. Stanovte 95% interval spolehlivosti pro podíl lidí z celé populace.

6. Testování hypotéz

- předmět testování hypotéz = ověřování úsudku o základním souboru
    - Je průměrná mzda víc než 12500?

Základní pojmy

- hypotéza
    - H_0 = testovaná hypotéza (nulová hypotéza) = tvrzení o základním souboru, jehož platnost ověřujeme (průměrná mzda je 15000, preference politické strany je 36,5%)
    - H_1 = alternativní hypotéza = obsahuje tvrzení, které popírá platnost H_0 (průměrná mzda není 15000, je to víc, je to míň)
- testové kritérium
    - vhodně zvolená statistika, která má při platnosti testované hypotézy známé pravděpodobnostní rozdělení, tj. já zvolím již vymyšlenou statistiku (vhodně) → budu znát její pravděpodobnostní rozdělení
- kritický obor
    - množina hodnot testového kritéria, pro kterou platí: jestliže testové kritérium do této množiny spadá, pak testovanou hypotézu zamítáme
- obor přijetí
    - množina hodnot testového kritéria, pro kterou platí: jestliže testové kritérium do této množiny spadá, pak testovanou hypotézu nezamítáme
- kritická hodnota
    - představuje takovou hodnotu, která odděluje kritický obor od oboru přijetí
    - jedná se o kvantil známého pravděpodobnostního rozdělení testového kritéria
- vztah mezi testovanou (nulovou) hypotézou a rozhodnutím
    - mohou nastat tyto situace

H_0	Správná hypotéza	Nesprávná hypotéza
nezamítáme	*správné rozhodnutí*	*chyba II. Druhu*
zamítáme	*chyba I. Druhu*	*správné rozhodnutí*

    - chyba I. druhu = jestliže tvrzení je správné, ale matematický aparát tvrdí zamítáme (= zamítnutí správné hypotézy)
    - chyba II. druhu = jestliže tvrzení je špatné, ale matematický aparát tvrdí nezamítáme (= nezamítnutí nesprávné hypotézy)
- α-P (chyby 1. druhu) – hladina významnosti (α) – je to pravděpodobnost chyby 1. druhu
    - hladinu významnosti si volíme sami dopředu
    - nejčastěji na úrovni 5 % (lze i 1 %, 10 %)
    - 5 % = 5% pravděpodobnost chyby (NE že se v 5% případů zmýlím!)
- β-P(chyby II. druhu)
    - 1-β – síla testu = pravděpodobnost zamítnutí nesprávné hypotézy H_0
- platí
    - pokles α – růst β = klesá pravděpodobnost chyby I. druhu, což má za následek růst chyby II. (klesá síla testu)

- zvýšením n lze snížit α i β = zvýšením n lze snížit pravděpodobnost chybného rozhodnutí

- formalizovaný postup
  - volba hladiny významnosti
  - stanovení hypotéz
  - výpočet testového kritéria
  - stanovení kritického oboru
  - závěr testu
    - ‚ručně': pokud testové kritérium spadne do kritického oboru – na hladině významnosti α zamítáme H_0
    - SW: (p-hodnota): pokud p-hodnota ≤ α – na hladině významnosti α zamítáme H_0

- test o střední hodnotě normálního rozdělení

H_0	H_1	Testové kritérium	Kritický obor
$\mu = \mu_0$	$\mu > \mu_0$ $\mu < \mu_0$ $\mu \neq \mu_0$	σ^2 známé $U = \frac{\bar{x} - \mu_0}{\sigma}\sqrt{n}$ $\quad U \sim N(0,1)$	$W_\alpha = \{U \geq u_{1-\alpha}\}$ $W_\alpha = \{U \leq -u_{1-\alpha}\}$ $W_\alpha = \{\lvert U\rvert \geq u_{1-\alpha/2}\}$
		σ^2 neznámé $t = \frac{\bar{x} - \mu_0}{s'_x}\sqrt{n}$ $\quad t \sim t(n-1)$	$W_\alpha = \{t \geq t_{1-\alpha}\}$ $W_\alpha = \{t \leq -t_{1-\alpha}\}$ $W_\alpha = \{\lvert t\rvert \geq t_{1-\alpha/2}\}$

  - H_0 – co má v indexu 0 = dosazování konkrétní hodnoty, kterou chceme testovat
  - H_1 – vybereme jednu alternativu ze 3, podle toho, co chceme prokázat (nárůst = 1, pokles = 2, prostě není = 3)
  - volba testového kritéria – musíme znát, jaké máme výchozí předpoklady (Známe nebo neznáme rozptyl? Máme dostatečně velký rozsah, aby platila centrální limitní věta?

- test o střední hodnotě lib. rozdělení při velkém výběru

H_0	H_1	Testové kritérium	Kritický obor
$\mu = \mu_0$	$\mu > \mu_0$ $\mu < \mu_0$ $\mu \neq \mu_0$	σ^2 neznámé ($n > 30$) $U = \frac{\bar{x} - \mu_0}{s'_x}\sqrt{n}$ $\quad U \sim N(0,1)$	$W_\alpha = \{U \geq u_{1-\alpha}\}$ $W_\alpha = \{U \leq -u_{1-\alpha}\}$ $W_\alpha = \{\lvert U\rvert \geq u_{1-\alpha/2}\}$

  - kritický obor stanovujeme podle alternativní hypotézy
  - nezáleží na rozdělení – musíme vědět, že je dostatečně velký soubor

- test o podílu (parametr π)
  - Dostane se politická strana do parlamentu? (= Jsou její preference větší než 5%?)

H_0	H_1	Testové kritérium	Kritický obor
$\pi = \pi_0$	$\pi > \pi_0$ $\pi < \pi_0$ $\pi \neq \pi_0$	$U = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$ $U \sim N(0,1)$	$W_\alpha = \{U \geq u_{1-\alpha}\}$ $W_\alpha = \{U \leq -u_{1-\alpha}\}$ $W_\alpha = \{\|U\| \geq u_{1-\alpha/2}\}$

- nezávislé výběry (výběry z normálního rozdělení)
  - máme-li porovnávat dvě střední hodnoty, musíme se rozhodovat nejprve podle toho, jestli máme závislé nebo nezávislé výběry (závislé – jedny hodnoty ovlivňují ty druhé (dvakrát testujeme jednu skupina = jednu skupinu lidí zvážíme, pošleme do fitka, podruhé je zvážíme, dvě skupinky ve vedlejších třídách = nezávislé výběry)

H_0	H_1	Testové kritérium	Kritický obor
$\mu_1 = \mu_2$ $\mu_1 - \mu_2 = 0$	$\mu_1 > \mu_2$ $\mu_1 < \mu_2$ $\mu_1 \neq \mu_2$	a) σ_1^2, σ_2^2 známé $U = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ $\quad U \sim N(0,1)$	$W_\alpha = \{U \geq u_{1-\alpha}\}$ $W_\alpha = \{U \leq -u_{1-\alpha}\}$ $W_\alpha = \{\|U\| \geq u_{1-\alpha/2}\}$
		σ_1^2, σ_2^2 neznámé, ale předpokládáme, že $\sigma_1^2 = \sigma_2^2$ $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{(n_1-1)s_1'^2 + (n_2-1)s_2'^2}{n_1+n_2-2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$ $t \sim t(n_1 + n_2 - 2)$	$W_\alpha = \{t \geq t_{1-\alpha}\}$ $W_\alpha = \{t \leq -t_{1-\alpha}\}$ $W_\alpha = \{\|t\| \geq t_{1-\alpha/2}\}$
		σ_1^2, σ_2^2 neznámé, ale předpokládáme, že $\sigma_1^2 \neq \sigma_2^2$ $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1'^2}{n_1} + \frac{s_2'^2}{n_2}}}$ $\quad t \sim t(\nu)$ $\nu = \frac{\left(\frac{s_1'^2}{n_1} + \frac{s_2'^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1'^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2'^2}{n_2}\right)^2}$	$W_\alpha = \{t \geq t_{1-\alpha}\}$ $W_\alpha = \{t \leq -t_{1-\alpha}\}$ $W_\alpha = \{\|t\| \geq t_{1-\alpha/2}\}$

- velké nezávislé výběr

H_0	H_1	Testové kritérium	Kritický obor
$\mu_1 = \mu_2$ $\mu_1 - \mu_2 = 0$	$\mu_1 > \mu_2$ $\mu_1 < \mu_2$ $\mu_1 \neq \mu_2$	σ_1^2, σ_2^2 neznámé $U = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1'^2}{n_1} + \frac{s_2'^2}{n_2}}}$ $\quad U \sim N(0,1)$	$W_\alpha = \{U \geq u_{1-\alpha}\}$ $W_\alpha = \{U \leq -u_{1-\alpha}\}$ $W_\alpha = \{\|U\| \geq u_{1-\alpha/2}\}$

- závislé výběry z normálního rozdělení (párový t-test)
    - založen na tom, že jednoho jedince testujeme 2x, za každého známe diferenci mezi těmi 2 testováními, diference zprůměrujeme za celou skupinu a vypočítáme směrodatnou odchylku

H_0	H_1	Testové kritérium	Kritický obor
$\mu_1 = \mu_2$ $\mu_1 - \mu_2 = 0$	$\mu_1 > \mu_2$ $\mu_1 < \mu_2$ $\mu_1 \neq \mu_2$	$t = \frac{\bar{d}}{s_d}\sqrt{n-1} \qquad t \sim t(n-1)$ $d_i = x_{1i} - x_{2i},\ i=1,2,..,n$	$W_\alpha = \{t \geq t_{1-\alpha}\}$ $W_\alpha = \{t \leq -t_{1-\alpha}\}$ $W_\alpha = \{\lvert t\rvert \geq t_{1-\alpha/2}\}$

- chí-kvadrát test dobré shody
    - používáme ve 2 případech
        - chceme ověřit shodu empirického rozdělení četností s rozdělením teoretickým
        - chceme ověřit, zda má základní soubor rozdělení určitého typu

H_0 a H_1	Testové kritérium	Kritický obor
H_0: $\pi_i = \pi_{0,i} \quad i = 1, .., k$ H_1: non H_0	$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n\pi_{0,i})^2}{n\pi_{0,i}} \qquad \chi^2 \sim \chi^2(k\text{-}1)$	$W_\alpha = \{\chi^2 \geq \chi^2_{1-\alpha}\}$

- podmínka užití testu = dostatečné zastoupení ve všech skupinách, tj. $n\pi_{0,i} > 0{,}5$

Příklad číslo 1

Současná teorie vychází z předpokladu, že průměrná výška studenta prvního ročníku vysoké školy je 170 cm. Na základě průzkumu náhodně vybraných 92 studentů prvního ročníku bylo zjištěno, že průměrná výška studenta je 174,8 cm a rozptyl výšky 69 cm2. Ověřte, zda toto zjištění není v rozporu s původním předpokladem. Použijte 5% hladinu významnosti.

