PSY117

Statistická analýza dat v psychologii

Přednáška 6 2015

Vlastnosti a využití korelace

Parciální korelace

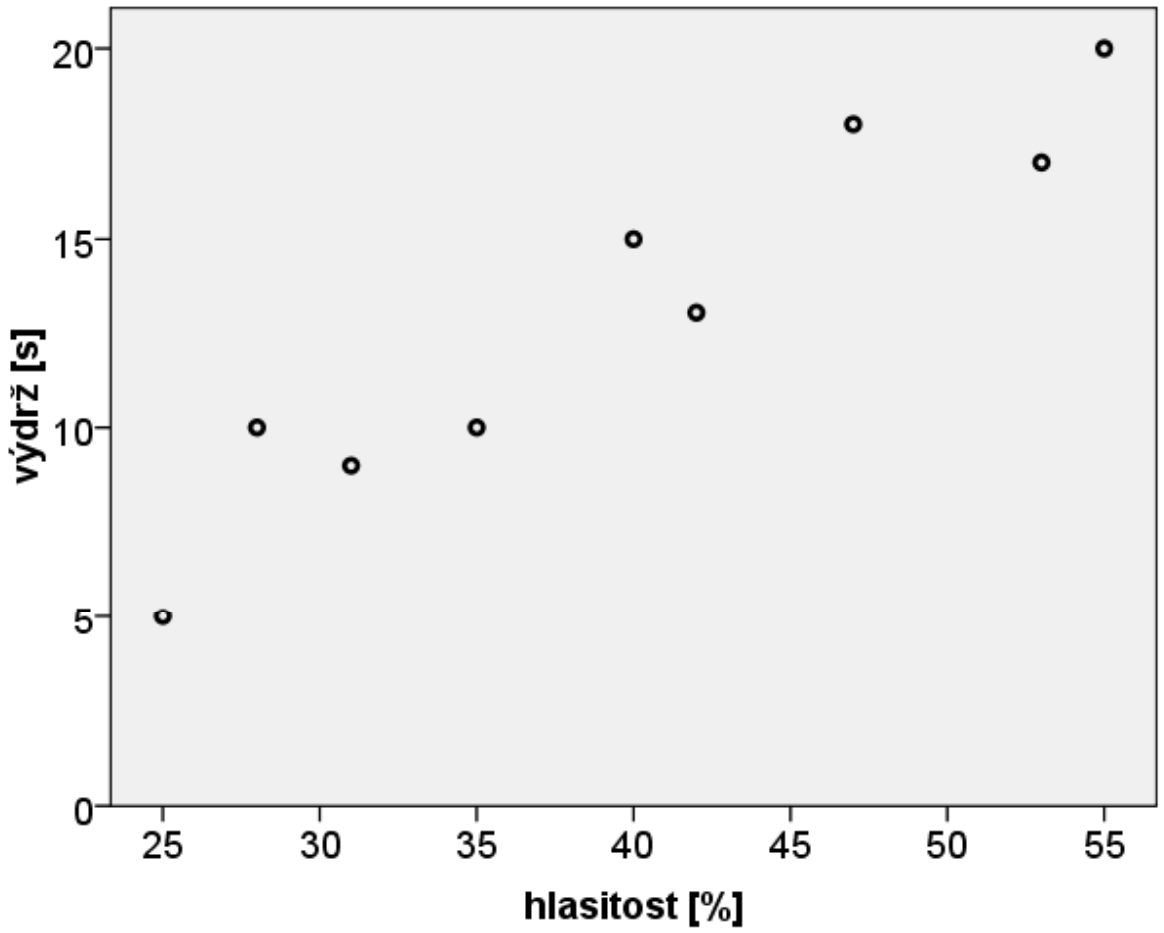
Pořadová korelace a nezávislost

Robustnost a resistence statistik

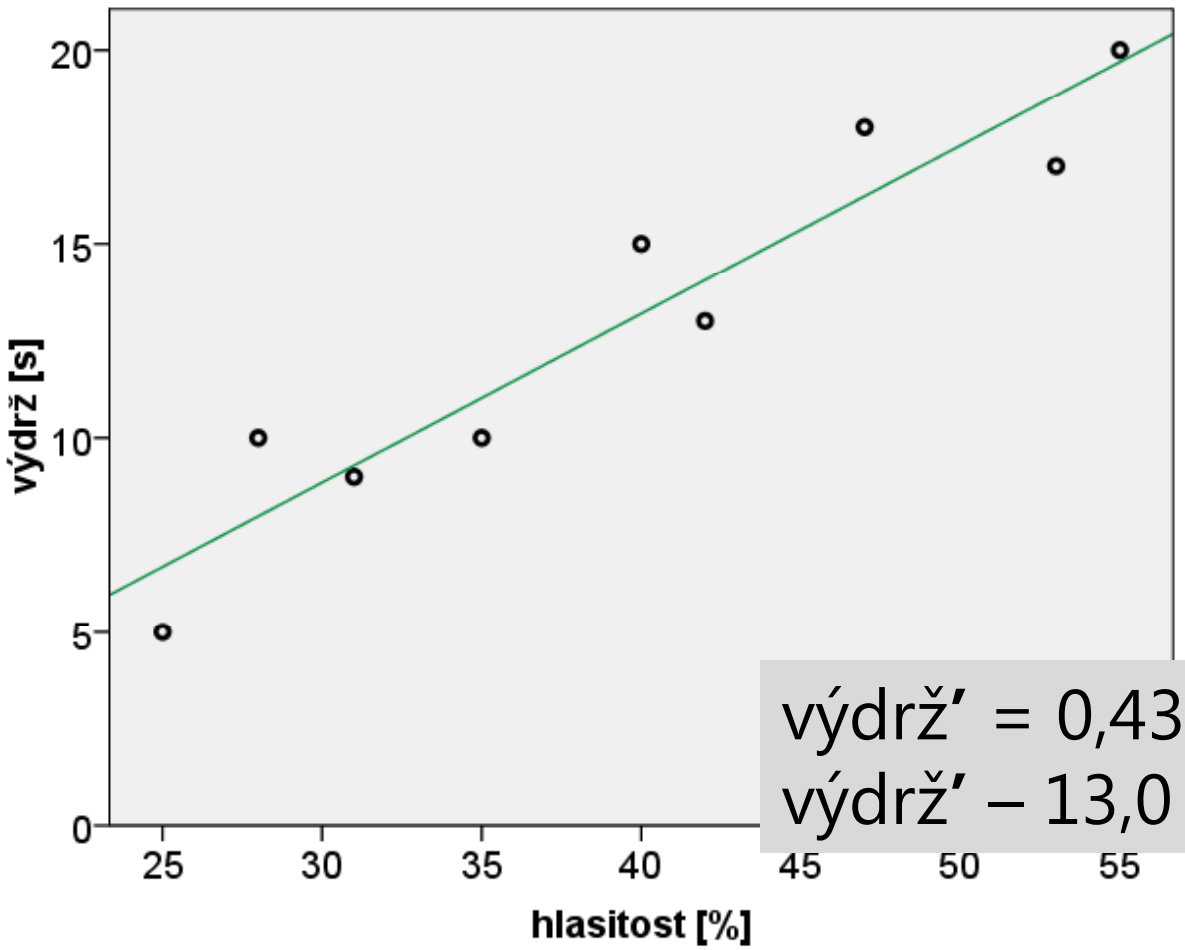
Statistics are like bikinis. What they reveal is suggestive, but what they conceal is vital.

Dlouhodobá adaptace sluchu

Souvisí **hlasitost** poslechu osobního přehrávače *[% z maxima přehrávače]* s **výdrží** snášení nepříjemného hlasitého zvuku?



