Seminarni prace

Popisná statistika, data nesmí být časovou řadou
Zkoumat můžeme třeba mzdy, obraty atd. (takže možná QA?)
Formát pdf, poslat nejpozději den před zkouškou.
Podrobnější informace jsou na ŠISU (soubor statistika_informace_ls_2014-2015.pdf)

2-3 stranky staci, dat nema byt 3 a nema jich byt pul milionu

k te seminarce

1. sehnat si jakakoliv data, ale ne casova rada (ze stataku, menove kurzy z narodni banky atd.) popiseme co jsme porovnavali a odkud jsme to vzali - aspon 10 udaju asi

2. vybereme vhodny prumer, median , modus, variacni odchylku, atd. variacni rozpeti
3. vhodny graf (histogram neni vhodny pro …)
4.

Statistika

Statistický soubor - množina prvků (každý z těchto prvků je statistickou jednotkou) které mají aspoň jednu společnou vlastnost. Těmito jednotkami mohou být lidé, firmy, domy atd.

> **Základní soubor** - obsahuje všechny existující jednotky.
> **Výběrový soubor** - obsahuje pouze vybranou část statistických jednotek

Populace - synonymum pro základní soubor.

Statistická jednotka - konkrétní prvek statistického souboru. Například, statistickou jednotkou může být jeden konkrétní člověk.

Statistický znak - vlastnost statistické jednotky kterou zkoumáme (hledání průměrného věku). Například pokud je statistickou jednotkou člověk, statistickým znakem může být plat, výška, věk atd.

Identifikační znak - společná vlastnost pro všechny jednotky souboru. Indentifikační znak umožňuje určit, zda prvek do statistického souboru patří, nebo nepatří.

Identifikačním znakem může být třeba u žáků, kteří jsou v tuto chvíli v naší třídě například: studovaný předmět, mísnost kde jsme, studovaný předmět, nebo nejvyšší dokončené vzdělání.

Jde o to, co máme společné!

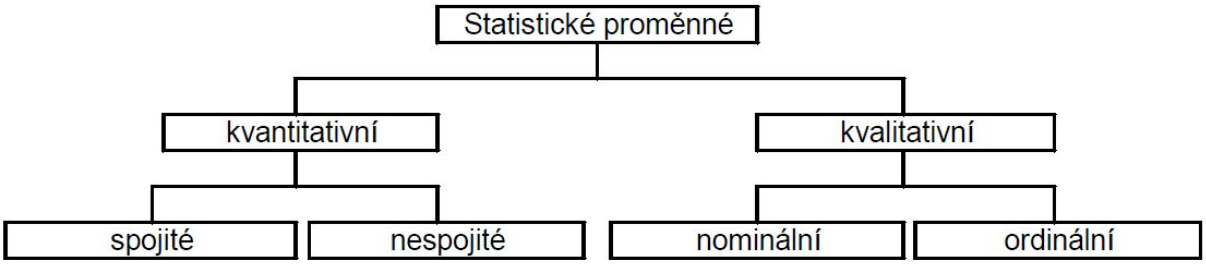
Statistická proměnná - u jednotlivých jednotek v souboru nabývá různých hodnot. Například u žáků ve třídě to může být: pohlaví, počet dětí, počet telefonnů atd.

Statistické proměnné

Slovní (kvalitativní)
> **nominální** - pohlaví, barva (nejde je seřadit)
> **ordinální** - velikosti: malý/střední/velký, vzdělání, hodnosti vojáků, známky ve škole (můžeme je seřadit podle stupně vlastnosti), datum narození, vzdělání

Číselné (kvantitativní)
> **spojité** - nabývají libovolného počtu hodnot - reálná čísla, obrat firmy, výsledky měření, například váha
> **nespojité (diskrétní)** - nabývají hodnot malého počtu jednoznačně izolovaných hodnot. Počet mobilních telefonnů, počet lidí ve třídě, počet dětí, počet vyrobených kazových výrobků.