- řešení:
- $H_0: \mu = 170$
- $H_1: \mu \neq 170$
- $U = \frac{174{,}8-170}{\sqrt{69}} \cdot \sqrt{92} = 5{,}5425$
- $W_{0,05} = \{U; |U| \geq 1{,}96\}$
  $\rightarrow zamítáme\ H_0\ na\ \alpha = 0{,}05$

Příklad číslo 2

O určitý výrobek mělo zájem 30 % zákazníků. Lze předpokládat, že o nově zaváděný výrobek bude stejný zájem, pokud z 200 respondentů by si nový výrobek koupilo 50 z nich? Předpokládejte 5% hladinu významnosti.

- řešení:
- $H_0: \alpha = 0{,}3$
- $H_1: \alpha \neq 0{,}3$
- $U = \frac{0{,}25-0{,}3}{\sqrt{\frac{0{,}3\cdot(1-0{,}3)}{200}}} = -1{,}543$
- $W_{0,05} = \{U; |U| \geq 1{,}96\}$
  $\rightarrow nezamítáme\ H_0\ na\ \alpha = 0{,}05$

Příklad číslo 3

Z pokladních účtenek, které obsahovaly nákup jednoho bochníku chleba, bylo náhodně vybráno 100. Šlo o chléb tří druhů, z toho prvního druhu se prodalo 24 bochníků, druhého 32 a třetího 44. Můžeme předpokládat, že zákazníci kupují od každého druhu stejně?

- řešení:

H_0: $\pi_1 = \pi_2 = \pi_3$,
H_1: non H_0.

Typ	Zjištěná četnost (n_i)	Očekávaná četnost ($n\pi_{i,0}$)	Rozdíl ($n_i - n\pi_{i,0}$)	$\frac{(n_i - n\pi_{i,0})^2}{n\pi_{i,0}}$
1	24	33,33	-9,33	2,61
2	32	33,33	-1,33	0,05
3	44	33,33	10,67	3,41
Σ	100	x	x	6,07

$$W_{0,05} = \{\chi^2; \chi^2 \geq 5{,}99\}$$

$$W_{0,01} = \{\chi^2; \chi^2 \geq 9{,}21\}$$

→ $zamítáme\ H_0\ na\ \alpha = 0{,}05$
→ $nezamítáme\ H_0\ na\ \alpha = 0{,}01$

Příklad číslo 4

U pacientů, kteří trpí určitou nemocí, se sleduje obsah jisté látky v krvi. U 13 náhodně vybraných pacientů sledována úroveň této látky v krvi. Těmto pacientům byl aplikován nový lék a bylo sledováno, jaká bude úroveň této látky v krvi po aplikaci tohoto léku. Na 5% hladině významnosti ověřte, zda aplikací tohoto léku došlo ke snížení obsahu látky v krvi.

Hodnoty před	Hodnoty po
2,5	2,0
2,7	2,1
2,9	2,4
2,7	2,2
2,8	2,3
3,2	2,6
2,9	2,4
2,8	2,3
2,7	2,2
2,8	2,1
2,5	2,0
2,9	2,4
3,1	2,5

- řešení:

$H_0: \mu_1 = \mu_2,$
$H_1: \mu_1 > \mu_2.$

Hodnoty před (x_{1i})	Hodnoty po (x_{2i})	d_i
2,5	2,0	0,5
2,7	2,1	0,6
2,9	2,4	0,5
2,7	2,2	0,5
2,8	2,3	0,5
3,2	2,6	0,6
2,9	2,4	0,5
2,8	2,3	0,5
2,7	2,2	0,5
2,8	2,1	0,7
2,5	2,0	0,5
2,9	2,4	0,5
3,1	2,5	0,6

$$\overline{d} = 0{,}538$$

$$s_d = 0{,}0625$$

$$t = \frac{0{,}538}{0{,}0625}\sqrt{13-1} = 31{,}067$$

$$W_{0,05} = \{t; |t| \geq 1{,}782\}$$

$\rightarrow$ *zamítáme* H_0 *na* $\alpha = 0{,}05$

7. Úvod do analýzy závislostí

Metody zkoumání závislostí podle typů proměnných

- X – kategoriální proměnná (slovní); Y – kategoriální proměnná (kontingenční tabulka)
- X – kategoriální proměnná; Y – číselná proměnná (analýza rozptylu); závislost je jednostranná – závisí pouze Y na X, ne obráceně
- X – číselná proměnná; Y – číselná proměnná (regresní a korelační analýza)

Kontingenční tabulka a chí-kvadrát test nezávislosti v kontingenční tabulce

- kontingenční tabulka = dvourozměrná tabulka četností
  - obsahuje četnosti
    - sdružené četnosti n_{ij}
    - marginální (okrajové – na okraji tabulky – vznik součtem jednotlivých řádků a sloupců) četnosti $n_{i.}$ a $n_{.j}$

$$n_{i.} = \sum_{j=1}^{s} n_{ij} \qquad n_{.j} = \sum_{i=1}^{r} n_{ij}$$

- kontingenční tabulka

$X_i \backslash Y_j$	Y_1	Y_2	...	Y_s	Σ
X_1	n_{11}	n_{12}	...	n_{1s}	$n_{1.}$
X_2	n_{21}	...	...	...	$n_{2.}$
X_3	...	...	...	...	...
....	...	...	...	...	...
X_r	n_{r1}	...	...	n_{rs}	$n_{r.}$
Σ	$n_{.1}$	$n_{.2}$	...	$n_{.s}$	n

- úkol
  - ověřit, zdali jsou proměnné X a Y nezávislé
    - tabulka skutečných (naměřených) četností
    - tabulka hypotetických (teoretických) četností
      - označení teoretických četností (teoretické četnosti nám říkají, kolik by v dané kombinaci proměnných X a Y muselo být respondentů, kdyby byly zcela nezávislé)
        - sdružené četnosti n'_{ij}
        - marginální četnosti
          - řádkové $n'_{i.}$
          - sloupcové $n'_{.j}$
      - princip konstrukce

$$n'_{ij} = \frac{n_{i.} n_{.j}}{n}$$

- tabulka teoretických (hypotetických) četností

$X_i \backslash Y_j$	Y_1	Y_2	...	Y_s	Σ
X_1	n'_{11}	n'_{12}	...	n'_{1s}	$n_{1.}$
X_2	n'_{21}	...	...	...	$n_{2.}$
X_3	...	...	...	...	...
....	...	...	...	...	...
X_r	n'_{r1}	...	...	n'_{rs}	$n_{r.}$
Σ	$n_{.1}$	$n_{.2}$	...	$n_{.s}$	n

- chí-kvadrát test nezávislosti v kontingenční tabulce

H_0	H_1	Testové kritérium	Kritický obor
$\pi_{ij} = \pi_{i.}\,\pi_{j}$ $1 \le i \le r$ $1 \le j \le s$ $\Rightarrow$ ***nezávislost X a Y***	non H_0	$G = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij} - n'_{ij})^2}{n'_{ij}}$ $G \sim \chi^2((r-1)(s-1))$	$W_\alpha = \{G \ge \chi^2_{1-\alpha}\}$

  - !!! hypotéza H_0 = nezávislost X a Y !!!
  - KO jde vždy směrem doprava
- koeficienty, které měří sílu (intenzitu) závislosti
  - Pearsonův

$$C = \sqrt{\frac{G}{n+G}} \quad C \in <0;1) \quad maxC = \sqrt{\frac{m-1}{m}} \quad m = \min(r,s)$$

  - Cramerův

$$V = \sqrt{\frac{G}{n(m-1)}} \quad V \in <0;1>$$

- podmínka užití testu: $n'_{ij} \ge 5$

Příklad číslo 1

U 130 studentů prvního ročníku vysoké školy bylo zaznamenáno jejich pohlaví a hmotnost vztažená k výšce (nízká, střední, vysoká). Údaje byly uspořádány do následující tabulky. Ověřte na 5% hladině významnosti, zdali závisí váhová kategorie studenta (ve vztahu k výšce) na jeho pohlaví. Vypočtěte koeficient, který vyjadřuje sílu závislosti.

Pohlaví\Hmotnost	Nízká	Střední	Vysoká
Muži	22	11	37
Ženy	42	10	8

- řešení:
  - tabulka skutečných četností

Pohlaví\Hmotnost	Nízká	Střední	Vysoká	Σ
Muži	22	11	37	70
Ženy	42	10	8	60
Σ	64	21	45	130

- tabulka teoretických četností

Pohlaví\Hmotnost	Nízká	Střední	Vysoká	Σ
Muži	34,4615	11,3077	24,2308	70
Ženy	29,5385	9,69231	20,7692	60
Σ	64	21	45	130

- výpočet testového kritéria G

Pohlaví\Hmotnost	Nízká	Střední	Vysoká	Σ
Muži	4,50618	0,00837	6,72918	11,2437
Ženy	5,25721	0,00977	7,85071	13,1177
Σ	9,76339	0,01814	14,5799	**24,3614**

- testové kritérium: G=24,36143
- kritická hodnota: $\chi^2_{0,95} = 5,991$
  - zamítáme ho na α = 0,05 (váhová kategorie studenta závisí na jeho pohlaví)
- koeficient, který měří sílu (intenzitu) závislosti
  - $C = \sqrt{\frac{24,36}{130+24,36}} = 0,3972$
  - $V = \sqrt{\frac{24,36}{130(2-1)}} = 0,4329$
  - $-\ ne\ příliš\ intenzivní\ závislost$

Analýza rozptylu (ANOVA)

- X – kategoriální proměnná (faktor) – vysvětlující proměnná
- Y – kvantitativní spojitá proměnná – vysvětlovaná proměnná
- úkol: ověřit zda proměnná Y není závislá na proměnné X (jednostranná závislost)
- princip ANOVA
  - rozklad celkové variability na jednotlivé složky
    - celková

      $$S_y = S_{y,M} + S_{y,V}$$

    - meziskupinová

      $$S_{y,M} = \sum_{i=1}^{k} (\bar{y}_i - \bar{y})^2 n_i$$

      - k = počet skupin, které vzniknou roztříděním podle slovního faktoru (požadavkem, aby se dala použít analýza rozptylu, je, že k musí být větší než 2 = alespoň 3)
      - $\bar{y}_i$ = průměry ve skupinách (skupinové průměry)
      - $\bar{y}$ = celkový průměr
      - n_i = počet hodnot (respondentů) ve skupině