Dlouhodobá adaptace sluchu



m_h=39,6

s_h = 10,7

m_v=13,0

s_v = 4,9

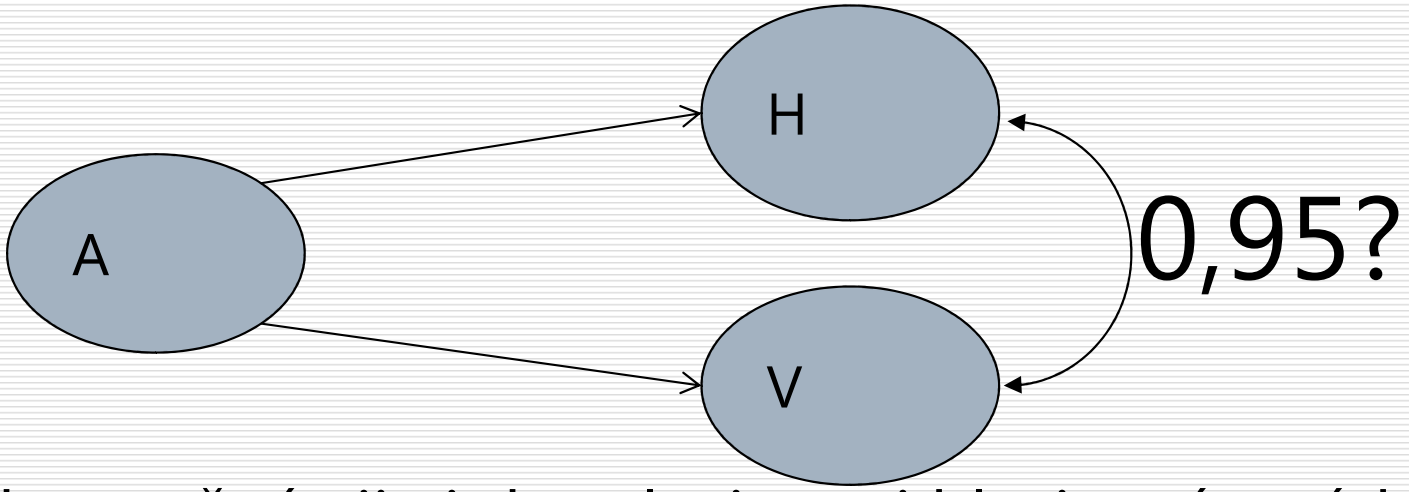
