Přednáška 1: **ZÁKLADNÍ STATISTICKÉ POJMY, CHARAKTERISTIKY DAT**

Mgr. Jiří Mazurek, Ph.D.

Informace o předmětu

- Vyučující: Mgr. Jiří Mazurek, Ph.D., A 407.
- Přednáška: čtvrtek 8:55 – 10:30, učebna A216.
- Seminář: čtvrtek 10:35 – 11:20, učebna A216.
- Počet kreditů: 5.
- Prezenční i kombinované studium.

Informace o předmětu

Cíle předmětu:

Poskytnout hlubší pohled na statistické metody vhodné ke zpracování vícerozměrných dat, ovládnout teoretický aparát vybraných metod a naučit se je aplikovat pomocí statistických programů na počítači.

- Materiály – e-learning, IS
- Opora – e-learning, IS

Zkouška

- Písemná, částečně za pomoci počítače.
  Pro úspěšné zvládnutí předmětu musíte mít alespoň 60 bodů ze 100.
- Ke zkoušce si můžete přinést jakékoliv studijní materiály v papírové formě.

Účast na seminářích

- Aktivní účast na seminářích je hodnocena body navíc (nezapočítávají se semináře, kde se píše test).
- 1 bod = 1x účast
- Maximum 10 bodů za účast.

Hodnocení

- **Celkem 110 bodů**
- 0 až 59: nedostatečně (F), 4
- 60 až 64: dostatečně (E), 3
- 65 až 69: uspokojivě (D), 2,5
- 70 až 79: dobře (C), 2
- 80 až 89: velmi dobře (B), 1,5
- 90 až 110: výborně (A), 1.

Plán semestru

Týden 1: 22. 9.

- **Přednáška:** Informace o podmínkách absolvování. Základní pojmy a metody ze statistiky.
  (Charakteristiky polohy, charakteristiky variability, šikmost, špičatost, statistický soubor se dvěma znaky, testy statistických hypotéz.)
- **Seminář:** Charakteristiky polohy, charakteristiky variability, šikmost, špičatost.

Týden 2: 29.9.

- **Přednáška:** Testování hypotéz – parametrické testy.
  (Marketingová případová studie, co přináší parametrické testování statistických hypotéz v marketingu, jednovýběrový t-test, dvouvýběrový t-test - nepárový a párový.)
- **Seminář:** Testy statistických hypotéz, jednovýběrový t-test, dvouvýběrový t-test – nepárový a párový.

Týden 3: 6. 10.

- **Přednáška:** Testování hypotéz - neparametrické testy.
  (Mediánový test (pro 1 výběr), chi-kvadrát test pro 1 výběr, dvouvýběrové testy, chi-kvadrát test pro 2 výběry, Mann-Whitneyův test, Wilcoxonův párový test.)
- **Seminář:** Chi-kvadrát test pro 1 výběr, chi-kvadrát test pro 2 výběry.

Týden 4: 13. 10.

- **Přednáška:** Regresní analýza
  (Podstata regresní analýzy, odhad regresních koeficientů, test významnosti regresních koeficientů, intervaly spolehlivosti regresních koeficientů, test vhodnosti regresního modelu.)
- **Seminář:** Odhad regresních koeficientů, test významnosti regresních koeficientů, intervaly spolehlivosti regresních koeficientů, test vhodnosti regresního modelu.

Týden 5: 20. 10.

- **Přednáška:** Metody prognózování
  (Analýza trendové složky, analýza sezónní složky, model konstantní sezónnosti, analýza náhodné složky, testování vlastností náhodné složky, prognózování, kauzální prognostické metody.)

Týden 6: 27. 10.

- **Přednáška:** Korelační analýza
  (Koeficient korelace, index korelace, Spearmanův koeficient (pořadové) korelace, vícenásobná lineární závislost - vztahy pro dvě vysvětlující proměnné.)
- **Seminář:** Analýza trendové složky, analýza sezónní složky, model konstantní sezónnosti, analýza náhodné složky, testování vlastností náhodné složky, prognózování, kauzální prognostické metody.