- vnitroskupinová
  - y_{ij} = konkrétní hodnoty, které máme od respondentů

$$S_{y,V} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

- postup ANOVA

H_0	H_1	Testové kritérium	Kritický obor
$\mu_1 = \mu_2 = ... = \mu_k$	non H_0	$F = \frac{\frac{\sum_{i=1}^{k} (\bar{y}_i - \bar{y})^2 n_i}{k-1}}{\frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}{n-k}} = \frac{\frac{S_{y,m}}{k-1}}{\frac{S_{y,v}}{n-k}}$ $F \sim F(k-1, n-k)$	$W_\alpha = \{F \geq F_{1-\alpha}\}$

- koeficient, který měří sílu (intenzitu) závislosti = poměr determinace
  - poměr determinace

$$P^2 = \frac{S_{y,M}}{S_y} \quad P^2 \in < 0; 1 >$$

    - říká, kolik % z celkové variability tvoří ta odlišnost meziskupinové variability
    - když bude meziskupinový rozptyl 0 (= průměry ve skupinách jsou všechny úplně identické) = jsou zcela nezávislé
    - 1 = meziskupinová variabilita se rovná celkové variabilitě (= vnitroskupinová variabilita = 0 = uvnitř skupiny se všechny (např.) mzdy rovnají, ale skupiny se odlišují) = jsou zcela závislé
    - čím je hodnota vyšší, tím je závislost silnější
- předpoklady ANOVA
  - hodnoty proměnné Y v každé z k skupin představují výběry z normálního rozdělení
    - pozn. porušení tohoto předpokladu normality proměnné Y nemá zásadní vliv na rozdělení statistiky F
  - tyto výběry jsou nezávislé (pokud jsou závislé, není možné použít analýzu rozptylu)
  - shoda všech skupinových rozptylů
    - lze ověřit pomocí tzv. Bartlettova testu
    - pozn. v případě, že jsou rozsahy skupin stejné, uvádí se, že nesplnění předpokladu o shodě skupinových rozptylů, nemá zásadní vliv na analýzu rozptylu

Příklad číslo 1

Na základě informací z tabulky rozhodněte, zda počet prodejen určité firmy závisí na okrese, ve kterém se nachází (uvažujte obvyklou hladinu významnosti). Určete koeficient vyjadřující sílu závislosti.

Okres	Počet prodejen
A	5,7,5,6,9
B	9,8,8,7,10
C	17,18,18,21,19

- řešení

y_{ij}	y_i	n_i	$\overline{y}_i$	$\overline{y}$	$(y_{ij}-\overline{y}_i)^2$	$(\overline{y}_i-\overline{y})^2 \cdot n_i$
5	1	5	6,4	11,13	1,96	112,02
7					0,36	
5					1,96	
6					0,16	
9					6,76	
9	2	5	8,4		0,36	37,36
8					0,16	
8					0,16	
7					1,96	
10					2,56	
17	3	5	18,6		2,56	278,76
18					0,36	
18					0,36	
21					5,76	
19					0,16	
Σ	-	**15**	-	-	**25,60**	**428,13**

  - výpočet testového kritéria

$$F = \frac{\frac{428{,}13}{3-1}}{\frac{25{,}6}{15-3}} = 100{,}34$$

  - kritická hodnota

$$F_{0{,}95}[2;12] = 3{,}88$$

    - zamítáme H_0, závislost existuje
  - koeficient, který měří intenzitu závislosti

$P^2 = \frac{428{,}13}{453{,}73} = 0{,}94$ (počet prodejen je velmi silně ovlivňován okresem)

- ukázka běžného standardizovaného výstupu ze SW

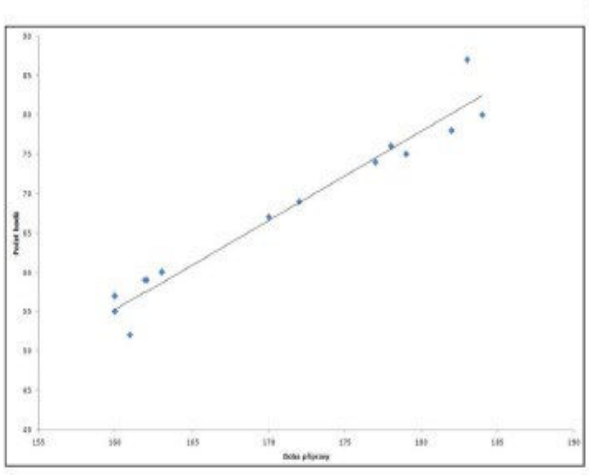
ANOVA Table for pocet_prodejen by okres

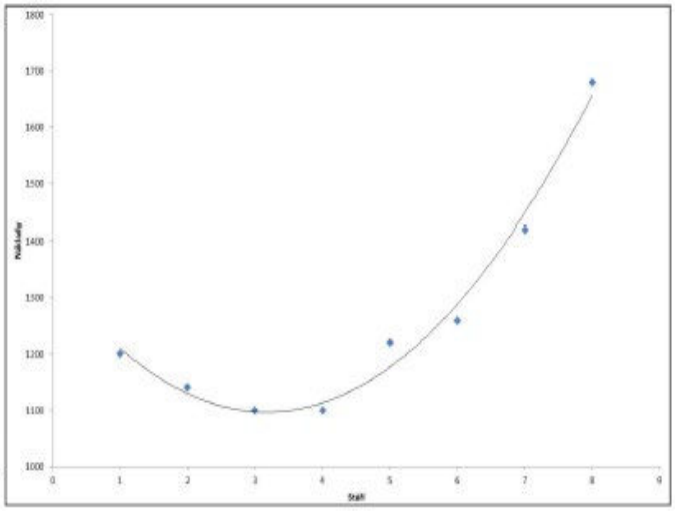
Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Between groups	428,133	2	214,067	100,34	0,0000
Within groups	25,6	12	2,13333		
Total (Corr.)	453,733	14			

Zkoumání závislosti mezi kvantitativními proměnnými

- druh závislosti
  - pevná (funkční) = změně jednoho znaku jednoznačně odpovídá změna druhého znaku (podle funkčního vztahu) – v ekonomických veličinách jich je minimum
  - volná (statistická) = změnám jedné veličiny odpovídají změny druhé veličiny tak, že změna jedné proměnné zvýší pravděpodobnost změny druhé proměnné
- směr závislosti
  - jednostranné závislosti → regresní analýza
  - oboustranné (vzájemné) závislosti → korelační analýza
- bodový diagram
  - slouží ke grafickému zobrazení závislosti
  - první graf obsahuje počet bodů a dobu přípravy
  - druhý stáří automobilu a vynaložené náklady





  - z grafu je možné usuzovat na
    - průběh závislosti (lineární, nelineární)
    - intenzitu závislosti (podle kolísání hodnot kolem křivky)

Regresní analýza

- zkoumá jednostrannou závislost: příčina – následek)
- slouží k popisu statistických závislostí
- cíl: pomocí hodnot jedné či více proměnných X (kvantitativní) odhadovat hodnoty proměnné Y (kvantitativní)
- jednoduchá regresní analýza
    - Y ... vysvětlovaná (závislá) proměnná (následek)
    - X ... vysvětlující (nezávislá) proměnná (příčina – ovlivňuje, ale zpětně není ovlivňovaná)
    - y=f(x)
- vícenásobná regresní analýza
    - Y ... vysvětlovaná (závislá) proměnná
    - $X_1, X_2, ..., X_k$ ... vysvětlující (nezávislé) proměnné
    - $y=f(x_1, x_2, ..., x_k)$
- teoretický regresní model = model, který je v populaci, neznáme ho, chceme ho odhadovat
    - popisuje vztah mezi proměnnými v populaci

      $$y_i = Y_i + \varepsilon_i \quad i = 1,2, \dots, n$$

        - teoretická regresní funkce:

      $$Y_i = f(x, \beta_0, \beta_+ \dots, \beta_k)$$

        - neznámé parametry regresních funkcí:

      $$\beta_0, \beta_1, \dots, \beta_k$$

        - nesystematická (náhodná) složka:

      $$\varepsilon_i$$

- předpoklady o náhodné složce

$$\varepsilon_i \sim \mathrm{N}(0; \sigma^2)$$

$$\mathrm{E}(\varepsilon_i) = 0$$

$$\mathrm{D}(\varepsilon_i) = \sigma^2 = \mathrm{const.}$$

$$\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$$

    - budeme chtít, aby byla normálně rozdělená (nulová střední hodnota a konstantní rozptyl – chceme, aby se změnou hodnot proměnné X nereagoval, byl konstantní, se neměnil rozptyl proměnné Y)= homoskedasticita
    - kovariance těch dvou složek se rovná 0
- základní typy regresních funkcí
    - lineární z hlediska regresních parametrů:
        - β_0 = konstanta
        - β_1 = směrnice/sklon

regresní přímka: $Y = \beta_0 + \beta_1 x$,

regresní rovina: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$,

regresní nadrovina: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K$,

regresní hyperbola: $Y = \beta_0 + \beta_1 \frac{1}{x}$,

regresní logaritmická funkce: $Y = \beta_0 + \beta_1 \ln x$,

regresní parabola: $Y = \beta_0 + \beta_1 x + \beta_2 x^2$.

