ZÁKLADY STATISTIKY

Doc. RNDr. Jiří Zháněl, Dr.

PRŮVODCE

STATISTICKÝM ZPRACOVÁNÍM

KVANTITATIVNÍCH DAT

Přednášky nk+np 2019

https://is.muni.cz/auth/el/1451/jaro2017/nk2019/um/

DOPORUČENÁ LITERATURA

Anděl, J. (1993). *Statistické metody*. Praha: Matfyzpress.

Cyhelský, L., Kahounová, J. & Hindls, R. (1996). *Elementární statistická analýza*. Praha: Management Press.

Gajda,V. & Zvolská, J. (1982). *Úvod do statistických metod*. PF Ostrava. [Skriptum].

Gibilisco, S. (2009). *Statistika bez předchozích znalostí*. Brno: Computer Press.

DOPORUČENÁ LITERATURA

Hendl, J. (2012). ***Přehled statistických metod zpracování dat. Analýza a metaanalýza dat.*** **Praha: Portál.**

Kovář, R. & Blahuš, P. (1989). ***Aplikace vybraných statistických metod v antropomotorice*. Praha: SPN. [Skriptum].**

Meloun M. & Militký, J. (1994). Statistické zpracování experimentálních dat. Praha: Plus.

DOPORUČENÁ LITERATURA

Meloun M. & Militký, J. (1996). ***Statistické zpracování experimentálních dat. Sbírka úloh*****. Pardubice: Univerzita Pardubice.**

Seger, J. & Hindls, R. (1993). ***Statistické metody v ekonomii*****. Praha: H & H.**

Seger, J. & Hindls, R. (1995). ***Statistické metody v tržním hospodářství*****. Praha: Victoria Publishing.**

A mnoho dalších ...

PROGRAM PŘEDNÁŠEK

1. ÚVOD

1.1 Historie statistiky, pojem a struktura statistiky, základní statistické pojmy

1.2 Teorie měření, měřící stupnice (škály), metodologické problémy měření

PROGRAM PŘEDNÁŠEK

2. DESKRIPTIVNÍ (POPISNÁ) STATISTIKA

2.1 Statistické třídění dat, zpracování a grafické znázornění

2.1.1 Jednorozměrné rozdělení četností

2.1.2 Jednorozměrné intervalové rozdělení četností

2.1.3 Grafické znázornění rozdělení četností

2.2 Míry polohy

2.3 Míry variability

2.3.1 Kvantilové míry variability

2.3.2 Momentové míry variability

PROGRAM PŘEDNÁŠEK

2.4 Standardní skóre

2.5 Míry závislosti

2.5.1 Závislost pevná, volná, statistická a korelační

2.5.2 Lineární korelace a lineární regrese

2.5.3 Součinová a pořadová korelace

3. ANALYTICKÁ STATISTIKA

3.1 Testování statistických hypotéz

1.1 HISTORIE STATISTIKY

Nejstarší písemné památky statistické povahy pocházejí ze Sumeru (nejstarší stát světa 3000 – 2000 př. n. l., Perský záliv).

Hliněné destičky obsahují záznamy o časových intervalech, počtech osob, domácího zvířectva, úrodě atd.

1.1 HISTORIE STATISTIKY

Pojem *statistika* pochází z latinského slova *status* (tj. postavení, stav), …

Počátky statistických postupů využívány již ve středověku ke zjišťování počtu obyvatelstva, velikosti majetku, území, obchodu, armády…

Statistika jako součást přednášek na středověkých univerzitách => *UNIVERZITNÍ STATISTIKA.*

1.1 HISTORIE STATISTIKY

V 17. století se Angličané John Graunt a William Petty zabývali zkoumání různých *hromadných společenských jevů* za pomocí číselných charakteristik skupin obyvatelstva např. *počty narozených a zemřelých osob, počtem obyvatel a složením rodin.*

Postupy byly nazvány *POLITICKÁ ARITMETIKA* (využitelné politicky, používány aritmetické postupy).

1.1 HISTORIE STATISTIKY

17. století: rozvoj *TEORIE PRAVDĚPODOBNOSTI*

- ***ve Francii*** **B. Pascal, P. de Fermat, de Moivre, de Laplace, Poisson;**
- ***v Holandsku*** **Ch. Huygens;**
- ***ve Švýcarsku*** **J. Bernoulli, Euler;**
- **v *Německu*** **C. F. Gauss**
- **v *Rusku*** **Čebyšev, Markov, Ljapunov.**

Postupná integrace

UNIVERZITNÍ STATISTIKY, POLITICKÉ ARITMETIKY a

TEORIE PRAVDĚPODOBNOSTI

docházelo v 19. století k utváření

MODERNÍ STATISTIKY.

V 19. století vznik vědecké teorie statistiky, a její aplikace do statistické praxe, do výzkumu o příčinných vztazích mezi hromadnými jevy (belgický matematik L. A. J. Quételet).

V pozdějších letech dochází k pronikání statistiky do přírodních a technických věd (Angličan Galton, Pearson a Fisher).

HISTORIE STATISTIKY V ČECHÁCH

(Český statistický úřad, www.czso.cz)

Nejstarší dochovaný zápis:

„Soupis majetku litoměřického kostela z roku 1058“

(součást zakládací listiny kapituly sv. Štěpána v Litoměřicích.

HISTORIE STATISTIKY V ČECHÁCH

VÝZNAMNÁ DATA

13. října 1753 … vydán patent císařovny Marie Terezie (1717 – 1780) o každoročním sčítání lidu,

6. března 1897 … zřízen Zemský statistický úřad Království českého, (první statistický úřad na území dnešní České republiky).

Rok 1909 … vyšla první „Statistická příručka království Českého“ (další v roce 1913).

HISTORIE STATISTIKY V ČECHÁCH

VÝZNAMNÁ DATA

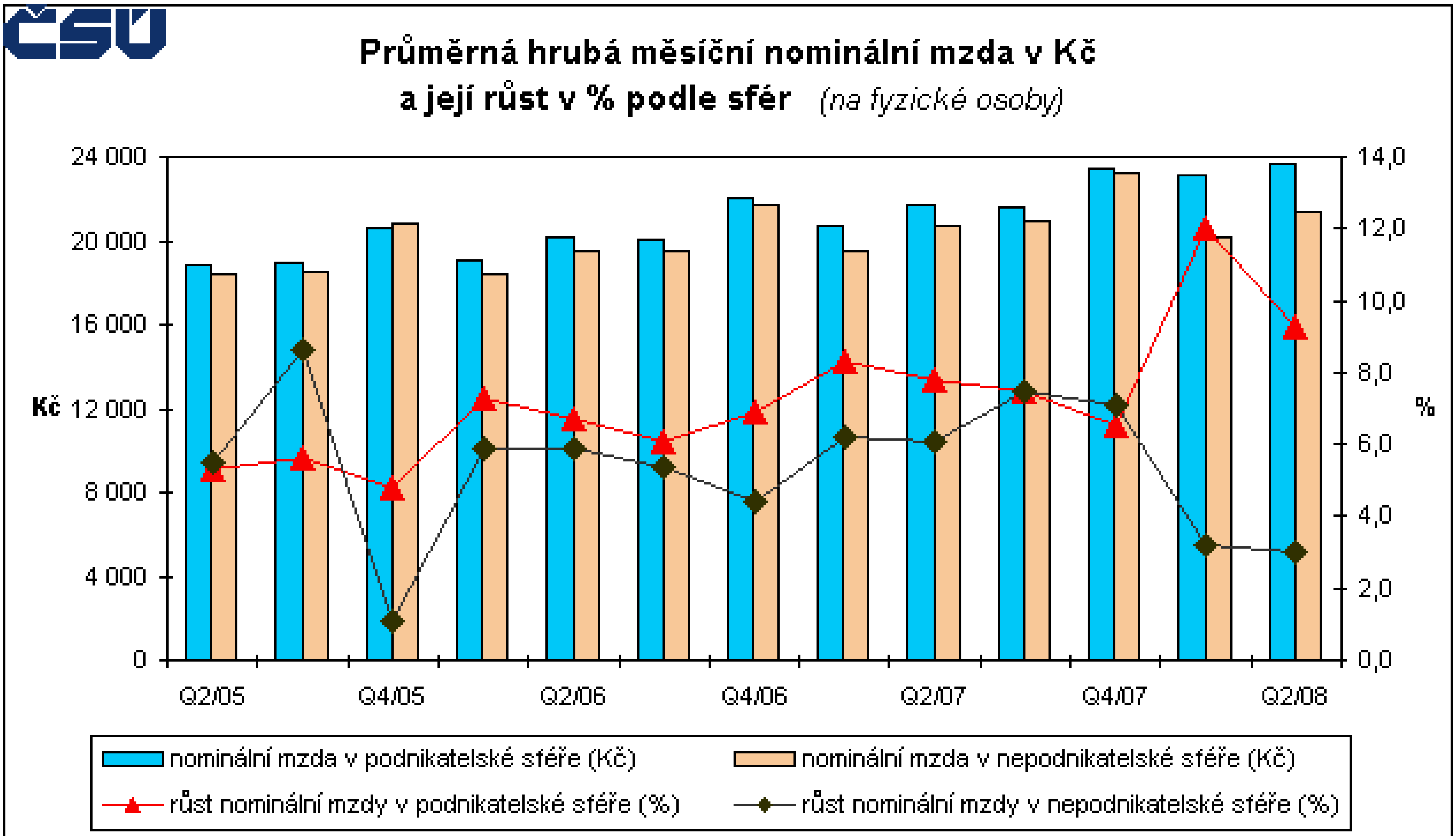
1918 (vznik samostatného Československa) … zákon č. 49 o organizaci statistické služby (1919).

1919 … založen Státní úřad statistický (SÚS) jako orgán pověřený celostátními statistickými šetřeními (např. sčítání lidu).

1.1.1993 (vznik ČR) převzal všechny kompetence Český statistický úřad (ČSÚ).

Český statistický úřad (ČSÚ)

Nejžádanější informace: inflace, makroekonomické údaje, obyvatelstvo, regiony, města, obce, ročenky, sčítání lidu, volební výsledky, základní údaje o ČR.



1.1.2 POJEM A STRUKTURA STATISTIKY

STATISTIKA OBECNĚ

Obor zabývající se zpracováním, rozborem a zveřejňováním informací, které kvantitativně charakterizují zákonitosti společenského života (Encyklopedický slovník, 1982).

1.1.2 POJEM A STRUKTURA STATISTIKY

MATEMATICKÁ STATISTIKA

- **matematický obor zabývající se zpracováním dat a rozborem statistických charakteristik popisovaného statistického souboru** (Encyklopedický slovník, 1982),
- **věda o formálních metodách vyhodnocení dat, které umožňují zpracovat, popsat a analyzovat shromážděná data** (Fleischer, 1988).

Předmět „Základy statistiky“ = opravdu jen ZÁKLADY!

Např. Pravděpodobnost a statistika (Friesl, 2004).

Náhodný jev (definice):

Je-li dána množina Ω (všech výsledků náhodného pokusu, tj. pokusu, jehož výsledek není jednoznačně určen podmínkami, za kterých je prováděn), pak náhodným jevem (v Ω) nazýváme každou podmnožinu množiny Ω.

Jev Ω nazýváme jistý, jev Ø nemožný. Jednotlivé výsledky ω patří do Ω se nazývají elementární jevy.

(MATEMATICKÁ) STATISTIKA

DESKRIPTIVNÍ

(popisná)

ANALYTICKÁ

(inferentní, induktivní, srovnávací)

DESKRIPTIVNÍ STATISTIKA se zabývá **zpracováním** a **popisem** dat.

Poskytuje metody umožňující přehledné a názorné zpracování dat, např.:

- **tabulky,**
- grafické znázornění **rozložení četností,**
- výpočet **statistických charakteristik** (např. **aritmetický průměr** nebo **korelační koeficient**).

ANALYTICKÁ (INFERENTNÍ) STATISTIKA

… vychází z výsledků Deskriptivní statistiky (zpracování dat), umožňuje nám data analyzovat tzn. vyhodnotit, např. stanovit, zda …

Např. střední hodnoty dvou tréninkových skupin A, B sportovců (nebo dvou skupin pacientů) vykazují nějaký významný rozdíl, který může být vysvětlován použitou tréninkovou metodou (nebo léčebným účinkem medikamentace u pacientů).

SYMBOLICKÉ ZNÁZORNĚNÍ FUNKCE STATISTIKY

STATISTIKA = ZPRACOVÁNÍ + POPIS + ANALÝZA DAT

1.1.3 ZÁKLADNÍ STATISTICKÉ POJMY

STATISTICKÝ SOUBOR
je souhrn (množina) statistických jednotek stejného druhu

Rozlišujeme pojmy základní soubor a výběrový soubor.

Rozsah základního souboru N, výběrového souboru n.

Základní soubor **(N) je soubor všech statistických jednotek, které teoreticky mohou být předmětem sledování.**

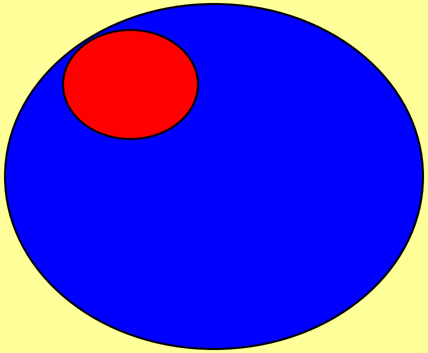
… např. všichni studenti oboru TV a sport = v ČR, Evropě, na světě, …

ZÁKLADNÍ SOUBOR (ZS)

ZS má zpravidla značný rozsah, zjištění zkoumaných vlastností všech prvků je buďto nemožné nebo je příliš časově a ekonomicky náročné.

Výzkumné šetření (zjištění) se proto provádí u vybraných jednotek ze základního souboru = výběrový soubor (n).

Výběrový soubor **je náhodnou podmnožinou prvků** ***základního souboru*** **a reprezentuje jej.**



VÝBĚROVÝ SOUBOR získáváme tzv. NÁHODNÝM VÝBĚREM,

kdy každý prvek základního souboru má stejnou možnost být vybrán.

O vybrání či nevybrání rozhoduje tedy pouze náhoda.

Na základě poznání vlastností *výběrového souboru* se usuzuje – při splnění určitých podmínek - na vlastnosti *základního souboru*.

STATISTICKÉ JEDNOTKY
jsou prvky statistického souboru, které mají alespoň jednu společnou vlastnost (znak)

Statistickými jednotkami mohou být např. osoby, věci, události, jejichž vlastnosti nás zajímají (student).

Zjišťujeme-li u každé statistické jednotky pouze *jeden statistický znak* (např. tělesnou výšku), hovoříme o *jednorozměrném statistickém souboru.*

Zjišťujeme-li *dva nebo více znaků,* hovoříme o dvourozměrném (výška a hmotnost), resp. o vícerozměrném statistickém souboru (3 a více znaků).

STATICKÝ ZNAK

je společná vlastnost jednotek statistického souboru

Statistické znaky **vyjadřují** ***vlastnosti statistických jednotek.***

Dělení statistických znaků:

1. KVALITATIVNÍ vyjádřeny slovně:

Např. plavec či neplavec; trenér I., II. či II. třídy),

- ***alternativní (binární)*** **nabývá-li znak pouze dvou variant např. muž-žena)**

- ***množné*** **(nabývá-li znak více než dvou variant např. držitel zlaté, stříbrné či bronzové medaile).**

Dělení statistických znaků:

2. **KVANTITATIVNÍ (vyjádřeny číselně, např. věk 37 let, tělesná výška 183 cm), které dále dělíme na …**

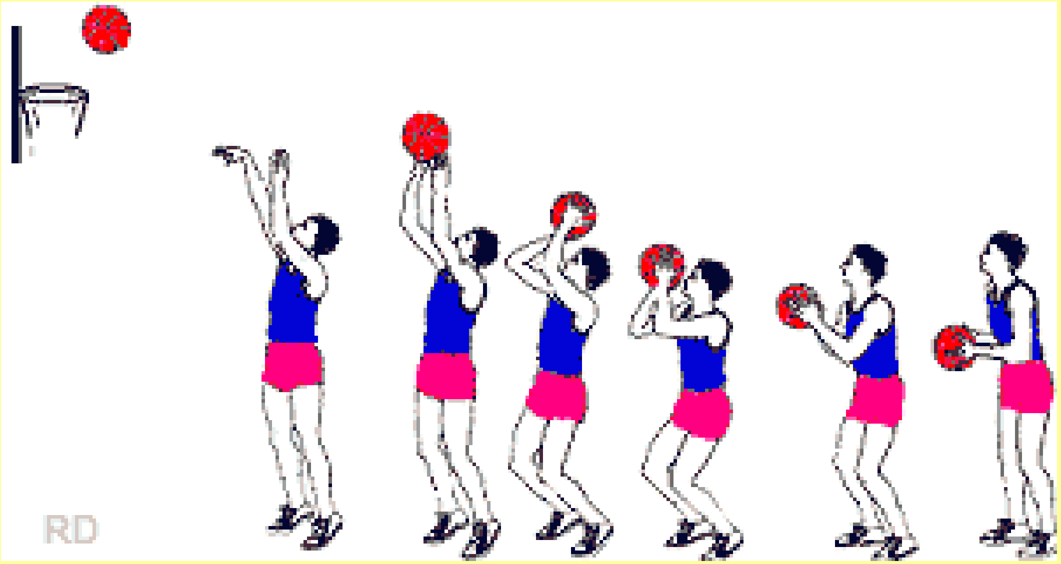
☯ ***spojité*** **neboli** ***kontinuální*** **(nabývají libovolných reálných číselných hodnot např. výsledek v běhu na 100 m, ve skoku vysokém, atd.),**

Dělení statistických znaků:

***2.* KVANTITATIVNÍ (vyjádřeny číselně),**

☯ ***nespojité* neboli *diskrétní* (nabývají pouze některých číselných hodnot, nejčastěji z oboru celých nezáporných čísel.**

např. počet úspěšných hodů na koš, leh-sedy, …).



1.2 TEORIE MĚŘENÍ, MĚŘÍCÍ STUPNICE (ŠKÁLY)

1.2.1 ZÁKLADNÍ POJMY TEORIE MĚŘENÍ

Měření … v průběhu historického vývoje lidské společnosti běžná každodenní procedura (užití hodinek, tachometru automobilu, atd.).

Historické počátky měření … porovnávání objektů s počtem prstů, délkou palce, délkou chodidla, lokte, paže tj. primitivní měřící způsoby.

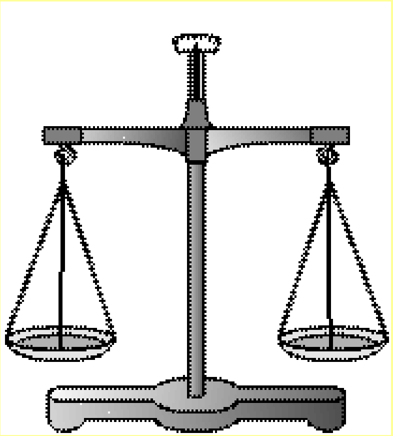
Rozvoj vědy a techniky složitých měřících přístrojích.

1.2 TEORIE MĚŘENÍ, MĚŘÍCÍ STUPNICE (ŠKÁLY)

1.2.1 ZÁKLADNÍ POJMY TEORIE MĚŘENÍ

Problematiku kvantifikace (měření) řeší obor nazývaný TEORIE MĚŘENÍ.

a) Měřitelnost *fyzikálních vlastností* (délka, čas, hmotnost),

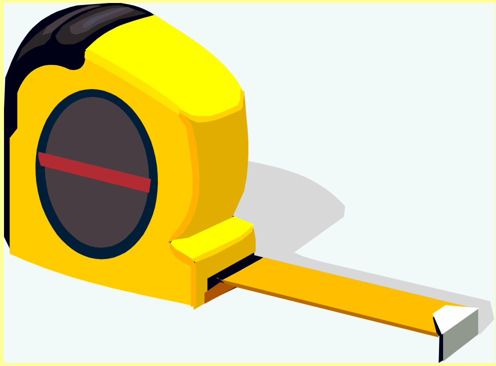


b) Měřitelnost *psychických vlastností* (inteligence, strach, postoje).

REPREZENTAČNÍ TEORIE MĚŘENÍ **(Campbell):**

… měření jako „přiřazování číslic k reprezentaci vlastností“.

… později doplněna (Stevens) o formulaci ***„…za měření lze považovat každé přiřazování číslic k objektům nebo událostem … podle pravidel.***



Klasická koncepce měření rozlišuje

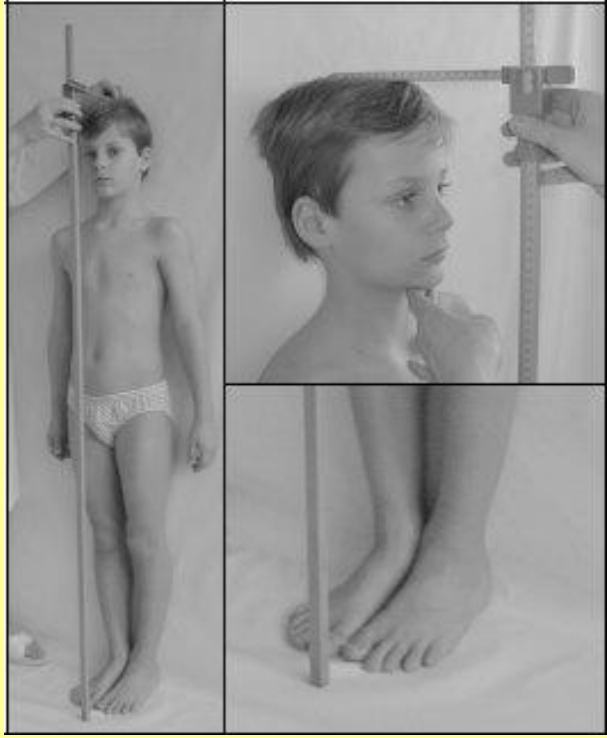
(1) fundamentální měření

(2) odvozené měření

Někteří autoři zmiňují ještě

(3) měření asociativní **(Berka, 1977) resp.** ***asociační*** **(Blahuš, 1996), označované rovněž jako měření** ***per fiat, per Definition, by fiat*** **či** ***měření na základě konvence.***

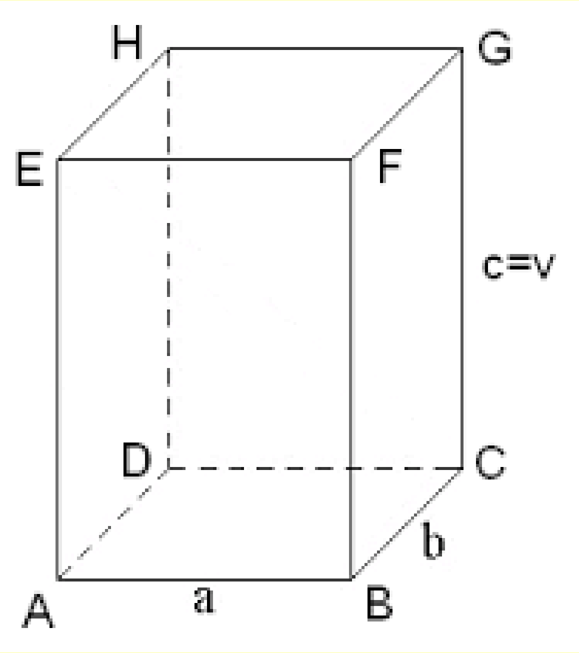
(1) FUNDAMENTÁLNÍ MĚŘENÍ



„se vztahuje na bezprostřední měření veličin“ a je to „každé měření, které nezahrnuje žádná předcházející měření“.

Příklad: měření tělesné výšky

(2) ODVOZENÉ MĚŘENÍ



„předpokládá jiná, dříve provedená měření, z nichž je odvozeno na základě vztahů“; a tedy „závisí na předcházejících měřeních“.

Příklad: měření objemu kvádru

(3) ASOCIATIVNÍ MĚŘENÍ (ASOCIAČNÍ)

je takové měření, kdy „je přímo měřená veličina asociována s nepřímo měřitelnou veličinou“.

Příklad 1: při měření teploty vycházíme ze závislosti změny objemu kapaliny na teplotě.

Bimetalový teploměr – využívá pásek složený ze dvou kovů s různými teplotními součiniteli délkové roztažnosti.

Kapalinový teploměr: využívá teplotní roztažnosti kapaliny (rtuť, líh apod.).

ASOCIATIVNÍ MĚŘENÍ je časté v oblasti sociálních věd – jedná se o *předpokládaný (hypotetický) vztah* mezi

(1)pozorovanými vlastnostmi, které nejsou měřitelné (např. *vytrvalost sportovce*) a

(2) měřitelnými veličinami *(počet uběhnutých metrů* při Cooperově testu).

Příklad 2: testování vytrvalostních schopností (např. Cooper test). *Průběh … vytrvalost x vzdálenost?*

Typickým příkladem asociativního měření je

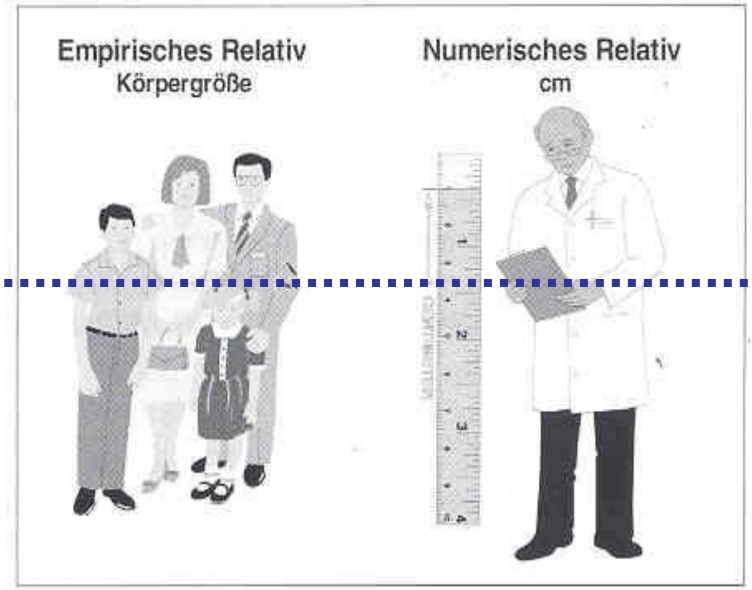
TESTOVÁNÍ MOTORICKÝCH SCHOPNOSTÍ.

1.2.2 MĚŘÍCÍ STUPNICE (ŠKÁLY)

Rozdíl v měření

Empirická proměnná **Numerická proměnná**

Tělesná výška → **cm**

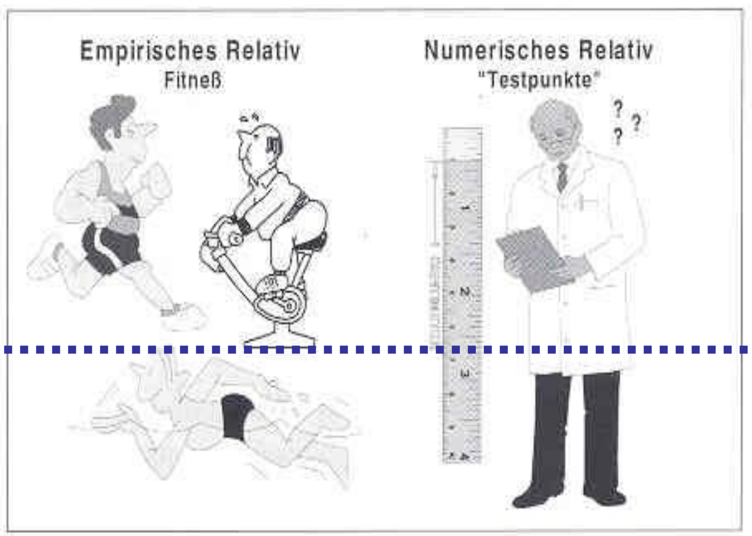


Zuordnungen von Körpergrößen (ER) zu Zahlenwerten (NR)

Rozdíl ve způsobu měření a přiřazení!

