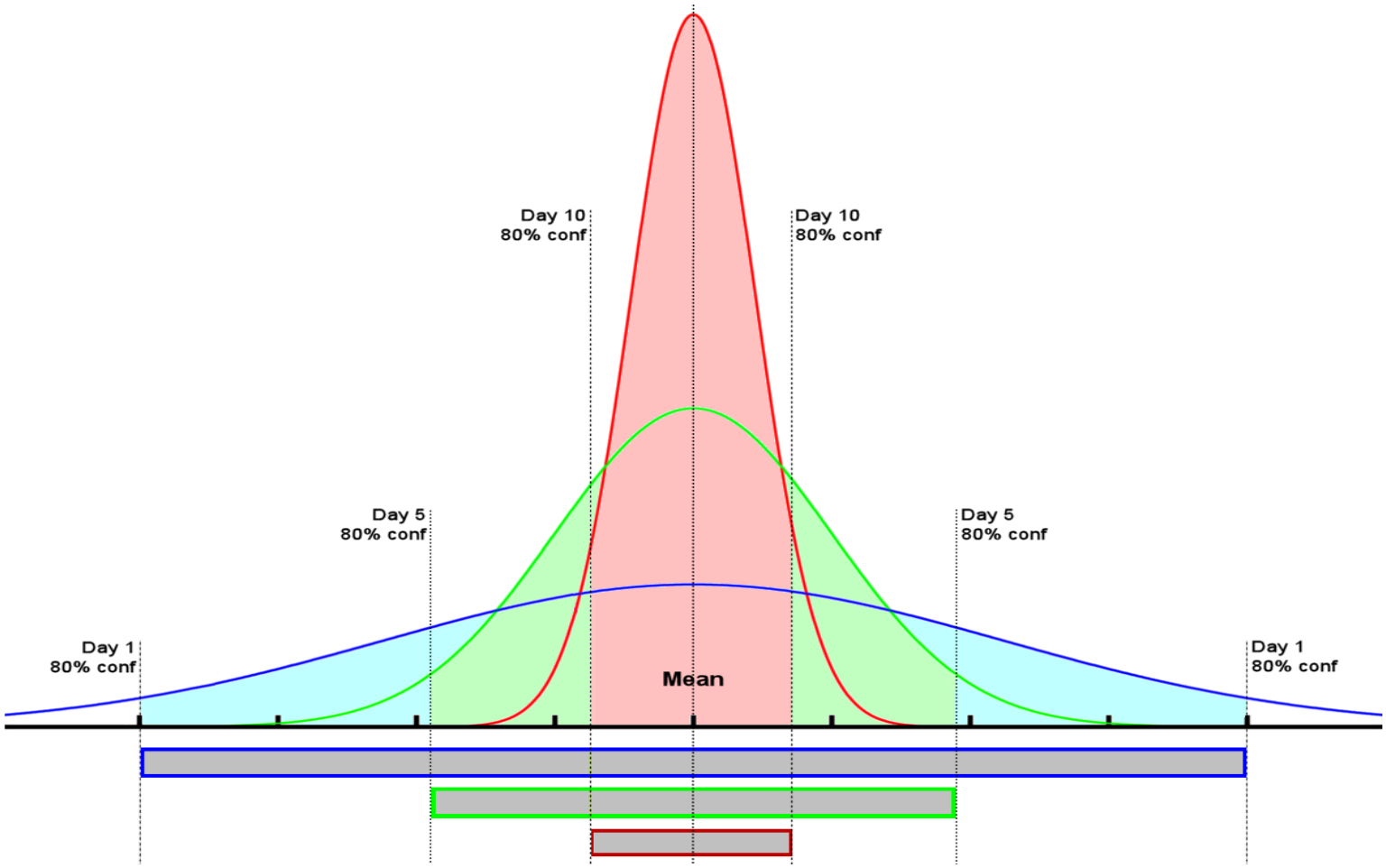
Pearsonův korelační koeficient r



Pearsonův korelační koeficient:

Výpočet korelačního koeficientu je další velmi častou metodou, používanou pro ověření určitých hypotéz ve studentských pracích. V případě t-testu jsme ověřovali, zda se liší nebo neliší mezi sebou dvě podskupiny respondentů v nějaké vlastnosti, postoji, míře určitých preferencí či výkonu.