Týden 7: 3. 11.

- **Přednáška:** Analýza rozptylu (ANOVA)
  (Jednofaktorová ANOVA, postup při analýze rozptylu s jedním faktorem, míra těsnosti závislosti.)
- **Seminář:** Výpočet koeficient korelace, Spearmanova koeficientu korelace, test statistické významnosti korelačního koeficientu.

Týden 8: 10. 11.

- **Přednáška:** Analýza rozptylu (ANOVA) : Dvojné třídění a Latinské čtverce
  (Dvojné třídění, trojné třídění (Latinské čtverce).)
- **Seminář:** Postup výpočtu při analýze rozptylu s jedním faktorem.

Týden 9: 17. 11. – státní svátek

Týden 10: 24. 11.

- **Přednáška:** Úplné a částečné faktorové plány
  (Základy experimentování a oblasti použití, experimentální procedura, efekt (vliv) faktoru, významnost efektu, test významnosti efektu, grafické hodnocení efektu faktoru, grafy interakcí, model experimentu.)
- **Seminář:** Postup výpočtu při analýze rozptylu se dvěma faktory.

Týden 11: 31. 11.

- **Přednáška:** Částečný faktorový experiment se dvěma úrovněmi (Poloviční plány, grafická metoda.)
- **Seminář:** Úplný faktorový plán: model experimentu, efekt (vliv) faktoru, významnost efektu, test významnosti efektu, grafické hodnocení efektu faktoru, grafy interakcí.

Týden 12: 8. 12.

- **Přednáška:** Taguchiho metody: ztrátová funkce (Definice a vlastnosti ztrátové funkce, ztrátová funkce pro různé typy tolerance.)
- **Seminář:** Ztrátová funkce pro různé typy tolerance, monitorování nákladů na jakost.

Týden 13: 15. 12.

- **Přednáška:** Taguchiho metody: celkové náklady na jakost (Monitorování nákladů na jakost, regulační diagramy.)

Kontakt

- mazurek@opf.slu.cz
- A407.

Základní statistické pojmy, charakteristiky dat

- Hlavním cílem statistiky je analyzovat jisté datové soubory.
- Daný soubor dat je obvykle vytvořen za jistým účelem – za účelem analýzy podoby či chování nějaké veličiny, které se říká *statistický znak*.

Populace versus výběr

- Množina všech hodnot, kterých znak může nabýt, se ve statistice nazývá *základní soubor* nebo také *populace.* Populace se vztahuje k danému statistickému pojmu a je to tedy v tomto smyslu relativní pojem.
- Statistici se nicméně častěji setkávají se situací, kdy základní soubor k dispozici není. V takovém případě jim nezbývá nic jiného než provést výběr z této populace a získat tzv. *výběrový soubor.*
- Ve statistice se nejčastěji požaduje *náhodný výběr*, což je datový soubor vznikající tak, že každý jeho prvek má stejnou pravděpodobnost, že bude vybrán.

Deskriptivní statistika

- Je-li k dispozici základní soubor, může být jedinou ambicí statistika tuto populaci popsat. Metody sloužící k tomuto účelu utvářejí *deskriptivní/popisnou statistiku*.
- Charakteristika je obecně údajem, který jistým způsobem shrnuje informaci o sledovaném datovém souboru.
- Charakteristiky využívané k popisu populace se logicky nazývají *populační charakteristiky*.
- V případě, že je k dispozici pouze výběrový soubor, užívají se k popisu tohoto výběru *výběrové charakteristiky.*
- Zvyklostí je užívat ke značení populačních charakteristik písmena řecké abecedy, zatímco pro výběrové charakteristiky se užívá obvykle latinka.

STATISTICKÝ SOUBOR S JEDNÍM ZNAKEM

- Nechť je dán základní soubor skládající se z hodnot $x_1, x_2 \ldots x_n$, kde *n* je přirozené a tedy konečné číslo (my budeme pracovat zejména se soubory konečné velikosti).
- Sledovaným statistickým znakem nechť je veličina *X*. Čísla $x_1, x_2 \ldots x_n$ jsou hodnoty, kterých tato veličina nabývá.
- Pokud bychom na tento soubor aplikovali náhodný výběr, můžeme na proměnnou *X* nahlížet jako na (diskrétní) náhodnou veličinu.