Příklady:
Známka ve škole (1-4) není číselná proměnná, ale slovní ordinální. S čísly je možné provádět matematické operace, což se známkami nedává smysl.
Datum výroby automobilu v technickém průkazu je také slovní, ordinální. ???



K těm známkám jsem dohledal tohle:

is 5. The set of possible values of school marks has a different structure from that of a set of natural numbers. For school marks the relations =, <, and > are defined, other operations which are meaningful with the set of numbers are not defined. In other words, school marks form a completely ordered set, and nothing more. An even weaker structure is that of tram route numbers. Unlike real numbers, a tram Nr 12 cannot be replaced by two trams numbered 6. Unlike a completely ordered set, the tram Nr 2 is no better than the tram Nr 3, but, at the same time, not a shred worse than the tram Nr 1. When

Popisná statistika

Četnost

absolutní četnost: kolikrát se něco vyskytlo
relativní četnost: v %

Tabulka rozdělení četností pro nespojitý znak

x_i (sledovaný znak) = měsíční obrat firmy (třeba v miliónech korun)
Poznámka: při opakovaných průzkumech je dobré nechat intervaly stejné.

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
SUM	n = 100	1,00		

x_i - interval - například x_1 jsou firmy s obratem od 0 do 20 miliónů korun, x_4 je čtvrtý interval (firmy s obratem od 61 do 80 miliónu korun)
n_i - počet firem v daném intervalu, například v druhém intervalu x_2 je 15 firem (15 firem má obrat mezi 21-40 milióny korun).
N_i - kumulativní četnost je součet prvků v dané skupině a prvků ve všech skupinách nižších (předchozích). Lze použít slova jako maximálně, nebo nejvýše.
p_i - relativní četnost - podíl četnosti v tomto řádku na celku
P_i - kumulativní relativní četnost je součet relativní četnosti v dané skupině a ve všech nižších (předchozích) skupinách, nebo lépe jako **N_i /n**.
n ve statistice vždy označuje četnost, třeba počet respondentů

Počty automobilů

n = 100, máme tedy 100 firem
z toho 15 firem má jedno auto
14 firem má tři auta

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
Σ	**100**	**1,00**	-	-

15 firem má 1 auto

76 firem má maximálně 2 auta

Četnost - vzorce

Poznámka: vzorce jsou jen pro ilustraci, nejsou moc užitečné.

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

Vzorec pro absolutní četnost říká, že když sečteme všechny absolutní četnosti n_i, získáme celkový součet všech prvků,který značíme n.

relativní četnost $p_i = \frac{ni}{n}$

Relativní četnosti p_i tedy získáme vydělením absolutní četnosti (n_i) součtem všech prvků, který značíme n.

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

Takže kumulativní absolutní četnost N_i je součet prvků v dané skupině (i-té skupině, takže sčítáme skupiny od 1 do i) a prvků ve všech skupinách nižších.

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

Kumulativní relativní četnost P_i je součet relativní četnosti v dané skupině a ve všech nižších (předchozích) skupinách, nebo lépe jako $\mathbf{N_i/n}$.

Poznámka: Totéž je vysvětleno i zde:
http://moodle.lfhk.cuni.cz/moodle2/mod/book/view.php?id=2113

Pravidlo pro stanovení počtu intervalů

Počet intervalů k se rovná odmocnině z počtu hodnot (n). Tedy: k = $\sqrt{n}$

Zaokrouhlování hranic intervalu je vždy nahoru. I když mi odmocnina vychází 5.1, zaokrouhluji nahoru na 6 (takže ne jako ve v matematice). Jde o to, aby nám nikdo nezbyl.

Pokud se to hodí, měly by být intervaly stejně široké. Nehodí se to například u věku, kde může být rozdělení na preproduktivní věk, produktivní věk a poproduktivní věk.

Intervaly se nesmí přesahovat. Příklad toho jak má vypadat interval (0-99, 100-199, …)

Bylo by pěkné, aby ty intervaly vypadaly podobně.

Charakteristiky polohy - popisují soubor z hlediska úrovně, jinak řečeno velikosti hodnot v souboru (střední hodnoty, kvantily atd.)