Výpočet korelačního koeficientu používáme, když chceme ověřit, zda navzájem souvisejí dvě různé proměnné (=dvě číselně vyjádřené otázky v dotazníku – měřící vlastnosti, postoje, výkony apod.).

- Např. zda souvisí tolerance k cizincům s věkem respondentů? Zde můžeme vytvořit třeba tuto hypotézu, ze s narůstajícím věkem klesá tolerance k cizincům.
- Jiným příkladem hypotézy by mohlo být, že s počtem pravidelně navštěvovaných volnočasových kroužků klesá u dospívajících množství užívaných návykových látek.

Nebo ověřujeme zájem o zdravý životní styl u vzdělanější části naší populace. Hypotézu můžeme do dotazníku (a následného ověřování) operacionalizovat takto: s narůstajícím ukončeným vzděláním stoupá i počet dodržovaných pravidel zdravého životního stylu. (Dosažené vzdělání respondentů lze převést na pořadovou proměnnou. Např. tak, že nejnižší číselnou hodnotu 1 přiřadíme základnímu vzdělání, ukončenému SOU dáme číslo 2, maturitě=3, Dis.=4, VŠ=5 apod.).

Výpočet korelací tedy používáme u hypotéz, kde **zjišťujeme vztah mezi dvěma proměnnými** (vyčíslenými jevy) **u jednoho souboru respondentů.**

Metoda pro kontinuální proměnné k vyhodnocení položek tohoto typu:

- **Výpověď** respondentů je vyjádřená **číselně**: věk; jak dlouho se léčí s nějakou nemocí; průměrný čas (v minutách), který tráví denně hraním počítačových her, …

- **Výkon** respondentů je vyjádřený **číselně** (známka z matiky; dosažený skór v IQ testu; skór v dotazníku úzkostnosti,…)

- **Odpovědi** na otázky, které mají vyjádřit **číslem**, např.:
  „Na škále od 1 do 7 vyznačte, jak moc se vám líbí na HTF.“

- Jednotlivé **odpovědi** u dané otázky vyjadřují **míru** něčeho – např. „míru“ dosaženého vzdělání (lze je seřadit podle míry; tj. pořadové proměnné), máme-li aspoň cca 100 respondentů.

= Jednotlivé varianty odpovědí na danou otázku odrážejí i vztah mezi sebou, **dají se seřadit podle kvantity nebo intenzity**, kterou vyjadřují.

– např. **LIKERTOVA ŠKÁLA** (jak moc souhlasím s určitým výrokem):

„Cizinci si za své problémy v naší zemi většinou mohou sami“

a) rozhodně ano	b) spíše ano	c) nevím	d) spíše ne	e) rozhodně ne
5	**4**	**3**	**2**	**1**

„Matematiku mám:“

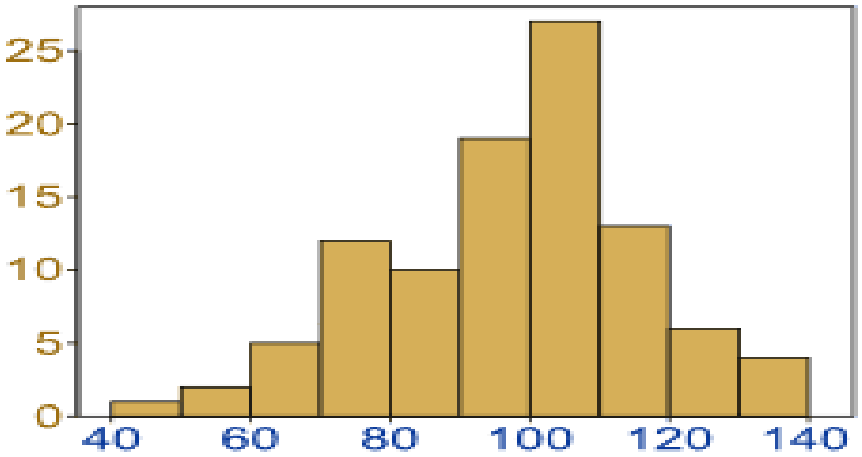
a) Hodně rád	b) někdy rád	c) jak kdy	d) spíše nerad	e) Nerad
5	**4**	**3**	**2**	**1**