Empirická proměnná **Numerická proměnná**

Kondice → **Testové skóre**



Zuordnung von Fitneßausprägungen (ER) zu Testpunkten (NR)

Teorii škál **= relativně samostatná část** ***teorie měření*.**

Pojem ***škála*** **resp.** ***měřítko*** **nebo** ***stupnice*.**

ZÁKLADNÍ DRUHY ŠKÁL (STUPNIC)

1. NOMINÁLNÍ škála
(jmenná, klasifikační)

NEMETRICKÉ

2. ORDINÁLNÍ škála
(pořadová)

INTERVALOVÁ

3. METRICKÉ škály

METRICKÉ

POMĚROVÁ

1. NOMINÁLNÍ ŠKÁLA

(jmenná, klasifikační)

… je škála založena na jakémkoliv *přiřazování číslic* ve smyslu pouhého *pojmenování*.

Jde vlastně o *pojmenování* osob či skupin *čísly*, o uspořádání do tříd, které se navzájem *vylučují*.

Např. pohlaví (M, Ž), kuřák, nekuřák, národnost, čísla hráčů, atd.

1. NOMINÁLNÍ ŠKÁLA

Třídění na znaky:

1. **alternativní (binární, dichotomické) = 2 možnosti**

   (plavec, neplavec; kuřák, nekuřák; muž, žena)

2. **množné (polytomické) = více než 2 možností**

Základní empirickou operací je *„určení rovnosti“*.

Možné relace =, ≠,

Zpracování znaků = neparametrické statistické metody

2. ORDINÁLNÍ ŠKÁLA (pořadová)

... předpokládá přirozené uspořádání objektů vzhledem k nějaké vlastnosti.

Stupnice umožňuje *uspořádání* objektů *do pořadí,*

je možno určit vztah větší či menší, těžší či lehčí, atd.

Nejsou známy odstupy (intervaly) mezi znaky (čísly) !!!

Př. školní známky, stupnice tvrdosti, pořadí v cíli.

Základní empirické operace (2):

„*určením rovnosti*“ a „*určením vztahu více nebo méně*“.

Relace =, ≠, >, <,

Zpracování znaků=neparametrické statistické metody.

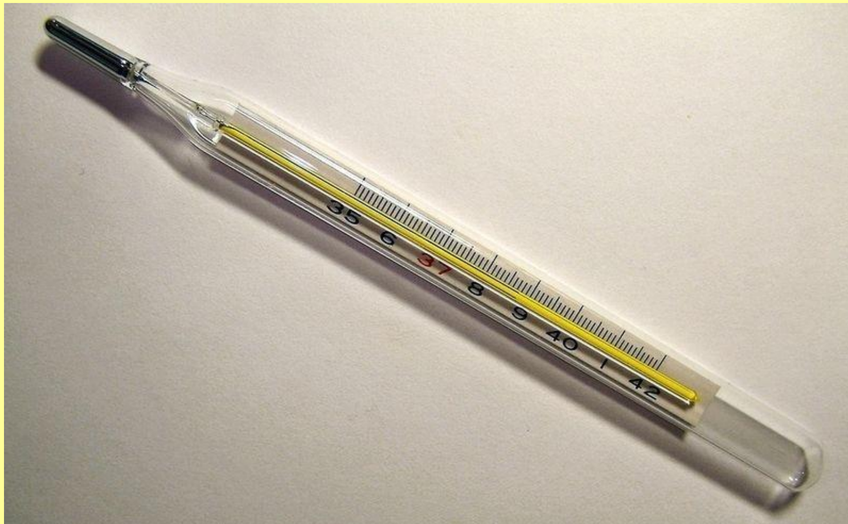
3. METRICKÉ ŠKÁLY (INTERVALOVÁ A POMĚROVÁ)

Nová vlastnost: je zavedena *jednotka měření,* tzn. jsou známy *odstupy (intervaly)* mezi hodnotami (čísly).

3. 1 INTERVALOVÁ ŠKÁLA

… vyžaduje stanovení *měrové jednotky a počátku,* jsou přípustné všechny aritmetické operace.

Nula je zvolená !!! => stanovení počátku dohodou.



Např. letopočet, teplota °C, …

3. METRICKÉ ŠKÁLY (INTERVALOVÁ A POMĚROVÁ)

3. 2 POMĚROVÁ ŠKÁLA

… z formálního hlediska vlastně intervalová *škála s přirozeným počátkem,* jsou přípustné všechny aritmetické operace.

Nula je absolutní … (nepřítomnost jevu).

Např. čas, věk, výška, hmotnost, teplotní stupnice dle Kelvina (v podstatě všechny fyzikální jednotky).

Statistické metody:

parametrické i neparametrické.

Přehled typů škál (upraveno, Bruhn, 1986; Roth, 1995)

TYP ŠKÁLY	NEMETRICKÉ ŠKÁLY		METRICKÉ ŠKÁLY	
	NOMINÁLNÍ	ORDINÁLNÍ	INTERVALOVÁ	POMĚROVÁ
Příklady	Číselné označení barev, psychologického typu, pohlaví, atd.	Školní známky, stupnice tvrdosti, služební pořadí, Richterova stupnice	Teplota ve °C, Fahrenheita, letopočet, inteligenční kvocient	Teplota °Kelvina, věk, váha, výška, velikost úhlu, čas
Operace	$=$, $\neq$	$=$, $\neq$, $>$, $<$	Navíc: intervaly, nula zvolená	Navíc: nula absolutní
Statistické charakteris.	Modus, absolutní a relativní četnosti	Navíc: medián, kvantily a kvantilové odchylky, procentily	Navíc: arit. Průměr, směrodat.odchylka, šikmost, špičatost	Navíc: koeficient variability, geometr. průměr
Testy Významnosti	χ^2 - test, McNemar test, Cochran test,...	Znaménkový test, Mann-Whitney U-test, Friedmanova pořadová analýza variance, aj.	Parametrické metody: F-test t-test (pro závislé či nezávislé soubory)	Parametrické metody: F-test t-test (pro závislé či nezávislé soubory)
Míry závislosti	Kontingenční a čtyřpolní koeficient	Navíc: pořadová korelace	Navíc: Pearsonova součinová korelace	Navíc: Pearsonova součinová korelace
Statistické metody	Některé neparametrické metody	Všechny neparametrické metody	Všechny neparametrické a parametrické metody	Všechny neparametrické a parametrické metody

x^2 - test

Jsou rodinný stav nevěsty a ženicha závislé znaky?

Původní rodinný stav ženicha	Původní rodinný stav nevěsty			Celkem
	svobodná	ovdovělá	rozvedená	
svobodný	75 564	824	3 463	79 851
ovdovělý	1 370	904	798	3 072
rozvedený	4 603	590	2 943	8 136
Celkem	81 537	2 318	7 204	91 059

McNemarův test

Ovlivňuje požití alkoholu správné projetí trati?

Před požitím alkoholu	Po požití alkoholu		Celkem
	bez chyby	chybně	
bez chyby	45	35	80
chybně	15	5	20
Celkem	60	40	100

Znaménkový test

Je zvýšení tlaku stejně pravděpodobné jako jeho pokles?

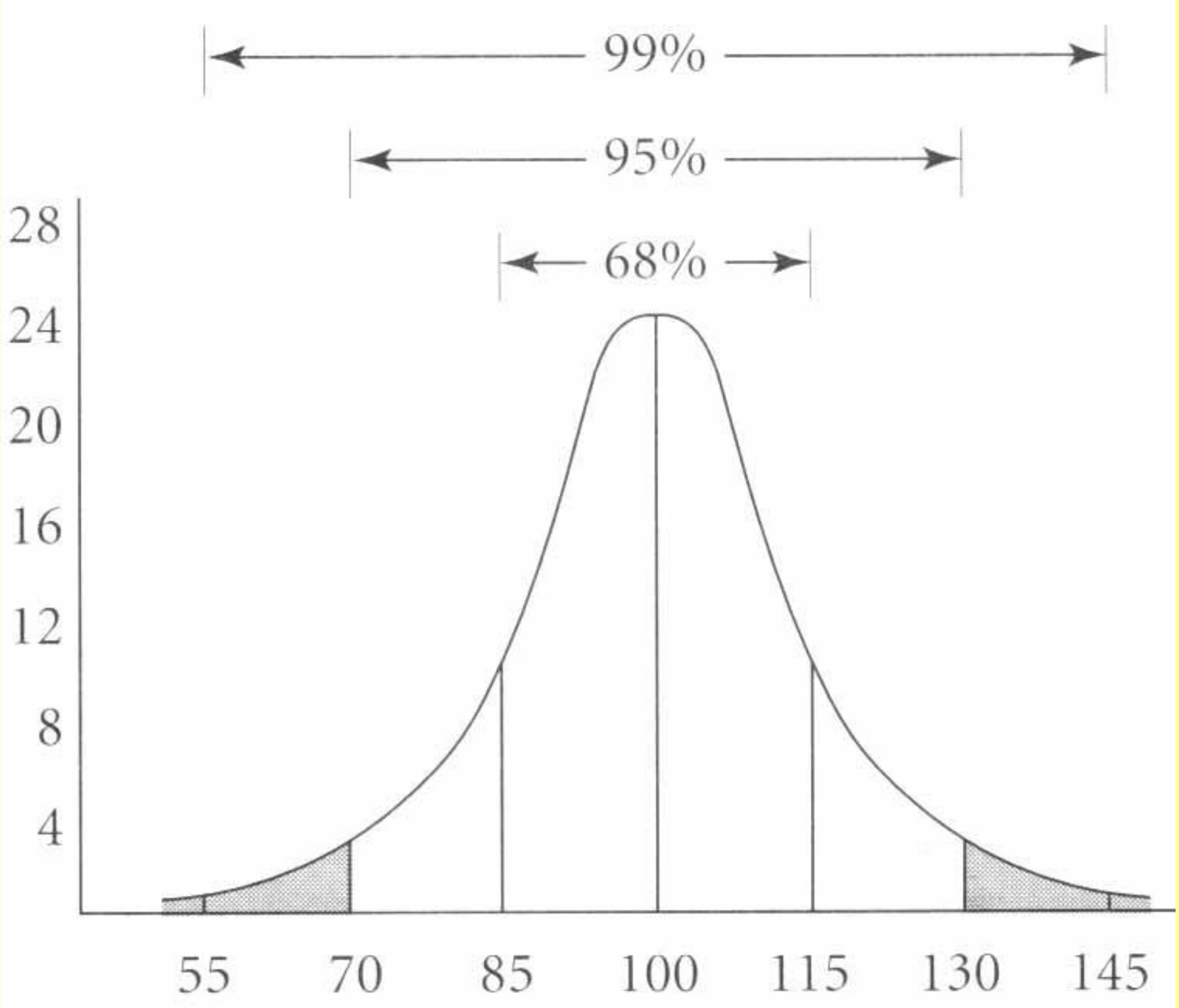
Osoba	A	B	C	D	E	F	G	H
tlak před pokusem	130	185	162	136	147	181	138	139
tlak po pokusu	139	190	175	135	155	175	158	149
Rozdíly	9	5	13	-1	8	-6	20	10

Friedmanův test

Vliv preparátů na srážlivost krve

Osoba	preparát			
	K	A	B	C
A	11,3	11,2	11,4	11,0
B	11,9	12,1	11,8	9,5
C	11,8	13,2	12,0	11,1
D	12,1	12,8	12,0	12,5
E	11,2	13,5	11,5	9,0
F	11,3	12,5	11,5	8,4
G	10,8	10,7	10,9	9,7

Inteligenční kvocient (IQ) je index inteligence, který má normální rozložení s průměrem 100 a standardní odchylkou 15.



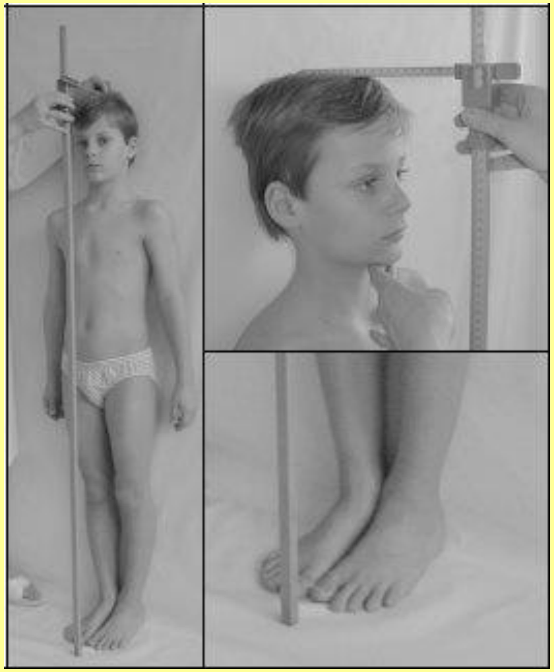
POSTUP PŘI URČENÍ TYPU ŠKÁLY: A. výška (cm)

1. Je známa *jednotka měření?* ANO Škály metrické

2. Je počátek *zvolený* nebo **absolutní?** absolutní **POMĚROVÁ**

3. Lze *stanovit pořadí?* Nemá smysl zjišťovat

4. Jedná se jen o *pojmenování znaků* čísly? Nemá smysl zjišťovat



Znaky ?

- Kvantitativní
- Výška = spojitý

POSTUP PŘI URČENÍ TYPU ŠKÁLY: B. dějepis (známka)

1. Je známa *jednotka měření?* NE => Nemohou být metrické

2. Je počátek *zvolený* nebo absolutní? Nemá smysl zjišťovat

3. Lze *stanovit pořadí?* ANO => ORDINÁLNÍ

4. Jedná se jen o *pojmenování znaků* čísly? Nemá smysl zjišťovat

Znak?

- Kvantitativní
- Dějepis = spojitý

Pozn. Známky jsou spojitými znaky, i když jsou měřeny pouze na ordinální škále.

Klasifikujte znaky obsažené v tabulce – správnou odpověď označte křížkem (X)

Znak	ŠKÁLA				ZNAK	
	Nominální (a)	Ordinální (b)	Intervalová (c)	Poměrová (d)	Spojitý (e)	Diskrétní (f)
1. Pohlaví						
2. Věk						
3. Počet sourozenců						
4. Známka z matematiky						
5. Inteligenční kvocient						
6. Hodnocení v krasobruslení						
7. Výkon ve skoku dalekém						

Pozn. Známky jsou spojitými znaky, i když jsou měřeny pouze na ordinální škále.

Znak	ŠKÁLA				ZNAK	
	Nominální (a)	Ordinální (b)	Intervalová (c)	Poměrová (d)	Spojitý (e)	Diskrétní (f)
1. Pohlaví	■					■
2. Věk				■	■	
3. Počet sourozenců				■		■
4. Známka z matematiky		■			■	
5. Inteligenční kvocient			■		■	
6. Hodnocení v krasobruslení				■	■	
7. Výkon ve skoku dalekém				■	■	

Řešení: 1. a, f; 2. d, e; 3. d, f; 4. b, e; 5. c, e; 6. d, e; 7. d, e.
Pozn. Známky jsou spojitými znaky, i když jsou měřeny pouze na ordinální škále.

ANALÝZA JEDNOROZMĚRNÉHO SOUBORU

METODY DESKRIPTIVNÍ STATISTIKY

1. URČENÍ TYPU ŠKÁLY (nominální, ordinální, metrické)

a) nominální + ordinální ⇒ *neparametrické stat. metody*

b) metrické ⇒ *parametrické statistické metody*

2. ROZLOŽENÍ ČETNOSTÍ ZNAKŮ (NORMÁLNÍ ČI JINÉ)

a) normální ⇒ *parametrické statistické metody*

b) jiné ⇒ *neparametrické statistické metody*

3. VÝPOČET ZÁKLADNÍCH STATISTICKÝCH CHARAKTERISTIK

a) míry centrální tendence

b) míry variability

c) míry závislosti

2. DESKRIPTIVNÍ STATISTIKA

2.1 STATISTICKÉ TŘÍDĚNÍ DAT

Výsledkem měření, testování (statistického šetření) jsou *neuspořádané, neroztříděné a nepřehledné* statistické údaje

Dynamometr: 31; 28; 27; 23; 18; 22; 36; 25; 19; 26

Chceme-li získat podrobnější informace, je třeba údaje uspořádat - tato činnost se nazývá *statistické zpracování (třídění) dat*. Nejjednodušším způsobem statistického zpracování dat je tzv. *tabulka rozdělení (rozložení) četností.*

2.1.1 JEDNOROZMĚRNÉ ROZDĚLENÍ ČETNOSTÍ

Je-li *JEDNA VLASTNOST* statistického souboru charakterizovaná *JEDNÍM STATISTICKÝM ZNAKEM*, hovoříme o *jednorozměrném statistickém souboru (o jednorozměrném rozdělení resp. rozložení četností).*

Konstrukce tabulky - postup vhodný pro:

(1) nespojité kvantitativní statistické znaky

(např. počet dětí v rodině, úspěšné koše),

(2) spojité statistické znaky s malým počtem výskytu

(např. pro statistické soubory s malým rozsahem).

PŘÍKLAD 1. **Při** ***dvakrát opakovaném testování střelby na koš*** **byly u deseti osob (n=10) zjištěny výsledky uvedené v tabulce (zaznamenán počet úspěchů z deseti pokusů při 1. resp. 2. testování).**

Tabulka (hrubé skóre)

Hráč	A	B	C	D	E	F	G	H	J	K
x_i	7	6	7	8	9	8	8	8	9	10
y_i	4	8	6	8	7	8	7	4	8	10

Pro znaky x_i sestavte (frekvenční) tabulku rozdělení četností.

Posouzení znaků x_i: ***kvantitativní, nespojité, poměrová*** **⇒ …**

⇒ … tabulka jednorozměrného rozdělení četností.

Frekvenční tabulka jednorozměrného rozdělení četností.

X_i	Čárkovací metoda	n_i	f_i	Kumulativní četnost N_i	Kumulativní četnost F_i
6	/	1	0.1	1	0.1
7	/ /	2	0.2	3	0.3
8	/ / / /	4	0.4	7	0.7
9	/ /	2	0.2	9	0.9
10	/	1	0.1	10	1.0
Σ		10	1.0	-	-

Vysvětlivky:

n…rozsah souboru x_i…hodnota znaku

n_i…absolutní četnost f_i…relativní četnost   ($f_i = n_i / n$)

Ni … absolutní kumulativní četnost

Fi … relativní kumulativní četnost

Absolutní četnost – vyjadřuje absolutní výskyt jednotlivých znaků, **relativní četnost** – vyjádření v procentech.

Kumulativní relativní četnost – vyjadřuje v % (po vynásobená stem) jaké procento rozsahu souboru má odpovídající variantu a menší hodnotu dané proměnné.

$F_i = 0{,}7$ => *70 % hráčů* dosáhlo výsledku *8 úspěšných pokusů a méně.*

X_i	Čárkovací metoda	n_i	f_i	Kumulativní četnost N_i	F_i
6	/	1	0.1	1	0.1
7	/ /	2	0.2	3	0.3
8	/ / / /	4	0.4	7	0.7
9	/ /	2	0.2	9	0.9
10	/	1	0.1	10	1.0
Σ		10	1.0	-	-

2. 1. 2 JEDNOROZMĚRNÉ INTERVALOVÉ (SKUPINOVÉ) ROZDĚLENÍ ČETNOSTÍ

Konstrukce tabulky *jednorozměrného intervalového rozdělení četností* je postup vhodný pro:

(1) spojité kvantitativní statistické znaky

(např. výsledky měření běhu na 100 m, tělesné výšky, skoku dalekého),

(2) nespojité statistické znaky s velkým počtem výskytů.

DOPORUČENÁ PRAVIDLA

pro konstrukci tabulky jednorozměrného intervalového rozložení četností

URČENÍ ŠÍŘKY A POČTU INTERVALŮ

Variační rozpětí (R) $R = x_{max} - x_{min}$

Šířka intervalu (h) $h = 0{,}08 \times R$ $h = (x_{max} - x_{min})/k$

Počet intervalů (k) $k = \sqrt{n}$

$k \leq 5.\ \log n$

$k \approx 1 + 3.3 \log n$

(Sturgesovo pravidlo)

Je-li n < 30 doporučuje se vytvořit ne více než 6 intervalů.

Je-li 30 < n < 100 doporučuje se vytvořit 7 až 10 intervalů.

POZOR !

Intervaly musí být vytvořeny tak,

aby jeden statistický znak

nemohl být současně zařazen

do dvou různých intervalů!!!

Intervaly na sebe musejí navazovat!!!

POZOR !

PŘÍKLAD 2. Pro znaky y_i sestavte tabulku *skupinového (intervalového) rozdělení četností.*

Hráč	A	B	C	D	E	F	G	H	J	K
x_i	7	6	7	8	9	8	8	8	9	10
y_i	4	8	6	8	7	8	7	4	8	10

Pomocné výpočty pro určení šířky (h) a počtu intervalů (k)

Variační rozpětí (R) $R = x_{max} - x_{min}$ $R = 10 - 4 = 6$

Šířka intervalu (h) $h = 0{,}08 \times R$ $h = 0.08 \times 6 = 0{,}48 \approx 1$ **(pokus)**

PŘÍKLAD 2. Pro znaky y_i sestavte tabulku

skupinového (intervalového) rozdělení četností.

Pomocné výpočty pro určení

šířky (h) a počtu intervalů (k)

Počet intervalů (k)

$k = \sqrt{n}$ $\quad$ $k \leq 5.\ \log n$ $\quad$ $k \approx 1 + 3.3 \log n$

$k = 3.16$ $\quad$ $k \leq 5$ $\quad$ $k \approx 4.3\ (\log 10 = 1)$

$\Rightarrow$ ***Doporučená šířka intervalu: 1***

$\Rightarrow$ ***Doporučený počet intervalů: 3 až 5***

Tabulka skupinového (intervalového) rozdělení četností (znak y_i).

Třída	Interval	Střed	n_i	f_i	N_i	F_i
1	4 – 5	4,5	2	0,2	2	0,2
2	6 – 7	6,5	3	0,3	5	0,5
3	8 – 9	8,5	4	0,4	9	0,9
4	10 –	10,5	1	0,1	10	1,0
Σ	-	-	10	1,0	-	-

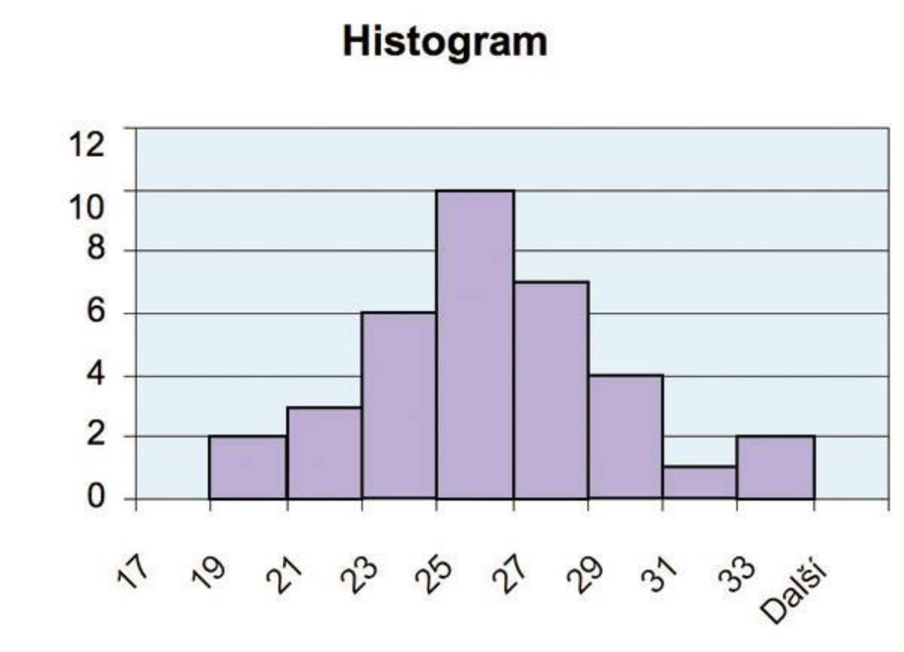
2. 1. 3 GRAFICKÉ ZNÁZORNĚNÍ ROZDĚLENÍ ČETNOSTÍ

Grafické znázornění = přehlednější a názornější forma znázornění *rozdělení četností.*

1) HISTOGRAM ČETNOSTÍ

(sloupkový diagram, sloupcový graf)

Histogram … jedna z nejčastěji užívaných forem grafického znázornění *rozdělení četností.*



Histogram je tvořen sloupci

… jejich šířka odpovídá *šířce třídního intervalu*,

… jejich výška odpovídá *absolutní četnosti*

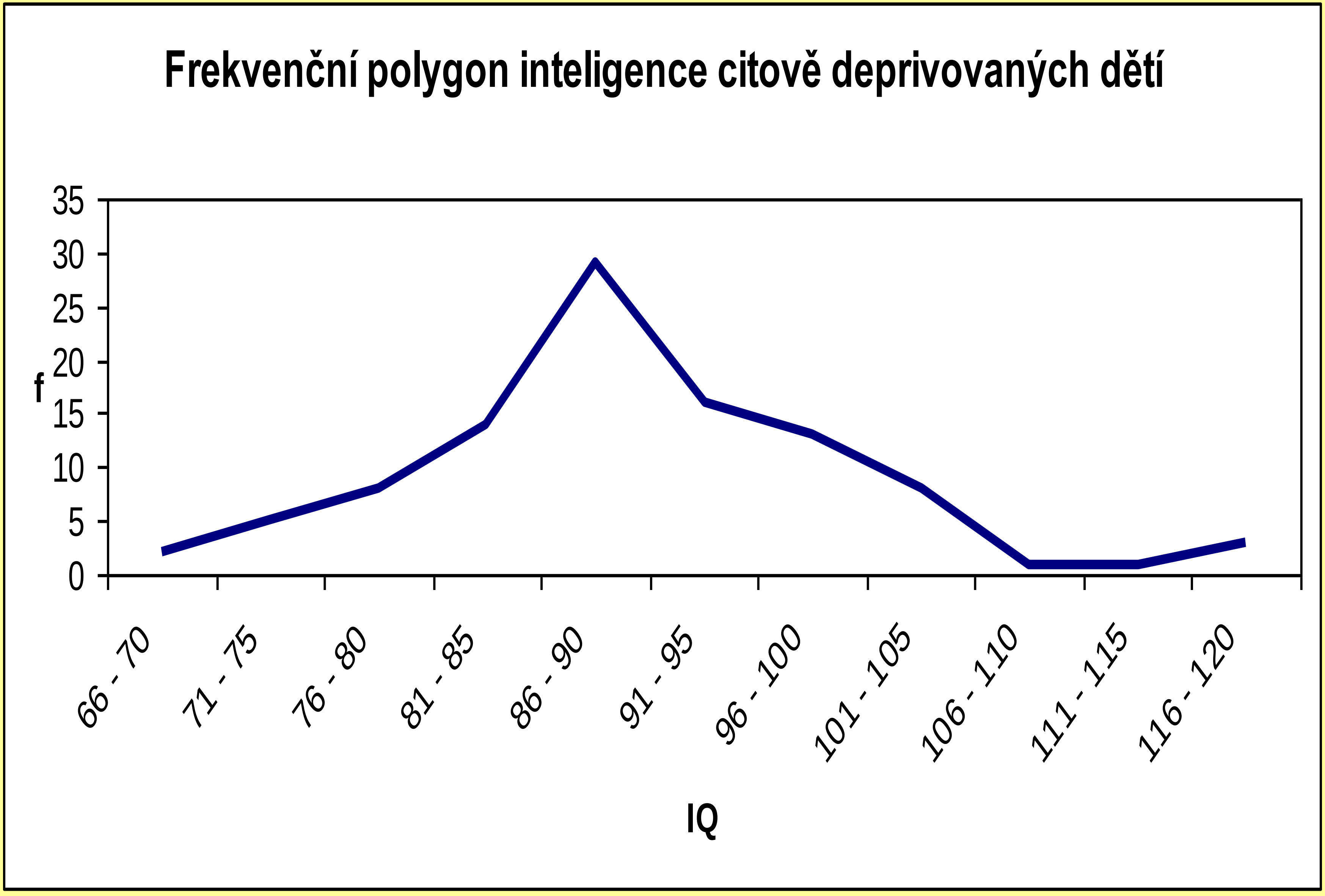
sledovaného statistického znaku.