Četnosti výskytu

- Přestože soubor obsahuje hodnoty $x_1, x_2 \ldots x_n$, některé z čísel se mohou opakovat. V takovém případě pak nabývá veličina *X* pouze k různých hodnot $x_1 *, x_2 * \cdots x_k *$.
- Hodnota $x_1 *$ se může v souboru vyskytovat f_1-krát a číslo f_1 pak nazýváme *absolutní četností* výskytu hodnoty $x_1 *$.
- Obdobně se hodnota $x_2 *$ vyskytuje v souboru f_2-krát, hodnota $x_3 *$ ... f_3 -krát a tak dále, až konečně číslo $x_k *$ je obsaženo v souboru f_k -krát.

Typy četností

- Kromě absolutních četností pracujeme také s jinými typy četností:
- a) s *relativní četností* výskytu hodnoty x_i* danou výrazem f_i/n kde n značí rozsah souboru.

- Pokud seřadíme hodnoty vzestupně, můžeme zavést také pojmy
- b) *absolutní kumulativní četnost* hodnoty
- c) *relativní kumulativní četnost* hodnoty.

- Uvedené druhy četností mohou být využity v souvislosti s populací i výběrovým souborem.

Příklad četností

Prodejce aut Bourák s.r.o. prodal každý den v únoru následující počet automobilů:
4,5,2,5,3,5,6,3,1,2,5,4,6,8,5,4,4,3,4,5,6,3,2,5,2,5,4,7.

Absolutní četnosti jednotlivých hodnot jsou:
Hodnota 1 má četnost 1 (vyskytuje se v souboru jednou), hodnota 2 má četnost 4 (vyskytuje se v souboru čtyři krát), hodnota 3 má četnost 4 , hodnota 4 má četnost 6, hodnota 5 má četnost 8, hodnota 6 má četnost 3, hodnota 7 má četnost 1 a hodnota 8 má četnost 1.

Podívejme se na hodnotu 2 (matematicky zapsáno: $x_2^* = 2, f_2 = 4$):

Absolutní četnost je 4.
Relativní četnost hodnoty je rovna 4/28= 0,143 = 14,3 %.
Kumulativní četnost je 4+1 = 5.
Relativní kumulativní četnost je 5/28 = 0,179 = 17,9 %

CHARAKTERISTIKY POLOHY

- Populační aritmetický průměr:
- $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Výběrový aritmetický průměr:
- $\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$ (typicky je m mnohem menší než n)
- Vážený aritmetický průměr:
- $\mu = \frac{1}{n}\sum_{i=1}^{n} w_i \cdot x_i$, kde w_i je váha hodnoty x_i.

- *Modus* $\hat{x}$
  - hodnota, která má v daném souboru dat nejvyšší absolutní četnost. (Tento popis neurčuje modus jednoznačně, a tak se může stát, že datový soubor bude mít více modů).
- Medián $\tilde{x}$
  - Prostřední hodnota (sudý počet hodnot vs lichý počet hodnot)

Příklad na charakteristiky polohy

Prodejce aut Bourák s.r.o. prodal každý den v únoru následující počet automobilů: 4,5,2,5,3,5,6,3,1,2,5,4,6,8,5,4,4,3,4,5,6,3,2,5,2,5,4,7.

Tento soubor budeme považovat za populační.

Populační aritmetický průměr: $\bar{x} = \frac{1}{28}\sum_{n=1}^{28} x_i = \frac{1}{28} \cdot (4 + 5 + 2 + \cdots + 7) = 4{,}21$

Modus: $\hat{x} = 5$

Medián: $\tilde{x} = 4$

Medián určíme takto: seřadíme hodnoty od nejmenší po největší:

1,2,2,2,2,3,3,3,3,4,4,4,4,4,4,5,5,5,5,5,5,5,5,5,6,6,6,7,8.