Charakteristiky variablity - popisují soubor z hlediska měnlivosti, různorodosti odlišnosti hodnot atd. (absolutní variability, relativní variability).

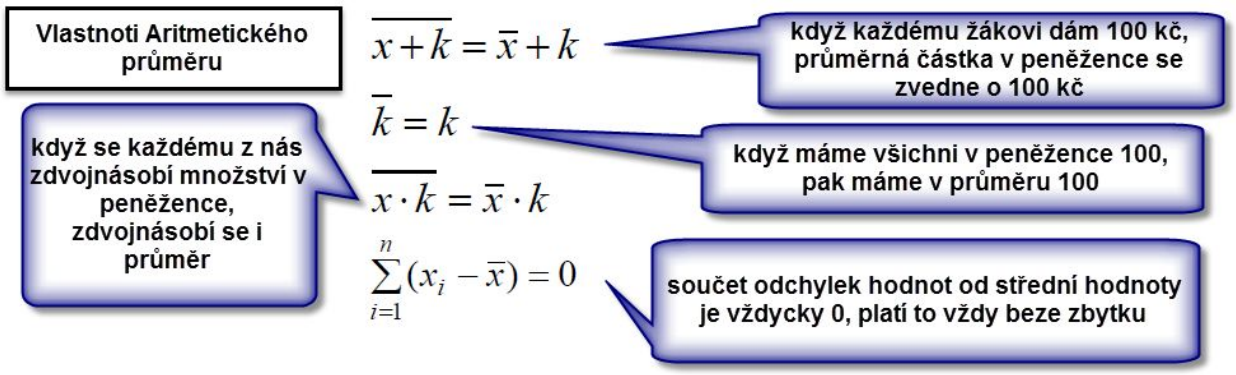
Prosté/vážené charakteristiky

Prosté charakteristiky v případě, že vstupní údaje nejsou uspořádány do **tabulky četností**.
Vážené charakteristiky použijeme tehdy, když máme vstupní údaje usporádány do **tabulky četností**.

Strední hodnoty

Aritmetický průměr prostý - součet hodnot dělený jejich počtem. $\overline{x} = \frac{\sum_{i=1}^{n} xi}{n}$
Vážený aritmetický průměr - součet součinu hodnot a jejich četnosti, vydělený celkovým počtem hodnot.

(V případě, že máme spojité hodnoty rozdělené do intervalů, použijeme středy intervalu.)



Aritmetický průměr se používá vždy s těmito výjimkami.

Harmonický průměr

prostý HP

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

Používá se:
Při výpočtu průměrné rychlosti (čím pomaleji na některém úseku jedu, tím více času na tom úseku strávím, a tím větší váhu ten úsek v průměru má).
(Pracnost, něco na jednotku, tam kde pracujeme s převrácenými hodnotami)
V rámci indexních analýz - při výpočtu cenových indexů, inflace atd.

Geometrický průměr

n-tá odmocnina ze součinu n hodnot

prostý GP

$$\bar{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$$

Je to tolikátá odmocnina kolik hodnot násobíme.
Takže v tabulce dole to bude 6 tá odmocnina. Hodnot je šest, počet roků nás nezajímá.

Příklad:

2000	2001	2002	2003	2004	2005	2006
-	1,1	1,2	1,05	1,1	1,1	1,3

$\sqrt[6]{1.1 * 1.2 * 1.05 * 1.1 * 1.1 * 1.3}$ = 1.14

V Pythonu:

```
>>> a =1.1 * 1.2 * 1.05 * 1.1 *1.1 *1.3
>>> a
```

```
2.1801780000000006
>>> import math
>>> math.pow(a, 1/6)
1.1387157445033058
```
(Nápověda zde: https://docs.python.org/2/library/math.html)

Existuje i vážený geometrický průměr, ale moc se nepoužívá.

GP se používá pro nelineárně roustoucí funkce (geometricé funkce) - při výpočtu průměrného koeficientu (tempa) růstu. Klasický příklad na geometrický průměr je **složené úročení**. Průměrné tempo růstu HDP, tržeb, nezaměstnosti atd. Cokoliv co sledujeme v relativním vyjádření.