2) (FREKVENČNÍ) POLYGON

Forma grafického znázornění rozdělení četností, kdy *místo sloupců* použijeme ke znázornění rozdělení četností *lomenou čáru.*

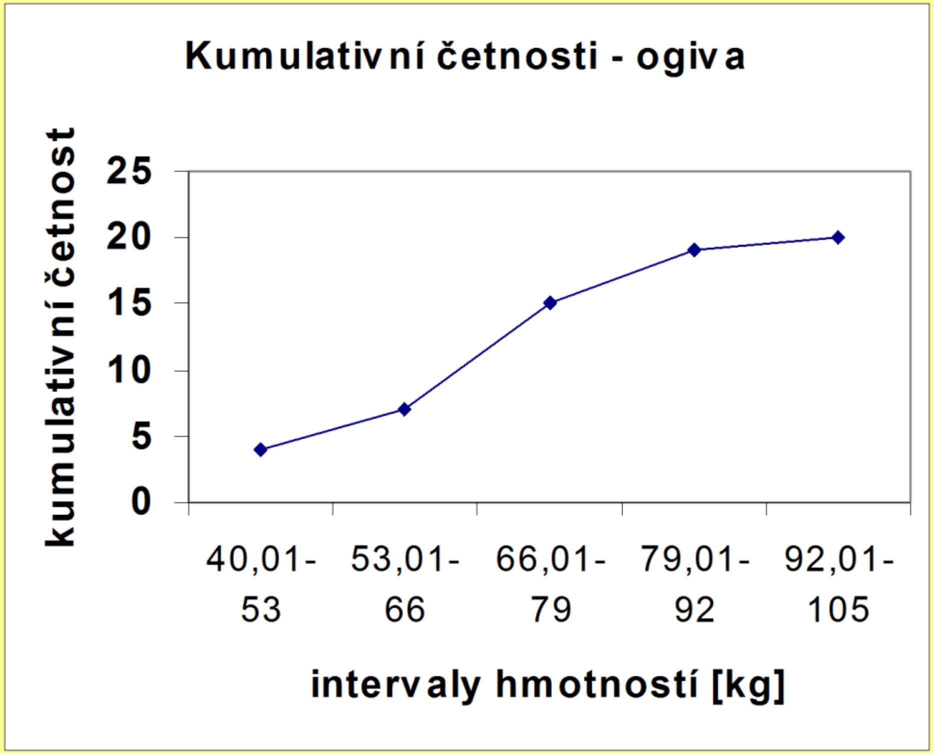
Tato lomená čára je *spojnice bodů* vytvořených v průsečících *středů intervalů* a *příslušných četností.*

2) (FREKVENČNÍ) POLYGON



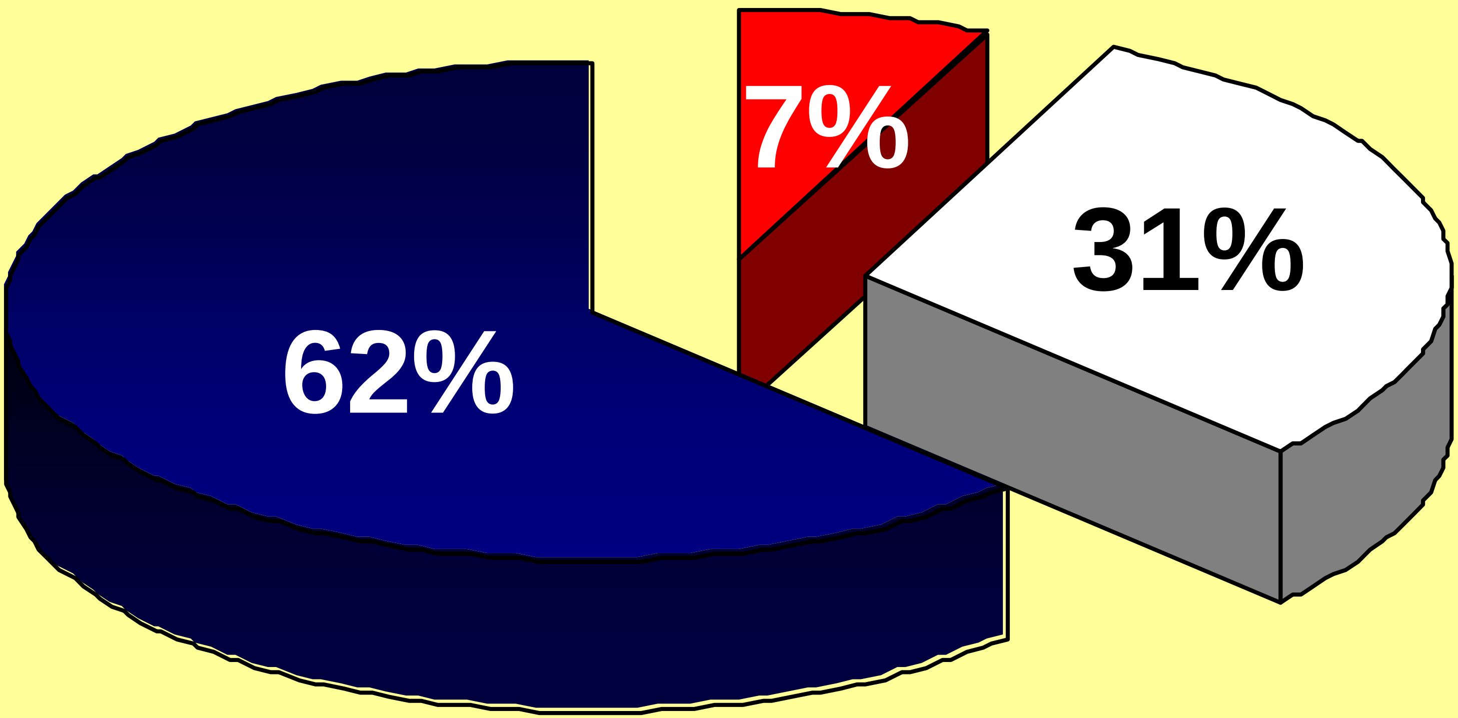
3) (GALTONOVA) OGIVA

Pojem *ogival* je v architektuře používán pro lomený oblouk, ve statistice tento pojem charakterizuje *esovitě lomenou křivku* znázorňující *kumulativní četnosti* *(absolutní nebo relativní).*

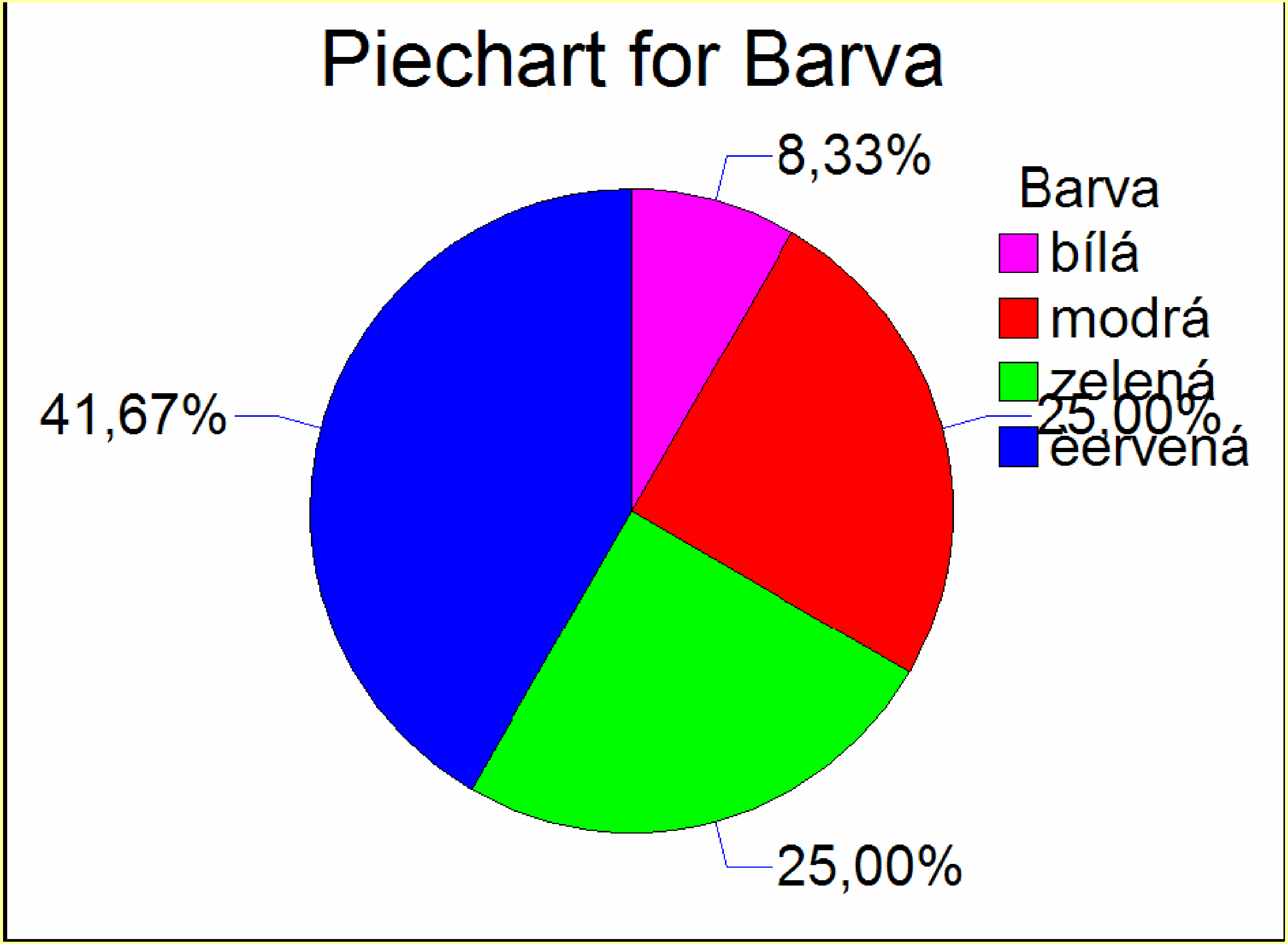


4) VÝSEČOVÝ (SEKTOROVÝ) GRAF

Jedná se o *kruhový graf*, vyjadřující *relativní četnosti* jako charakteristiku struktury daného souboru (nejčastěji v %).



4) VÝSEČOVÝ (SEKTOROVÝ) GRAF

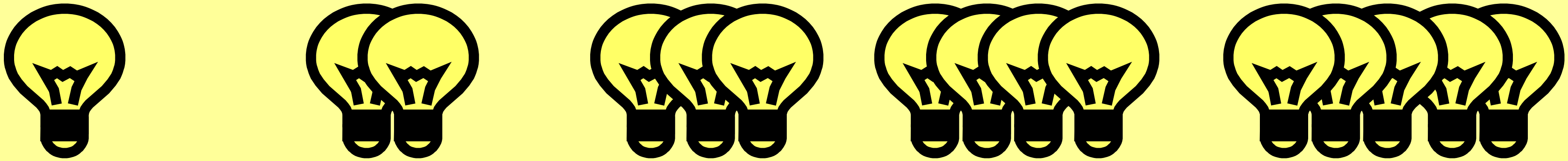


5) PIKTOGRAM

Piktogram = grafický znak znázorňující *pojem* nebo *sdělení* obrazově (např. dopravní značky), též piktograf. Vyjadřuje absolutní četnosti bez nároků na přesnost, má spíše informativní charakter a používá obrazových symbolů (např. lokomotiva, váček s penězi, postava vojáka).

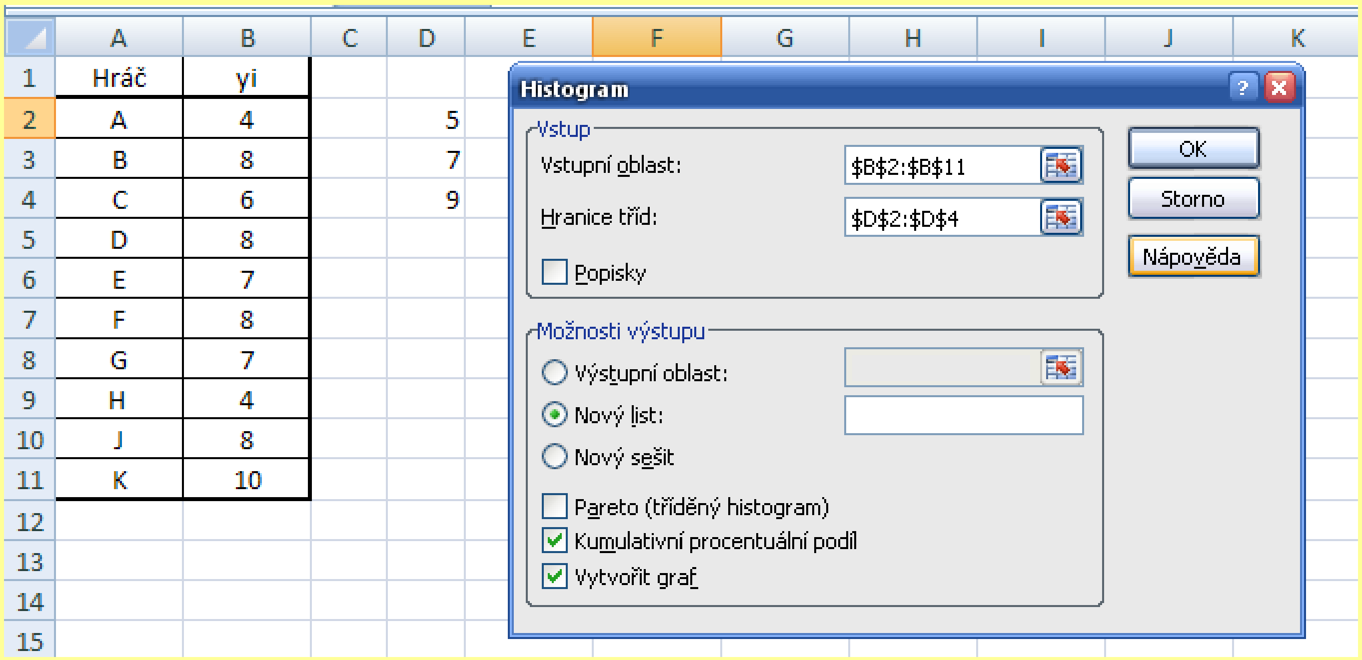
Spotřeba energie v městě X v letech

1960	1970	1980	1990	2000



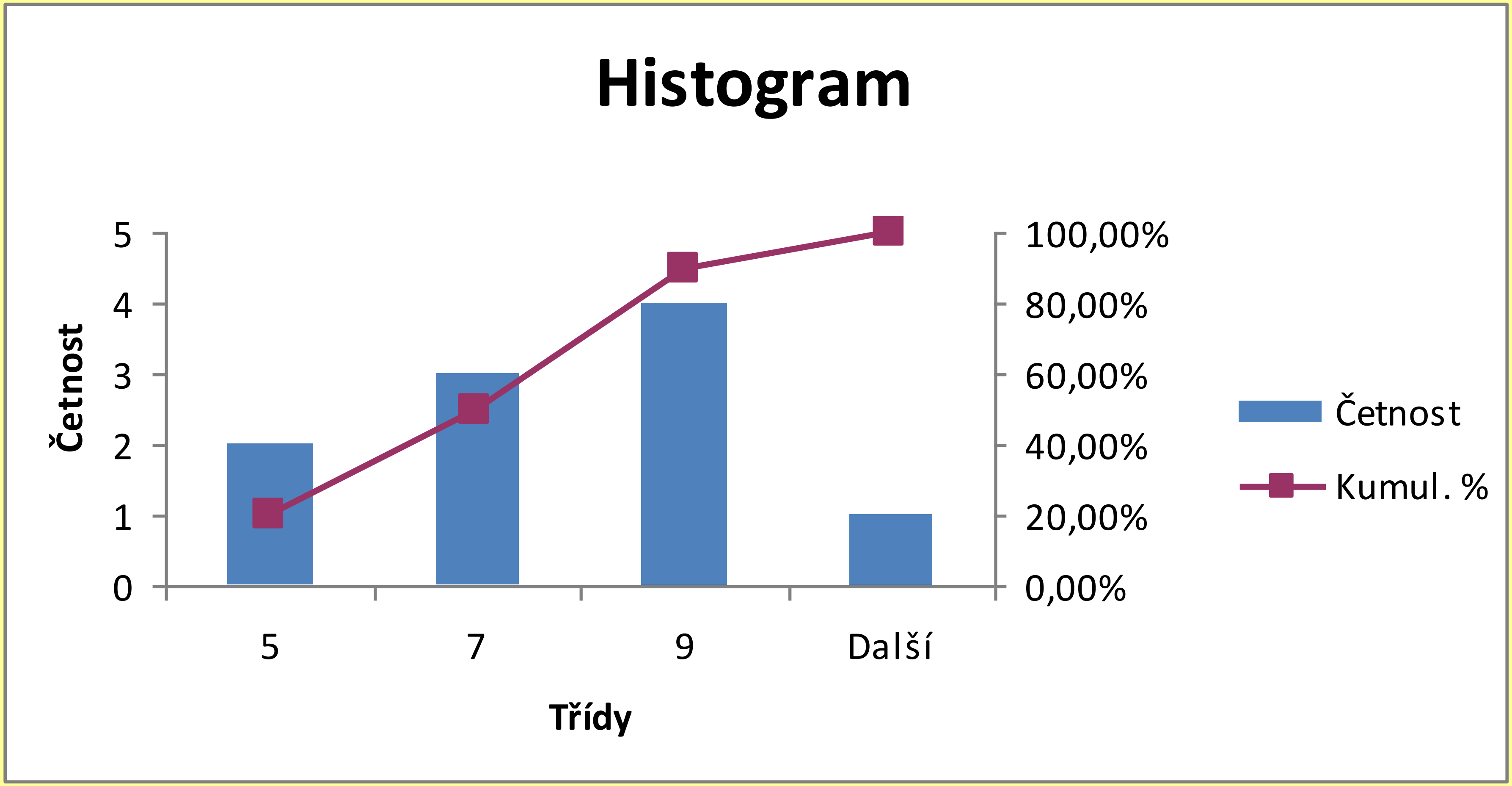
10 MW	22 MW	28 MW	43 MW	52 MW

Pro znaky yi sestavte tabulku skupinového (intervalového) rozdělení četností.

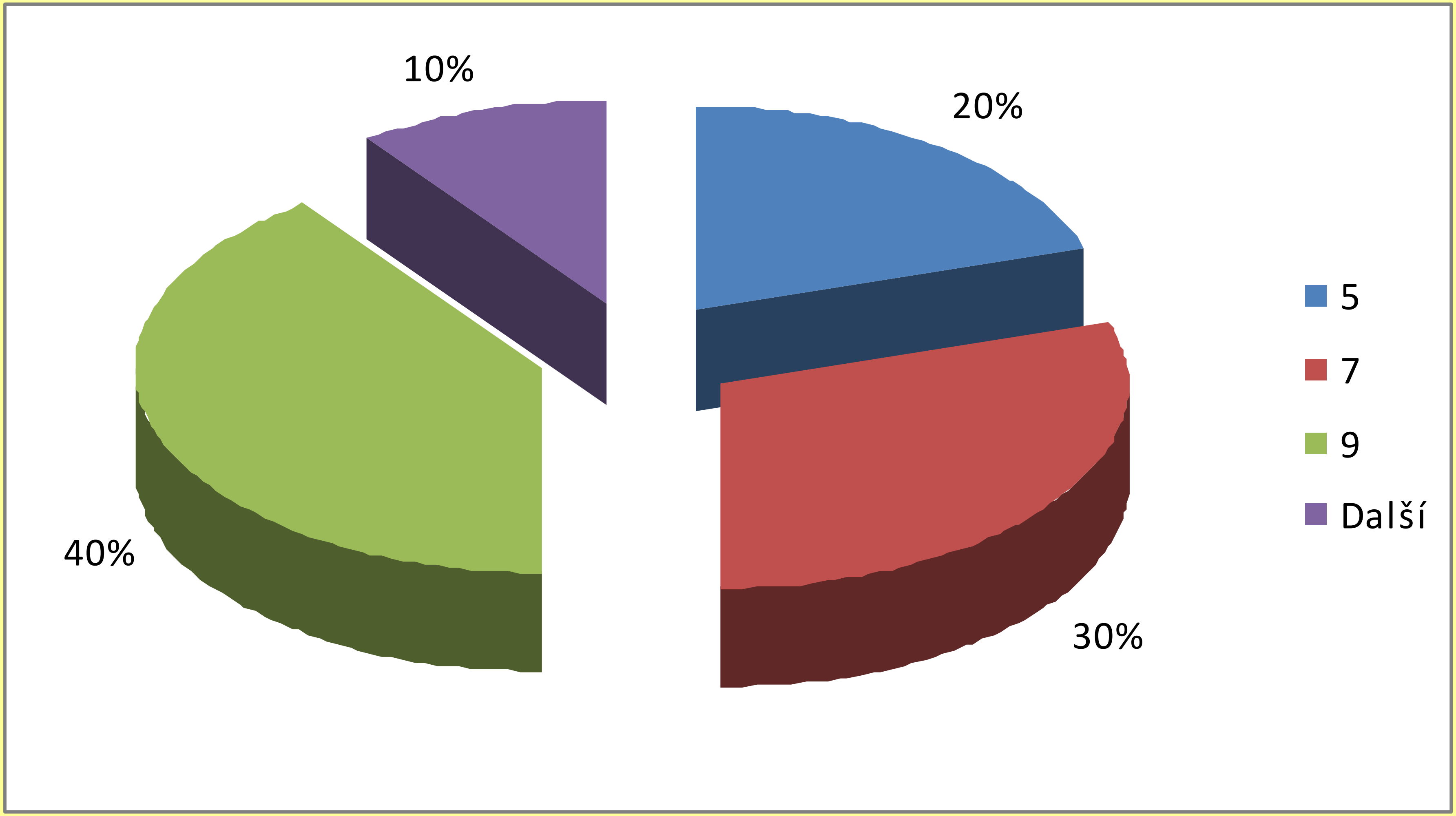


Třídy	*Četnost*	*Kumul. %*
5	2	20,00%
7	3	50,00%
9	4	90,00%
Další	1	100,00%

Histogram četností



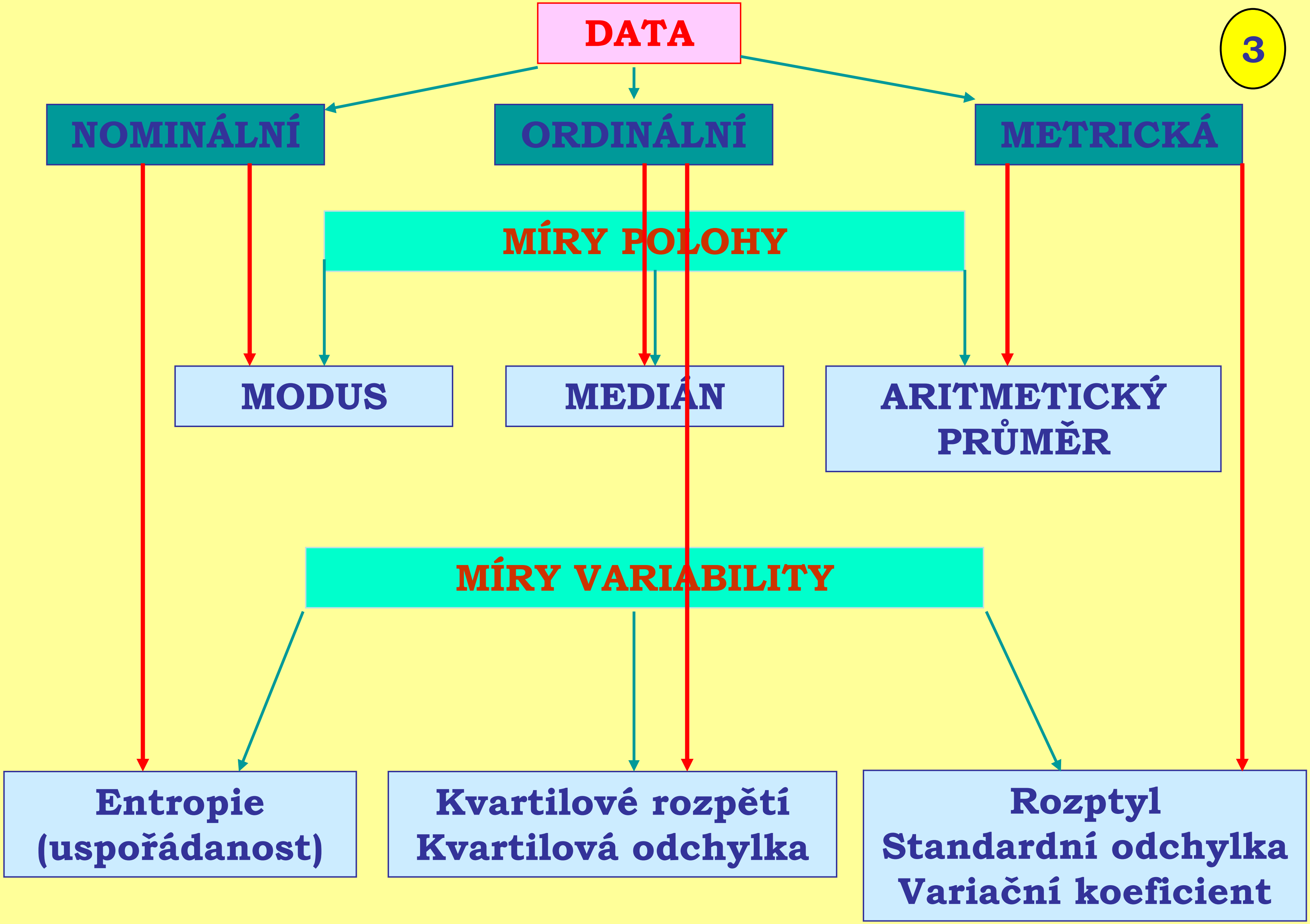
Výsečový (sektorový) graf



ZÁKLADNÍ STATISTICKÉ CHARAKTERISTIKY

DATA

NOMINÁLNÍ, ORDINÁLNÍ, METRICKÁ


3
DATA
NOMINÁLNÍ
ORDINÁLNÍ
METRICKÁ
MÍRY POLOHY
MODUS
MEDIÁN
ARITMETICKÝ PRŮMĚR
MÍRY VARIABILITY
Entropie (uspořádanost)
Kvartilové rozpětí
Kvartilová odchylka
Rozptyl
Standardní odchylka
Variační koeficient

2. 2 MÍRY POLOHY

Míry polohy **(neboli *míry centrální tendence*)**

…charakterizují *úroveň* statistického souboru z hlediska *polohy jeho střední hodnoty,*

…zevšeobecňují, zastupují, reprezentují jednotlivé hodnoty sledovaného statistického znaku,

…umožňují srovnání polohy dvou či více rozdělení četností, srovnání střední úrovně dvou či více souborů.