Máme dvě prostřední hodnoty, obě jsou 4, medián je tedy roven čtyřem.

CHARAKTERISTIKY VARIABILITY

- Populační rozptyl: $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$
- Výběrový rozptyl: $s^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2$

- Populační směrodatná odchylka: σ
- Výběrová směrodatná odchylka: s
- Variační rozpětí: R = max – min
- Populační variační koeficient: $V = \frac{\sigma}{|\mu|}$
- Výběrový variační koeficient: $V = \frac{s}{|\bar{x}|}$.

Příklad na charakteristiky variability dat

Prodejce aut Bourák s.r.o. prodal každý den v únoru následující počet automobilů:
4,5,2,5,3,5,6,3,1,2,5,4,6,8,5,4,4,3,4,5,6,3,2,5,2,5,4,7.

Tento soubor budeme považovat za populační.

Populační rozptyl: $\sigma^2 = 2{,}597$
Populační směrodatná odchylka: 1,612
Variační rozpětí: R = 7
Variační koeficient: V = 0,382

CHARAKTERISTIKY KONCENTRACE DAT

- Ukazatele, které v jistém slova smyslu odrážejí míru seskupení hodnot tvořících analyzovaný datový soubor
- Charakteristika šikmosti *Sk* (anglicky skewness)

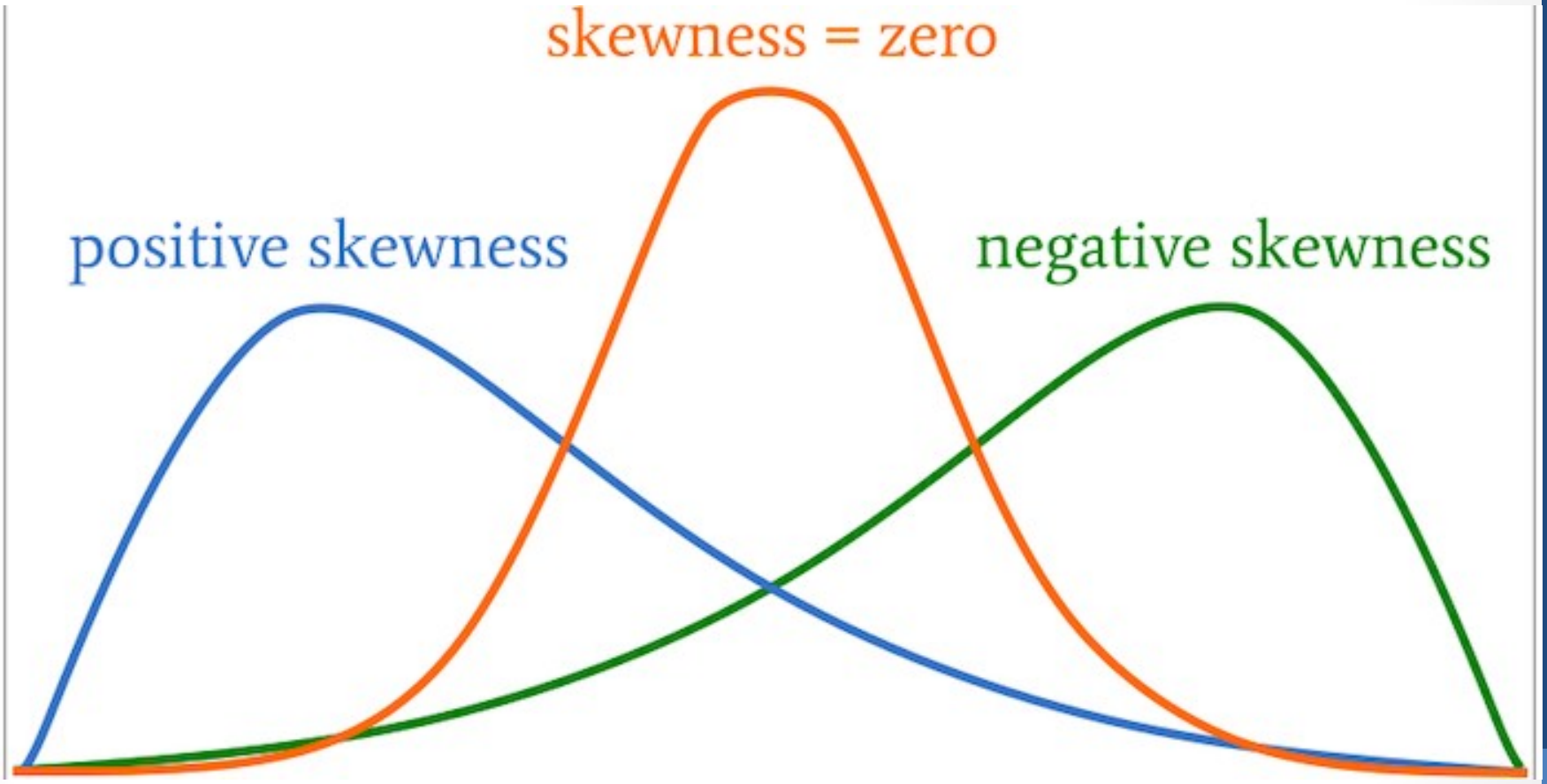
$$Sk = \frac{\sum_{i=1}^{n}(x_i-\mu)^3}{n\sigma^3},$$