– nebo **BOGARDOVA ŠKÁLA** SOCIÁLNÍCH VZDÁLENOSTÍ, např.:

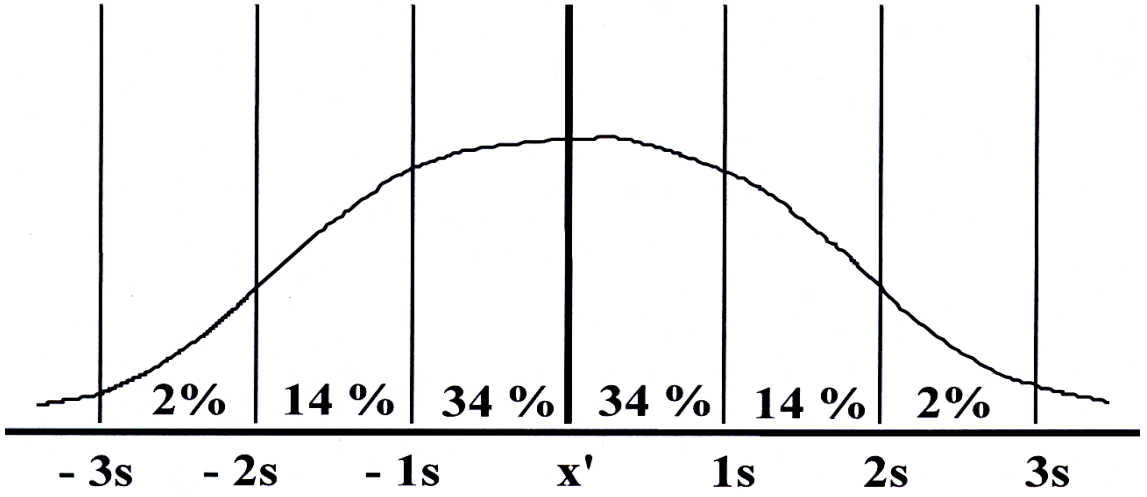
„Nevadilo by mi, kdyby nějaký cizinec nebo imigrantská rodina:“

a)	**Bydleli v naší ulici**	**1**
b)	**Bydleli v našem domě**	**2**
c)	**Přátelili se mými dětmi**	**3**
d)	**Byli mými blízkými příbuznými**	**4**

Pro pochopení principu předpokládejme ideální případ Gaussovy křivky či histogramu s dobrým rozložením odpovědí



GAUSSOVA KŘIVKA



Tentokrát však neporovnáváme dvě skupiny respondentů,
ale jednu skupinu **ve dvou číselně vyjádřených odpovědích na dvě různé otázky**

NAPŘ: *Existuje souvislost mezi výší vzdělání a tolerancí?*

Existuje vztah mezi věkem a jazykovými dovednostmi v angličtině?

Existuje vztah mezi velikostí nabídky služeb v různých pobytových zařízeních pro seniory a jejich osobní spokojeností?

- Máme jednu skupinu respondentů **(jedno společné N)**
- dvě otázky a z nich dvě vyčíslené průměrné odpovědi **(x‘** a **y‘)**
- dvě směrodatné odchylky **(tedy dvě různě široké Gaussovy křivky - SD_x** a **SD_y)**

Pearsonův korelační koeficient:

Výpočet Pearsonova korelačního koeficientu vychází ze stejného principu jako výpočet t-testu.