Dynamometrie: 18; 19; 22; 23; 25; 26; 27; 28; 31; 36

NEJČASTĚJI POUŽÍVANÉ MÍRY POLOHY

1. NOMINÁLNÍ STUPNICE (DATA)

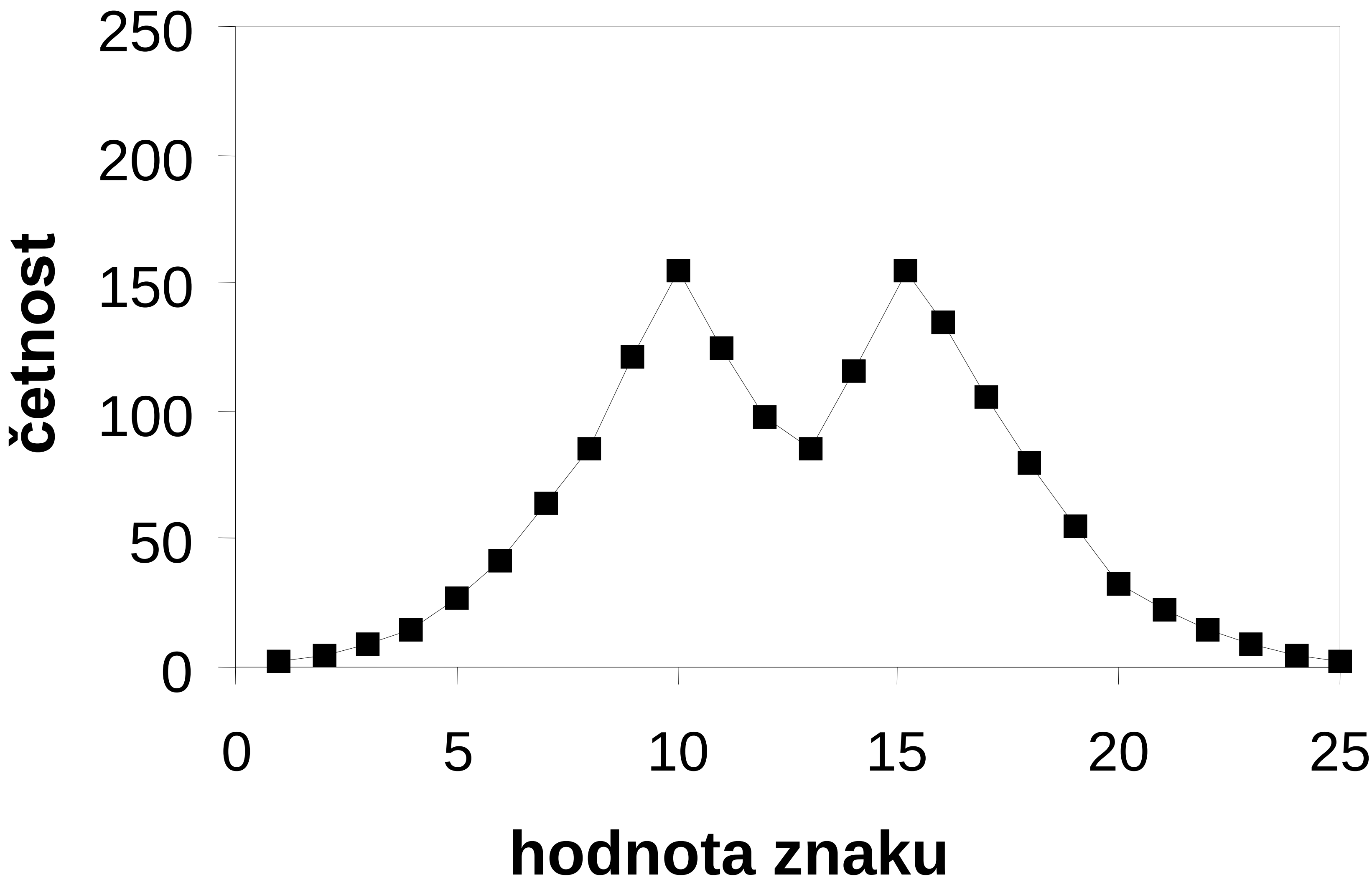
MODUS (Mo) … označuje *nejčastěji se vyskytující* hodnotu statistického souboru (hodnota s největší četností).

Modus … je nejsnáze zjistitelná míra polohy.

Soubor může mít jeden či více modů (soubor *bimodální*, soubor *trimodální*).

Modus je použitelný pro nominální stupnice (a všechny vyšší).

Rozdělení bimodální



2. ORDINÁLNÍ STUPNICE (DATA)

MEDIÁN (Me) … označuje *prostřední člen variační řady* (dělí výsledky seřazené podle velikosti na polovinu).

Medián … není citlivý na velikost extrémních (krajních) hodnot.

Medián … je použitelný pro ordinální stupnici (a vyšší).

Ukázka výpočtu pro sudý a lichý počet dat:

xi : 6 7 7 8 8 8 8 9 9 10 (sudý počet) **Mo = 8 Me = 8**

xi : 6 7 7 8 8 8 8 9 9 10 10 (lichý počet) **Mo = ? Me = ?**

18; 19; 22; 23; 25; 26; 27; 28; 31; 36

3. METRICKÉ STUPNICE (DATA)

ARITMETICKÝ PRŮMĚR … nejpoužívanější míra polohy. $\bar{x}$

… je použitelný (pouze!) pro metrické škály (stupnice).

… je to součet všech hodnot statistického souboru dělený rozsahem souboru (n).

a) aritmetický průměr prostý (jednoduchý)

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

X - statistický znak **n - rozsah souboru**

X_i - hodnota statistického znaku - pozorování

b) vážený aritmetický průměr

- užívá se u početnějších souborů, výpočet vychází z rozdělení četností,
- ***vážený*** se nazývá proto, že jednotlivým hodnotám znaku je přisuzována ***váha*** odpovídající počtu výskytů.

$$\bar{x} = \frac{x_1 w_1 + x_2 w_2 + \ldots + x_m w_m}{w_1 + w_2 + \ldots + w_m} = \frac{\sum_{i=1}^{m} x_i w_i}{\sum_{i=1}^{m} w_i}$$

w_i ... váha (počet výskytů)

n ... rozsah souboru (počet hodnot).

$$n = \sum_{i=1}^{m} w_i$$

b) aritmetický průměr vážený - příklad

Přijímací řízení FTK 1991 – 1993

Výsledky testu běh na 100m

1991 (n = 350)	AP = 13,0
1992 (n = 230)	AP = 12,5
1993 (n = 120)	AP = 12,0

Jaký je průměrný výkon v běhu na 100 m v letech 1991 – 1993?

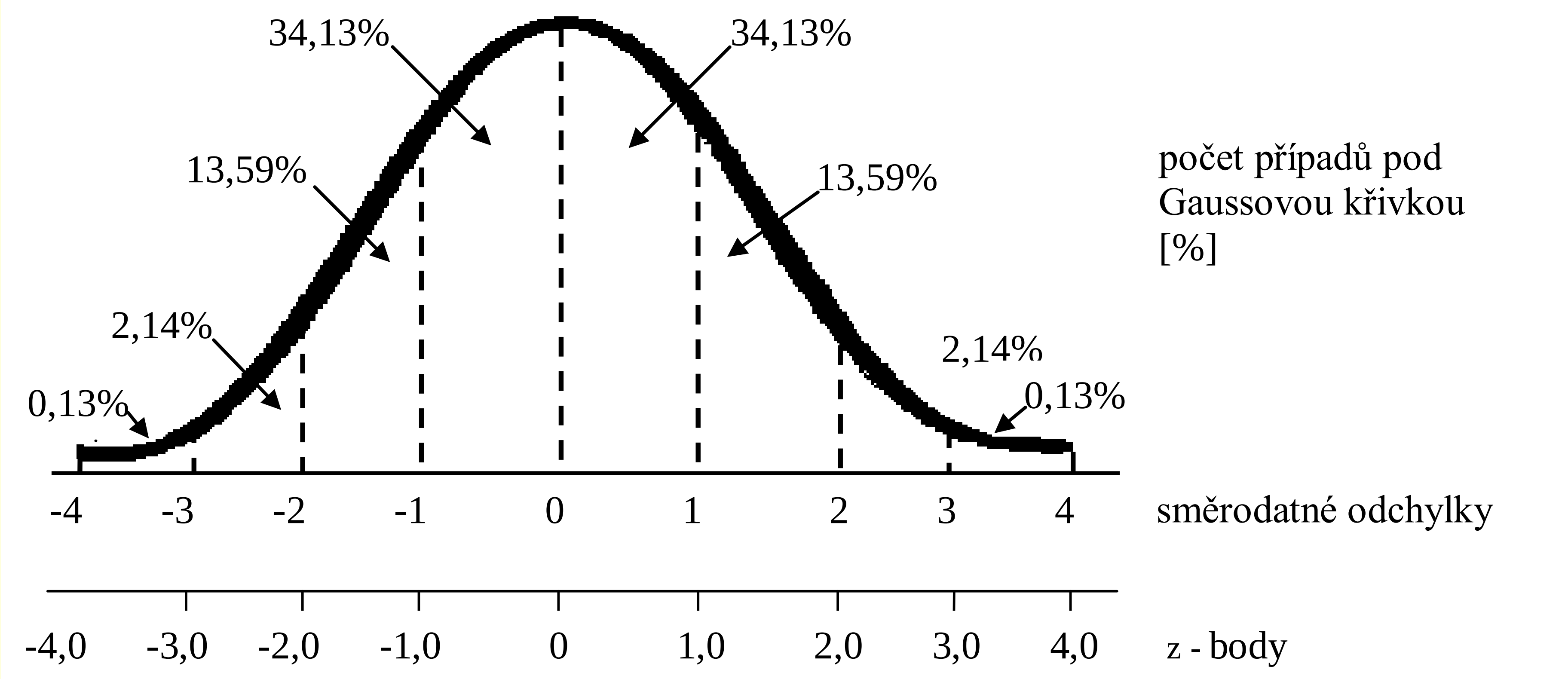
??? 13,0 + 12,5 + 12,0 = 37,5/3 = 12,5 ???

$$\bar{x} = \frac{x_1 w_1 + x_2 w_2 + \ldots + x_m w_m}{w_1 + w_2 + \ldots + w_m} = \frac{\sum_{i=1}^{m} x_i w_i}{\sum_{i=1}^{m} w_i}$$

$$AP = \frac{13{,}0 \times 350 + 12{,}5 \times 230 + 12{,}0 \times 120}{350 + 230 + 120} = 12{,}66$$

Poznámky

- v tzv. **Gaussově normálním rozložení** (rozdělení) **četností** (symetrickém) jsou hodnoty aritmetického průměru, modu a mediánu **stejně velké.**
- čím více se od sebe tyto **hodnoty liší**, tím více je rozdělení četností **asymetrické** (hovoříme o porušené normalitě rozložení četností).



PŘÍKLAD 3

Výpočet: modus, medián, aritmetický průměr.

Hráč	A	B	C	D	E	F	G	H	J	K
x_i	7	6	7	8	9	8	8	8	9	10
y_i	4	8	6	8	7	8	7	4	8	10

Variační řada znaku xi: 6, 7, 7, 8, 8, 8, 8, 9, 9, 10

Mo = ? **Me = ?** **AP = ?** **Vážený AP = ?**

Mo = 8 **Me = 8** **AP = 8** **Vážený AP = 8**

SAMI: Variační řada znaku yi: 4, 8, 6, 8, 7, 8, 7, 4, 8, 10

Mo = ? **Me = ?** **AP = ?** **Vážený AP = ?**

Mo = 8 **Me = 7,5** **AP = 7** **Vážený AP = 7**

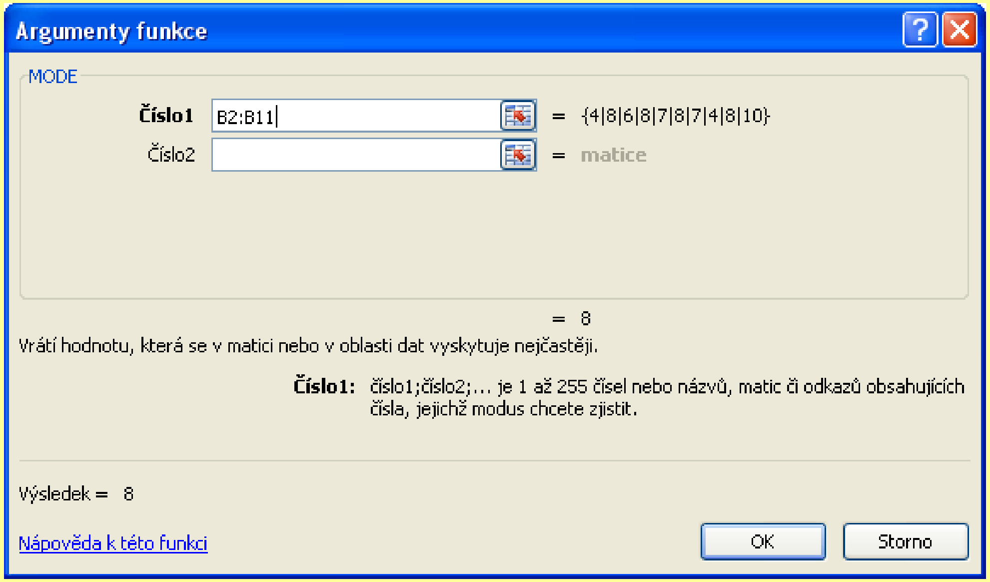
Pomocí Excelu – Statistické funkce
Výpočet: modus, medián, aritmetický průměr.

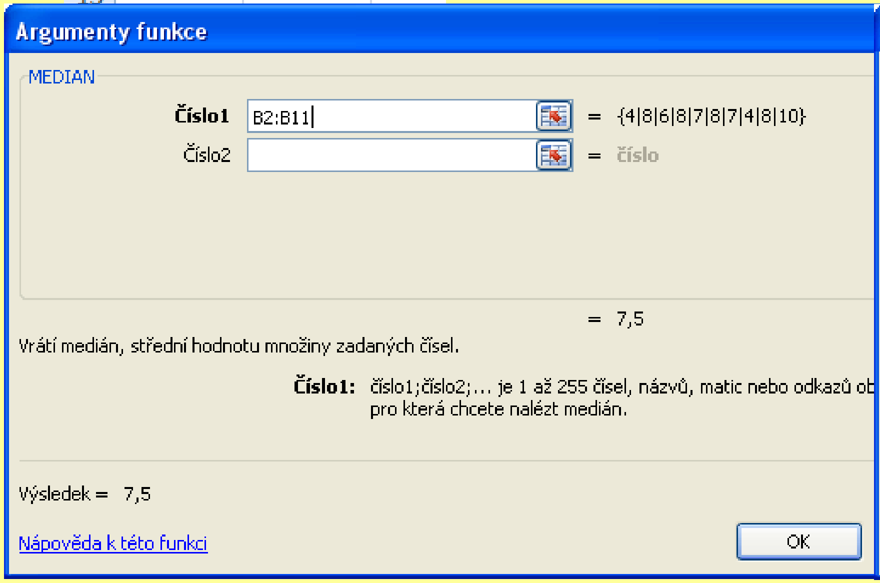
	A	B	C
1	Hráč	yi	
2	A	4	
3	B	8	
4	C	6	
5	D	8	
6	E	7	
7	F	8	
8	G	7	
9	H	4	
10	J	8	
11	K	10	
12			
13			
14			
15			

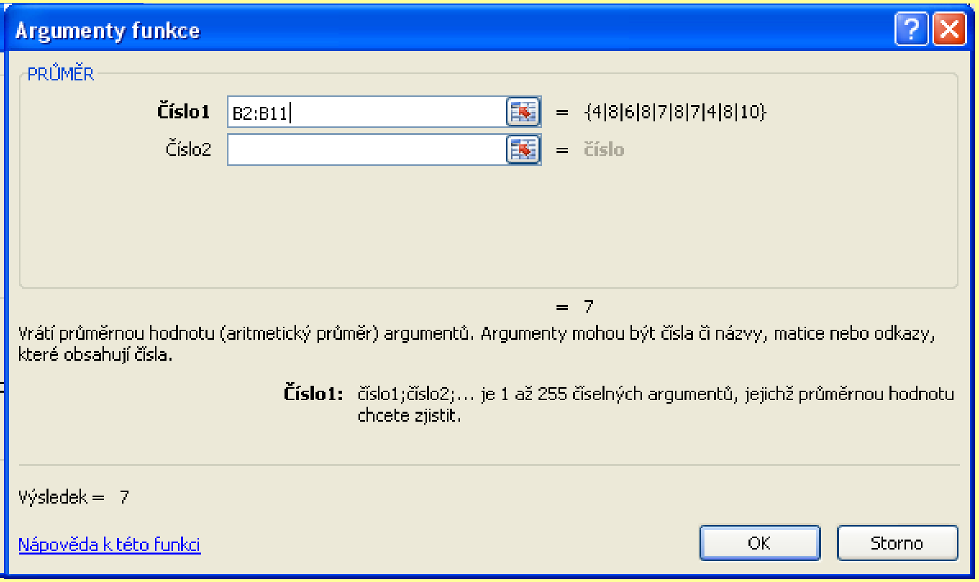
MODE

MEDIAN

PRŮMĚR



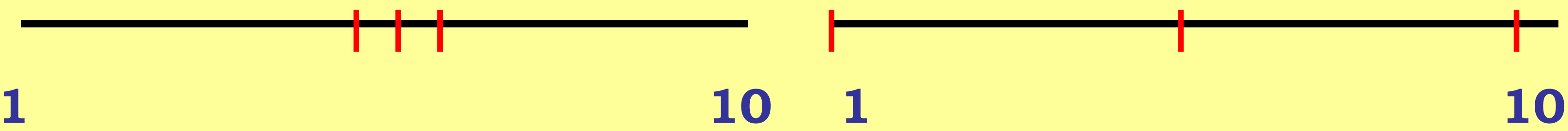


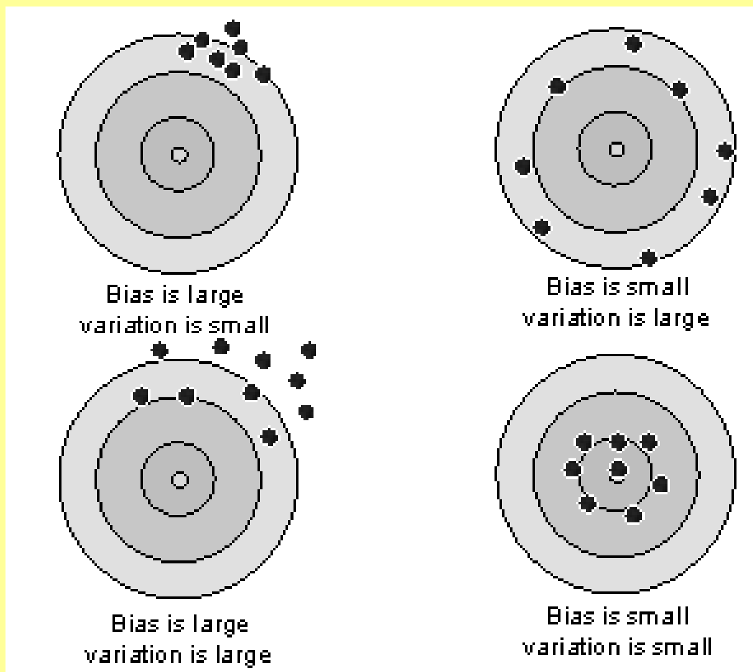


2. 3 MÍRY VARIABILITY

Popis statistického souboru pomocí měr polohy (tedy určení středních hodnot) není dostačující - viz příklad!

Př. 1: 4, 5, 6 ⇒ 15/3=5 (AP 1) Př. 2: 1, 5, 9 ⇒ 15/3=5 (AP 2)





Accuracy versus Quality of an Estimator Using Bias and Variation as Measurable Quantities Respectively

MÍRY VARIABILITY

Míry variability, **tedy přinášejí další důležité informace o vlastnostech souboru, v odborné literatuře jsou též označovány jako** ***míry variace, rozptýlení, měnlivosti.***

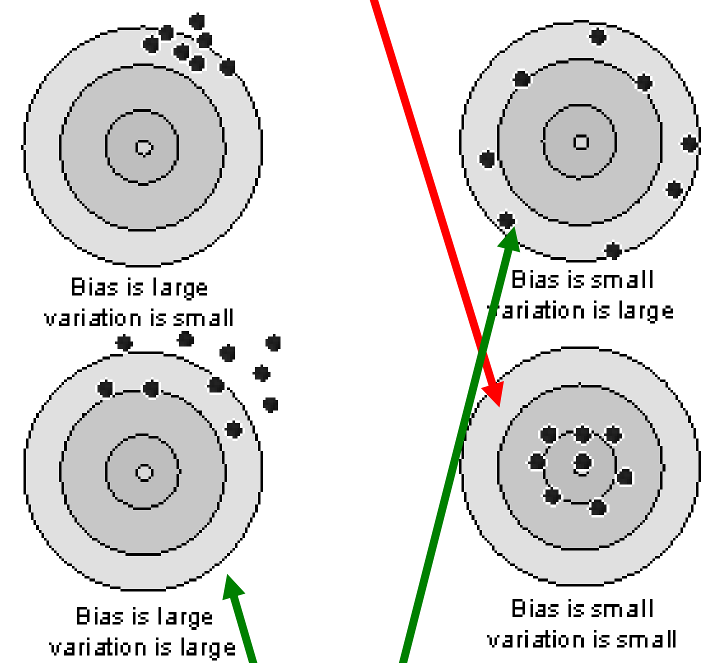
Vlastnosti variability:

- **charakterizují vyrovnanost jednotek souboru,**

- **ukazují, jak jsou hodnoty souboru rozptýleny, tedy jak se jednotlivé hodnoty znaku vzájemně liší.**

- **umožňují posoudit, do jaké míry je sledovaný soubor** ***homogenní*** **(stejnorodý) resp.** ***heterogenní*** **(nestejnorodý, různorodý).**

Soubor homogenní



Accuracy versus Quality of an Estimator Using Bias and Variation as Measurable Quantities Respectively

Soubor heterogenní

2. 3. 1 KVANTILOVÉ MÍRY VARIABILITY

KMV jsou využitelné pro stupnice ordinální a dále pro stupnice metrické v případech, kdy nelze prokázat normalitu rozložení četností dat (*proč ne pro nominální stupnice?*).

NEJČASTĚJI POUŽÍVANÉ KVM

VARIAČNÍ ŘADA = znaky statistického souboru seřazené podle velikosti.

VARIAČNÍ ROZPĚTÍ =diference mezi největší a nejmenší hodnotou znaku statistického souboru tj. $R = x_{max} - x_{min}$

KVANTIL=hodnota kvantitativního statistického znaku, která rozděluje (láme) variační řadu na jisté části.

PŘÍKLAD 4 Výpočet: variační řada, variační rozpětí.

Hráč	A	B	C	D	E	F	G	H	J	K
x_i	7	6	7	8	9	8	8	8	9	10
y_i	4	8	6	8	7	8	7	4	8	10

VARIAČNÍ ŘADA znaků $x_i \Rightarrow$ 6, 7, 7, 8, 8, 8, 8, 9, 9, 10

VARIAČNÍ ROZPĚTÍ $R = x_{max} - x_{min} \Rightarrow R = 10 - 6 = 4$

Sami totéž pro znaky y_i

VARIAČNÍ ŘADA znaků y_i ...4,4,6,7,7,8,8,8,8,10

VARIAČNÍ ROZPĚTÍ $R = x\ max - x\ min \Rightarrow R = 10 - 4 = 6$

DRUHY KVANTILŮ (kvartil, decil, percentil)

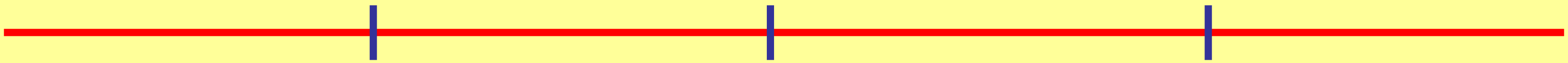
1. KVARTIL (Y) … kvartily **rozdělují variační řadu na čtvrtiny**, na 4 skupiny.

KOLIK MÁME KVARTILŮ?

Dolní kvartil (Q_1, x_{25})

Horní kvartil (Q_3, x_{75})

(Střední kvartil) = **medián**



VÝPOČET KVANTILŮ

$$z_p = \frac{n \cdot p}{100} + 0{,}5$$

z_p - pořadí kvantilu x_p

n - rozsah souboru

Příklad : Urči dolní kvartil x_{25} , přičemž počet hodnot v souboru je n = 40.

$$z_p = \frac{40 \cdot 25}{100} + 0{,}5 = 10{,}5$$

tzn., že 10,5 hodnota (resp. průměr desáté a jedenácté hodnoty) v uspořádaném souboru představuje dolní kvartil x_{25}

$$x_{25} = \frac{x_{(10)} + x_{(11)}}{2}$$

2. DECIL

… decily **rozdělují variační řadu na desetiny**, na 10 skupin o 10% rozsahu souboru.

Označují se x_{10}, x_{20}, …x_{90}

3. PERCENTIL (PROCENTIL)

…percentily rozdělují **variační řadu na setiny**, na 100 skupin o 1% rozsahu. Označují se x_1, x_2, …x_{99}

DALŠÍ KVANTILOVÉ CHARAKTERISTIKY VARIABILITY

KVANTILOVÉ ROZPĚTÍ

- kvartilové rozpětí $x_{75} - x_{25}$
- decilové rozpětí $x_{90} - x_{10}$
- percentilové rozpětí $x_{99} - x_{1}$

KVANTILOVÉ ODCHYLKY

a) kvartilová odchylka

$$Q = \frac{x_{75} - x_{25}}{2}$$

Měří varianci lépe než variační rozpětí, není ovlivněna extrémy.

b) decilová odchylka

$$Q = \frac{x_{90} - x_{10}}{2}$$

Je osminou rozpětí krajních decilů, záleží tedy na rozpětí prostředních 80% prvků souboru.

c) percentilová odchylka

$$Q = \frac{x_{99} - x_{1}}{2}$$

Je devadesátiosminou rozpětí krajních percentilů.

2. 3. 2 MOMENTOVÉ MÍRY VARIABILITY

Předchozí (kvantilové) míry variability udávají jen rozpětí, v němž se znaky pohybují.

Vhodnějšími mírami variability jsou momentové míry variability (pokud je ovšem možné je vypočítat).

Jsou to takové charakteristiky, které udávají ...

(1) variaci ve smyslu vzájemné odlišnosti jednotlivých hodnot znaku (mezi sebou),

(2) variaci ve smyslu odlišnosti jednotlivých hodnot znaku od průměru.

NEJČASTĚJI POUŽÍVANÉ MOMENTOVÉ MÍRY VARIABILITY

1. ROZPTYL $s^2 = \dfrac{\sum (x_i - AP)^2}{n}$

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

(pro rozsáhlé soubory)

AP…aritmetický průměr

x_i …hodnota znaku

Rozptyl (s^2) … je aritmetickým průměrem ze čtverců odchylek jednotlivých hodnot znaku od jejich aritmetického průměru.