- Charakteristika špičatosti *Ku* (z anglického kurtosis)

$$Ku = \frac{\sum_{i=1}^{n}(x_i-\mu)^4}{n\sigma^4}$$

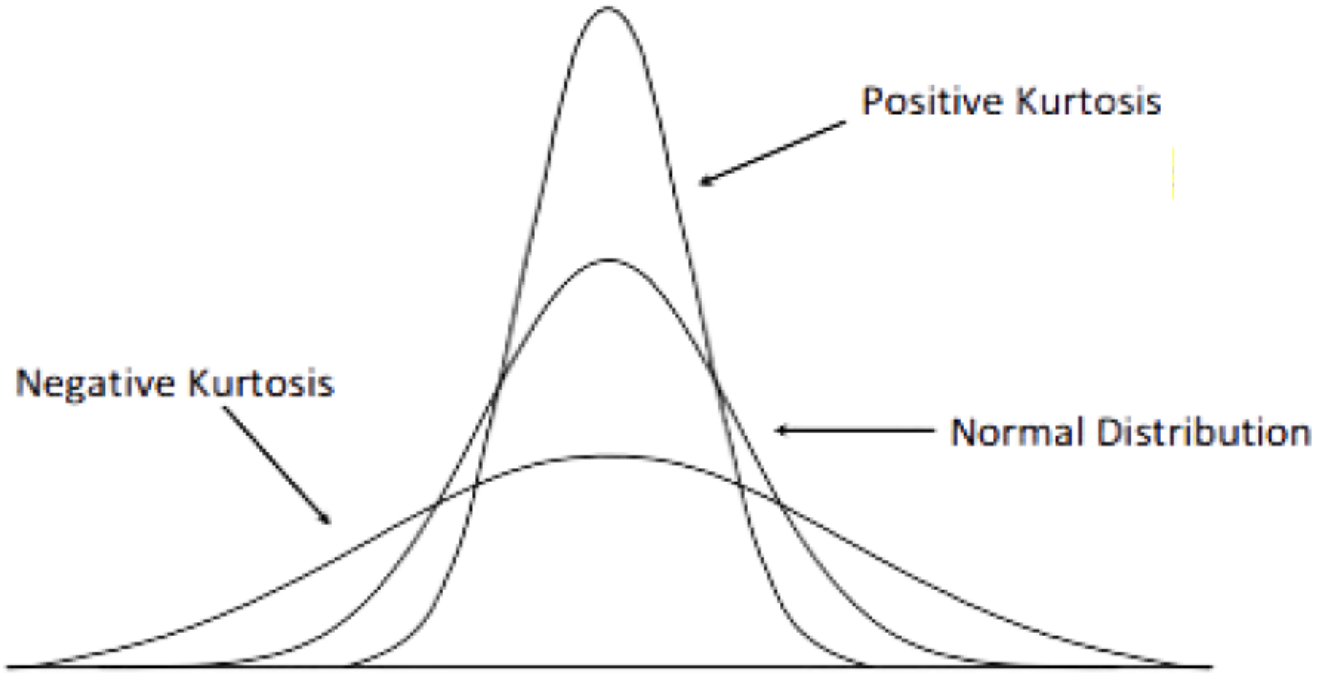
Šikmost

- Jak vyplývá z definičních vzorců, šikmost může nabývat libovolné reálné hodnoty.
- V případě, že ukazatel vychází nula, poukazuje tento výsledek na symetrické rozdělení četností hodnot v daném datovém souboru. Koncentrace malých hodnot je stejná jako koncentrace velkých hodnot v daném souboru.
- Pokud vychází šikmost kladně, má rozdělení četností hodnot z daného souboru kladné zešikmení (zešikmení doprava) a koncentrace malých hodnot je v takovém souboru vyšší než koncentrace velkých hodnot.