- nelineární z hlediska regresních parametrů (převoditelné)

regresní mocninná funkce: $Y = \beta_0 x^{\beta_1}$,

regresní exponenciální funkce: $Y = \beta_0 \beta_1^{\ x}$

- $\hat{y}_i = f(\boldsymbol{x}, b_0, b_1, ..., b_k)$ => **výběrová** (empirická) regresní funkce
- b_0, b_1, ..., b_k = výběrové regresní parametry, které budou představovat odhady populačních regresních parametrů

$$y_i = \hat{y}_i + e_i$$

- e_i = reziduum = nevysvětlený zbytek, rozdíl mezi skutečnou a modelovou hodnotou

$$e_i = y_i - \hat{y}_i$$

- metoda k odhadu regresních parametrů
  - princip metody nejmenších čtverců

$$\sum_{i=1}^{n} e_i^{\ 2} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min.$$

  - princip metody nejmenších čtverců pro přímku

$$\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 = \min.$$

    - odhad parametru β_1

$$b_1 = b_{yx} = \frac{n\sum y_i x_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{\overline{xy} - \overline{xy}}{\overline{x^2} - \bar{x}^2}$$

    - odhad parametru β_0

$$b_0 = \frac{\sum y_i \sum x_i^2 - \sum y_i x_i \sum x_i}{n\sum x_i^2 - (\sum x_i)^2} = \bar{y} - b_1 \bar{x}$$

    - směrnice říká, jak se v průměru změní hodnota závislé proměnné Y, jestliže se hodnota nezávislé proměnné změní o jednu svoji jednotku
    - říká nám, jaký bude sklon přímky (úrovňová konstanta říká, jaké bude její posunutí)

- rozklad celkového součtu čtverců
  - celkový

    $$S_y = S_{y,R} + S_{y,T}$$
  - reziduální – vznikne jako rozdíl skutečné a modelové hodnoty, umocněno na druhou a sečteno; představuje nevysvětlený zbytek variability, to co ten model není schopen zachytit

    $$S_{y,R} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
  - teoretický (modelový) – od modelových hodnot odečteme průměr a umocníme na druhou; popisuje schopnost modelu vysvětlit variabilitu vysvětlované proměnné

    $$S_{y,T} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$
- koeficient k hodnocení kvality modelu – pokud jej vynásobíme 100, tak říká, kolik % variability vysvětlované proměnné se podařilo vysvětlit pomocí zvoleného regresního modelu

$$I^2 = R^2 = \frac{S_{y,T}}{S_y} \qquad I^2 \in \langle 0;1 \rangle$$

$$I^2_{upr.} = 1 - (1 - I^2)\frac{n-1}{n-p}$$

- jestliže porovnáváme kvalitu modelů s různým počtem regresních parametrů – upravený index determinace
- model je tím lepší, čím vyšší hodnoty tento koeficient nabývá
- celkový F-test (test o modelu)
  - hodnotí model jako celek
  - H_0 říká, že model jako celek je úplně špatně – ani jedna z proměnných nepůsobí na námi vysvětlovanou proměnnou
  - testové kritérium je založeno na podílu součtu čtverců
  - p = počet parametrů dané regresní funkce

H_0	H_1	Testové kritérium	Kritický obor
$\beta_0 = c$ $\beta_1 = 0$ ... $\beta_k = 0$	non H_0	$F = \frac{\frac{S_T}{p-1}}{\frac{S_R}{n-p}}$ $\quad F \sim F(p-1, n-p)$	$W_\alpha = \{F \geq F_{1-\alpha}\}$

- dílčí t-testy (testy o regresních parametrech)
  - dílčí t-testy slouží k testování významnosti jednotlivých vysvětlujících proměnných a konstanty v modelu samostatně, bude jich celkem p

H_0	H_1	Testové kritérium	Kritický obor
$\beta_i = 0$	$\beta_i \neq 0$	$t = \frac{b_i}{s_{b_i}} \qquad t \sim t(n-p)$	$W_\alpha = \{\lvert t\rvert > t_{1-\alpha/2}\}$

- $s(b_i)$ = směrodatná chyba odhadu

- korelační koeficient = číslo z intervalu od -1 do 1, které vyjadřuje míru lineární závislosti mezi proměnnými, kladné hodnoty znamenají přímo úměrnou závislost, záporné hodnoty nepřímo úměrnou závislost, hodnota 0 lineární nezávislost (čím jsou hodnoty bližší krajním hodnotám tím je závislost silnější)

$$r_{yx} = r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}} = \frac{\overline{xy} - \bar{x}\bar{y}}{\sqrt{(\overline{x^2} - \bar{x}^2)(\overline{y^2} - \bar{y}^2)}}$$

$$r_{xy} \in \langle -1; 1 \rangle$$

  - s_{xy} = kovariance = míra, která vyjadřuje závislost (hodnoty od mínus nekonečna, do nekonečna, nejsme schopni přesně identifikovat míru)
- výpočet R^2 z hodnoty korelačního koeficientu
  - $R^2 = r_{xy}^2$
  - znaménko korelačního koeficientu – podle parametru β_1
- test o korelačním koeficientu
  - ρ = korelační koeficient v celé populaci
  - testovaná hypotéza říká, že proměnné jsou lineárně nezávislé
  - když zamítneme testovanou hypotézu = proměnné jsou korelované, je mezi nimi statisticky významná závislost
  - testové kritérium je založené na hodnotě vypočteného korelačního koeficientu

H_0	H_1	Testové kritérium	Kritický obor
$\rho_{yx} = 0$	$\rho_{yx} \neq 0$	$t = \frac{r_{yx}\sqrt{n-2}}{\sqrt{1-r_{yx}^2}} \qquad t \sim t(n-2)$	$W_\alpha = \{\lvert t\rvert \geq t_{1-\alpha/2}\}$

- uvažujeme regresní přímky (v případě vzájemné závislosti)

$$Y = \beta_0 + \beta_1 x \qquad \hat{y} = b_0 + b_{yx} x$$

$$X = \beta_0 + \beta_1 y \qquad \hat{x} = b_0 + b_{xy} y$$

- přímky X a Y se nazývají sdružené regresní přímky
- koeficienty b_{yx} a b_{xy} se nazývají sdružené regresní koeficienty (nemůže nastat situace, kdy by sdružené regresní koeficienty měli různá znaménka)

$$b_{xy}b_{yx} = r_{xy}^2 \qquad r_{yx} = r_{xy} = \sqrt{b_{yx}b_{xy}}$$

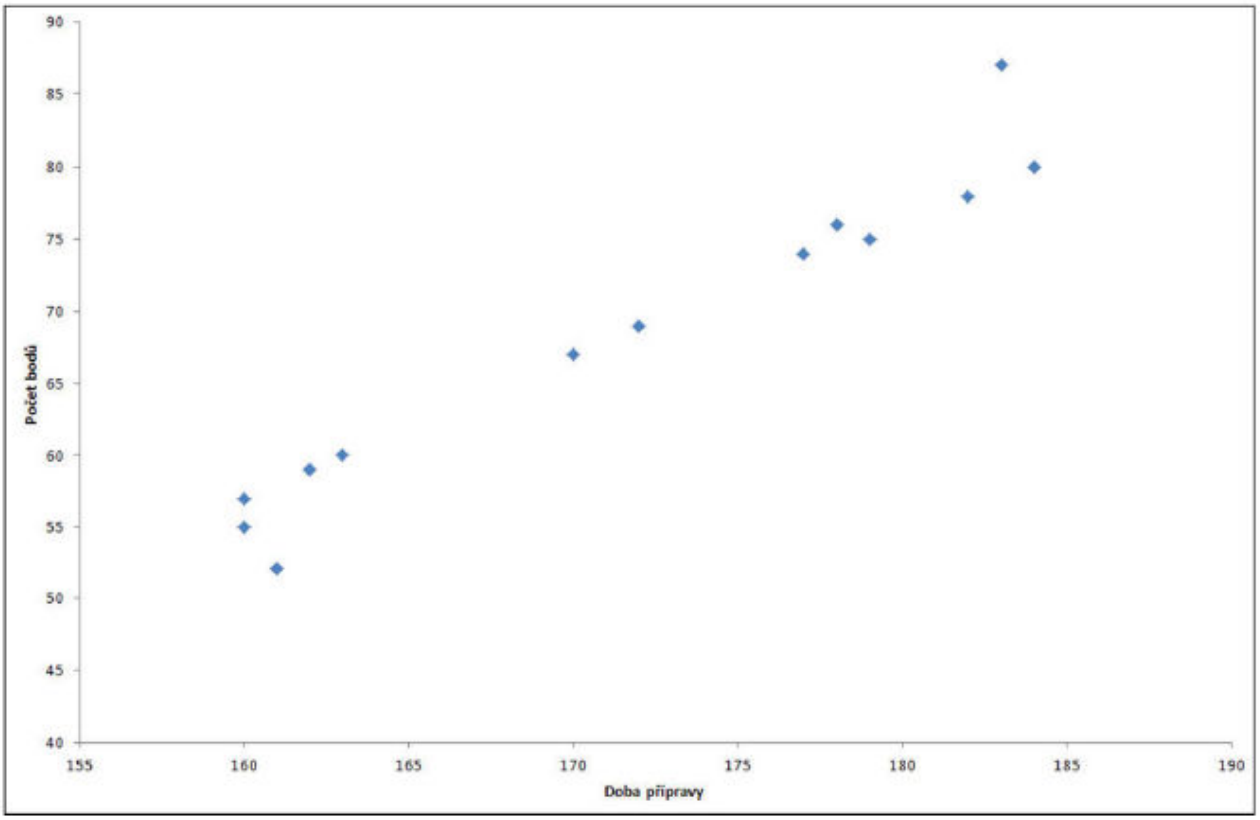
Sdružené regresní koeficienty		*Korelační koeficient* r_{xy}
b_{yx}	b_{xy}	
+	+	+
-	-	-

Příklad číslo 1

U 13 náhodně vybraných studentů byla pomocí experimentu zjišťována doba přípravy na určitý test (v minutách) a počet dosažených bodů. Pomocí regresní přímky vyjádřete závislost počtu bodů na době přípravy studenta. Zhodnoťte kvalitu regresního modelu, vyjádřete intenzitu závislosti počtu bodů na době přípravy. Odhadněte střední hodnotu počtu bodů studenta, který se připravoval 182 minut.

Doba př.	160	160	162	163	161	170	172	177	179	178	182	184	183
Počet b.	57	55	59	60	52	67	69	74	75	76	78	80	87