Rozptyl „měří“ variaci ve smyslu odlišnosti jednotlivých hodnot znaku *od průměru* i ve smyslu vzájemné odlišnosti *jednotlivých hodnot znaku.*

2. SMĚRODATNÁ (STANDARDNÍ) ODCHYLKA (s)

Symbolický tvar

s = √ s² (var x)

$$s = \sqrt{\frac{\sum (x_i - AP)^2}{n}}$$

$$s = \sqrt{s^2} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

(pro rozsáhlé soubory)

Směrodatná odchylka (s)

… je kvadratickým průměrem odchylek jednotlivých hodnot znaku od aritmetického průměru.

3. VARIAČNÍ KOEFICIENT

$$V = \frac{s}{|AP|}$$

$$V = 100 \times \frac{s}{|AP|}$$

Variační koeficient ...

- **umožňuje provést *srovnání variability dvou či více souborů,* jejichž znaky jsou měřeny v různých jednotkách (cm, kg, sekundy, ...),**

- **udává *poměr směrodatné odchylky k aritmetickému průměru,* přesněji řečeno udává, *kolik % aritmetického průměru tvoří směrodatná odchylka.***

PŘÍKLAD 5

Výpočet: rozptyl, směrodatná odchylka, variační koeficient

Hráč	A	B	C	D	E	F	G	H	J	K
x_i	7	6	7	8	9	8	8	8	9	10
y_i	4	8	6	8	7	8	7	4	8	10

(1) Rozptyl

AP=8

$$s^2 = \frac{(7-8)^2+(6-8)^2+(7-8)^2\ldots\ldots(8-8)^2+(9-8)^2+(10-8)^2}{10}$$

$$= \frac{1+4+1+0+1+0+0+0+1+4}{10} = \frac{12}{10} = \mathbf{1{,}20}$$

(2) Směrodatná odchylka $s = \sqrt{s^2} = 1{,}09 = 1{,}1$

Sami – směrodatná odchylka znaků y_i ...

Variační koeficient V_1

$$V_1 = \frac{s}{AP} = \frac{1{,}09}{8} = 0{,}14 \quad \text{resp.} \quad V_1 = \frac{s}{AP} \cdot 100 = 14\ \%$$

Sami - variační koeficient V_2 tj. znaků y_i ...

$V_2 = 0{,}26$ resp. $V_2 = 26\ \%$ $\Rightarrow$ $V_1 < V_2$

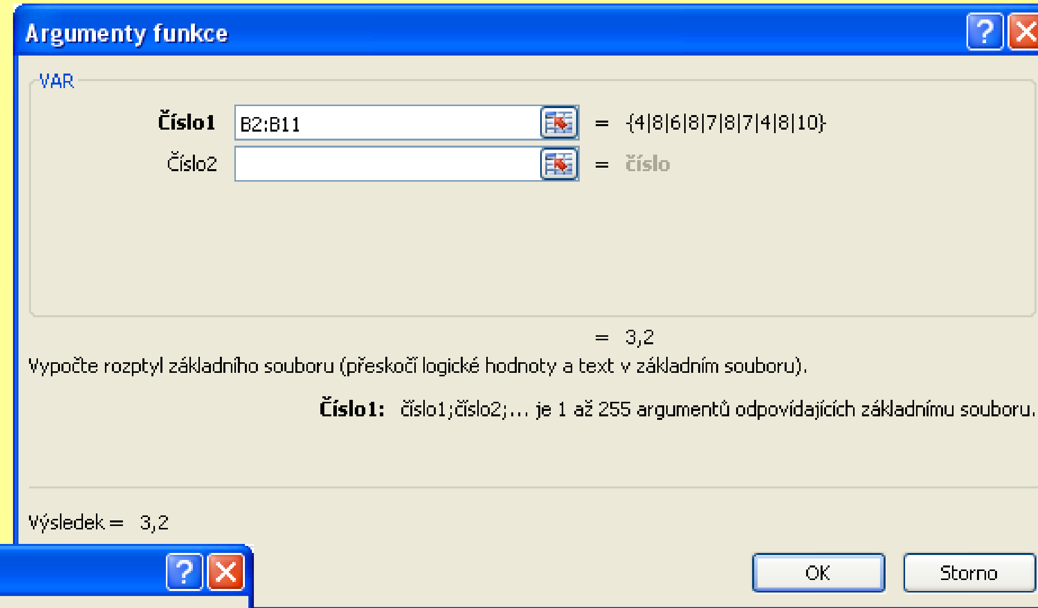
Interpretace ...

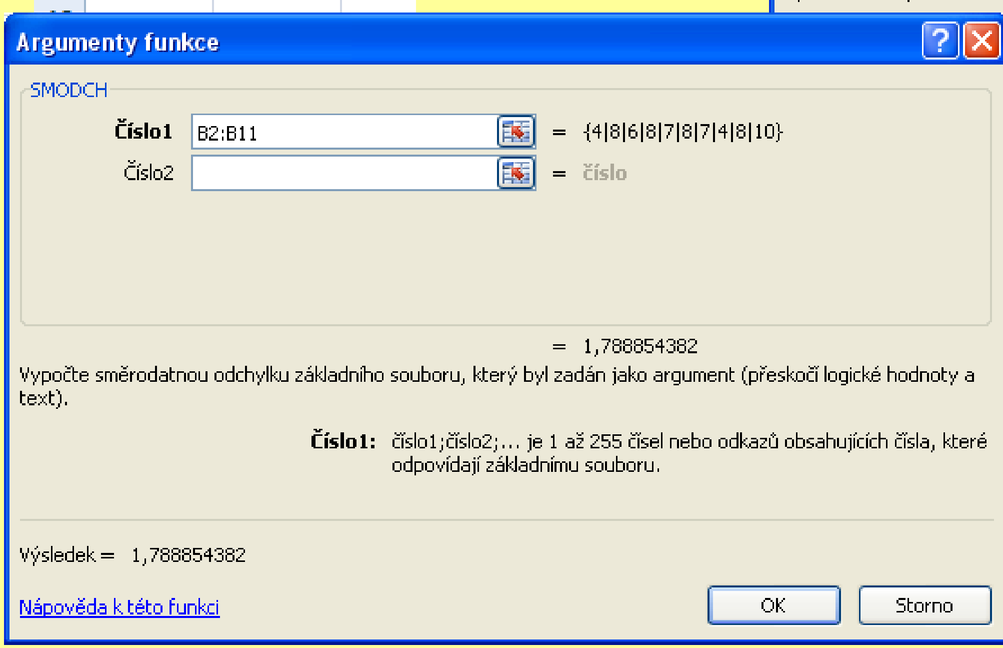
Pomocí Excelu – Statistické funkce
Výpočet: rozptyl, směrodatná odchylka, variační koeficient

	A	B	C
1	Hráč	yi	
2	A	4	
3	B	8	
4	C	6	
5	D	8	
6	E	7	
7	F	8	
8	G	7	
9	H	4	
10	J	8	
11	K	10	

VAR

SMODCH



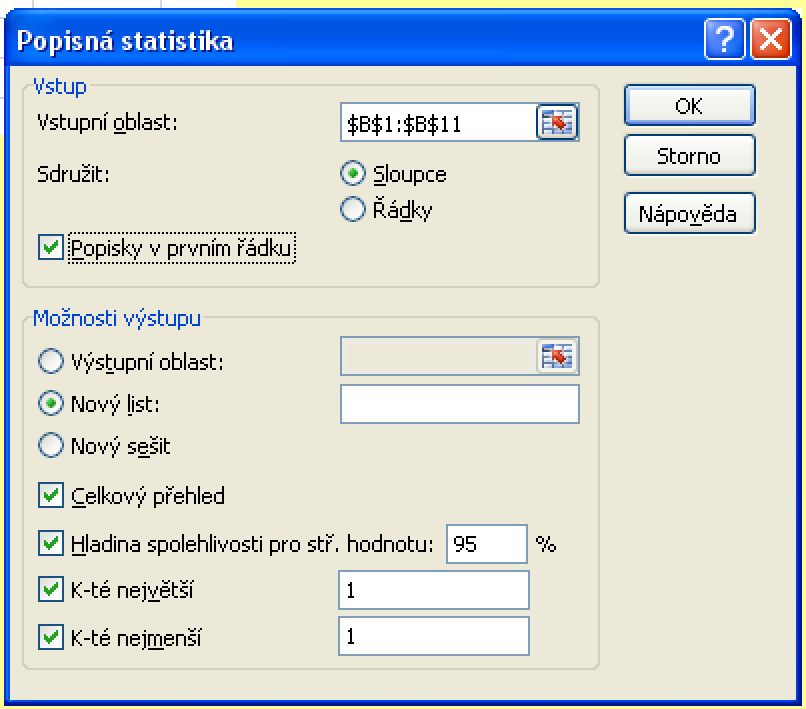


VAR.VÝBĚR
vypočte rozptyl výběru

SMODCH.VÝBĚR
vypočte směrodatnou odchylku výběru

Pomocí Excelu – Analýza dat – Popisná statistika

	A	B	C
1	Hráč	yi	
2	A	4	
3	B	8	
4	C	6	
5	D	8	
6	E	7	
7	F	8	
8	G	7	
9	H	4	
10	J	8	
11	K	10	
12			
13			
14			
15			



	yi
Stř. hodnota	7
Chyba stř. hodnoty	0,596285
Medián	7,5
Modus	8
Směr. odchylka	1,885618
Rozptyl výběru	3,555556
Špičatost	-0,05776
Šikmost	-0,49718
Rozdíl max-min	6
Minimum	4
Maximum	10
Součet	70
Počet	10
Největší (1)	10
Nejmenší (1)	4
Hladina spolehlivosti (95,0%)	1,34889

ANALÝZA JEDNOROZMĚRNÉHO SOUBORU

METODY DESKRIPTIVNÍ STATISTIKY

1. URČENÍ TYPU ŠKÁLY (nominální, ordinální, metrické)

a) nominální + ordinální ⇒ *neparametrické stat. metody*

b) metrické ⇒ *parametrické statistické metody*

2. ROZLOŽENÍ ČETNOSTÍ ZNAKŮ (NORMÁLNÍ ČI JINÉ)

a) normální ⇒ *parametrické statistické metody*

b) jiné ⇒ *neparametrické statistické metody*

3. VÝPOČET ZÁKLADNÍCH STATISTICKÝCH CHARAKTERISTIK

a) míry centrální tendence

b) míry variability

c) míry závislosti

2.5 MÍRY ZÁVISLOSTI

2.5.1 ZÁVISLOST PEVNÁ, VOLNÁ, STATISTICKÁ A KORELAČNÍ

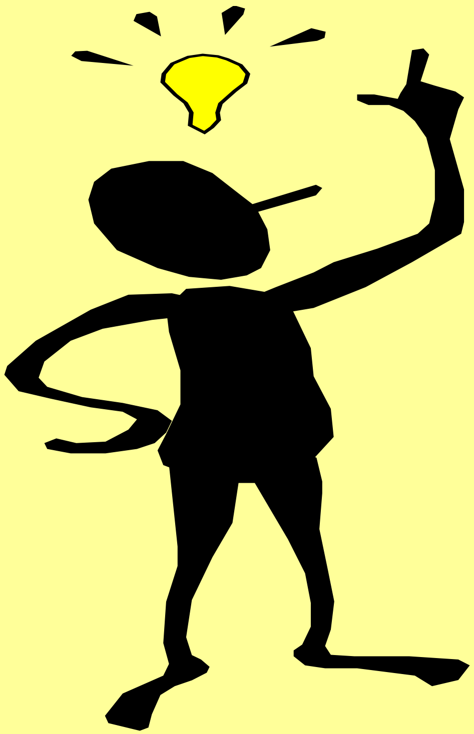
Jednorozměrné soubory **- charakterizovány jednotlivými statistickými znaky (výkon ve skoku do dálky, v běhu na 100m, tělesná výška, tělesná hmotnost, ...**

Možné *souvislosti mezi znaky*:

- **úspěšnost střelby 1. a 2. pokus,**
- **tělesná výška x hmotnost,...**

2.5.1 ZÁVISLOST PEVNÁ, VOLNÁ, STATISTICKÁ A KORELAČNÍ

Hledáním, zkoumáním a hodnocením souvislostí (závislostí) mezi dvěma (či více) statistickými znaky se zabývají míry závislosti.



Závislosti znaků, věcí a jevů mohou být velmi rozmanité:

- **některé jsou nepodstatné (náhodné)**
- **jiné jsou výrazem určité vnitřní nutnosti – hovoříme o tzv. *příčinné (kauzální) závislosti.***

Příčinná (kauzální) závislost **je taková závislost, kdy daný jev či několik jevů (příčina)** ***nutně vyvolává*** **za určitých podmínek jiný jev (následek, účinek).**

Nejjednodušší formy ***kauzálních závislostí*** **u přírodních jevů např. ….**

… při ***zahřívání tělesa*** **(elementární příčina) za konstantních podmínek dochází ke** ***zvětšování jeho objemu*** **(elementární účinek) tj. princip teploměru.**

Z hlediska metody zkoumání rozlišujeme ***závislosti pevné a volné.***

PEVNÁ ZÁVISLOST

Pevná závislost **= případ, kdy výskytu jednoho jevu NUTNĚ ODPOVÍDÁ výskyt druhého jevu.**

Tedy *každé hodnotě jedné proměnné odpovídá jen jedna hodnota jiné proměnné (funkční závislost).*

Např. zahříváme-li těleso 5 min, vzroste teplota o 10 ° C, zahříváme-li těleso 10 min, vzroste teplota o 20 ° C, atd.

PEVNÁ ZÁVISLOST

Pevná závislost

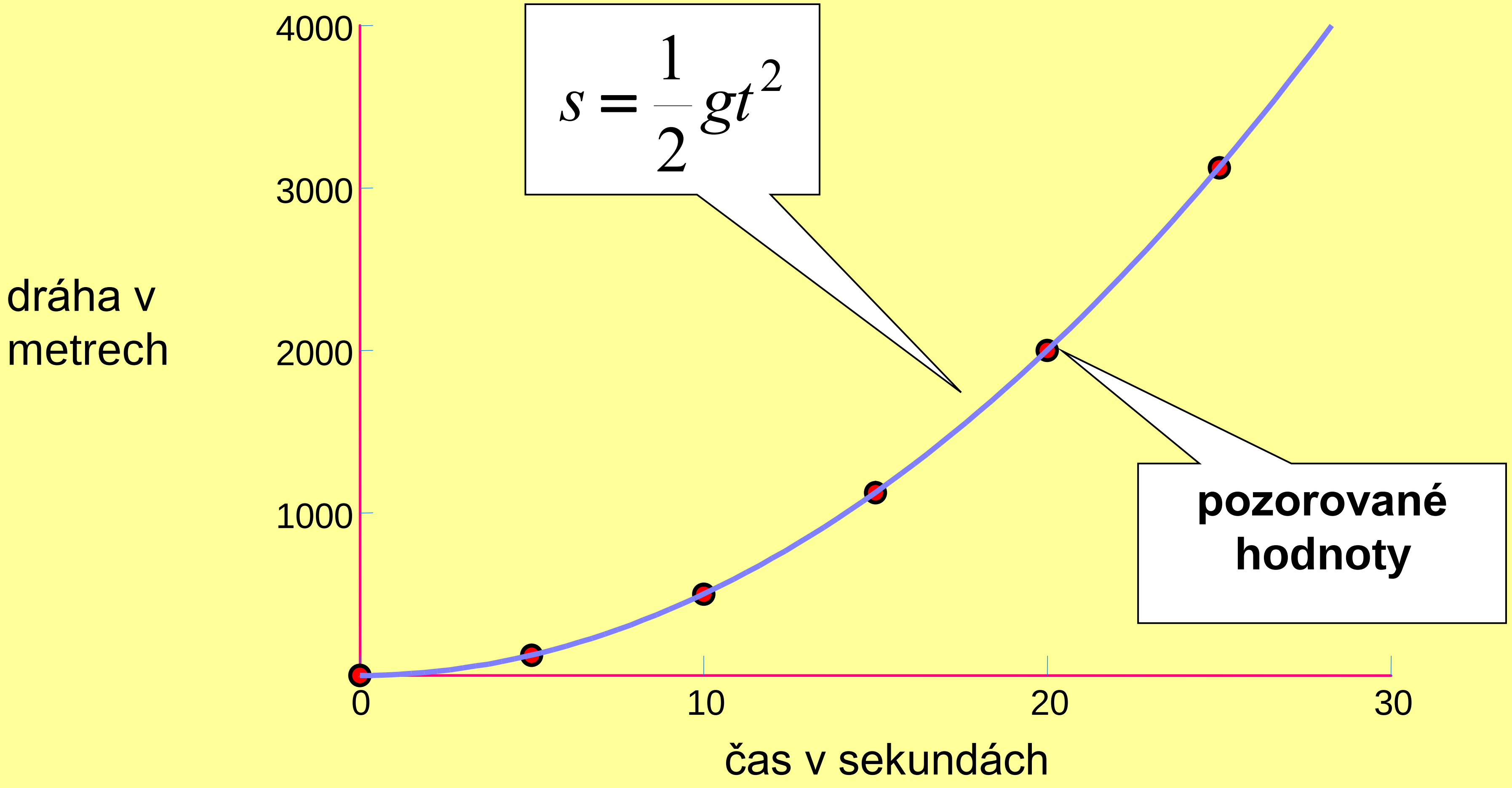
... se opakuje ve všech jednotlivých případech (při dodržení standardních podmínek).

...může být tedy charakterizována jediným pozorováním (větší počet pozorování slouží k ověření výsledků a vyloučení chyb).

...setkáváme se s ní při formulování zákonitostí vztahů mezi proměnnými (např. Newtonův či Ohmův zákon).

PEVNÁ ZÁVISLOST

volný pád



VOLNÁ ZÁVISLOST

Volná závislost **= případ, kdy výskyt jednoho jevu OVLIVŇUJE výskyt druhého jevu (NE nutně odpovídá).**

Tedy *každé hodnotě jedné proměnné odpovídají různé hodnoty jiné proměnné (statistická závislost).*

Volnou závislost

...lze zkoumat pouze na základě mnoha pozorování, jediné pozorování může přinést naprosto nahodilý výsledek,

VOLNÁ ZÁVISLOST

...při jejím zkoumání hraje důležitou roli volba *vhodných statistických znaků* (tedy takových, které postihují jevy co nejpřesněji),

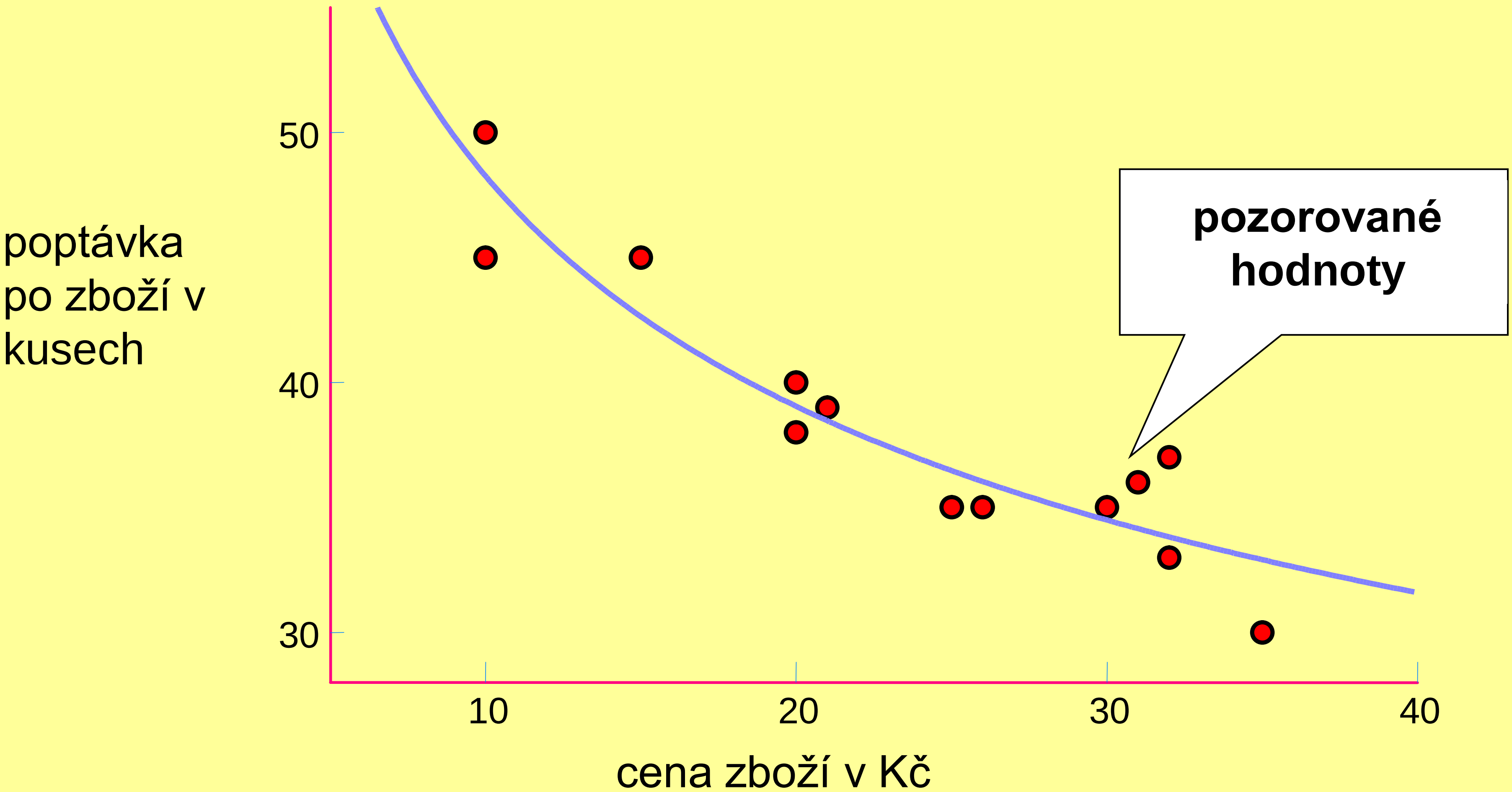
...a *dostatečný rozsah souboru* (u malých souborů se může projevit vliv náhodných a vedlejších činitelů).

Při zkoumání společenských jevů se většinou nesetkáváme s pevnou závislostí; nejčastěji určitá příčina vede k různým účinkům, jedná se tedy o *volné závislosti*.

Např. skok daleký ... rychlost x délka skoku ...

VOLNÁ ZÁVISLOST

tržní poptávka



2.5.2 REGRESNÍ A KORELAČNÍ ANALÝZA

Metody regresní a korelační analýzy **slouží k poznání a matematickému popisu statistických závislostí; jsou souhrnně označované jako** ***korelační počet.***

Hlavní úkoly korelačního počtu

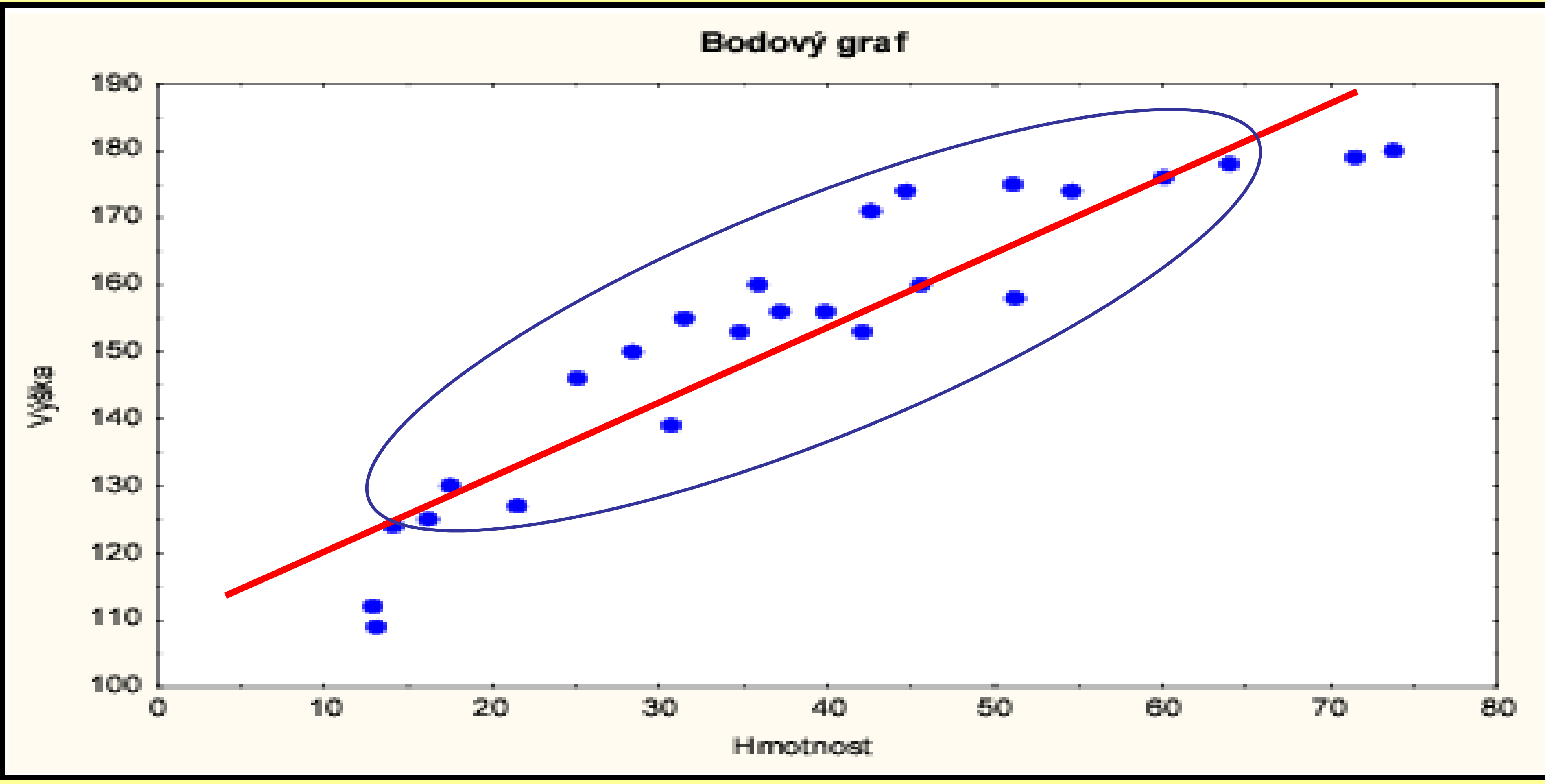
1. postižení povahy **korelační závislosti (regresní analýza),**

2. měření těsnosti **korelační závislosti (korelační analýza).**

Hlavní úkoly korelačního počtu

1. *postižení povahy* korelační závislosti (regresní analýza),

2. *měření těsnosti* korelační závislosti (korelační analýza).