r = 0,95

výdrž' = 0,43.hlasitost – 4,15

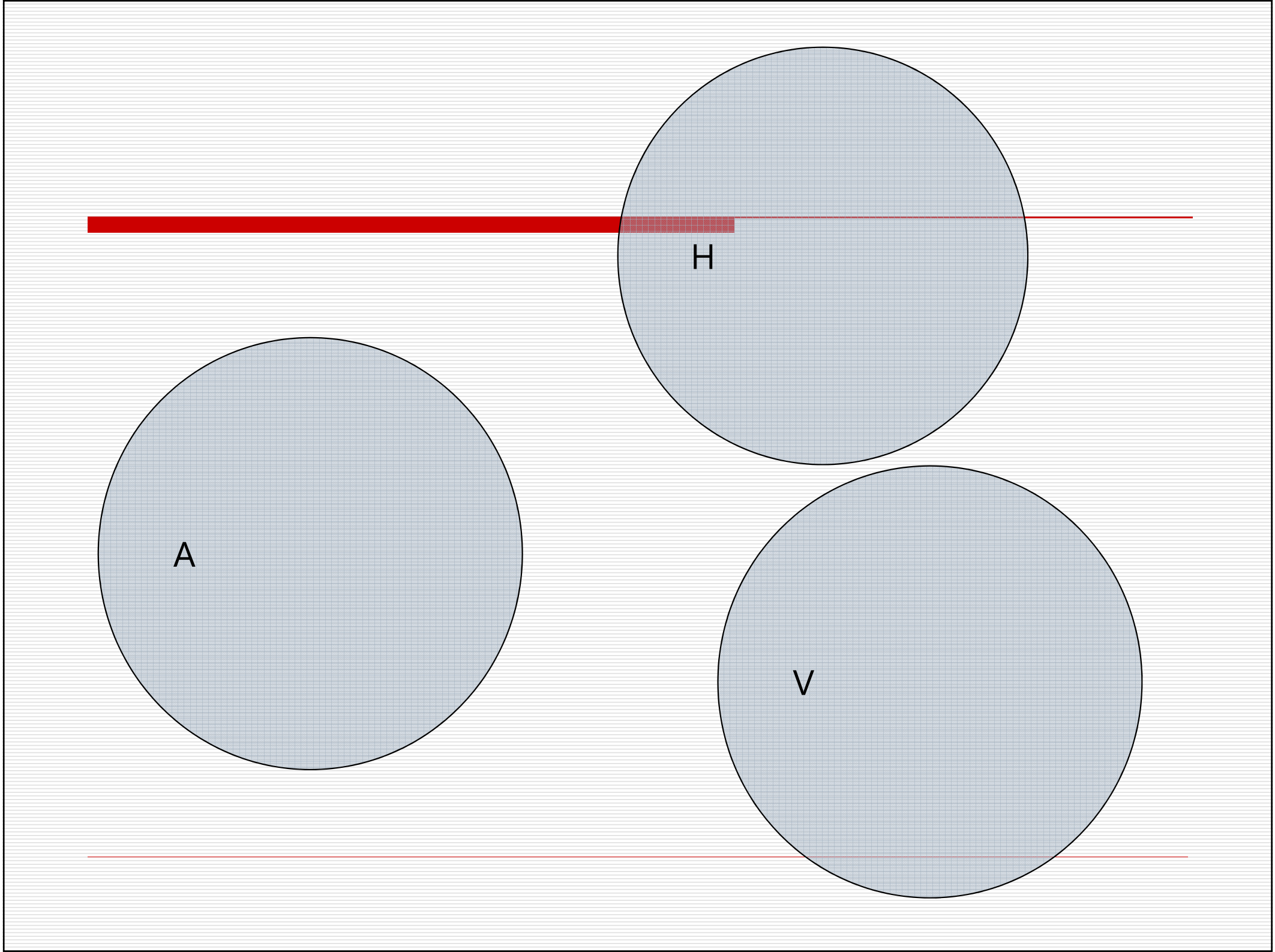
výdrž' – 13,0 = 0,43(hlasitost – 39,6)