Předpokládáme existenci dvou (Gaussových) křivek: pro odpovědi na první a na druhou otázku, které chceme vzájemně korelovat.

N je počet respondentů v souboru

x‘ je průměr vypočítaný z číselně vyjádřených odpovědí na první otázku

y‘ je průměr vypočítaný z číselně vyjádřených odpovědí na druhou otázku

$\mathbf{SD_x}$ je směrodatná odchylka vypočítaná z odpovědí u první otázky

$\mathbf{SD_y}$ je směrodatná odchylka vypočítaná z odpovědí u druhé otázky

Z těchto proměnných vychází výpočet Pearsonova korelačního koleficientu.

- Je to tedy stejné jako u t-testu, pouze máme jeden stejný počet respondentů (společné N).

Výpočet **Pearsonova korelačního koeficientu *r***

(sledujeme **vztah mezi dvěma proměnnými [„otázkami“] u jedné skupiny respondentů)**

tj. vztah dvou „Gaussových křivek“, avšak v jedné skupině:

$$r = \frac{[(x_1 - x') \cdot (y_1 - y')] + [(x_2 - x') \cdot (y_2 - y')] + \ldots + [(x_n - x') \cdot (y_n - y')]}{N \cdot SD_x \cdot SD_y}$$

Pearsonův korelační koeficient (výpočet):

Ten vzoreček si jen prohlédněte. Není třeba se ho učit, ale může nám pomoci, když pochopíme jeho princip. Pokud chcete (je to čistě dobrovolné):

*V první kulaté závorce je uvedena číselná odpověď prvního respondenta (**x_1**), kterou odečítáme od průměru odpovědí na tuto první otázku za celou skupinu (**x‘**).*

*Ve druhé kulaté otázce je totéž, ale týká se to druhé otázky: odpověď na druhou otázku (**y_1**) se odečítá od průměru u druhé otázky za celou skupinu (**y‘**).*

*Uvnitř první hranaté závorky **[ ]** se oba výsledky násobí. Výsledek této závorky je kladný, když mají obě kulaté závorky kladnou, nebo obě zápornou hodnotu. Znamená to, že první respondent odpověděl v obou otázkách buď vyšší hodnotou než průměrnou, nebo naopak nižší hodnotou než je průměr skupiny.*

Pokud však respondent odpověděl v jedné otázce nad průměrem a ve druhé pod průměrem, vyjde v první kulaté závorce číslo kladné a ve druhé číslo záporné. Jejich vynásobením vyjde záporné číslo.

(Druhá hranatá závorka obsahuje stejný výpočet pro druhého respondenta: jeho odpovědi na první a druhou otázku se odečítají od průměrné odpovědi za celou skupinu.)