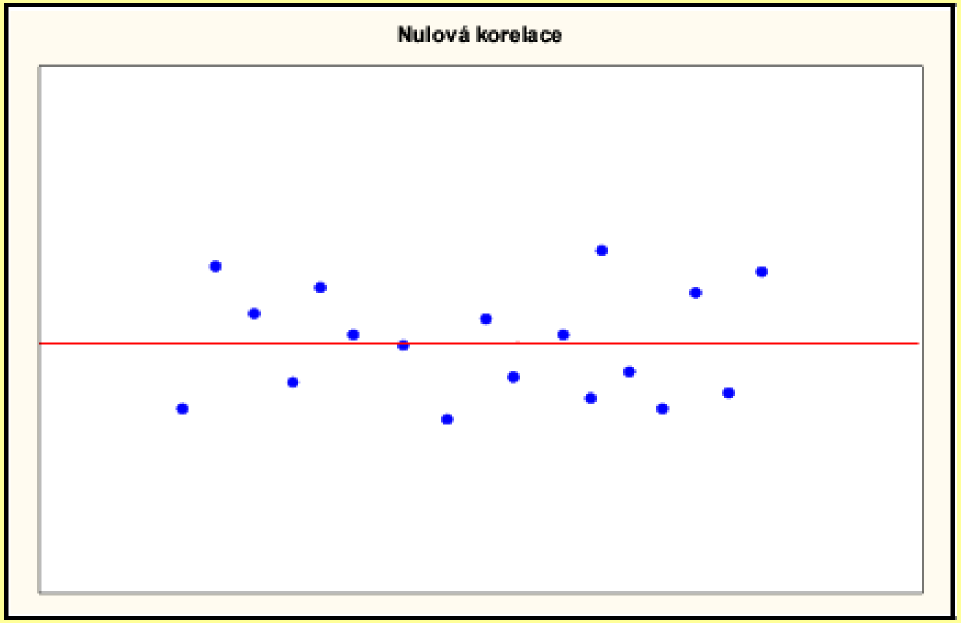


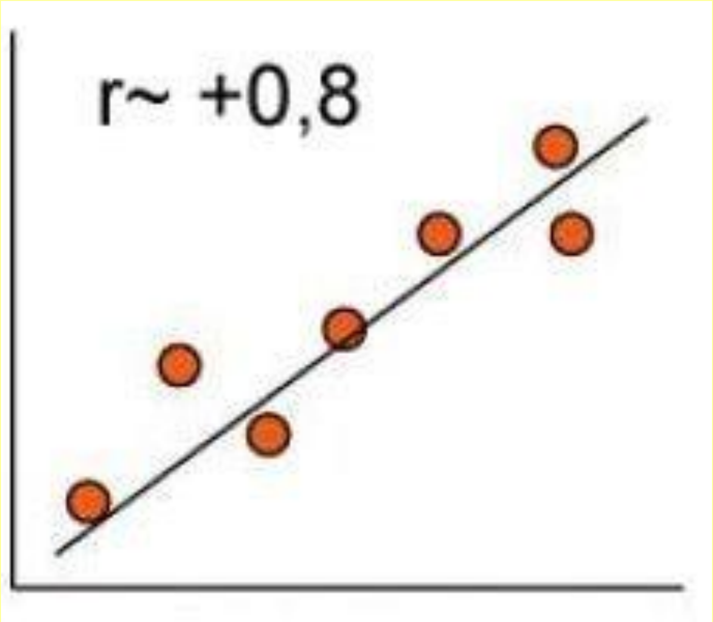
HLAVNÍ ÚKOLY KORELAČNÍHO POČTU

1. postižení povahy **korelační závislosti umožňuje odhady neznámých hodnot *závisle proměnné y* při známých hodnotách *nezávisle proměnné x* - hovoříme o *regresi.***

Povaha korelační závislosti je nejčastěji vyjadřována matematickou funkcí - hovoříme o *regresní funkci.*

Tento úkol korelačního počtu nazýváme *regresní analýza.*



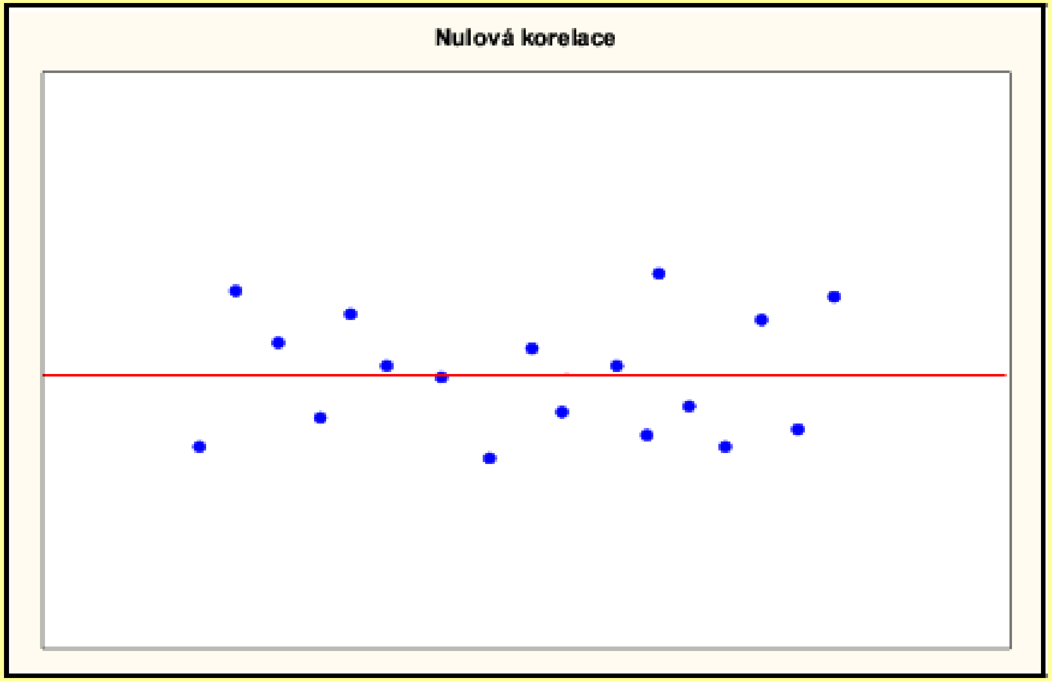


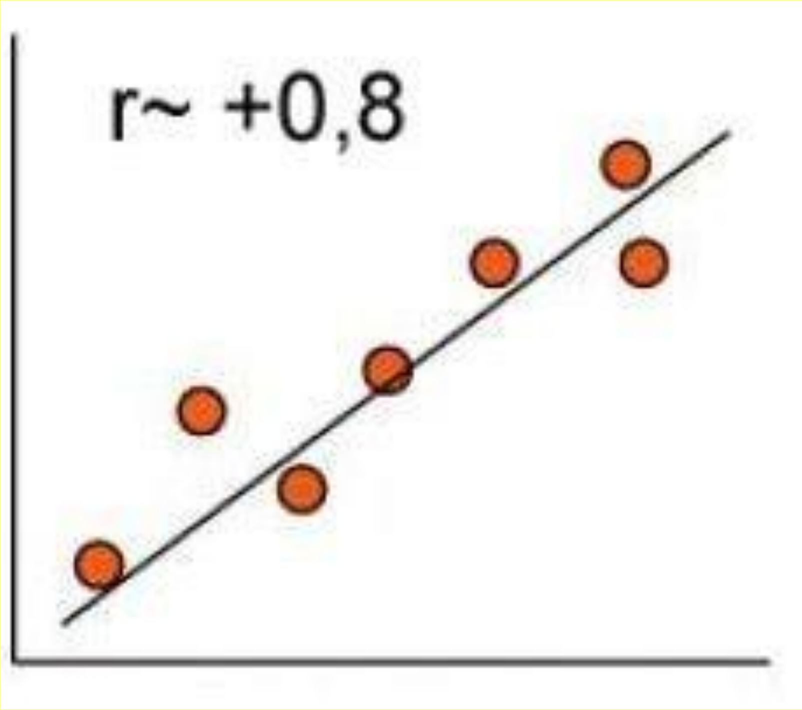
HLAVNÍ ÚKOLY KORELAČNÍHO POČTU

2. měření těsnosti **korelační závislosti umožňuje posuzovat *přesnost regresních odhadů* a *míru korelační závislosti* - hovoříme o *vlastní korelaci.***

Korelace je vyjadřována tzv. *korelačním koeficientem.*

Tento úkol korelačního počtu nazýváme *korelační analýza.*





2.5.3 REGRESNÍ ANALÝZA (LINEÁRNÍ)

1. ÚKOL „POSTIŽENÍ POVAHY KORELAČNÍ ZÁVISLOSTI“

Regresní analýza **umožňuje postihnout povahu závislosti pomocí *matematické funkce,* která by co nejlépe vyjadřovala charakter zkoumané závislosti.**

Hledaná matematická funkce se nazývá regresní funkce a je vyjádřena regresní rovnicí.

Regresní funkce **může nabývat mnoha typů:**

- ***lineární regrese***
- ***kubická regrese***
- ***hyperbolická regrese***
- ***kvadratická regrese***
- ***polynomická regrese***
- ***logaritmická regrese***

…a mnohé jiné…

Nejjednodušší z nich je LINEÁRNÍ REGRESNÍ FUNKCE, která má ve své empirické podobě tvar

$$\boxed{Y = a + b \,.\, x}$$

Pro vyjádření regresní funkce konkrétní závislost (např. tělesné výšky a hmotnosti) je třeba určit tzv. regresní koeficienty a, b, přičemž vycházíme z empirických údajů (měřených znaků) sledované závislosti.

Pro výpočet regresních koeficientů a, b je výhodné použít následující vzorce

$$b = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - \left(\sum x_i\right)^2} \qquad a = \frac{\sum y_i - b \sum x_i}{n}$$

2.5.4 KORELAČNÍ ANALÝZA (LINEÁRNÍ)

2. ÚKOL „MĚŘENÍ TĚSNOSTI KORELAČNÍ ZÁVISLOSTI“

Pojem *korelace* pochází z latiny (*co – relation* = souvztažnost), korelaci nejčastěji označujeme symbolem „ r “.

Korelace je nejobecněji definována jako…

…volná kvantitativní závislost dvou či více jevů.

Korelace vyjadřuje míru (těsnost, stupeň) závislosti a je charakterizována číslem, tzv. korelačním koeficientem (r), který..

…„měří“ těsnost závislosti popsané lineární regresní funkcí.

VZORCE PRO VÝPOČET KORELAČNÍHO KOEFICIENTU

Symbolická podoba vzorce pro výpočet korelačního koeficientu

kovariance

$$r = \frac{s_{x,y}}{s_x \cdot s_y} = \frac{\mathrm{cov}\,(X, Y)}{\sqrt{\mathrm{var}\,X \cdot \mathrm{var}\,Y}}$$

součin obou směrodatných odchylek

Korelace je tedy matematicky *podíl kovariance a součinu obou směrodatných odchylek* (viz dále).

Metrická data ⇒

Pearsonův koeficient součinové korelace (vzorec)

$$r = \frac{\sum (x_i - \bar{x}).(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Pearsonův koeficient součinové korelace - výpočtový tvar

$$r = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{[n \sum x_i^2 - (\sum x_i)^2][n \sum y_i^2 - (\sum y_i)^2]}}$$

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

Pro metrická data užíváme k výpočtu míry těsnosti korelační závislosti tzv. Pearsonův koeficient součinové korelace.

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{\left[n\sum x_i^2 - \left(\sum x_i\right)^2\right]\left[n\sum y_i^2 - \left(\sum y_i\right)^2\right]}}$$

Pro ordinální data užíváme k výpočtu míry těsnosti korelační závislosti tzv. Spearmanův koeficient pořadové korelace.

$$r_{xy} = 1 - \frac{6 . \sum (i_x - i_y)^2}{n(n^2 - 1)} = 1 - \frac{6}{n(n^2 - 1)} \sum (R_i - Q_i)^2$$

VLASTNOSTI KORELACE

1. VELIKOST KORELACE

Korelační koeficient r nabývá hodnot z intervalu <-1 ; 1>

Význam hodnot 0, 1, -1

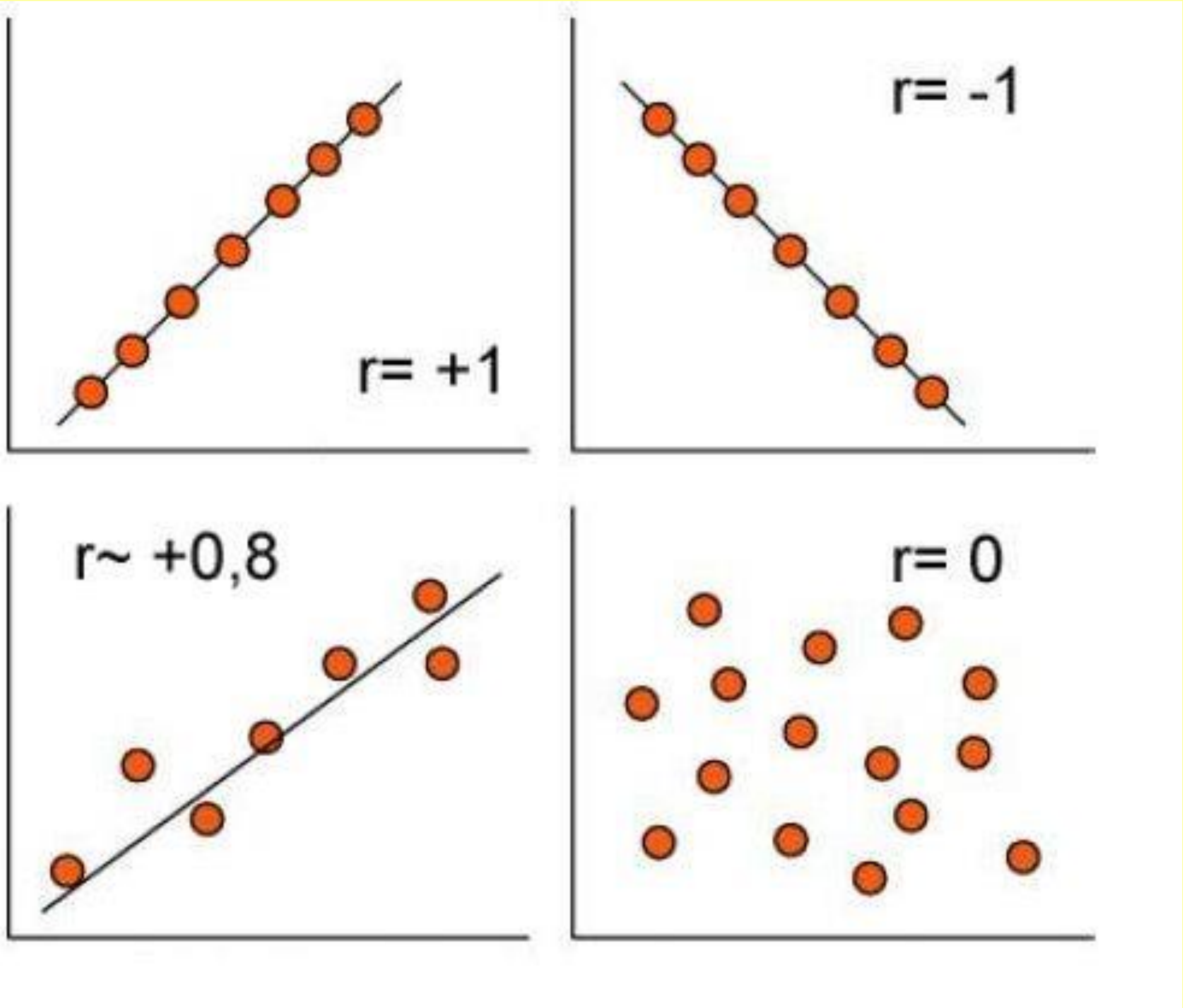
r = 0 ⇐ lineární nezávislost proměnných

r = 1 ⇐ úplná (*funkční)* pozitivní lineární závislost

r = -1 ⇐ úplná (*funkční)* negativní lineární závislost

Čím více se r blíží *hodnotě 1*, tím považujeme *závislost za silnější,*

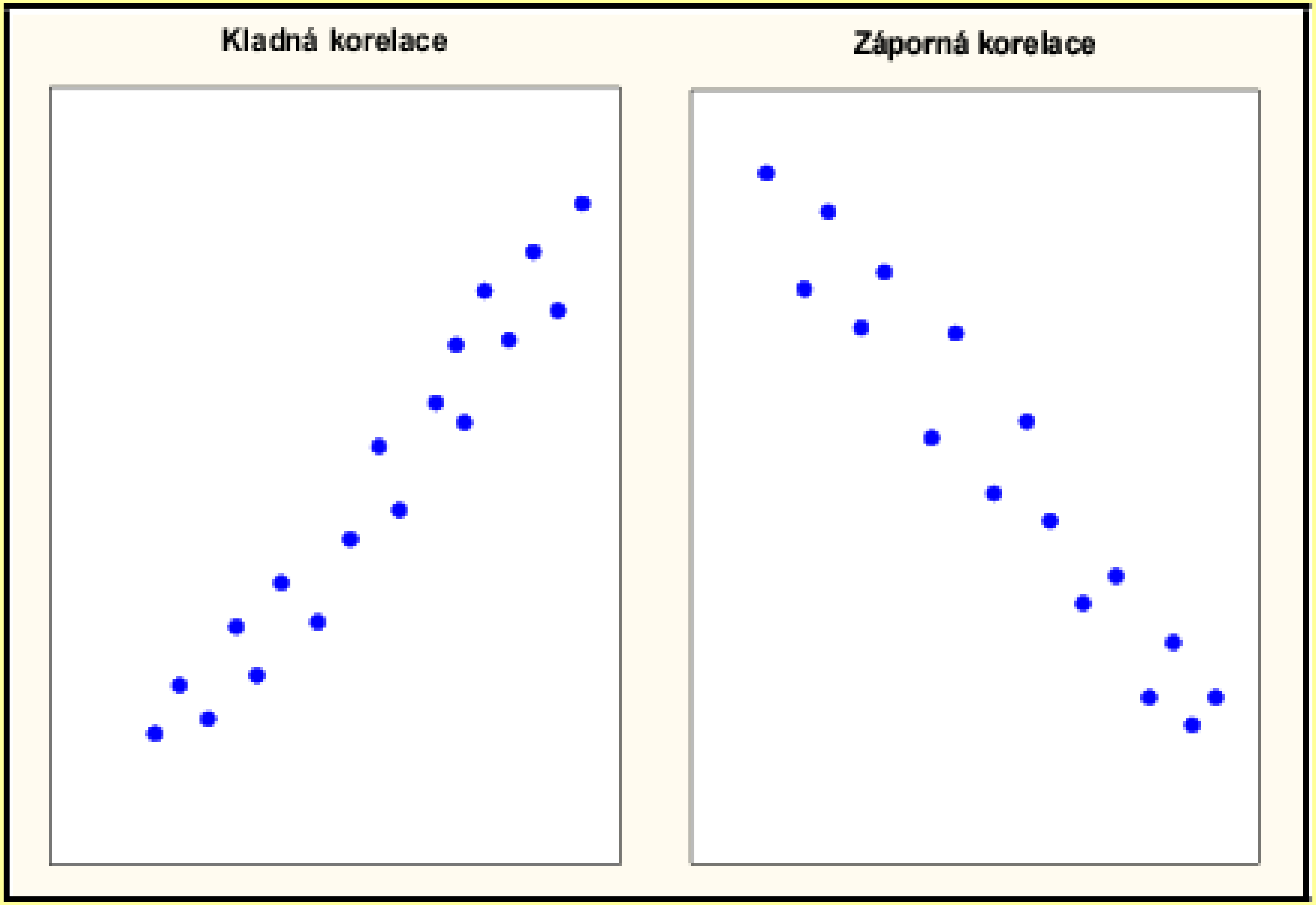
čím více se r blíží *hodnotě 0*, tím považujeme *závislost za slabší.*


r= +1
r= -1
r~ +0,8
r= 0

2. SMĚR KORELACE

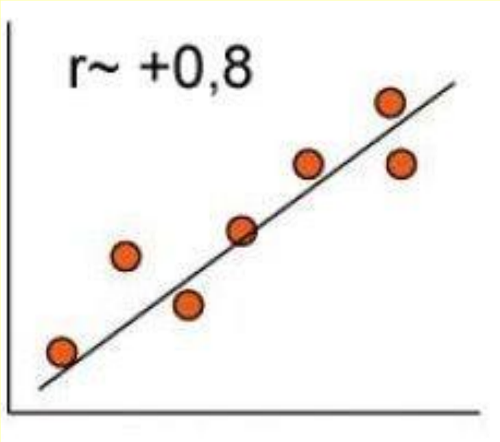
a) **kladná (pozitivní) <0;1>**

b) **záporná (negativní) <-1;0>**

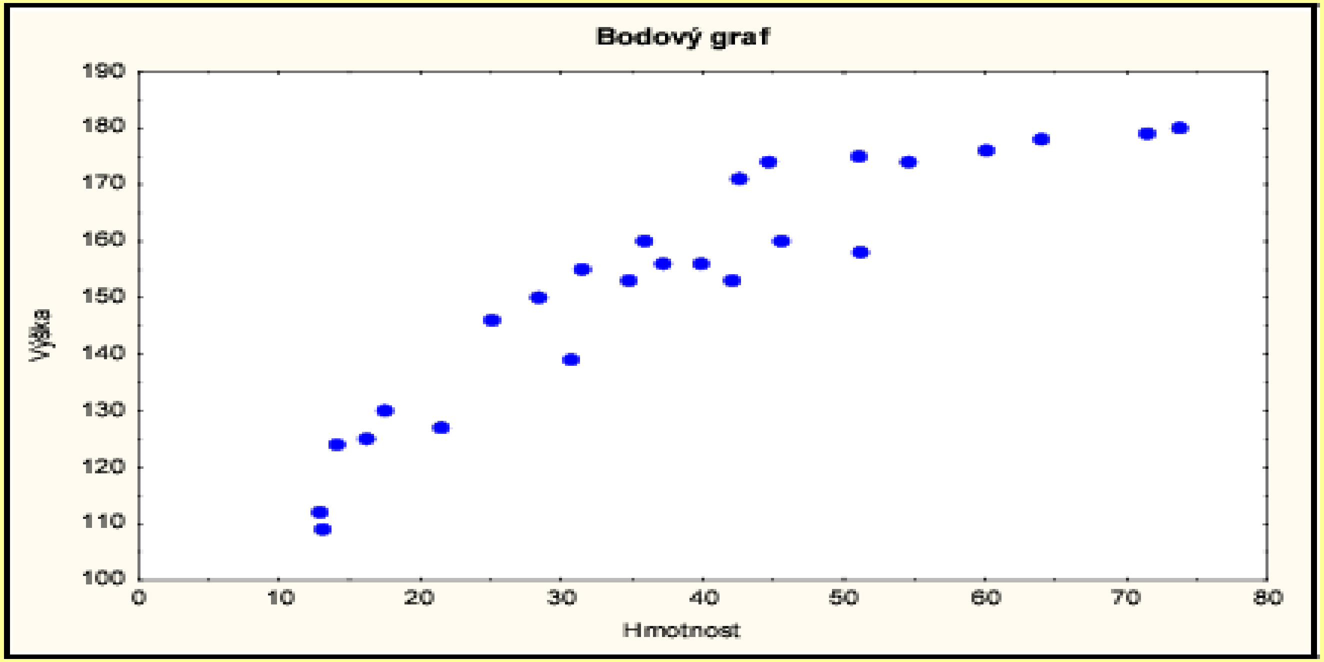


3. TVAR KORELACE

a) lineární (lze dosti dobře proložit přímku)



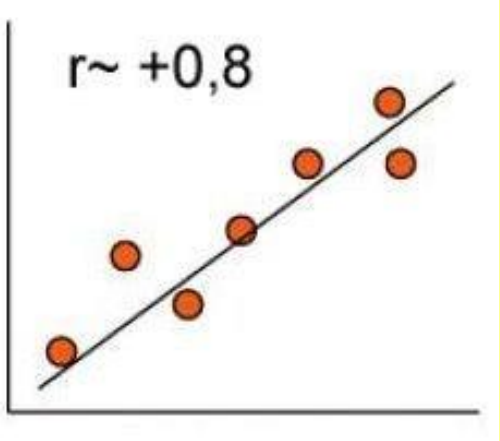
b) nelineární (nelze proložit přímku)



POZNÁMKY KE KORELACÍM

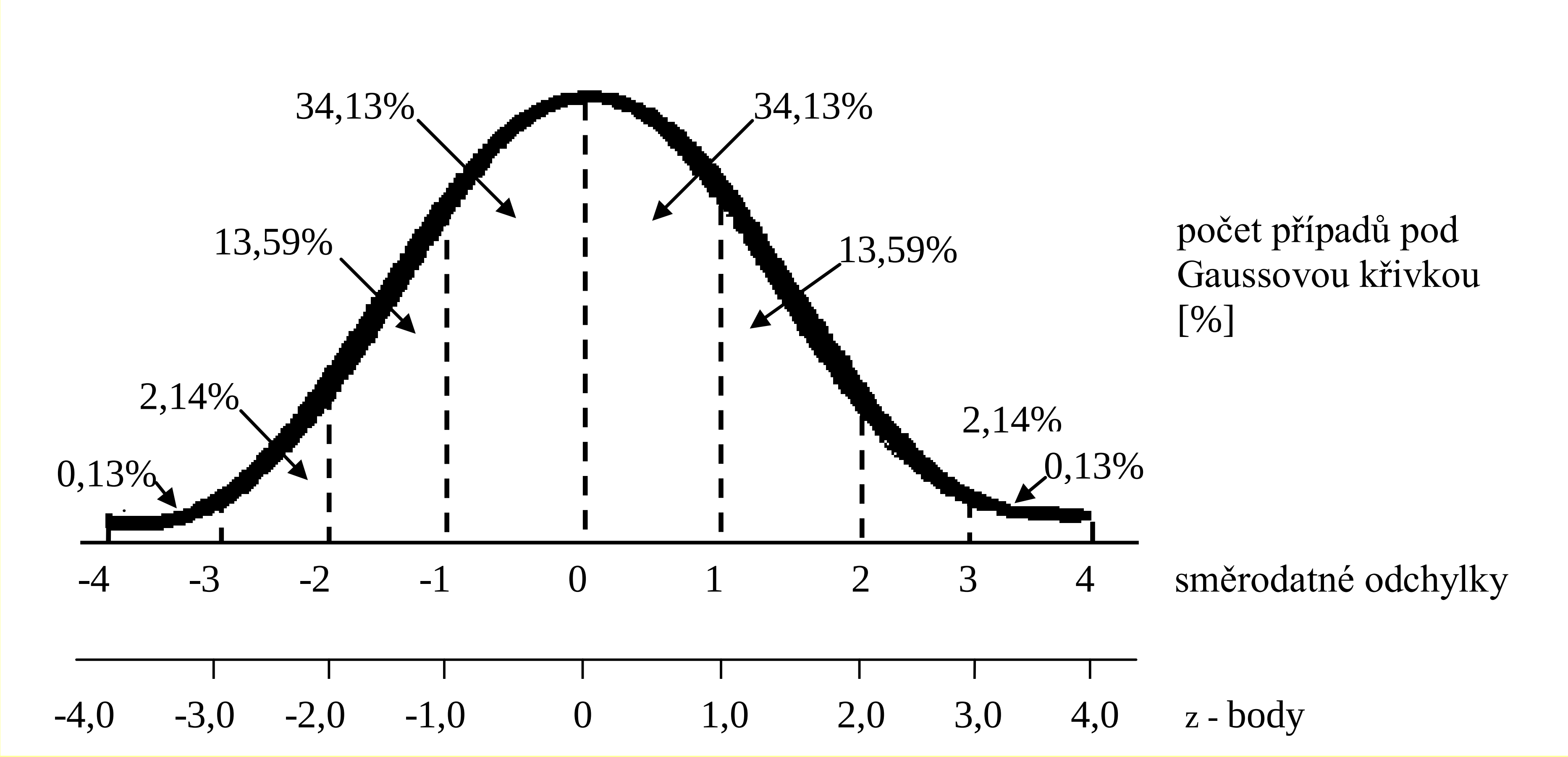
1. Matematicko-statistické předpoklady výpočtu korelačního koeficientu:

***a) linearita** (korelačním polem lze dosti dobře proložit přímku),*



1. Matematicko-statistické předpoklady výpočtu korelačního koeficientu:

b) normalita (dvojrozměrné normální rozložení četností),



1. Matematicko-statistické předpoklady výpočtu korelačního koeficientu:

c) dostatečný rozsah souboru (n=200, n=100, n=30)