Modus

modus - značí se jako x se stříškou

$$\hat{x}$$

Modus je nejčastěji se vyskytující varianta sledovaného znaku v souboru.
Modus se nepočítá, modus se hledá!

Kvantil je hodnota znaku, pro kterou platí, že nejméně p-procent prvků má hodnotu menší nebo rovnu xp a zbytek (tedy 100 -p procent) prvků je větších nebo rovno xp.

p% kvantil je tedy číslo, které rozdělí soubor **uspořádaný** podle velikosti na dvě části tak, že p% hodnot je menších než p% kvantil a zbytek (100 - p %) je větších než p% kvantil.

Poznámka: to že soubor musí být uspořádaný je zásadní

Takže třeba výška 168 cm, může být 25% kvantil u mužů. To znamená, že 25% mužů je nižších a 100-25% (tzn 75%) mužů je vyšších.

p% kvantil označujeme jako x s vlnovkou a číslem p. Například když je p = 40, jde o 40% kvantil.

$$\tilde{x}_p$$

Pojmenované kvantily

Medián je 50% kvantil - je to přesný prostředek
Kvartily, decily, percentily atd. - je to podobná věc. Jsou to stále kvantily, ale k rozdělení dochází podle nějaké vvýznamné hranice. Samozřejmě že uspořádaný soubor stále rozdělujeme na dvě části.

Decily:
první 10/90
druhý 20/80
atd.

Percentil je zbytečné, ale důležité synonymum ke kvantilu.

Variační rozpětí - je to rozdíl mezi kvantily. Nejdůležitější je kvartilové rozpětí, rozdíl mezi 75 a 25% kvartilem.

Charakteristiky variability

Nejjednodušší charakteristika variability je variační, například kvartilové rozpětí.

- ***variační rozpětí*** $R = x_{\max} - x_{\min}$
- ***kvartilové rozpětí*** $R_Q = \tilde{x}_{75} - \tilde{x}_{25}$

Průměrná absolutní odchylka

Je to součet absolutních odchylek vydělený počtem. Je to správně, dává to smysl, ale nepoužívá se to. V praxi se místo absolutních hodnot používá druhá mocnina, čímž získáme rozptyl.

prostý tvar $$d_x = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

Rozptyl
Rozptyl je průměrný čtverec odchylek (součet druhých mocnin hodnot odchylek vydělených počtem hodnot)

Výsledek vyjde v druhých mocninách korun, litrů atd, takže se to pak odmocňuje a vzniká směrodatná odchylka.

Prostý tvar rozptylu

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

Průměr obsahu čtverců vzdáleností jednotlivých hodnot od jejich průměru, neboli průměrná velikost čtverce.

Vážený tvar rozptylu

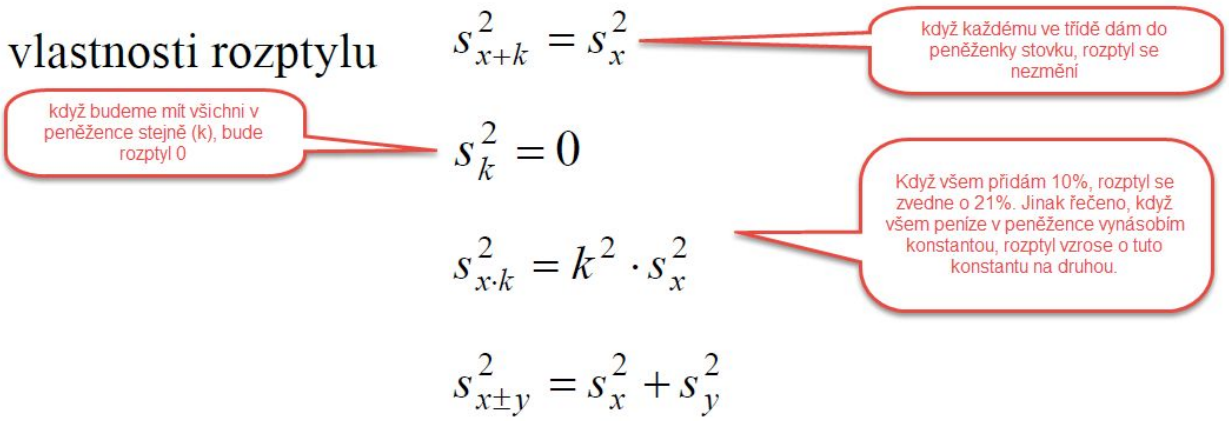
$$s_x^2 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 n_i}{\sum_{i=1}^{k} n_i}$$