- řešení:



i	x_i	y_i	$x_i y_i$	x_i^2
1	160	57	9 120	25 600
2	160	55	8 800	25 600
3	162	59	9 558	26 244
4	163	60	9 780	26 569
5	161	52	8 372	25 921
6	170	67	11 390	28 900
7	172	69	11 868	29 584
8	177	74	13 098	31 329
9	179	75	13 425	32 041
10	178	76	13 528	31 684
11	182	78	14 196	33 124
12	184	80	14 720	33 856
13	183	87	15 921	33 489
Σ	2 231	889	153 776	383 941

$$b_1 = \frac{13 \cdot 153776 - 2231 \cdot 889}{13 \cdot 383941 - 2231^2} = 1{,}13387$$

$$b_0 = \frac{889}{13} - 1{,}13387 \cdot \frac{2231}{13} = -126{,}204$$

$$Y = -126{,}204 + 1{,}13387 \cdot x$$

- růst doby přípravy o 1 minutu vyvolá v průměru růst počtu bodů o 1,13

i	x_i	y_i	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$	$(y_i - \bar{y})^2$
1	160	57	55,2152	3,1855	129,6095
2	160	55	55,2152	0,0463	179,1479
3	162	59	57,4829	2,3015	88,0710
4	163	60	58,6168	1,9132	70,3018
5	161	52	56,3491	18,9144	268,4556
6	170	67	66,5539	0,1990	1,9172
7	172	69	68,8216	0,0318	0,3787
8	177	74	74,4910	0,2411	31,5325
9	179	75	76,7587	3,0931	43,7633
10	178	76	75,6249	0,1407	57,9941
11	182	78	80,1603	4,6671	92,4556
12	184	80	82,4281	5,8956	134,9172
13	183	87	81,2942	32,5560	346,5325
Σ	2 231	889	-	73,1853	1 445,0769

$$Y = -126{,}204 + 1{,}13387 \cdot x$$

$$S_y = 1445{,}0769$$

$$S_{y,R} = 73{,}1853$$

$$S_{y,T} = 1445{,}0769 - 73{,}1853 = 1371{,}8994$$

$$R^2 = \frac{1371{,}8994}{1445{,}0769} = 1 - \frac{73{,}1853}{1445{,}0769} = 0{,}9494$$

$$r_{yx} = \sqrt{R^2} = \sqrt{0{,}9494} = 0{,}9743$$

- test o modelu

$$H_0: \beta_0 = c;\ \beta_1 = 0$$
$$H_1: \text{non } H_0.$$

$$F = \frac{\frac{1371{,}8994}{2-1}}{\frac{73{,}1853}{13-2}} = 206{,}20$$

$$F_{0,95}(1;11) = 4{,}84$$

  - na 5% hladině významnosti zamítáme H_0 – model jako celek nemusí být špatný
- testy o regresních parametrech

$$H_0: \beta_0 = 0 \qquad H_0: \beta_1 = 0$$
$$H_1: \text{non } H_0. \qquad H_1: \text{non } H_0.$$

$$t = \frac{-126{,}204}{13{,}57} = -9{,}30 \qquad t = \frac{1{,}13387}{0{,}0789} = 14{,}3597$$

$$t_{0,975}(11) = 2{,}20$$

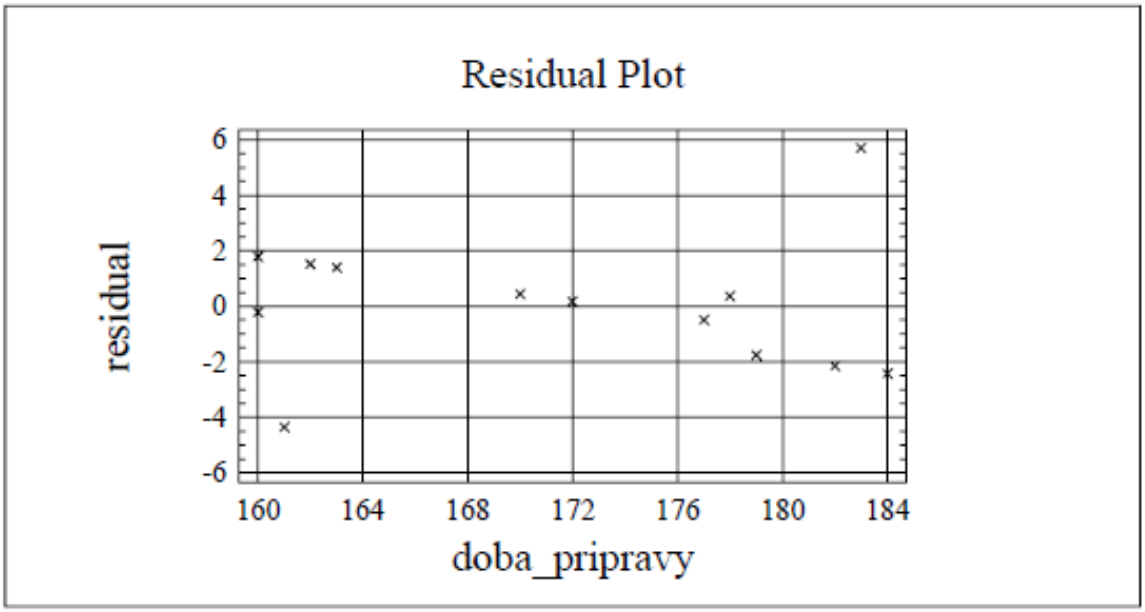
  - na 5% hladině významnosti zmítáme obě H_0
- standardizovaný výstup ze SW

```
Regression Analysis - Linear model: Y = a + b*X
-------------------------------------------------------------------------------
Dependent variable: pocet_bodu
Independent variable: doba_pripravy
-------------------------------------------------------------------------------
                             Standard          T
Parameter       Estimate         Error       Statistic        P-Value
-------------------------------------------------------------------------------
Intercept       -126,204          13,57       -9,30028         0,0000
Slope            1,13387      0,0789619        14,3597         0,0000
-------------------------------------------------------------------------------


                          Analysis of Variance
-------------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square    F-Ratio       P-Value
-------------------------------------------------------------------------------
Model                      1371,89      1      1371,89     206,20        0,0000
Residual                   73,1853     11      6,65321
-------------------------------------------------------------------------------
Total (Corr.)              1445,08     12

Correlation Coefficient = 0,974349
R-squared = 94,9355 percent
```

- graf reziduí

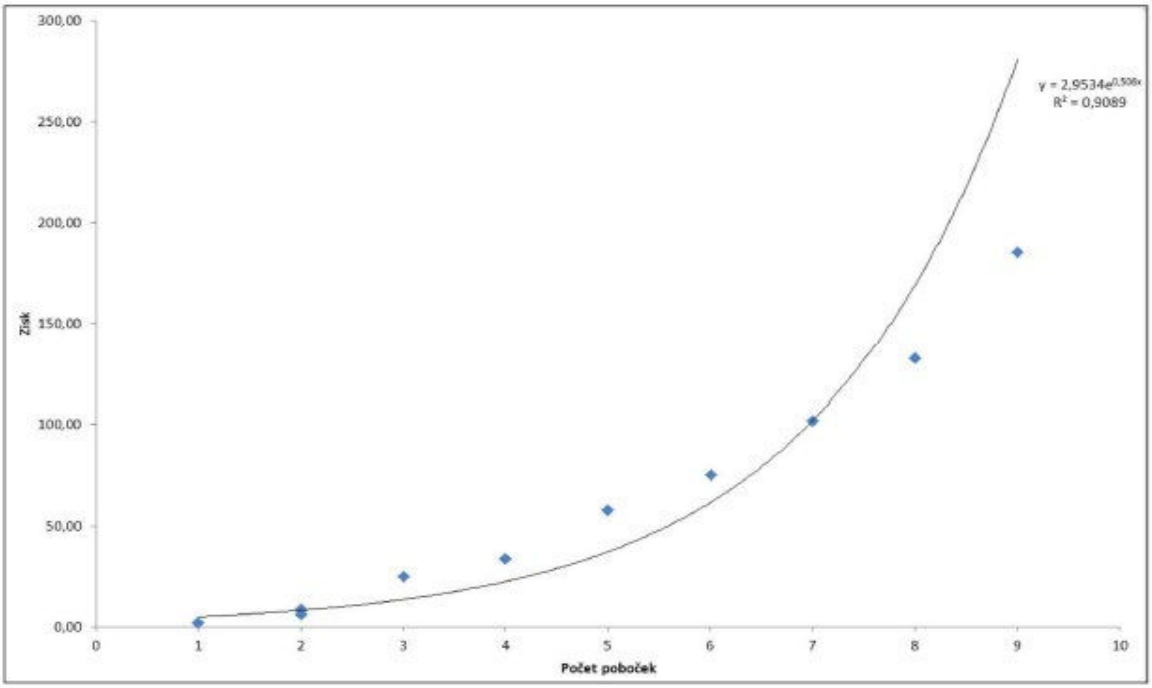


- funkce nelineární z hlediska regresních parametrů

$$Y = \beta_0 \beta_1^{\,x}$$

  - regresní exponenciální funkce
    - je třeba zlinearizovat
      $\ln Y = \ln \beta_0 + x\beta_1$
    - po úpravě lze zapsat
      $Y^* = \beta_0^* + x\beta_1^*$
    - odlogaritmování parametrů
  - $\beta_0 = e^{\beta_0^*} \quad \beta_1 = e^{\beta_1^*}$

- graf: příklad regresní exponenciální funkce



- regresní parabola

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2$$

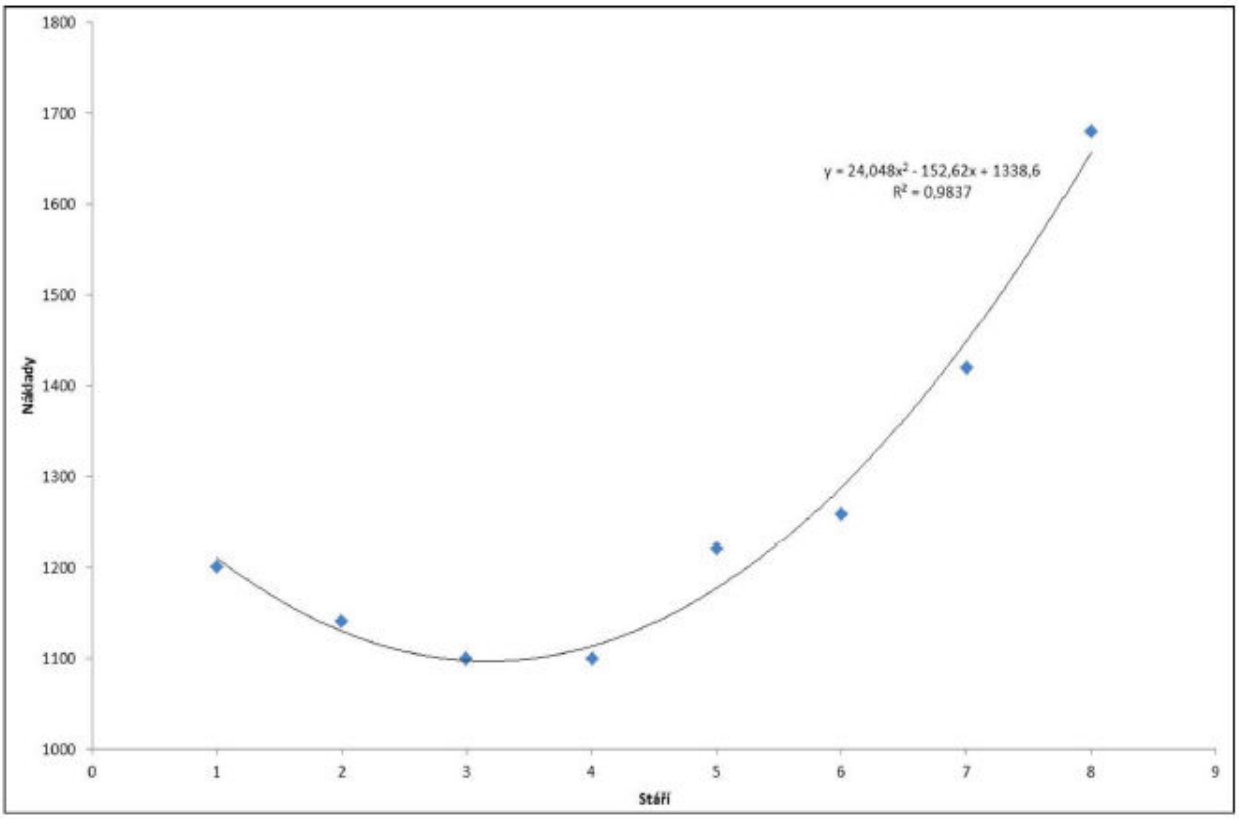
  - odhad parametrů MNČ
- příklad: výstup ze SW

```
Polynomial Regression Analysis
----------------------------------------------------------------------------
Dependent variable: Naklady
----------------------------------------------------------------------------
                                     Standard          T
Parameter                Estimate        Error      Statistic        P-Value
----------------------------------------------------------------------------
CONSTANT                 1338,57       41,5655       32,2039          0,0000
Stari                   -152,619       21,1918      -7,20179          0,0008
Stari^2                  24,0476       2,29857        10,462          0,0001
----------------------------------------------------------------------------

                          Analysis of Variance
----------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square     F-Ratio      P-Value
----------------------------------------------------------------------------
Model                     268162,0      2     134081,0      151,06       0,0000
Residual                    4438,1      5      887,619
----------------------------------------------------------------------------
Total (Corr.)             272600,0      7

R-squared = 98,3719 percent
R-squared (adjusted for d.f.) = 97,7207 percent
```