2. Věcný a formální smysl znaménka korelačního koeficientu

Např. vypočítaná korelační závislost výsledků studentů FTK (n=185) v běhu na 100m …

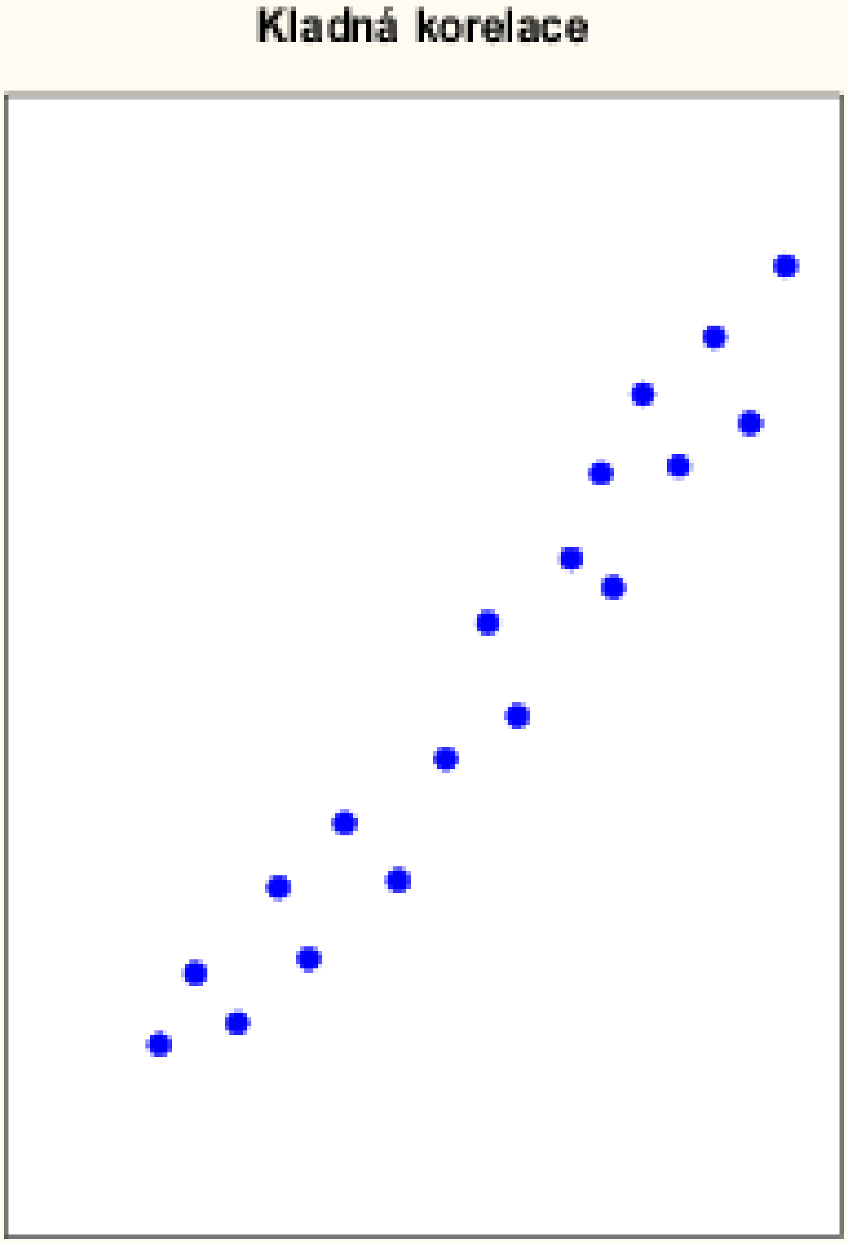
… a ve skoku dalekém je

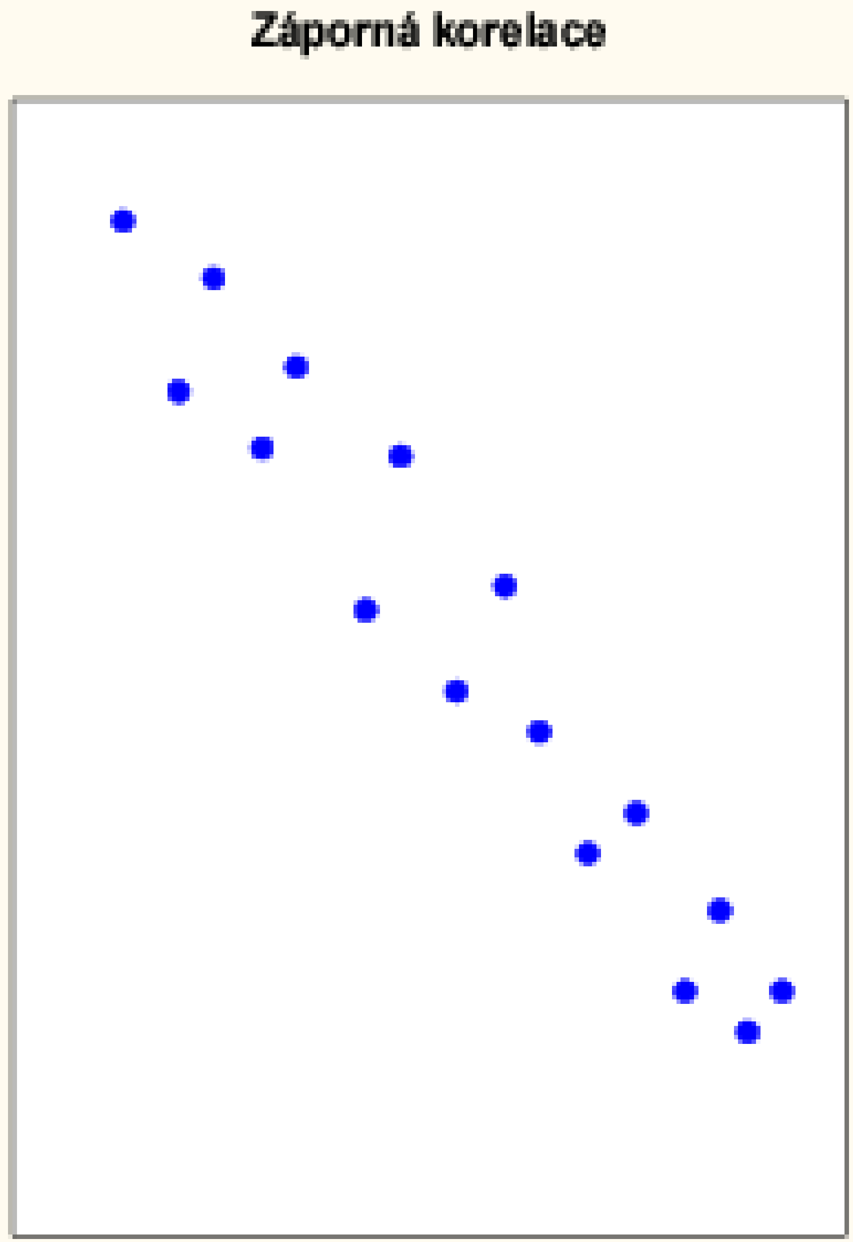
$$r = -0{,}8$$

<-1 ; 1>

Co to znamená z hlediska interpretace?

a) kladná (pozitivní) <0;1> **b) záporná (negativní) <-1;0>**

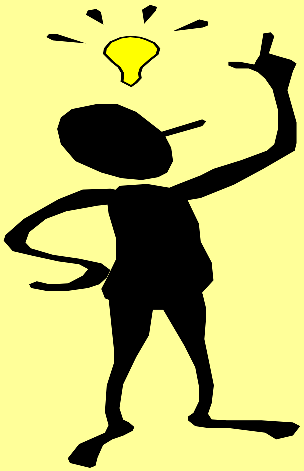




To by ovšem znamenalo, že kdo je <u>rychlejší</u> v běhu na 100m, ten dosahuje <u>horších</u> výsledků ve skoku dalekém.

To je ovšem…

…odborně i věcně NESMYSL!



PROČ ???

PROTOŽE…

…jakou „hodnotu“ má výsledek v běhu na 100 m

10,7 s versus 12,3 s?

…jakou „hodnotu“ má výsledek ve skoku dalekém

570 cm versus 430 cm?

3. Koeficient determinace r^2

… určuje *jaká část rozptylu výkonu* ***v jednom testu*** *je dána proměnlivostí výkonů* ***v druhém testu.***

Např. výše uvedená korelační závislost výsledků studentů FTK (n=185) ***v běhu na 100m a ve skoku dalekém*** *r = 0,8* **znamená, že…**

Koeficient determinace $r^2 = 0,64$ *(64 %).*

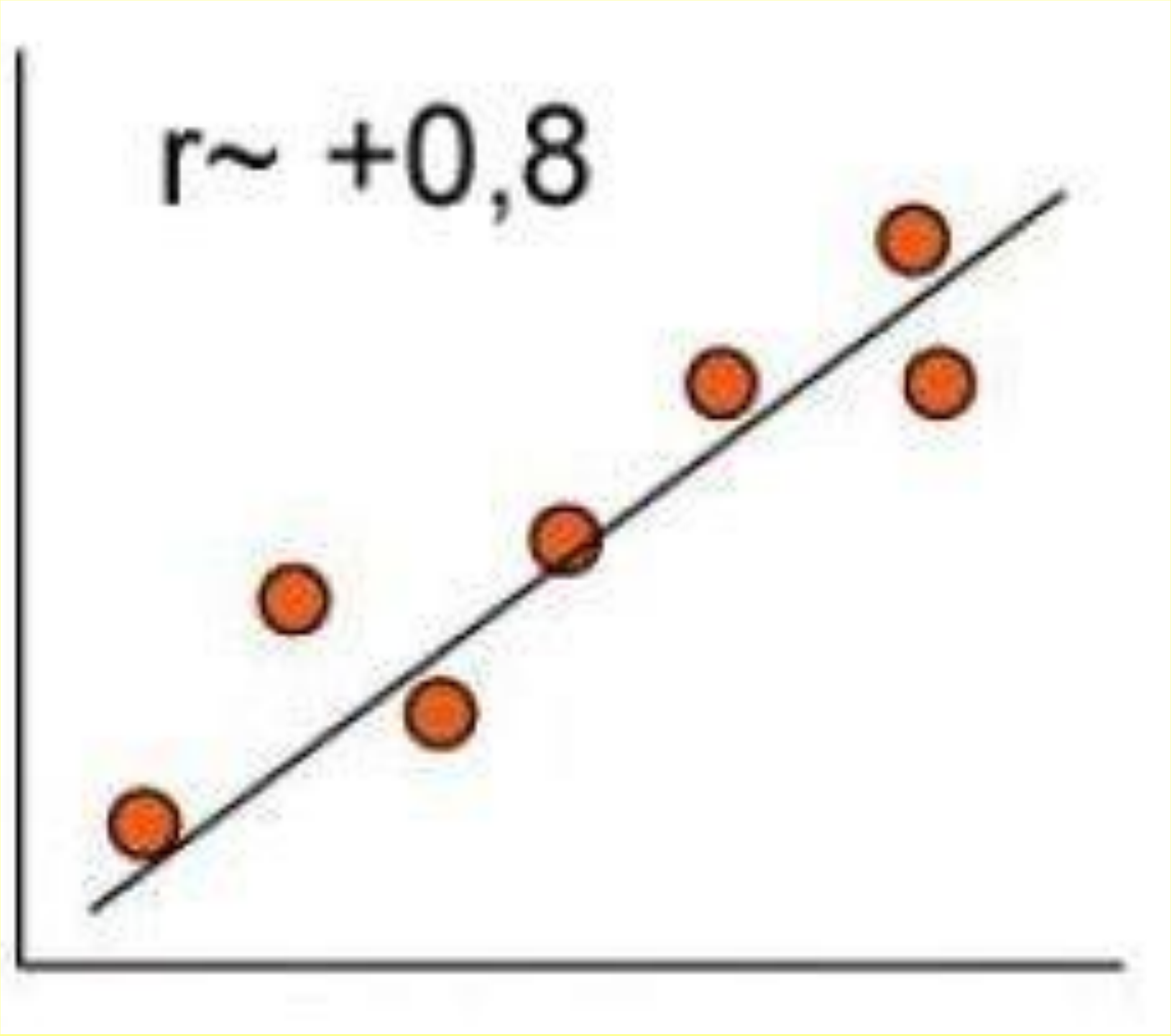
Tedy 64 % rozptylu výkonu ***ve skoku dalekém*** *je ovlivněno (determinováno) proměnlivostí výkonů* ***v běhu na 100m.***

REGRESNÍ ANALÝZA (1. úkol korelačního počtu)

6

PŘÍKLAD 7. Výpočet - koeficientů regresní přímky

Regresní přímka Y = a + b . x



REGRESNÍ ANALÝZA (1. úkol korelačního počtu)

PŘÍKLAD 7. Výpočet - koeficientů regresní přímky

Hráč	A	B	C	D	E	F	G	H	J	K
x_i	7	6	7	8	9	8	8	8	9	10
y_i	4	8	6	8	7	8	7	4	8	10

Regresní přímka $Y = a + b \cdot x$

POMOCNÁ TABULKA

Hráč	X_i	Y_i	X_i^2	Y_i^2	$X_i \cdot Y_i$

REGRESNÍ ANALÝZA (1. úkol korelačního počtu)

PŘÍKLAD 7. Výpočet - koeficientů regresní přímky

POMOCNÁ TABULKA

Hráč	X_i	Y_i	X_i^2	Y_i^2	$X_i \cdot Y_i$
A	7	4	49	16	28
B	6	8	36	64	48
C	7	6	49	36	42
D	8	8	64	64	64
E	9	7	81	49	63
F	8	8	64	64	64
G	8	7	64	49	56
H	8	4	64	16	32
J	9	8	81	64	72
K	10	10	100	100	100
Σ	80	70	652	522	569

Pořadí osob	X_i	Y_i	X_i^2	Y_i^2	$X_i \,.\, Y_i$
Σ	80	70	652	522	569

Statistické charakteristiky: $AP_x = 8$ $AP_y = 7$ $s_x = 1,1$ $s_y = 1,8$

Tento obrázek nyní nelze zobrazit.

$$b = \frac{10.\,569 - 80.70}{10\,.\,652 - (80)^2} = \frac{90}{6520 - 6400} = \frac{90}{120} = 0,75$$

$$a = \frac{70 - 0,75\,.\,80}{10} = \frac{10}{10} = 1$$

$$Y = a + b\,.\,x = 1 + 0,75\,.\,x$$

Konstrukce regresní přímky za pomocí regresní rovnice

$$Y = a + b \,.\, x = 1 + 0{,}75 \,.\, x$$

Pozn. x…nezávisle proměnná   y…závisle proměnná

Volba x $\Rightarrow$

X	8	10

$Y_1 = 1 + 0{,}75 \,.\, 8 = 7$ $Y_2 = 1 + 0{,}75 \,.\, 10 = 8{,}5$

Výpočet y $\Rightarrow$

Y	7	8,5

Z_1 (8; 7) Z_2 (10; 8,5)

2) $X = a + b \,.\, Y$ **SAMI !!!**

Pozn. y…nezávisle proměnná   x…závisle proměnná

Pořadí osob	X_i	Y_i	X_i^2	Y_i^2	$X_i . Y_i$
$\sum$	80	70	652	522	569

Statistické charakteristiky: $AP_x = 8$ $AP_y = 7$ $s_x = 1{,}1$ $s_y = 1{,}8$

Regresní přímka X = a + b . y

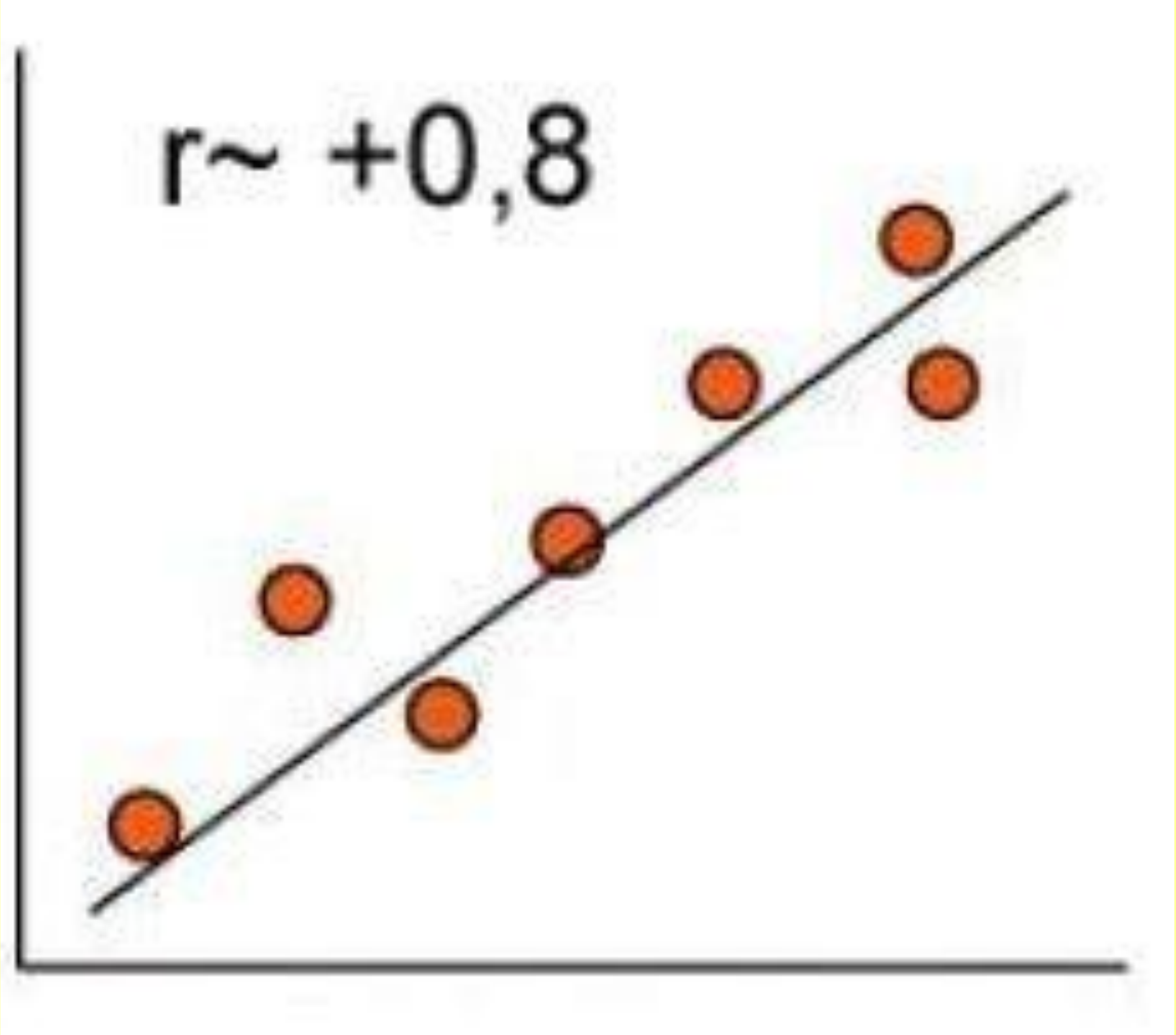
$$b = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum y_i^2 - (\sum y_i)^2} \qquad a = \frac{\sum x_i - b . \sum y_i}{n}$$

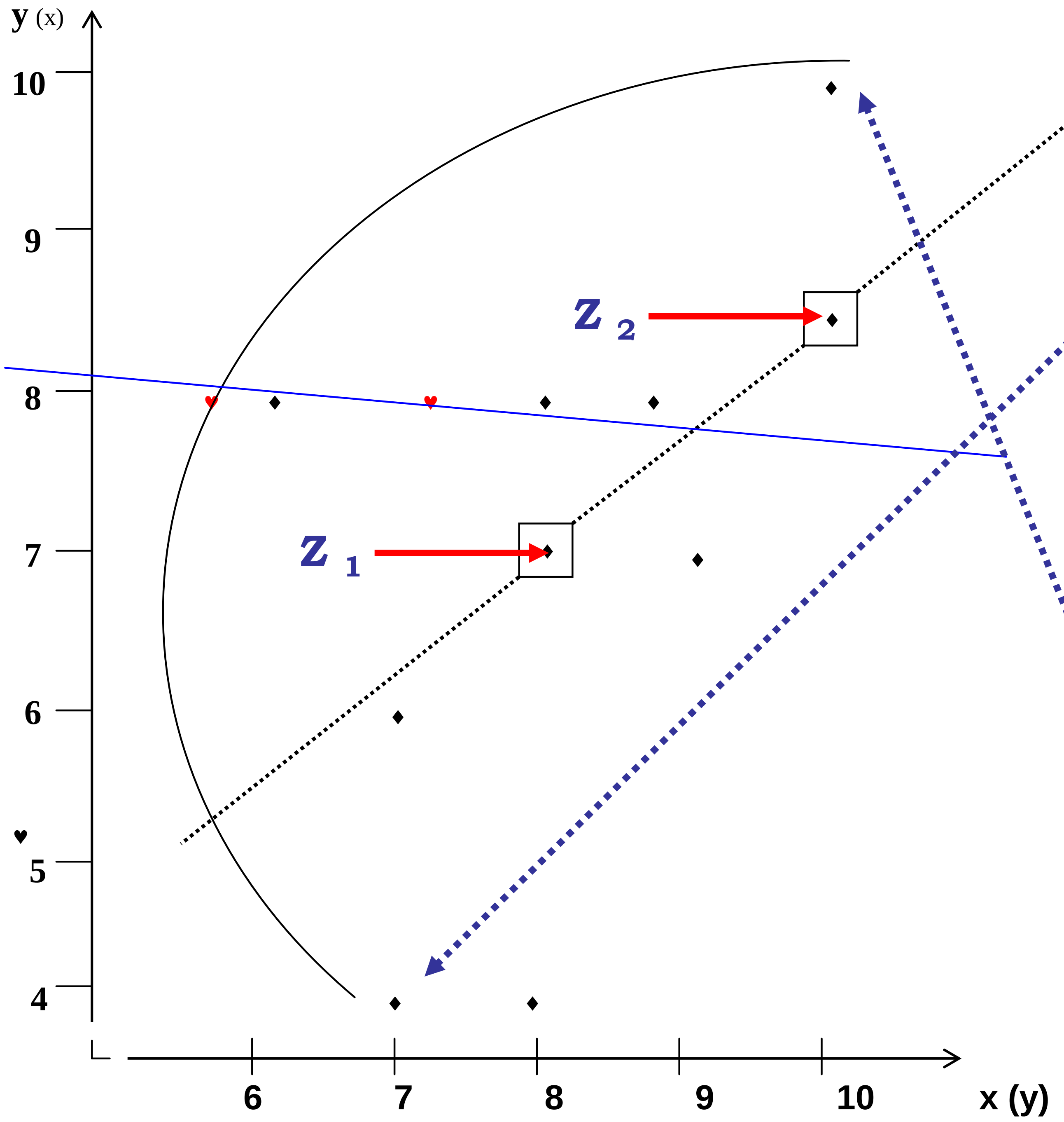
$$b = \frac{10.569 - 80.70}{10.522 - (70)^2} = \frac{5690 - 5600}{5220 - 4900} = \frac{90}{320} = 0{,}28$$

$$a = \frac{80 - 0{,}28.70}{10} = \frac{60{,}4}{10} = 6$$

$$X = a + b . y = 6 + 0{,}28 . y$$

Graf korelační závislosti (= korelogram) - konstrukce



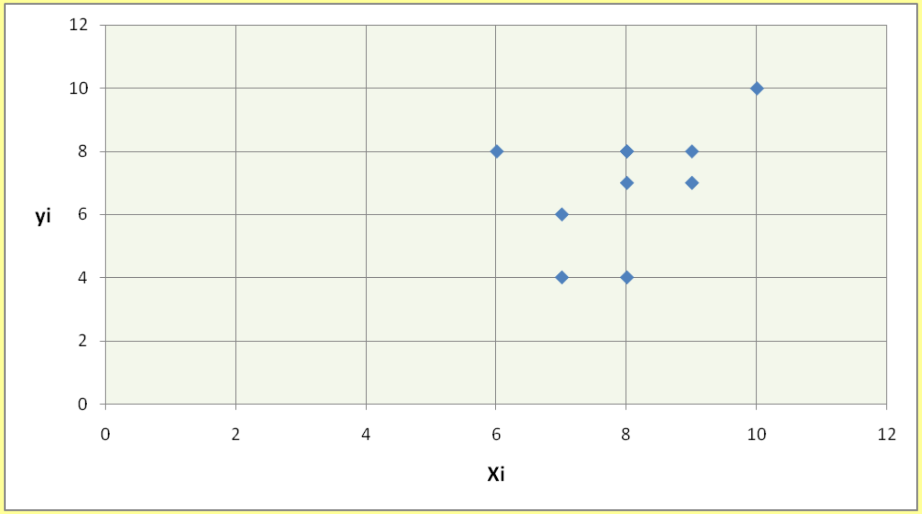


	x_i	y_i
A	7	4
B	6	8
C	7	6
D	8	8
E	9	7
F	8	8
G	8	7
H	8	4
J	9	8
K	10	10

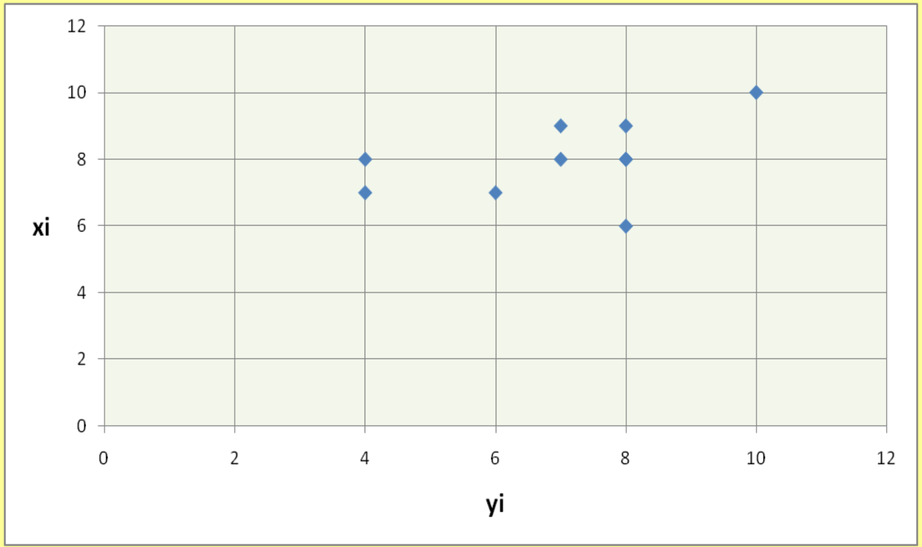
Pomocí Excelu – Statistické funkce
Výpočet koeficientů regresní přímky

	A	B	C	D
1	Hráč	xi	yi	
2	A	7	4	
3	B	6	8	
4	C	7	6	
5	D	8	8	
6	E	9	7	
7	F	8	8	
8	G	8	7	
9	H	8	4	
10	J	9	8	
11	K	10	10	
12				