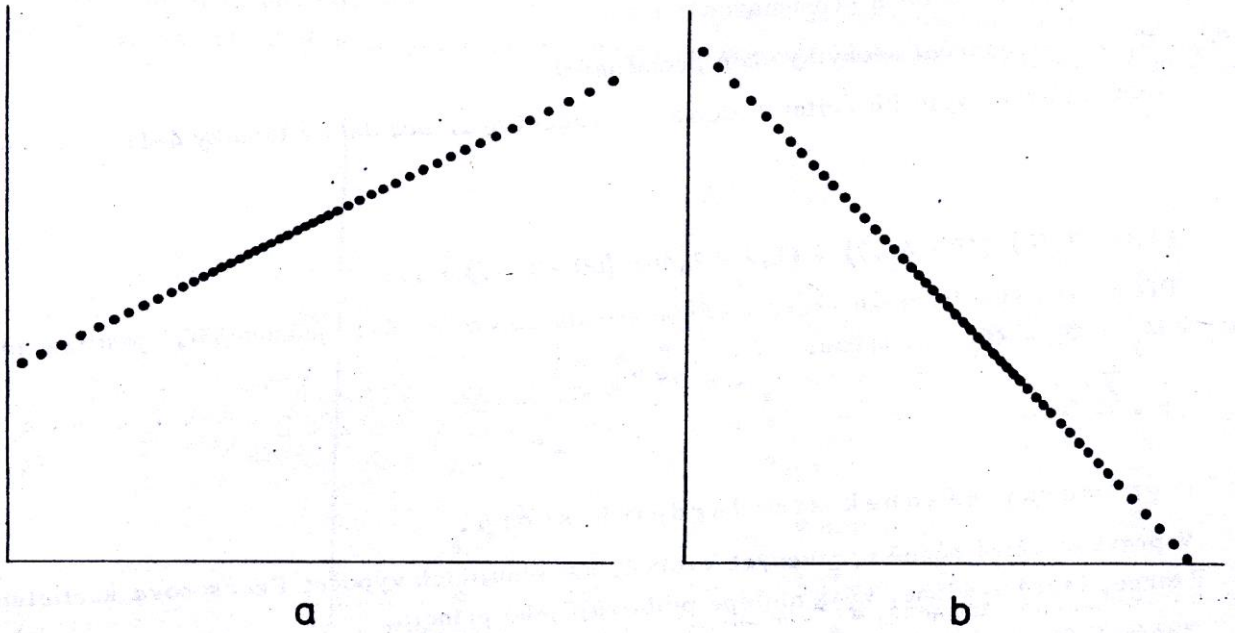
*Odpovědi každého respondenta se v hranatých závorkách takto odečítají a násobí (od prvního až po posledního, n-tého respondenta – viz **x_n** a **y_n** na konci vzorečku).*

*Jak bylo řečeno, když jsou obě odpovědi u každého respondenta (**x_{1-n}** a **y_{1-n}**) nad průměrem skupiny (nebo naopak obě odpovědi pod průměrem skupiny), jejich vynásobením získáme kladné číslo. Kladné číslo se k ostatním číslům přičítá a zvyšuje hodnotu celkové korelace.*

Když je rozdíl odpovědi od průměru u respondenta na jednu otázku kladný a na druhou záporný, jejich násobením vyjde záporné číslo, které se od ostatních čísel odečítá a tím snižuje hodnotu celkové korelace.

Výpočet **Pearsonova korelačního koeficientu *r***

ideální rozložení naměřených případů
(absolutní přímá souvislost [r = 1,0]) a absolutní nepřímá souvislost [r = -1,0])



Obr. 4-4
Korelační grafy pro extrémní hodnoty Pearsonova korelačního koeficientu:
/a/ r = 1,00 /b/ r = -1,00

Pearsonův korelační koeficient (ideální, nereálné případy):

Obrázek (a) znázorňuje absolutní přímou úměrnost mezi dvěma jevy (přibližně např. čím jsou děti starší, tím jsou také vyšší). Věk dětí by mohl být vynesený třeba na vodorovné ose a výška dětí na svislé.