- graf: příklad regresní paraboly



- vícenásobná regresní analýza

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K$$

  - odhad parametrů MNČ
- příklad: výstup ze SW

```
Multiple Regression Analysis
----------------------------------------------------------------------------
Dependent variable: cena
----------------------------------------------------------------------------
                                     Standard          T
Parameter               Estimate       Error       Statistic        P-Value
----------------------------------------------------------------------------
CONSTANT                 421,655      17,8932        23,5651          0,0000
pocet_km                -1,02969     0,293674       -3,50623          0,0027
stari                   -3,28191     0,392052        -8,3711          0,0000
----------------------------------------------------------------------------

                          Analysis of Variance
----------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square    F-Ratio       P-Value
----------------------------------------------------------------------------
Model                   284650,0       2     142325,0      100,22        0,0000
Residual                 24142,1      17      1420,12
----------------------------------------------------------------------------
Total (Corr.)           308792,0      19

R-squared = 92,1817 percent
R-squared (adjusted for d.f.) = 91,262 percent
```

- vyhodnocení
  - celkový F-test
    - hypotézu H_0 celkového F-testu zamítáme – model nemusí být špatný
    - existuje alespoň 1 vysvětlující proměnná, která má význam
  - dílčí t-testy
    - beta 0 – je významná, je nenulová
    - beta 1 – také významná
    - beta2 – také významná
  - pro srovnání funkcí s různým počtem vysvětlujících proměnných, je nutné využití upraveného koeficientu determinace!!!!
- multikolinearita
  - vzájemná lineární závislost (která je škodlivá) mezi vysvětlujícími proměnnými X
  - čím vyšší závislost – tím vyšší problém
  - pokud r_{xixj} (mezi libovolnou dvojicí vysvětlujících proměnných) > 0,75 (0,8)
  - je třeba vyřadit tu proměnnou, jejíž korelační koeficient s vysvětlovanou proměnnou je nižší
  - postup
    - vypočítat r_{xixj} – identifikace multikolinearity
    - vypočítat r_{yxi} – identifikace vyřazované proměnné
    - vyřadím tu proměnnou, jejíž KK ve vztahu k proměnné je menší
- může nastat situace kdy celkový F-test a dílčí t-testy jsou v rozporu – téměř ve všech případech je to důsledkem multikolinearity → spočítat korelační koeficient

8. Časové řady

- značení časové řady
  - y_t
- definice časové řady
  - posloupnost hodnot sledovaného ukazatele, která je jednoznačně uspořádána z hlediska času
- typy časových řad
  - intervalové a okamžikové
    - intervalová časová řada obsahuje řadu hodnot sledovaného ukazatele za určitý interval, tj. obsahuje tokové veličiny (př. HDP/rok, průměrný příjem/měsíc, zisk, tržby/rok,měsíc)
    - okamžikové časové řada obsahuje řadu hodnot sledovaného ukazatele k určitému okamžiku, tj. stavové veličiny (př. stav bankovnímu účtu k určitému dni, počet zaměstnanců k určitému dni)
  - krátkodobé a dlouhodobé
    - krátkodobá řada je řada hodnot sledovaného ukazatele, kde je perioda mezi dvěma po sobě jdoucími záznamy kratší než 1 rok
    - dlouhodobá řada je řada hodnot sledovaného ukazatele, kde je perioda mezi dvěma po sobě jdoucími záznamy roční a delší
- stanovení průměrné hodnoty ČR
  - intervalové časové řady – aritmetický průměr

$$\bar{y} = \sum_{t=1}^{n} y_t \,/n$$

  - okamžikové časové řady – chronologický průměr

$$\bar{y} = \frac{\frac{y_1+y_2}{2} + \frac{y_2+y_3}{2} + \ldots + \frac{y_{n-1}+y_n}{2}}{n-1} = \frac{\frac{1}{2}y_1 + \sum_{t=2}^{n-1} y_t + \frac{1}{2}y_n}{n-1} \qquad \bar{y} = \frac{\frac{y_1+y_2}{2}d_1 + \frac{y_2+y_3}{2}d_2 + \ldots + \frac{y_{n-1}+y_n}{2}d_{n-1}}{d_1+d_2+\ldots+d_{n-1}}$$

- základní míry dynamiky (pohybu v časové řadě)
  - první diference (absolutní přírůstek)
    - absolutní = ve stejných měrných jednotkách (relativní v %)
    - rozdíl 2 po sobě jdoucích záznamů

$$\Delta y_t = y_t - y_{t-1}$$

  - průměrná první diference (průměrný absolutní přírůstek)
    - aritmetický průměr – sečteme diference a vydělíme je jejich počtem

$$\overline{\Delta} = \frac{\sum_{t=2}^{n} \Delta y_t}{n-1} = \frac{(y_2-y_1)+(y_3-y_2)+\ldots+(y_n-y_{n-1})}{n-1} = \frac{y_n-y_1}{n-1}$$

- koeficient růstu – podíl dvou po sobě jdoucích hodnot, koeficient růstu nám po vynásobení 100 říká, k jaké procentuální změně sledovaného ukazatele v roce t oproti roku t-1 došlo

$$k_t = \frac{y_t}{y_{t-1}}$$

- průměrný koeficient růstu

$$\bar{k} = \sqrt[n-1]{k_2 k_3 \ldots k_n} = \sqrt[n-1]{\frac{y_n}{y_1}}$$

- relativní přírůstek – specifická interpretace koeficientu růstu, která říká, o kolik procent došlo ke změně

$$\delta_t = k_t - 1$$

- průměrný relativní přírůstek – průměrná meziroční změna (o kolik procent docházelo ke změně)

$$\bar{\delta} = \bar{k} - 1$$

- dekompozice (rozklad časové řady)
  - aditivní model (součtový model) = jednotlivé složky sčítáme

$$y_t = T_t + S_t + C_t + \varepsilon_t$$

  - multiplikativní model (součinový model) = jednotlivé složky násobíme

$$y_t = T_t \, S_t \, C_t \, \varepsilon_t$$

    - T_t = trendová složka (trend) = vyjadřuje dlouhodobé změny ve vývoji chování sledovaného ukazatele, udává tendenci (směr) vývoje (klesající, konstantní, rostoucí)
    - S_t = sezónní složka (sezónnost) = představuje pravidelně se opakující kolísání v časové řadě okolo trendu, které je způsobeno střídáním ročních období nebo zvyky ve společnosti; vyskytuje se pouze v krátkodobých časových řadách
    - C_t = cyklická složka (cyklus) = představuje nepravidelné kolísání v časové řadě okolo trendu, které je (způsobeno) ovlivněno zejména jednotlivými fázemi hospodářského cyklu; délka jednoho cyklu je delší než 1 rok
    - ε_t = nesystematická (náhodná) složka = nepravidelný zbytek v rámci kolísání v časové řady, bez systematického charakteru; na jeho pravděpodobnostní chování si budeme klást nějaké požadavky

Trendová složka

- regresní (klasický) přístup k modelování tendru
  - $T_t = f(t)$ = trend je matematickou funkcí času
    - vysvětlovaná proměnná = čas = umělá posloupnost od 1 do n

- základní trendové funkce

  lineární trend: $T_t = \beta_0 + \beta_1 t$,

  kvadratický trend: $T_t = \beta_0 + \beta_1 t + \beta_2 t^2$,

  exponenciální trend: $T_t = \beta_0 \beta_1^{\,t}$,

  logaritmický trend: $T_t = \beta_0 + \beta_1 \ln t$,

  hyperbolický trend: $T_t = \beta_0 + \beta_1 \frac{1}{t}$.