1) $Y = a + b \, . \, X$

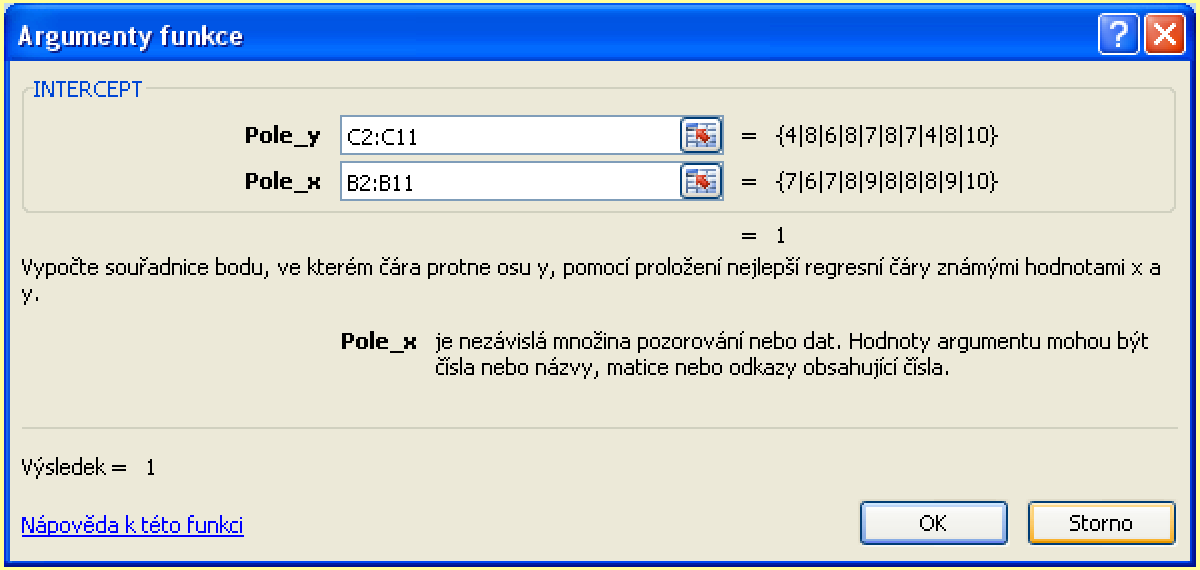


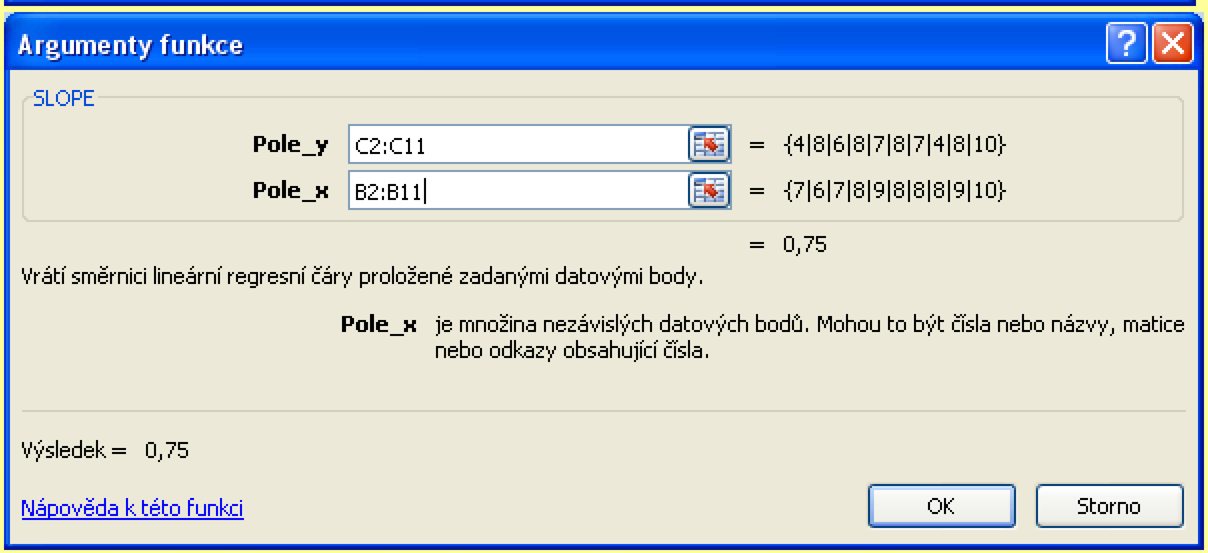
2) $X = a + b \, . \, Y$



Pomocí Excelu – Statistické funkce

Výpočet koeficientů regresní přímky





INTERCEPT
odhad parametru a

SLOPE
odhad parametru b

1) $Y = a + b \cdot X$

$Y = 1 + 0{,}75 \cdot X$

2) $X = a + b \cdot Y$

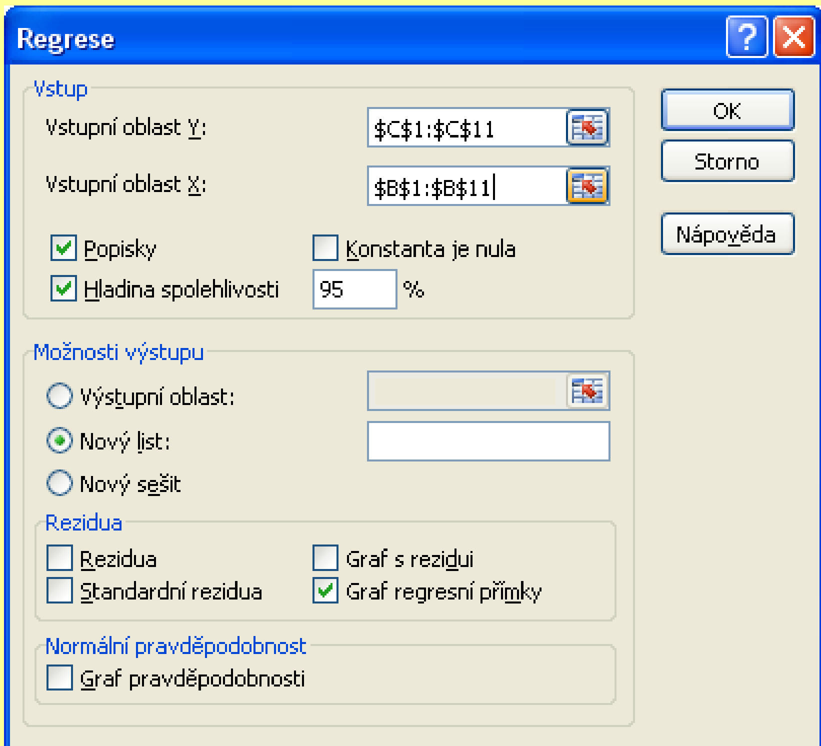
$X = 6 + 0{,}28 \cdot Y$

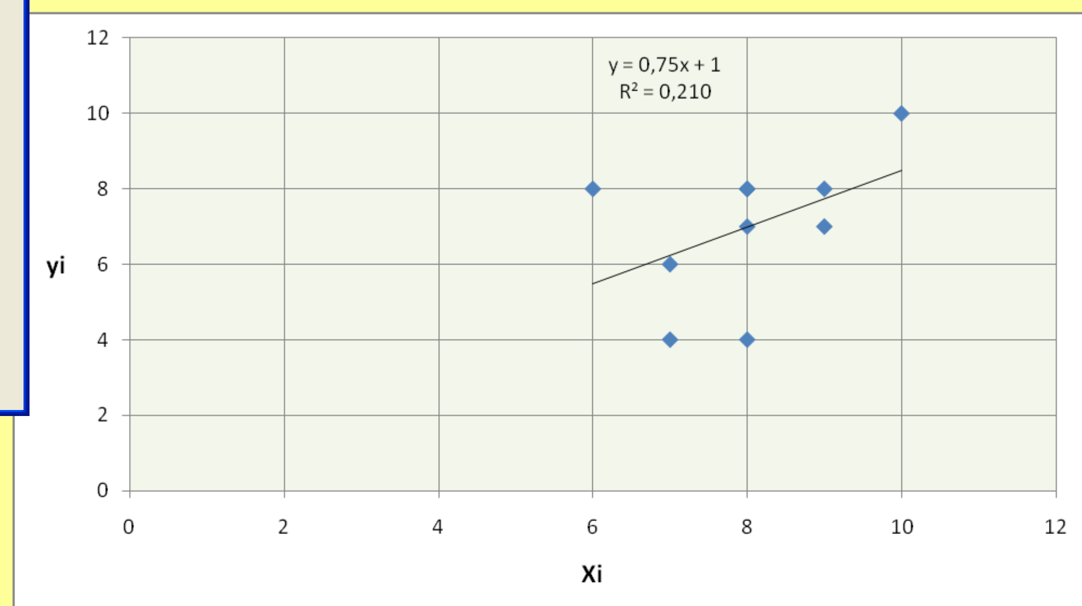
Pomocí Excelu – Analýza dat – Regrese
Výpočet koeficientů regresní přímky

1) $Y = a + b \cdot X$

$$Y = 1 + 0{,}75 \cdot X$$

	A	B	C	D
1	Hráč	xi	yi	
2	A	7	4	
3	B	6	8	
4	C	7	6	
5	D	8	8	
6	E	9	7	
7	F	8	8	
8	G	8	7	
9	H	8	4	
10	J	9	8	
11	K	10	10	
12				





Pomocí Excelu – Analýza dat – Regrese

Výpočet koeficientů regresní přímky

VÝSLEDEK

Regresní statistika	
Násobné R	0,459279327
Hodnota spolehlivosti R	0,2109375
Nastavená hodnota spolehlivosti R	0,112304688
Chyba stř. hodnoty	1,7765838
Pozorování	10

korelační koeficient
koeficient determinace

Významnost F < α = 0,05 →

model je statisticky vhodný

ANOVA

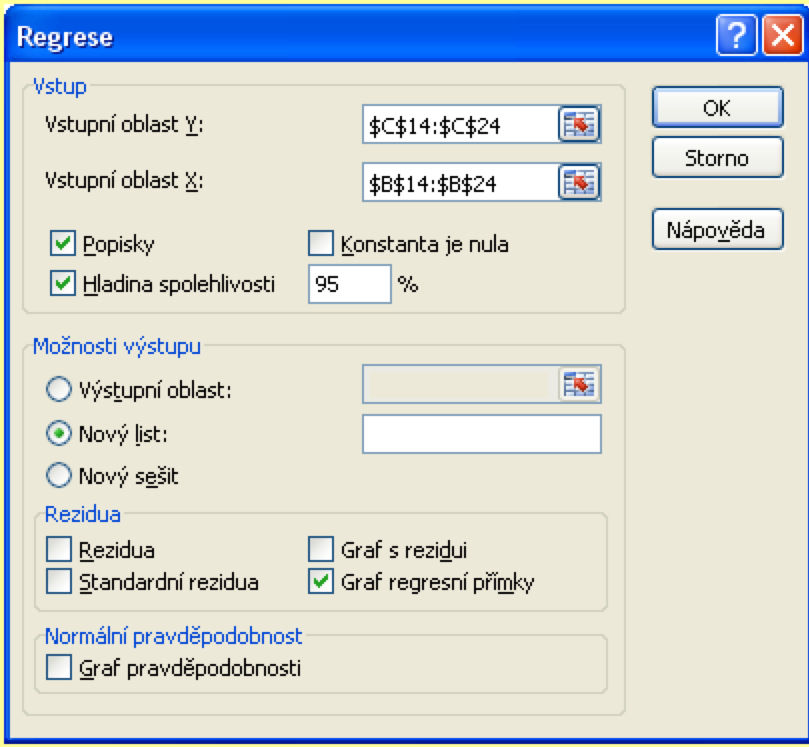
	Rozdíl	*SS*	*MS*	*F*	*Významnost F*
Regrese	1	6,75	6,75	2,138613861	0,181775314
Rezidua	8	25,25	3,15625		
Celkem	9	32			

		Koeficienty	*Chyba stř. hodnoty*	*t Stat*	*Hodnota P*	*Dolní 95%*	*Horní 95%*
Hranice	**a**	1	4,141130079	0,241479978	0,815257536	-8,549463079	10,54946308
xi	**b**	0,75	0,512855568	1,462400035	0,181775314	-0,432647059	1,932647059

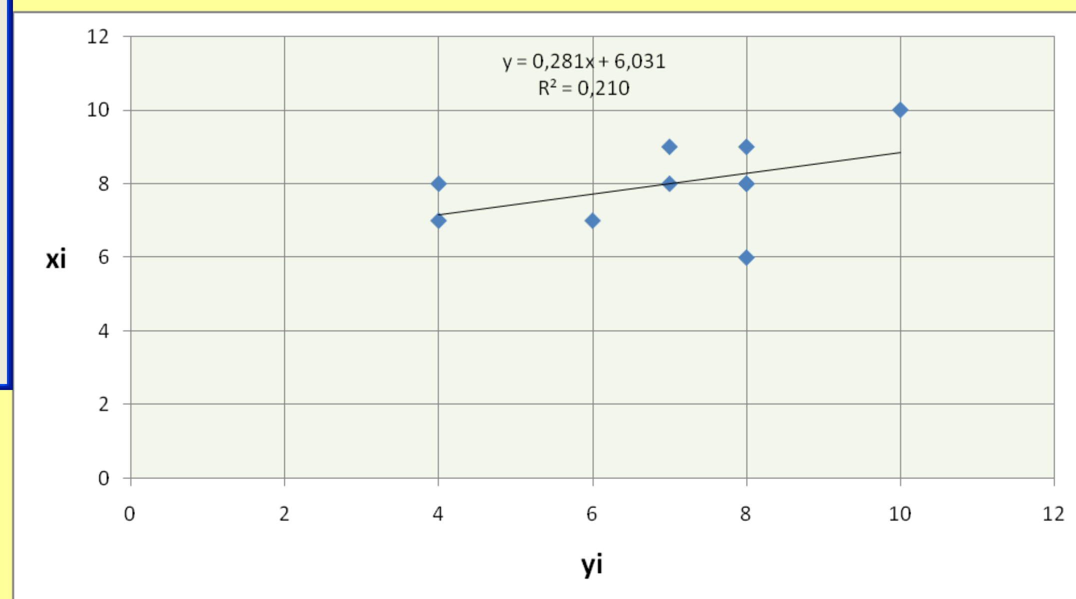
Pomocí Excelu – Analýza dat – Regrese
Výpočet koeficientů regresní přímky

2) X = a + b . Y

X = 6 + 0,28 . Y



	Hráč	yi	xi
15	A	4	7
16	B	8	6
17	C	6	7
18	D	8	8
19	E	7	9
20	F	8	8
21	G	7	8
22	H	4	8
23	J	8	9
24	K	10	10



Pomocí Excelu – Analýza dat – Regrese
Výpočet koeficientů regresní přímky

VÝSLEDEK

Regresní statistika	
Násobné R	0,459279327
Hodnota spolehlivosti R	0,2109375
Nastavená hodnota spolehlivosti R	0,112304688
Chyba stř. hodnoty	1,087930949
Pozorování	10

korelační koeficient
koeficient determinace

Významnost F < α = 0,05 →

model je statisticky vhodný

ANOVA

	Rozdíl	*SS*	*MS*	*F*	*Významnost F*
Regrese	1	2,53125	2,53125	2,138613861	0,181775314
Rezidua	8	9,46875	1,18359375		
Celkem	9	12			

		Koeficienty	*Chyba stř. hodnoty*	*t Stat*	*Hodnota P*	*Dolní 95%*	*Horní 95%*
Hranice	**a**	6,03125	1,389509735	4,340559729	0,002476521	2,827034807	9,235465193
yi	**b**	0,28125	0,192320838	1,462400035	0,181775314	-0,162242647	0,724742647

KORELAČNÍ ANALÝZA (2. úkol korelačního počtu)

PŘÍKLAD 7. Výpočet (Pearsonova) korelačního koeficientu

Pořadí osob	X_i	Y_i	X_i^2	Y_i^2	$X_i . Y_i$
$\sum$	80	70	652	522	569

Výpočtový tvar

$$r_{x,y} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{[n \sum x_i^2 - (\sum x_i)^2][n \sum y_i^2 - (\sum y_i)^2]}}$$

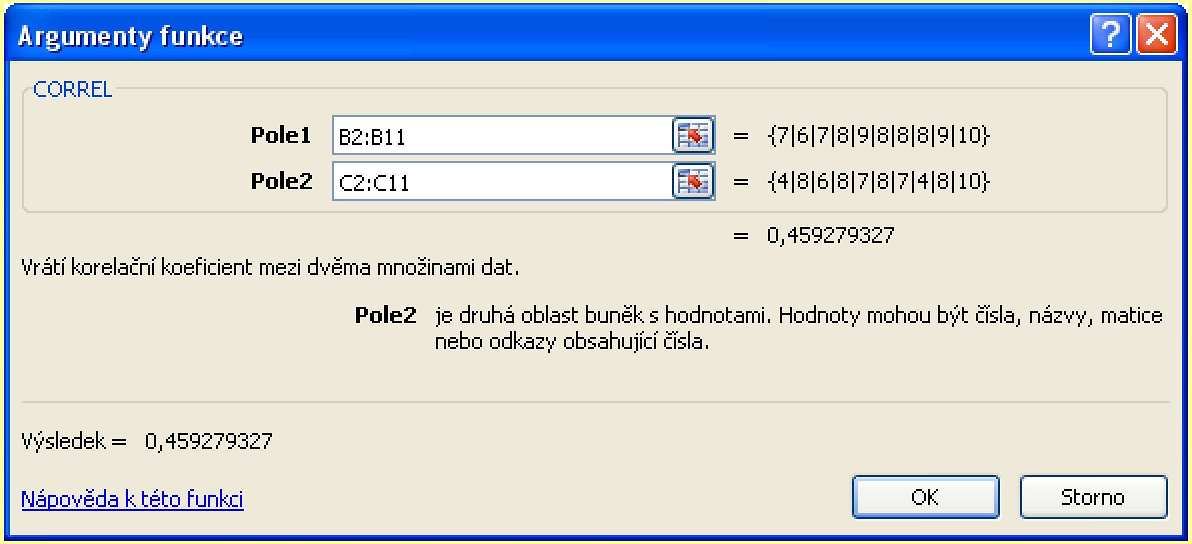
$$r = \frac{10.\ 569 - 80.70}{\sqrt{(10.\ 652 - 6400)\ .\ (10.\ 522 - 4900)}} = 0{,}46$$

Pomocí Excelu – Statistické funkce
Výpočet (Pearsonova) korelačního koeficientu

	A	B	C	D
1	Hráč	xi	yi	
2	A	7	4	
3	B	6	8	
4	C	7	6	
5	D	8	8	
6	E	9	7	
7	F	8	8	
8	G	8	7	
9	H	8	4	
10	J	9	8	
11	K	10	10	
12				

CORREL

výpočet korelačního koeficientu



r = 0,459

Jak „těsná“ je korelační závislost r = 0,46 ?

Vzhledem k intervalu <0;1> resp. <-1;0> se jedná o *střední míru závislosti (kladnou).*

POSOUZENÍ A INTERPRETACE ZÁVISLOSTI

1. Korelační závislost (vyjádřená korelačním koeficientem) platí pouze pro konkrétní výběr (soubor) s konkrétními osobami, nelze tedy považovat tento vztah za obecně platný!

2. Chceme-li zobecnit platnost vypočítané závislosti „r“ na základní soubor (populaci), musíme ověřit (testovat) hypotézu o statistické významnosti korelačního koeficientu.

3. Při testování významnosti „r“ (odlišnost od nuly), zjišťujeme, zda je výběrový korelační koeficient statisticky významný (s ohledem na rozsah souboru).

4. Zamítnutí (či nezamítnutí) nulové hypotézy provádíme s určitou pravděpodobností na tzv. hladině významnosti (obvykle bývá volena *p = 0,05 resp. p = 0,01).*

PRO NÁŠ PŘÍKLAD, kdy *r = 0,46; n = 10*

...zjistíme v tabulce kritických hodnot koeficientu součinové korelace, ...

Tabulka kritických hodnot

Počet dvojic	Kritické hodnoty (na α=0,05, α=0,01)	
n	α=0,05	α=0,01
9	0,666	0,798
10	0,632	0,765
11	0,602	0,735
30	0,361	0,463

... že „náš" korelační koeficient r = 0,46 je pro obě hladiny významnosti menší, než tzv. kritická hodnota, je tedy

STATISTICKY NEVÝZNAMNÝ.

Závěr: mezi výsledky 1. a 2. pokusů nebyla zjištěna závislost, nelze tvrdit, že... CO? Interpretace!

Ale pro n=30?

Příklad. **Výpočet (Pearsonova) korelačního koeficientu**

Testujte hypotézu, zda <u>výběrový korelační koeficient</u> je statisticky významný (s ohledem na rozsah souboru).

Test1	Test2
8	7
5	5
4	4
6	4
7	5
6	4
5	5
7	6

Tab. XVIII.5. Kritické hodnoty $r(\alpha)$ korelačního koeficientu r;
$P\{|r| \geqq r(\alpha)\} = \alpha$

Rozsah výběru n	α	
	0,05	0,01
3	0,996 9	0,999 9
4	0,950 0	0,990 0
5	0,878 3	0,958 7
6	0,811 4	0,917 2
7	0,754 5	0,874 5
8	0,706 7	0,834 3
9	0,666 4	0,797 7
10	0,631 9	0,764 6
11	0,602 1	0,734 8
12	0,576 0	0,707 9
13	0,552 9	0,683 5
14	0,532 4	0,661 4
15	0,514 0	0,641 1
16	0,497 3	0,622 6
17	0,482 2	0,605 5
18	0,468 3	0,589 7
19	0,455 5	0,575 1
20	0,443 8	0,561 4
21	0,432 9	0,548 7
22	0,422 7	0,536 8
23	0,413 2	0,525 6
24	0,404 4	0,515 1
25	0,396 1	0,505 2

Příklad. Výpočet (Pearsonova) korelačního koeficientu

Pořadí osob	X_i	Y_i	X_i^2	Y_i^2	$X_i \cdot Y_i$
$\sum$	48	40	300	208	247

$$r_{x,y} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{[\, n \sum x_i^2 - (\sum x_i)^2]\,[n \sum y_i^2 - (\sum y_i)^2]}}$$

$$r = \frac{8.\,247 - 48.40}{\sqrt{(8.\,300 - 2304)\,.\,(8.\,208 - 1600)}} = \frac{56}{78} = 0{,}71$$

r = 0,71 > 0,7067 pro α = 0,05

Rozsah výběru n	α 0,05	α 0,01
3	0,996 9	0,999 9
4	0,950 0	0,990 0
5	0,878 3	0,958 7
6	0,811 4	0,917 2
7	0,754 5	0,874 5
8	0,706 7	0,834 3
9	0,666 4	0,797 7
10	0,631 9	0,764 6

Na hladině α = 0,05 zamítáme nulovou hypotézu. Koeficient je statisticky významný.

SPEARMANŮV KOEFICIENT POŘADOVÉ KORELACE

Spearmanův koeficient pořadové korelace se používá pro výpočet těsnosti závislosti:

- **u znaků získaných na ordinální stupnici (ordinálních znaků)**
- **u souborů o nevelkém rozsahu (n menší než 20)**
- **jestliže znaky nemají (či nelze prokázat) normální rozložení četností**

Vzorec pro výpočet Spearmanova koeficientu pořadové korelace:

$$r_{xy} = 1 - \frac{6 \cdot \sum (i_x - i_y)^2}{n\,(n^2 - 1)}$$

kde i_x resp. i_y je *index pořadí znaků* x resp. y

Příklad. **Výpočet Spearmanova koeficientu pořadové korelace**

Pořadí	x_i	y_i	i_x	i_y	$(i_x - i_y)^2$
1	7 2,5.	4 1.5	2,5	1,5	1
2	6 1.	8 7,5.	1	7,5	42,25
3	7 2,5.	6 3	2,5	3	0,25
4	8	8 7,5.	5,5	7,5	4
5	9	7 4,5.	8,5	4,5	16
6	8	8 7,5.	5,5	7,5	4
7	8	7 4,5.	5,5	4,5	1
8	8	4 1.5	5,5	1,5	16
9	9	8 7,5.	8,5	7,5	1
10	10 10.	10 10.	10	10	0
$\sum$	-	-	-	-	85,5

$$r = 1 - \frac{6 \cdot \sum (i_x - i_y)^2}{n(n^2 - 1)} = 1 - \frac{6 \cdot 85{,}5}{10\,(100 - 1)} = 1 - \frac{513}{990} = 0{,}48$$

Spearmanův koeficient pořadové korelace

r = 0,48

Pearsonův koeficient součinové korelace

r = 0,46

POSOUZENÍ A INTERPRETACE ZÁVISLOSTI

...viz Pearsonův koeficient součinové korelace

Tab. XVIII.6. Kritické hodnoty $r_S(\alpha)$ Spearmanova korelačního koeficientu r_S; $P\{|r_S| > r_S(\alpha)\} \leq \alpha$

n	α 0,05	α 0,01
5	0,900 0	—
6	0,828 6	0,942 9
7	0,745 0	0,892 9
8	0,690 5	0,857 1
9	0,683 3	0,816 7
10	0,636 4	0,781 8
11	0,609 1	0,754 5
12	0,580 4	0,727 3
13	0,554 9	0,697 8
14	0,534 1	0,674 7
15	0,517 9	0,653 6
16	0,500 0	0,632 4
17	0,485 3	0,615 2
18	0,471 6	0,597 5
19	0,457 9	0,582 5
20	0,445 1	0,568 4
21	0,435 1	0,554 5
22	0,424 1	0,542 6
23	0,415 0	0,530 6
24	0,406 1	0,520 0
25	0,397 7	0,510 0
26	0,389 4	0,500 2
27	0,382 2	0,491 5
28	0,374 9	0,482 8
29	0,368 5	0,474 4
30	0,362 0	0,466 5

Příklad. **Výpočet Spearmanova koeficientu pořadové korelace**

	A	B	C
1	VÝROBEK	ODBORNÍCI	LAICI
2	1	7	8
3	2	9	9
4	3	8	7
5	4	10	10
6	5	6	6
7	6	5	4
8	7	3	5
9	8	4	3
10	9	2	2
11	10	1	1
12			

Tab. XVIII.6. Kritické hodnoty $r_S(\alpha)$ Spearmanova korelačního koeficientu r_S; $P\{|r_S| > r_S(\alpha)\} \leq \alpha$

n	α	
	0,05	0,01
5	0,900 0	—
6	0,828 6	0,942 9
7	0,745 0	0,892 9
8	0,690 5	0,857 1
9	0,683 3	0,816 7
10	0,636 4	0,781 8

F	G	H	I
ix	iy	ix - iy	(ix - iy)^2
7	8	-1	1
9	9	0	0
8	7	1	1
10	10	0	0
6	6	0	0
5	4	1	1
3	5	-2	4
4	3	1	1
2	2	0	0
1	1	0	0

$$r = 1 - \frac{6 \cdot \sum (i_x - i_y)^2}{n(n^2 - 1)} = 1 - \frac{6 \cdot 8}{10\,(100 - 1)} = 1 - \frac{48}{990} = 0{,}95$$

Kritické hodnoty z tabulek α = 0,05 0,6364

α = 0,01................ 0,7818

Hypotézu H_0 : $\rho = 0$ o nezávislosti zamítáme

ANALÝZA JEDNOROZMĚRNÉHO SOUBORU

METODY DESKRIPTIVNÍ STATISTIKY

1. URČENÍ TYPU ŠKÁLY (nominální, ordinální, metrické)

a) nominální + ordinální ⇒ *neparametrické stat. metody*

b) metrické ⇒ *parametrické statistické metody*

2. ROZLOŽENÍ ČETNOSTÍ ZNAKŮ (NORMÁLNÍ ČI JINÉ)

a) normální ⇒ *parametrické statistické metody*

b) jiné ⇒ *neparametrické statistické metody*

3. VÝPOČET ZÁKLADNÍCH STATISTICKÝCH CHARAKTERISTIK

a) míry centrální tendence

b) míry variability

c) míry závislosti

Děkuji za pozornost