Vztah mezi třemi proměnnými
Parciální a semiparciální korelace

Zjistili jsme, že účastníci našeho experimentu se nám opili. To nám vadí, protože opilost snižuje citlivost na podněty a zvyšuje obě naše proměnné.



Bylo by možné zjistit korelaci mezi hlasitostí a výdrží, bez vlivu alkoholu, tj. kdyby nikdo nepil?


H
A
V

Jak ale rozdělovat ty rozptyly?

Regrese dělí proměnnou na sdílený rozptyl a reziduální rozptyl….

Parciální korelace $r_{VH.A}$

- Uděláme regresi výdrže na alkohol – reziduum výdrže bez alkoholu
- Uděláme regresi hlasitosti na alkohol – reziduum hlasitosti bez alkoholu
- Korelace dvou reziduí je PARCIÁLNÍ KORELACE

$$r_{VH.A} = \frac{r_{VH} - r_{VA}r_{HA}}{\sqrt{1-r_{VA}^2}\sqrt{1-r_{HA}^2}}$$

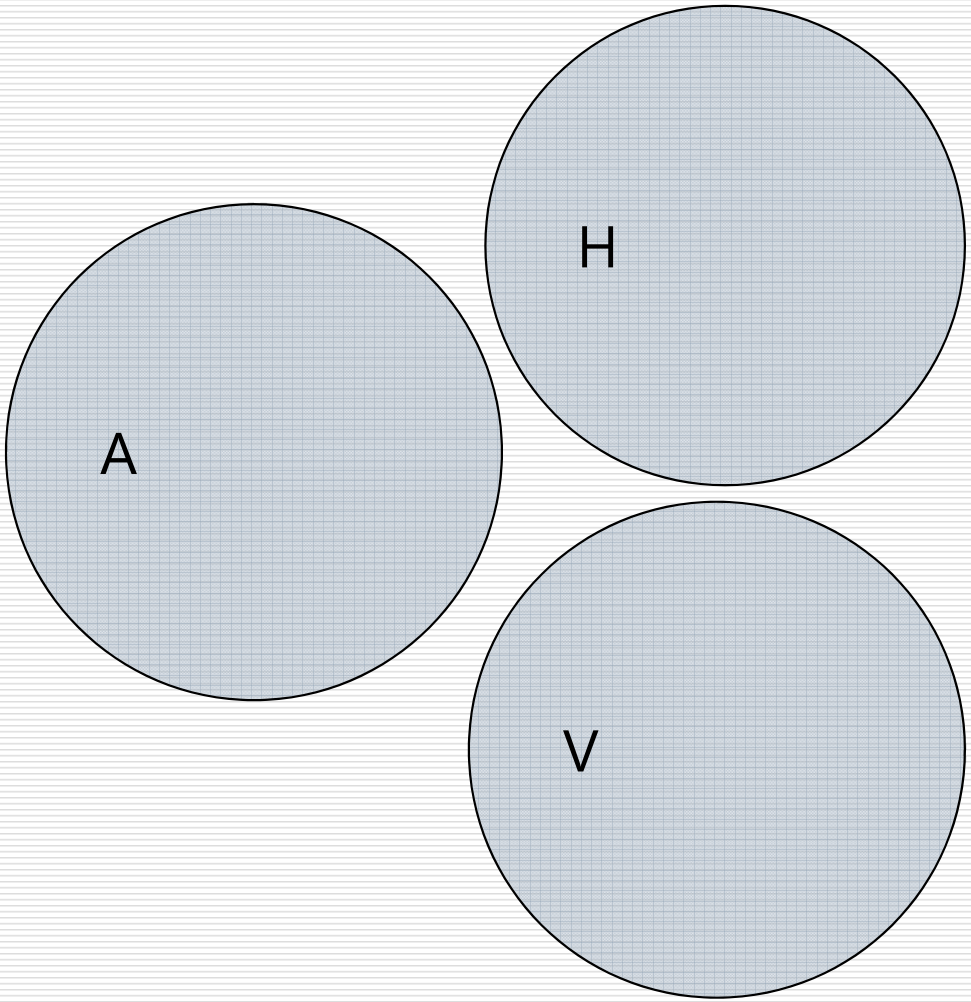
Semiparciální korelace $r_{V(H.A)}$

- Korelace rezidua (H.A) se závislou proměnnou (V)