- metody k odhadu regresních parametrů trendové funkce
  - MNČ (viz regresní analýza)
    - vysvětlující proměnná je čas – t = 1, 2, ..., n
- volba vhodného trendu
  - věcně ekonomická kritéria
  - analýza grafu hodnot časové řady
  - analýza měr dynamiky
    - např. pro lineární trendovou funkci $\Delta y_t = const.$
  - kritéria „úspěšnosti“ zvolené trendové funkce

$$ME = \sum_{t=1}^{n} (y_t - \hat{T}_t)\,/n$$

$$MSE = \sum_{t=1}^{n} (y_t - \hat{T}_t)^2 /n$$

  - kritéria z regresní analýzy

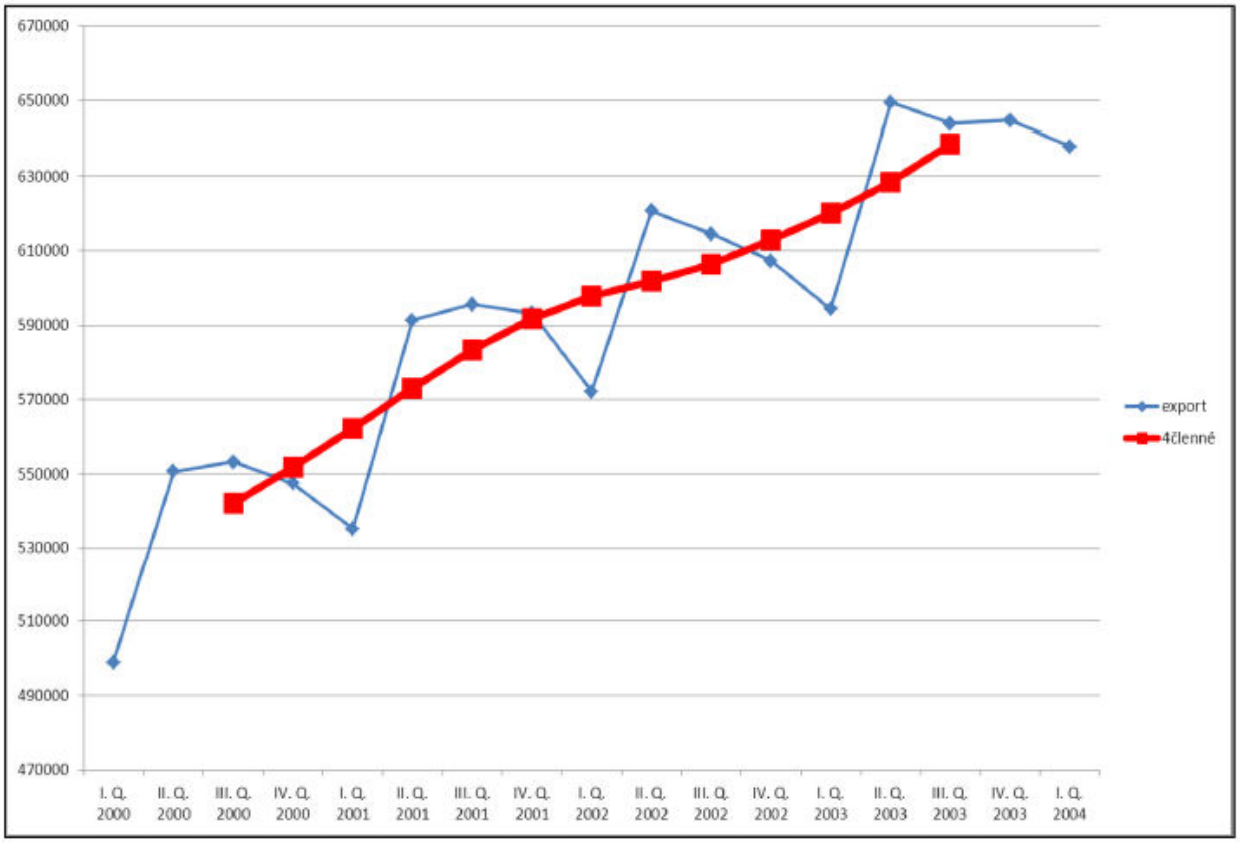
Adaptivní přístupy k modelování trendu

- metoda klouzavých průměrů
  - prosté klouzavé průměry
    - v případě, že lze předpokládat, že na jednotlivých klouzavých částech je definován lineární trend

$$\hat{T}_t = \frac{\sum_{i=-p}^{p} y_{t+i}}{m} = \frac{y_{t-p} + y_{t-p+1} + \ldots + y_t + \ldots + y_{t+p}}{m}$$

    - délka klouzavého průměru je označena m – volí se podle počtu sezón nebo cyklických výkyvů v časové řadě
    - pro sudé m – tzv. centrování = spočívá v tom, že výsledné hodnoty ještě jednou zprůměrujeme (po 2) – aby se ty odhady daly jednoznačně přiřadit k období
  - vážené klouzavé průměry
    - v případě, že lze předpokládat, že na jednotlivých klouzavých částech je definován parabolický trend

- graf: příklad 4-členné klouzavé průměry



- exponenciální vyrovnání (= další přístup k modelování trendu – jako KP) ≠ exponenciální regresní funkcí!!!
    - při standardní MNČ při odhadování trendu je každému pozorování časové řady přiřazena stejná váha (význam)
    - znamená, že každému pozorování časové řady je přiřazena určitá váha, která klesá exponenciálně do minulosti
    - nechť α – tzv. vyrovnávací konstanta, α (0;1)
    - odhady parametrů regresní funkce pomocí vážené metody nejmenších čtverců

$$\sum_{k=0}^{n-1}(y_{n-k}-T_{n-k})^2 w_k = \min.$$

$$\Rightarrow \text{kde } k = 0, 1, \ldots, n-1 \text{ je stáří a } w_k \text{ je váha, pro kterou platí}$$

$$w_k = \alpha^k$$

    - pokud předpokládáme, že trend je konstantní
        - jednoduché exponenciální vyrovnání
    - pokud předpokládáme, že trend je lineární
        - dvojité exponenciální vyrovnání
    - pokud předpokládáme, že trend je kvadratický
        - trojité exponenciální vyrovnání

Sezónní složka

- empirické sezónní odchylky = rozdíl mezi reálnou hodnotou časové řady a trendem
  - říká, o kolik měrných jednotek se v průměru mění hodnota časové řady oproti trendu díky sezónnosti
  - předpokládáme konstantní sezónnost
  - předpokládáme aditivní dekompozici

$$SO = y_t - \hat{T}_t$$

- empirické sezónní indexy = podíl reálné hodnoty časové řady a trendu
  - říká, o kolik se v průměru mění hodnota časové řady oproti trendu; v relativním vyjádření (v %)
  - v případě, že předpokládáme proporcionální sezónnost (mění svůj průběh v závislosti na trendu; pokud roste trend, roste sezónnost)
  - předpokládáme multiplikativní dekompozici

$$SI = \frac{y_t}{\hat{T}_t}$$

- regresní přístup k modelování sezónnosti
  - předpokládáme určitý tvar trendové funkce (např. přímka)
  - do regresního modelu přidáme umělé proměnné
  - $x_{jt} = 1$ – pro j-tou sezónu
  - $x_{jt} = 0$ – jinak
  - je jich o 1 méně, než je počet sezón (u čtvrtletní ČR – 3)
  - vytvoříme model (předpokládáme lineární trend)
    - $y_t = T_t + S_t + \varepsilon_t = \beta_0 + \beta_1 t + \alpha_1 x_{1t} + \alpha_2 x_{2t} + \cdots + \alpha_{r-1} x_{r-1t}$
    - α_j = j-tý sezónní faktor
    - odhad všech parametrů (α_j,β_i) MNČ
  - odhad empirických sezónních odchylek
    - součet sezónních odchylek musí být vždy nula

      $$\sum_{i=1}^{r} S_i = 0$$

    - $\bar{a} = \frac{a_1 + \cdots + a_{r-1}}{r} \quad S_i = a_i - \bar{a} \quad S_r = -\bar{a}$
    - r = počet sezón; i = 1, ..., r-1
  - úpravy konstanty u trendu
    - $\hat{T}_t = (b_0 + \bar{a}) + b_1 t$
  - pro čtvrtletní ČŘ (lineární trend)
    - $y_t = \beta_0 + \beta_1 t + \alpha_1 x_{1t} + \alpha_2 x_{2t} + \alpha_3 x_{3t} + \varepsilon_t$
    - $\bar{a} = \frac{a_1 + a_2 + a_3}{4} \quad S_1 = a_1 - \bar{a} \quad S_4 = -\bar{a}$
    - α = populační parametr
    - a = odhad aplikací MNČ
- extrapolace (předpověď) v časových řadách
  - určení odhadu budoucí hodnoty časové řady na základě znalosti vývoje v minulosti
  - časová řada musí být dostatečně dlouhá
  - délka předpovědi musí být přiměřená
  - co nejjednodušší funkci (pozor na její průběh)

- vychází z deterministického přístupu, že analyzovaná ČŘ do budoucna nemění své chování

Příklad číslo 1

Stanovte základní míry dynamiky sledované časové řady za celé sledované období, tj. absolutní, relativní, průměrný absolutní a průměrný relativní přírůstek.

t	y_t
2002	10
2003	12
2004	13
2005	9
2006	16

- řešení

t	y_t	Δ	$\overline{\Delta}$	k_t	$\overline{k}$	δ_t	$\overline{\delta}$
2002	10	-	1,50	-	1,125	-	12,5%
2003	12	2		1,200		20,0%	
2004	13	1		1,083		8,3%	
2005	9	-4		0,692		-30,8%	
2006	16	7		1,778		77,8%	

Příklad číslo 2

Předpokládejte lineární trend. Stanovte empirické sezónní odchylky následující časové řady:

Období	export
I. Q. 2000	498 859
II. Q. 2000	550 590
III. Q. 2000	553 159
IV. Q. 2000	547 450
I. Q. 2001	535 196
II. Q. 2001	591 285
III. Q. 2001	595 546
IV. Q. 2001	593 228
I. Q. 2002	572 149
II. Q. 2002	620 720
III. Q. 2002	614 504
IV. Q. 2002	607 296
I. Q. 2003	594 377
II. Q. 2003	649 415
III. Q. 2003	643 836
IV. Q. 2003	644 760
I. Q. 2004	637 921

- řešení:

Multiple Regression Analysis

Dependent variable: yt

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	518308,0	4446,32	116,57	0,0000
t	7987,53	314,403	25,4054	0,0000
x1t	-22495,6	4229,85	-5,31828	0,0002
x2t	20794,1	4490,57	4,63061	0,0006
x3t	11565,3	4457,43	2,59461	0,0235

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	2,94799E10	4	7,36997E9	186,39	0,0000
Residual	4,74475E8	12	3,95396E7		
Total (Corr.)	2,99543E10	16			

R-squared = 98,416 percent
R-squared (adjusted for d.f.) = 97,888 percent

$$\bar{a} = \frac{-22495{,}6 + 20794{,}1 + 11565{,}3}{4} = 2\,465{,}95$$

$$S_1 = -22\,495{,}6 - 2\,465{,}95 = -24\,961{,}55 \qquad S_2 = 20\,794{,}1 - 2\,465{,}95 = 18\,328{,}15$$

$$S_3 = 11\,565{,}3 - 2\,465{,}95 = 9\,099{,}35 \qquad S_4 = -2\,465{,}95$$

9. Indexní analýza

- definice indexu a diference
  - index = bezrozměrné číslo, které vyjadřuje změnu sledovaného ukazatele v relativním vyjádření
  - diference = číslo, které vyjadřuje změnu sledovaného ukazatele ve stejných jednotkách
- bazický index a řetězový index
  - řada bazických indexů srovnává vývoj sledovaného ukazatele tak, že všechny hodnoty v čase porovnáváme s jedním předem zvoleným obdobím, které nazýváme báze, neboli základ (většinou báze = první období)
  - řada řetězových indexů zkoumá vývoj sledovaného ukazatele tak, že období srovnání se podle předem stanoveného klíče mění
- individuální indexy – použijeme tehdy, jestliže máme jeden druh výrobku (p = cena, q = množství, objem, Q = tržba; pxq=Q)
  - jednoduché
    - zkoumáme jeden výrobek na jedné pobočce
  - složené
    - zkoumáme jeden výrobek na více pobočkách (za více poboček dohromady)
- souhrnné – použijeme tehdy, jestliže máme více druhů výrobků dohromady