Postup výpočtu

1. vypočítám aritmetický průměr x s čarou (budu ho potřebovat ve vzorečku)
2. ode všech hodnot odečtu průměr (takže v Excelu budu mít sloupeček s těmito rozdíly)

3. umocním odchylky na druhou (další sloupeček)
4. vynásobím tyto druhé mocniny četnostmi (další sloupeček)
5. sečtu hodnoty ze sloupce, kde mám druhé mocniny násobené četnostmi
6. celé to vydělím n (n je celkový počet hodnot)

Vlastnosti rozptylu



Směrodatná odchylka

Směrodatná odchylka, podobně jako rozptyl, určuje jako moc jsou hodnoty rozptýleny či odchýleny od průměru hodnot. **Směrodatná odchylka je rovna odmocnině z rozptylu.**

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

Míry relativní variability

Variační koeficient

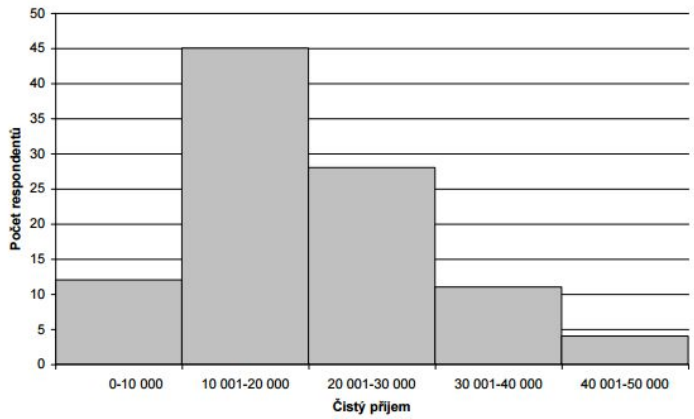
$$v_x = \frac{s_x}{x}$$

Variační koeficient se používá k porovnávání variability v různých souborech.

Když vynásobíme variační keoficient stovkou, získáme variabilitu v procentech a používá se pro srovnání variability různých souborů.
(Pozor, variační koeficient se plete s variačním rozpětím!)

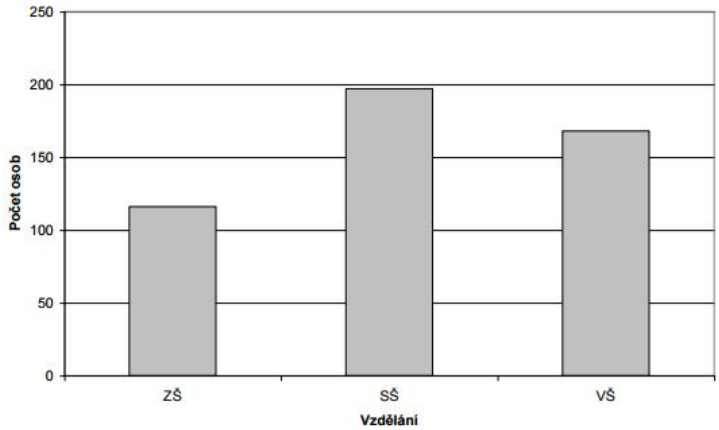
Grafy

Histogram četností



Pozor, sloupce se dotýkají, takže v Excelu je třeba nastavit mezera 0.
To že se intervaly (takže i sloupce) dotýkají je podmínka histogramu, jinak by to nebyl histogram. Je vhodný pro spojité proměnné. Šířka sloupce znamená šířku intervalu.