- Pokud vychází šikmost záporně, má rozdělení četností hodnot z daného souboru kladné zešikmení (zešikmení doleva) a koncentrace malých hodnot je v takovém souboru naopak menší než koncentrace velkých hodnot.
- V případě nenulové šikmosti hovoříme také o asymetrickém rozdělení četností.


skewness = zero
positive skewness
negative skewness

Špičatost

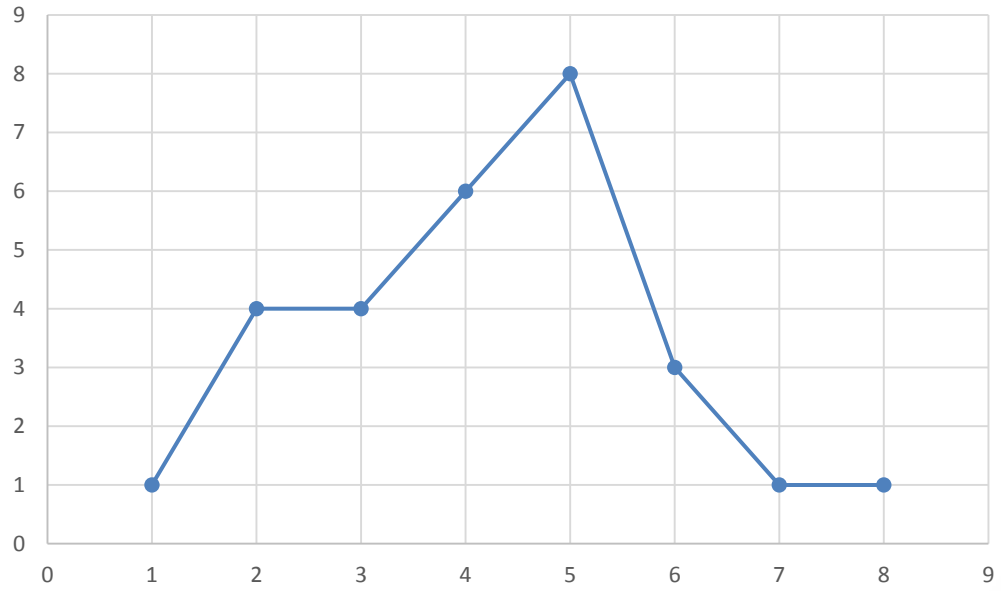
- Vyšší hodnota tohoto ukazatele vyjadřuje vyšší špičatost, tj. vyšší koncentraci hodnot blízkých prostřední hodnotě ve srovnání s ostatními hodnotami daného statistického znaku.
- Pokud špičatost nabývá kladných hodnot, znamená to, že graf daných hodnot je špičatější než normální (Gaussovo) rozdělení.
- Naopak, pokud je špičatost záporná, znamená to, že graf vytvořený ze zadaných hodnot je plošší než normální rozdělení, viz následující obrázek.