$$r_{V(H.A)} = \frac{r_{VH} - r_{VA}r_{HA}}{\sqrt{1-r_{HA}^2}}$$

AJ: partial correlation, part (semi-partial) correlation

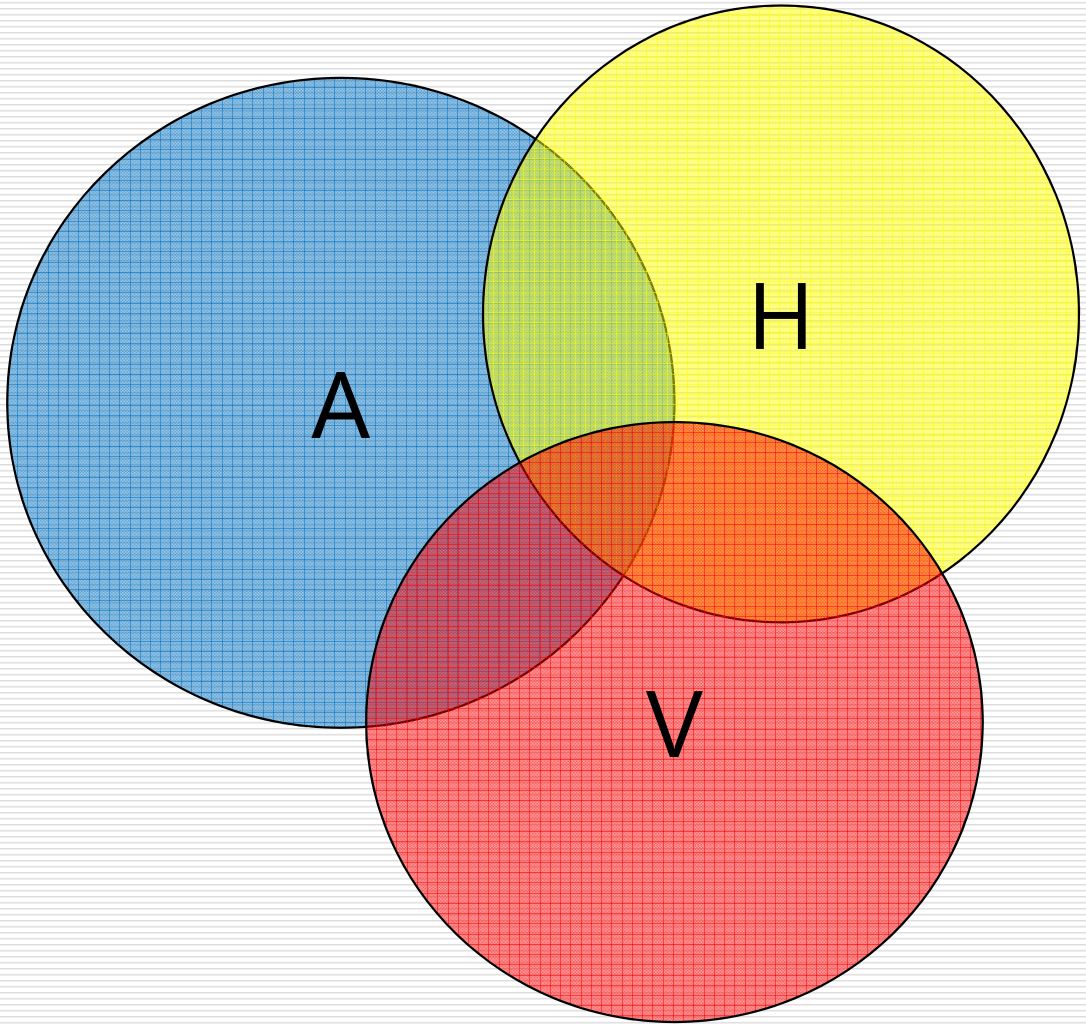
Korelace mezi hlasitostí a výdrží , **kontrolujeme-li statisticky*** alkohol je…



	hlasitost	vydrz	alkohol
hlasitost	1,000	,949	,864
vydrz	,949	1,000	,902
alkohol	,864	,902	1,000

$r_{VH.A} = 0{,}78$

* Též, „pokud by alkohol byl konstantní“



Korelace

(oranžová+hnědá)/(červená+ fialová oranžová+hnědá) ≈ r^2_{VH}

Parciální korelace

$r^2_{VH.A}$ ≈ or/(or+čer)

Semiparciální korelace

$r^2_{V(H.A)}$ ≈ or/(or+čer+fial)

Vždy nás zajímá vysvětlený rozptyl závislé proměnné – zde **Výdrž**

Vlastnosti Pearsonova korelačního koeficientu

- Jde o **momentový** koeficient korelace, a tedy je nutná intervalová a vyšší úroveň měření
- Je vhodný pro popis normálně rozložených proměnných (nebo alespoň **stejně rozložených**)
- Vyjadřuje **sílu(těsnost) lineárního** vztahu, tj. jak moc připomíná tvar scatteru štíhlou elipsu, čáru

Co když tyto podmínky nejsou splněny?

AJ: Pearson's product-moment correlation

Pořadová korelace

- Řeší mnohá omezení Pearsonovy *r*
- Čím víc, tím víc/míň nahrazuje ideou shody **pořadí**

Vysoká pozitivní (negativní) korelace pak znamená:

Má-li jeden člověk v jedné proměnné vyšší hodnotu než druhý člověk (tj. nižší pořadí), pak by i v druhé proměnné měl mít ten první vyšší (nižší) hodnotu než druhý.