Soupcový graf



Používá se pro všechna data kromě kromě kvantitativních (číselných) spojitých.
Takže pro data kvalitativní ordinální, kvalitativní nominální, kvantitativní nespojitá.

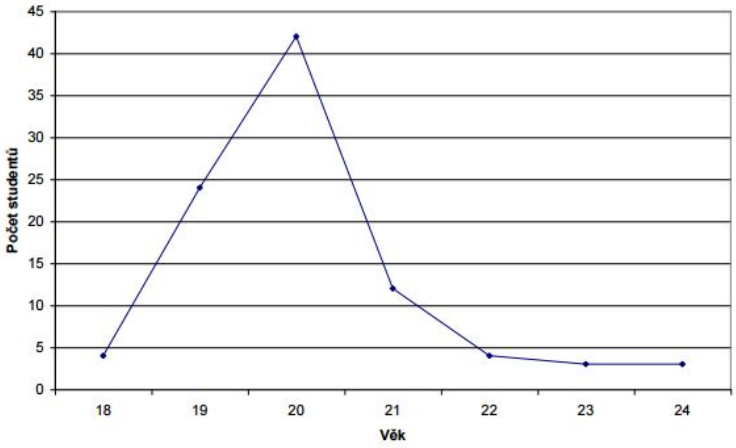
Kvalitativní nominální data - pořadí si musíme nějak vybrat. Pokud budeme průzkum opakovat, musíme použít pořadí sloupců odminule.

Koláčový graf

je vhodný pro nominální kvalitativní data
je nevhodný pro spojité hodnoty

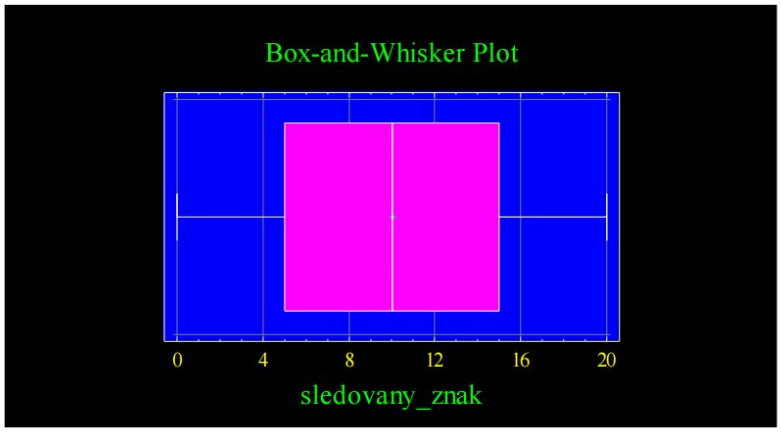
Poygon četností

Vznikne propojením jednotlivých bodů. Na ose x jsou alternativy, na ose y jsou četnosti.
Pro který druh proměnných je vhodný?



Krabičkový graf

Může být horizontální, nebo vertikální.



úplně vlevo je minimální hodnota (na obrázku tedy 0)
úplně vpravo je maximální hodnota (na obrázku 20)
levé hrana růžové krabice - 25% kvantil
pravá hrana 75 procentní kantil
čára uprostřed je medián
půlky růžové krabice ukazují kvartilové rozpětí
plus nebo hvězdička uprostřed značí aritmetický průměr

(více zde https://plot.ly/python/box-plots/)

Zkoumání závislosti mezi proměnnými

(Inflace/nezaměstnanost, cena objem prodeje atd.)

Druhy závislosti:

Pevná (funkční) zavislost - mám jasně danný funkční předpis, podle kterého se dvojce s jistotou chová. Taková závislost se obvykle vyskytuje ve fyzice, v ekonomii jen výjimečně.