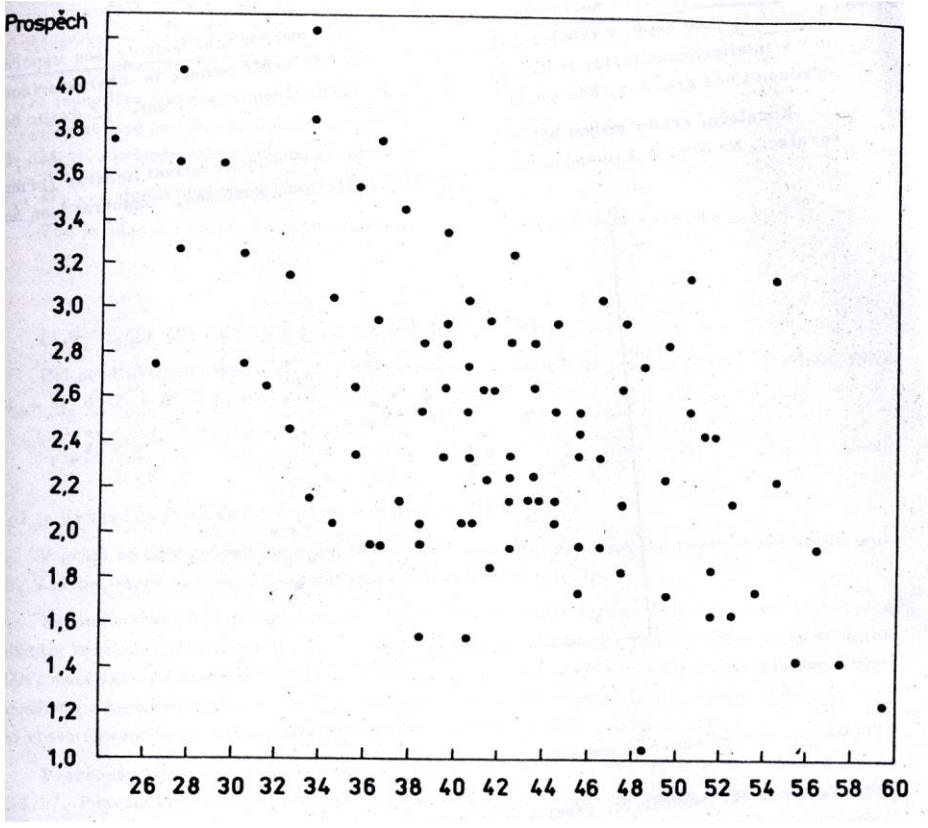
Obrázek (b) znázorňuje absolutní nepřímou úměrnost mezi dvěma jevy (přibližně např. čím víc se věnujeme ve volném čase nějakému koníčku, tím méně volného času nám zbývá na cokoliv jiného). Doba strávená s určitým koníčkem by byla vynesená třeba na vodorovné ose a zbývající čas na vše ostatní na svislé ose.

Absolutní přímá a nepřímá úměrnost se vyskytuje se spíš v přírodních vědách. Ve společenských vědách je vzácná.

Výpočet **Pearsonova korelačního koeficientu *r***

Obvyklá variabilita naměřených případů (vztah mezi počty bodů v IQ testu a prospěchem)

Jedná se o náznak přímé, nebo nepřímé úměrnosti?
Bude korelační koeficient „r“ kladné, nebo záporné číslo?



Pearsonův korelační koeficient (případy obvyklé ve společ. vědách):

Mnohem častěji se ve společenských vědách setkáváme s takovouto nepřesnou podobou korelačního vztahu mezi dvěma korelovanými proměnnými. Tečky jsou jednotliví respondenti ze zkoumaného souboru.

– Na vodorovné ose jsou počty bodů, kterých dosáhli v nějakém znalostním testu.

– Na svislé ose jsou jejich školní známky.

Pokud se díváme na tvar, který tečky jako celek naznačují, ukazuje se, že u většiny aspoň do jisté míry platí, že čím víc bodů v testu získali, tím lepší mají prospěch. Je to tedy také nějaká přímá úměrnost, i když vůbec ne tak přesná jako na předchozím snímku.

(Je však znázorněná jinak než na předchozím obrázku, protože lepší prospěch znamená nižší klasifikační číselný stupeň. Směr většiny teček je tedy obráceně než u předchozího snímku).

Výpočet **Pearsonova korelačního koeficientu *r***

- Rozsah korelačního koeficientu je vždy **v rozmezí +1 až -1**.
  - obr. (a) na snímku č. 10 by vyjadřoval korelační koeficient $r = 1$
  - obr. (b) na snímku č. 10 by vyjadřoval výsledek korelací $r = -1$
  - obr. na snímku 12 by mohl mít hodnotu korelací přibližně $r = 0{,}6$

- **Čím vyšší je jeho absolutní hodnota**, tím vyšší je korelační vztah (např. se dá předpokládat, že čím větší je nabídka služeb v DD, tím je větší spokojenost jejich obyvatel).