Positive Kurtosis
Negative Kurtosis
Normal Distribution

Příklad na koncentraci dat

Prodejce aut Bourák s.r.o. prodal každý den v únoru následující počet automobilů:
4,5,2,5,3,5,6,3,1,2,5,4,6,8,5,4,4,3,4,5,6,3,2,5,2,5,4,7.

Šikmost: Sk = 0,111
Špičatost: -0,106



OBECNÉ MOMENTY

- Obecné momenty jsou charakteristiky, které nahlížejí na strukturu dat z trochu jiného úhlu pohledu.
- Existuje několik důvodů, proč se s nimi pracuje.
- Jedním z těchto důvodů je skutečnost, že za jistých podmínek si rozdělení četností a momenty vzájemně jednoznačně odpovídají: datové soubory se stejnými momenty budou mít stejné rozdělení četností a naopak.
- Nás nicméně zajímá zejména druhý důvod práce s momenty, a tím je jejich vhodnost pro systematičtější výpočet některých charakteristik

Obecný moment

- Pro základní soubor dat definujeme *k-tý obecný moment Mk* předpisem

$$M_k = n^{-1} \cdot \sum_{i=1}^{n} x_i^k \, , k = 1,2, \ldots$$

- Jde tedy o průměr k-tých mocnin původních hodnot.

Užitečné vztahy

$$M_1 = \mu,$$
$$M_2 - M_1^2 = \sigma^2,$$
$$\sigma^{-3} \cdot (M_3 - 3M_1M_2 + 2M_1^3) = Sk,$$
$$\sigma^{-4} \cdot (M_4 - 4M_3M_1 + 6M_2M_1^3 - 3M_1^4) = Ku.$$

STATISTICKÝ SOUBOR SE DVĚMA ZNAKY

- Máme-li statistický soubor takový, že pro každé přirozené číslo i = 1, 2, ..., *m* a *j* = 1, 2, ..., *n* obsahuje soubor jistou dvojici (x_i, y_j) hodnot nebo i více dvojic s těmito hodnotami, hovoříme o *statistickém souboru se dvěma znaky* (též *argumenty*). Počet výskytů dvojice hodnot se nazývá *sdruženou četností* dvojice a značí se f_{ij}.

Kontingenční tabulka

- Rozdělení sdružených četností se zapisuje do dvourozměrné tabulky, která se nazývá *kontingenční tabulka*
- Do záhlaví tabulky se zapisují různé možné obměny obou sledovaných znaků, vnitřek tabulky obsahuje sdružené četnosti výskytu různých kombinací těchto znaků.

	y1	y2	y3
x1	4	7	2
x2	1	2	8
x3	5	5	8
x4	2	3	2

Populační charakteristiky

- Předpokládáme-li, že uvedená tabulka představuje celou populaci, můžeme při zavedené symbolice vypočítat základní dvě charakteristiky znaků *X* a *Y* – populační průměr, respektive střední hodnotu, a populační rozptyl, a to podle následujících vzorců

Populační průměry

$$\mu_X = \frac{1}{r} \sum_i x_i \sum_j f_{ij},$$

$$\mu_Y = \frac{1}{r} \sum_j y_j \sum_i f_{ij}.$$

Populační rozptyly

$$\sigma_X^2 = \frac{1}{r} \sum_i (x_i - \mu_X)^2 \sum_j f_{ij}$$

$$\sigma_Y^2 = \frac{1}{r} \sum_j (y_j - \mu_Y)^2 \sum_i f_{ij}$$

Výběrové charakteristiky

- Pokud by tabulka reprezentovala výsledek náhodného výběru, počítali bychom výběrové průměry a výběrové rozptyly podle vzorců