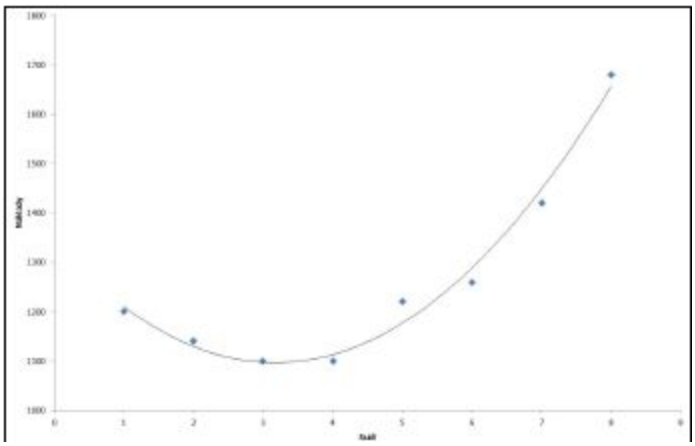
Volná (neboli statisticka) zavislost - změna proměnné způsobí změnu druhé proměnné s určitou pravděpodobností.

Směr závislosti:

jednostranné závislosti - jedna věc ovlivňuje druhou, ale ne obráceně. Jednostranné závislosti popisuje regresní analýza.

obousranné (vzájemné) závislosti - tím se zabývá korelační analýza

Pro závislosti dvou kvantntativních (číselných) proměnných používáme bodové grafy



Regresní analýza

Regresní analýza slouží k popisu volných (statistických) závislostí (u těch funkčních závislostí ji nepotřebujeme). Cílem je pomocí hodnot jedné proměnné odhadovat chování druhé proměnné (třeba jak se chová nezaměstnanost v závislosti na inflaci atd.). Obě proměnné musí být kvantitativní (číselné).

Y vysvětlovaná (závislá) proměnná.
X vysvětlující (nezávislá) proměnná.

Jednoduchá regresní analýza
Pokud máme jednu vysvětlovanou proměnnou Y a jednu vysvětlující proměnnou X, jde o **jednoduchou regresní analýzu**.

Vícenásobná regresní analýza
Jedna vysvětlovaná proměná - třeba výdaje domácnosti
Vysvětlujících proměnných je několik (X_1, X_2, X_3 ….X_n) - může to být počet dětí, počet aut, počet členů domácnosti atd. Hledáme mnotiny vysvětlujících faktorů.

Metoda nejmenších čtverců

(Tohle vypadá nadějně: http://www.kloudak.eu/metoda-nejmensich-ctvercu/)
(??? http://mathworld.wolfram.com/LeastSquaresFitting.html)
(Jestli je tohle k něčemu netuším http://www.wikiskripta.eu/index.php/M%C4%9B%C5%99en%C3%AD_z%C3%A1vislosti,_korelace_a_regrese)

Korelační keoficient - číslo od mínus jedné do jedné, které udává závislost mezi nezávislou a závislou proměnnou.

Záporné hodnoty korelačního koeficientu -> nepřímá úměra
Kladný korelační koeficient -> přímo úměrná závislost
Korelační koeficient = 0 -> lineárně nezávislé hodnoty (není tam žádná závislost)

Čím jsou hodnoty korelačního koeficientu bližší krajní hodnotě (plus/mínus) jedné, tím je závislost silnější.

Například, pokud je korelační koeficient -0.9, pak jde o silnou závislost a nepřímou úměru.

Druhou mocninou korelačního koeficientu je keoficient determinace R^2. Č

z korelacniho koeficinetu muzeme vypozitat koeficient determinace Rˆ2 a vypocita se jako durha mocicna - hodnoty od nuly do jedne. Cim je blizsi jedne, tim ma dana krivka ma vetsi schopnost (v %) ma krivka zachytit vztah tech promennych.

smernice b1 v tomto pripade je prumerna zmena y pri zmene x o jednicku

z keeficientu determinace nepozname znamenko korelacniho koeficientu, ale to pozname ze smernice

Zakladni miry dynamiky casovych rad

mam n obdobi, ale pocitat budu s n - 1, protoze zmen je o jednu mene nez obdobi

5.

co je korelacni koeficien a k cemu slouzi
kdy neni vhodne pouzit prumer