- Paascheho – fixuje druhou veličinu v běžném období
- Laspayresův – fixuje druhou veličinu v základním období
- Fishrerův – geometrický průměr Paascheho a Laspayresova
- pozn. běžné období značíme 1, základní 0

- individuální indexy jednoduché
  - cenový – změna ceny u jednoho druhu výrobku v jedné pobočce

$$i_p = \frac{p_1}{p_0}$$

  - množstevní (objemový) – změna objemu u jednoho druhu výrobku v jedné pobočce

$$i_q = \frac{q_1}{q_0}$$

  - hodnotový – součin i_p a i_q – změna tržby u jednoho výrobku v jedné pobočce

$$i_Q = \frac{Q_1}{Q_0}$$

- individuální indexy složené
  - množstevní (objemový) – změna prodávaného množství jednoho výrobku za všechny pobočky dohromady

$$I_q = \frac{\sum q_1}{\sum q_0}$$

  - hodnotový – změna tržby u jednoho výrobku za všechny pobočky dohromady

$$I_Q = \frac{\sum Q_1}{\sum Q_0}$$

  - cenový – změna průměrné ceny výrobku za všechny pobočky dohromady (aplikace váženého aritmetického průměru)

$$I_p = \frac{\overline{p}_1}{\overline{p}_0} = \frac{\frac{\sum p_1 q_1}{\sum q_1}}{\frac{\sum p_0 q_0}{\sum q_0}} = \frac{\frac{\sum Q_1}{\sum q_1}}{\frac{\sum Q_0}{\sum q_0}}$$

- souhrnné indexy
  - cenové indexy
    - Laspeyresův – fixuje množství v základním období, v čitateli fiktivní tržba – kolik bychom utržili, kdybychom loňské objemy prodali za letošní ceny

$${}_L I_p = \frac{\sum p_1 q_0}{\sum p_0 q_0}$$

    - Paascheho – fixuje množství v běžném období

$${}_P I_p = \frac{\sum p_1 q_1}{\sum p_0 q_1}$$

    - Fisherův – geometrický průměr Laspeyresova a Paascheho

$${}_F I_p = \sqrt{{}_L I_p \cdot {}_P I_p}$$

- množstevní (objemové) indexy
  - Laspeyresův – fixuje cenu v základním období; v čitateli fiktivní tržba, kolik bychom utržili, kdybychom letošní objemy prodali za loňské ceny

$$_L I_q = \frac{\sum p_0 q_1}{\sum p_0 q_0}$$

  - Paascheho – fixuje cenu v běžném období

$$_P I_q = \frac{\sum p_1 q_1}{\sum p_1 q_0}$$

  - Fisherův – geometrický průměr Laspeyresova a Paascheho

$$_F I_q = \sqrt{_L I_q \cdot {}_P I_q}$$

- stanovení hodnot potřebných k výpočtu souhrnných indexů

$$p_1 q_0 = i_p Q_0 = \frac{Q_1}{i_q}$$

$$p_0 q_1 = i_q Q_0 = \frac{Q_1}{i_p}$$

Index spotřebitelských cen

- založen na spotřebitelském koši

Cenové indexy

- indikátor změn v národním hospodářství
- index spotřebitelských cen – k měření tempa inflace
- cenový index
  - pro vybrané výrobky (reprezentanty)
  - u vybraných zpravodajských jednotek (náhodný výběr)
  - v určité periodicitě (měsíční, čtvrtletní)

1. homogenní skupiny výrobků; potravinářské zboží, nepotravinářské (odívání) a služby
2. v každé skupině reprezentant (typický výrobek – předpoklad – cenový vývoj nepodléhá nečekaným výkyvům) – 700

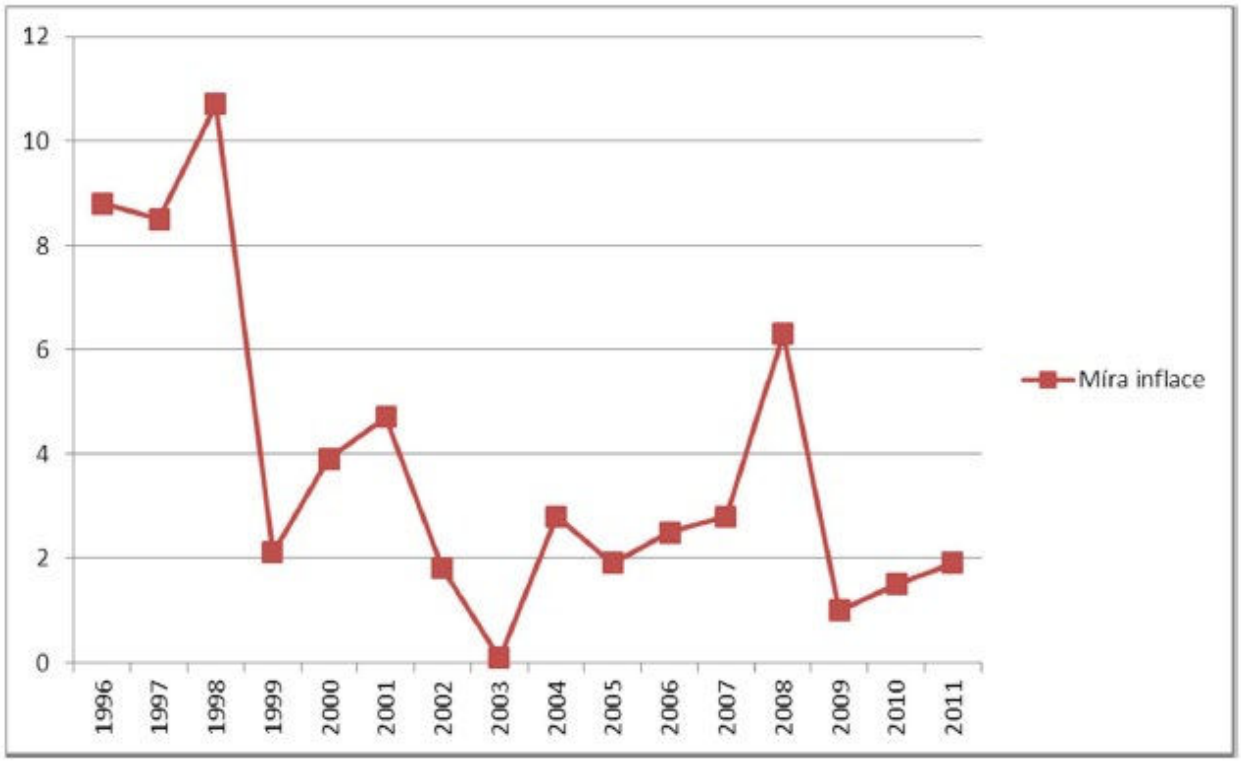
- váhový systém (například na základě údajů statistiky rodinných účtů – platnost cca 5 let)
- soubor reprezentantů + váhový systém = spotřební koš

- indexy cen spotřeby
  - index spotřebitelských cen
  - index životních nákladů
- indexy cen ve výrobě
  - indexy cen průmyslových výrobců
  - indexy cen stavebních prací a stavebních objektů
  - indexy cen zemědělských výrobců

Inflace

- změna cenové hladiny vyjádřená v procentech (relativně)

1. průměrná roční inflace vyjadřuje procentní změnu průměrné cenové hladiny (indexu spotřebitelských cen) za 12 posledních měsíců oproti průměru 12-ti předchozích měsíců
    - tato míra inflace je vhodná při úpravách nebo posuzování průměrných veličin, bere se v úvahu zejména při propočtech reálných mezd, důchodů apod.
    - graf: vývoj průměrné roční míry inflace (v %) v ČR v letech 1996 – 2011: (zdroj dat: ČSÚ)



2. meziroční míra inflace vyjadřuje procentní změnu cenové hladiny ve vykazovaném měsíci daného roku proti stejnému měsíci předchozímu roku
    - jedná se tedy o dosaženou cenovou úroveň, která vylučuje sezónní vlivy tím, že se porovnávají vždy stejné měsíce
    - tato míra inflace je vhodná ve vztahu ke stavovým veličinám, které měří změnu stavu mezi začátkem a koncem období bez ohledu na průběh vývoje během tohoto období
3. měsíční míra inflace vyjadřuje procentní změnu cenové hladiny sledovaného měsíce oproti předchozímu měsíci

Příklad číslo 1

V následující tabulce jsou uvedeny hodnoty HDP (v Kč)v běžných cenách v letech 1995 – 2006. Vypočtěte řadu bazických indexů se základem v roce 1995 a řadu řetězových indexů za celé období.

Ukazatel/Rok	HDP
1995	1466,5
1996	1683,3
1997	1811,1
1998	1996,5
1999	2080,8
2000	2189,2
2001	2352,2
2002	2464,4
2003	2577,1
2004	2814,8
2005	2987,7
2006	3231,6

- řešení
    - řetězové a bazické indexy

Index/Rok	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
$I_{t/t-1}$	-	1,148	1,076	1,102	1,042	1,052	1,074	1,048	1,046	1,092	1,061	1,082
$I_{t/1995}$	1,000	1,148	1,235	1,361	1,419	1,493	1,604	1,680	1,757	1,919	2,037	2,204

Příklad číslo 2

Máme k dispozici následující informace o 3 skupinách výrobků, tržbách za rok 2005 a 2006. Dále máme k dispozici individuální cenové indexy za každý výrobek. Určete, jak se změnila cenová hladina a prodávané množství za všechny výrobky dohromady.

Výrobek	Q_0	Q_1	i_p
A	400	600	1,2
B	400	500	1,0
C	200	300	1,5

- řešení

Výrobek	Q_0	Q_1	i_p	i_p*Q_0	Q_1/i_p
A	400	600	1,2	480	500
B	400	500	1,0	400	500
C	200	300	1,5	300	200
Σ	**1000**	**1400**	-	**1180**	**1200**

$$_{L}I_p = \frac{1180}{1000} = 1{,}18$$

$$_{P}I_p = \frac{1400}{1200} = 1{,}167$$

$$_{F}I_p = \sqrt{1{,}18 \cdot 1{,}167} = 1{,}173$$

$$_{L}I_q = \frac{1200}{1000} = 1{,}20$$

$$_{P}I_q = \frac{1400}{1180} = 1{,}186$$

$$_{F}I_q = \sqrt{1{,}20 \cdot 1{,}186} = 1{,}193$$