1. Výběrové průměry

$$\bar{x} = \frac{1}{r}\sum_i x_i \sum_j f_{ij},$$

$$\bar{y} = \frac{1}{r}\sum_j y_j \sum_i f_{ij}.$$

2. Výběrové rozptyly

$$s_X^2 = \frac{1}{r-1}\sum_i (x_i - \mu_X)^2 \sum_j f_{ij},$$

$$s_Y^2 = \frac{1}{r-1}\sum_j (y_j - \mu_Y)^2 \sum_i f_{ij}.$$

Kovariance

- Pracujeme-li se dvěma znaky jako v našem případě daném výše uvedenou kontingenční tabulkou, definujeme také další důležitou charakteristiku zvanou *kovariance*. Populační kovarianci znaků *X* a *Y* definujeme vzorcem :

$$\operatorname{cov}(X,Y) = \frac{1}{r}\sum_i\sum_j (x_i - \mu_X)(y_j - \mu_Y) f_{ij} = \frac{1}{r}\sum_i\sum_j x_i y_j f_{ij} - \mu_X \cdot \mu_Y .$$

- Pokud budeme pracovat s výběrovými daty o rozsahu větším než 2, definujeme výběrovou kovarianci vztahem

$$c_{XY} = \frac{1}{r-1}\sum_i\sum_j (x_i - \bar{x})(y_j - \bar{y}) f_{ij} .$$

Příklad

Znak X nabývá celých hodnot: 3, 5, 4, 6, 7, 9. Pro tyto hodnoty (v uvedeném pořadí) byly zjištěny následující hodnoty druhého znaku Y: 1, 2, 7, 9, 11, 13, tj. číslu 3 odpovídá hodnota 1 druhého znaku, číslu 5 odpovídá hodnota 2 druhého znaku, apod. Pro oba znaky platí, že každá hodnota má vždy absolutní četnost svého výskytu rovnu jedné. Spočtěme populační kovarianci.

ŘEŠENÍ

Využijeme vzorce 1-15, v němž jsou všechny četnosti rovny jedné. Průměrné X má hodnotu 5,66, průměr Y je roven 7,16. Dostáváme

$$\operatorname{cov}(X,Y) = \frac{1}{r}\sum_i\sum_j (x_i - \mu_X)(y_j - \mu_Y) f_{ij} = \frac{(3-5{,}66)\cdot(1-7{,}16) + \ldots + (9-5{,}66)\cdot(13-7{,}16)}{6} = 7{,}55.$$

Kovariance – poznámky

- Kovariance se využívá k vyjádření závislosti mezi znaky *X* a *Y* ve tvaru přímky, tj. k vyjádření jejich lineární závislosti.
- Lze říci, že pokud vychází kovariance kladně, existuje mezi oběma znaky do jisté míry závislost ve tvaru přímé úměry. Přímá úměra značí, že s růstem hodnoty jednoho znaku úměrně roste i hodnota druhého znaku.
- Vychází-li kovariance naopak záporná, signalizuje to existenci jisté míry nepřímé úměry: stoupne-li hodnota jednoho znaku, úměrně tomu klesne hodnota druhého znaku.
- Nulová kovariance naznačuje, že lineární závislost mezi oběma znaky neexistuje. Jak je vidět, u kovariance nás zajímá především její znaménko.
- Aby však tato charakteristika mohla posloužit lépe jako ukazatel lineární závislosti, převádí se její hodnota na škálu, resp. interval [-1,1], který je vhodnější referencí pro měření intenzity lineární závislosti. Výsledkem tohoto převodu je koeficient párové korelace, a to buď populační, pracujeme-li s populací, nebo výběrový, je-li k dispozici pouze výběrový soubor.

Korelační koeficient

- Populační koeficient párové korelace $\rho = \frac{\mathrm{cov}(X,Y)}{\sigma_X \cdot \sigma_Y}$,
- Výběrový koeficient párové korelace $r = \frac{c_{XY}}{s_X \cdot s_Y}$,
- Populační i výběrový koeficient korelace mohou nabývat pouze hodnot z intervalu [-1,1].
- Vyjde-li populační párová korelace jedna, znamená to, že mezi oběma znaky existuje přesná funkční závislost v podobě přímé úměry (rostoucí přímky).
- Vyjde-li populační korelace naopak minus jedna, existuje mezi oběma znaky přesná funkční závislost v podobě nepřímé úměry (klesající přímky).
- Pokud je populační korelace nulová, říkáme, že znaky *X* a *Y* jsou nezkorelované (nikoliv nezávislé!!).

Děkuji za pozornost