- **Čím víc se blíží nule**, tím je tento vztah mezi dvěma proměnnými [otázkami] míň významný. Např.: pravděpodobně neexistuje příliš silný vztah mezi délkou vlasů a výsledkem v IQ testu.

- **Pokud je číslo záporné**, jde o vztah nepřímé úměrnosti (např. čím je respondent starší, tím méně je tolerantní k minoritám). Kladný i záporný korelační vztah mezi dvěma proměnnými tedy může (nebo nemusí) být statisticky významný.

Jak poznáme, že je korelační koeficient významný?

- I zde jsou používány nám již známé dvě hladiny významnosti, pětiprocentní a jednoprocentní (tj. $p < 0,05$ nebo $p < 0,01$).

- I zde jako v t-testu platí, že na tom, zda bude určitá hodnota *r* statisticky významná nebo nevýznamná, je dáno i počtem respondentů:

  - korelační koeficient ***r* = 0,3** ještě nemusí být statisticky významný při 25 respondentech,

  - ale při počtu respondentů N=200 už významná nejspíš bude.

To, zda bude tato hodnota významná, nebo nevýznamná pro konkrétní výpočet, však souvisí nejen s velikostí ***r***, ale také s počtem respondentů (**N**) ve skupině (viz snímek).

Pro zjištění výsledku bohužel nemáme přesnou tabulku s hraničními hodnotami. **Určité odhady nám však poskytuje tabulka, kterou máte v PDF v další příloze.**

Co ještě můžeme vyčíst z hodnoty korelačního korelační koeficientu ***r***? - *nepovinné*

Z vypočítaného korelačního koeficientu můžeme dopočítat **koeficient determinace R** pomocí vzorce **R = 100 . r^2**

(koeficient determinace se vyjadřuje v procentech, proto se r^2 násobí hodnotou 100).

Například vypočítáme, že korelační koeficient vztahu mezi výškou a váhou studentů HTF je ***r* = 0,6**.

Koeficient determinace: **R** = $0{,}6^2$. 100 **R** = 36 %

Můžeme říci, že pokud známe výšku studenta, máme také 36 % informací o jeho váze. (Nebo naopak, známe-li jeho váhu, máme i 36 % informací o jeho výšce.)

Co ještě můžeme vyčíst z hodnoty korelačního korelační koeficientu ***r***? – *nepovinné – II.*

Na předchozím a tomto snímku jsou informace jen pro zájemce, kteří chtějí vědět, čeho se vlastně v korelačním výpočtu dopočítali (není to povinné).

Pokud jste si snímek přečetli, možná vás napadne, že mít 36 % informací místo celých 100 % informací zase není tak moc.

<u>Co těch chybějících 64 % informací</u>?

Tyto chybějící informace vypovídají o tom, že vztah mezi výškou a váhou není příliš silný. Váha tedy nesouvisí jen s výškou, ale mnohem víc nějakými jinými okolnostmi. Mohlo by jít např. o rozložitost postavy, o množství svaloviny, tuků a kosterní hmoty, o to, zda jde o ženy nebo muže aj.

(Složitější statistické metody berou ve výpočtu v potaz více proměnných najednou a zjišťují, kolik informací vysvětluje každá z nich. Budeme se jim trošku věnovat v příštím roce v navazujícím metodologickém semináři, který je nepovinný).

Informace pro uklidnění a povzbuzení:
Výpočty t-testu, směrodatných odchylek, průměrů ani korelací nepočítáme „ručně“

Počítáme je **pomocí Excelu nebo jiného statistického programu** (např. na katedře učitelství je pro naše studenty k dispozici program SPSS).

- Viz pokračování přednášky