Kendallův koeficient pořadové korelace tau

známka M	obvod hlavy	pořadí v M	pořadí v obv. h.	pořadí v M	pořadí v obv. h.	K+, D-
3	**48**	**3**	**3**	**1**	**5**	----
2	**43**	**2**	**2**	**2**	**2**	++-
1	**50**	**1**	**5**	**3**	**3**	+-
4	**49**	**4**	**4**	**4**	**4**	-
5	**40**	**5**	**1**	**5**	**1**	

τ **= (K-D) / [*N* (*N* -1)/2]** = (3-7)/(5.4/2) = -4/10 = -0,4

AJ: Kendall tau (-b,-c), ranks, rank ordering

Kendallův koeficient pořadové korelace tau

- τ = přeškálovaná pravděpodobnost, že dva náhodní lidé budou podle obou proměnných shodně(opačně) seřazeni
- τ <1;-1>
- τ zachycuje i monotonní nelineární vztah
- τ díky pořadovému základu není ovlivněno outliery
- τ kromě pořadové úrovně měření nepředpokládá nic

Modifikace τ_b a τ_c řeší problém shody pořadí (ties).
Podobné: (Goodmanova a Kruskallova) γ a Sommerovo *d*

Spearmanův koeficient pořadové korelace r_s

známka M	obvod hlavy	pořadí v M	pořadí v obv. h.
3	**48**	**3**	**3**
2	**43**	**2**	**2**
1	**50**	**1**	**5**
4	**49**	**4**	**4**
5	**40**	**5**	**1**

r_s = Pearsonova r spočítaná na transformovaných proměnných = -0,6

Spearmanovo r_s (ρ, ró, rho)

r_s – tak na půl cesty mezi r a τ

- Je pořadový a nepředpokládá striktně lineární vztah, ale zohledňuje **velikost odchylek** od ideálního pořadí
- Počítá se jako Pearsonova korelace, ale na pořadích
- Používá se obvykle jako rezistentnější varianta Pearsonovy *r*, která zachytí i monotónní nelineární vztahy.
  - Je-li $r_s > r$, je možné, že vztah není lineární
- Lze interpretovat r_s^2
- *Vychází obvykle numericky vyšší než tau, ovšem to by nikdy nemělo hrát roli ve vašem rozhodování. V obou případech jde o jiný typ vztahu.*

AJ: Spearman (rank-order correlation) rho,

Vztahy na nominální úrovni

- =rozdíly řádkových/sloupcových relativních četností v kontingenční tabulce
- =rozdíly pravděpodobností/šancí – poměry šancí, poměry rizik
- Lze vyjádřit jako *korelační koeficienty* založené na hodnotě χ^2
  - Kvůli neexistenci směru mají koeficienty rozsah od 0 (žádný vztah) do 1 (maximálně těsný vztah)
- Větší množství koeficientů se specializovaným užitím
  - Pearsonův kontingenční koeficient
  - Cramerovo *V*
  - r_ϕ – koeficient fí (phi)

Těmto vztahům se budeme věnovat později.

AJ: odds ratio, risk ratio

		A kterou z následujících, ne až tak nadpřirozených schopností byste nejvíc chtěli?				
		1 Neomezeně jíst, aniž by mi hrozilo přibírání na váze	2 Eidetickou paměť	3 Nikdy se neunavit	4 Být všemi oblíben(a)	Celkem
1 muž	n	0	8	13	2	23
	% z pohlavi	0,0%	34,8%	56,5%	8,7%	100,0%
2 žena	n	10	6	10	4	30
	% z pohlavi	33,3%	20,0%	33,3%	13,3%	100,0%
Celkem	n	10	14	23	6	53
	% z pohlavi	18,9%	26,4%	43,4%	11,3%	100,0%

Př. poměr šancí OR: O(paměť|muž)/O(paměť|žena)=(8/15)/(6/24)=2,1
Je asi 2násobná šance volby *paměť* u mužů oproti ženám

Konstrukce psychologických škál

Need for structure = součet 10 položek

	M	SD
Žít dobře uspořádaný život s pravidelným denním rozvrhem mi prostě sedne.	2,96	0,98
Stanovit si pevný režim mi pomáhá více si užívat života.	3,15	1,03
Líbí se mi jasný a uspořádaný způsob života.	3,11	1,09
Nerad(a) se vystavuji situacím, o nichž dopředu nevím, co mohu očekávat.	3,19	1,11
Nerad(a) trávím čas ve společnosti lidí, kteří jsou schopni jednat nepředvídatelně.	2,44	0,97
Nemám rád(a) nepředvídatelné situace.	2,59	1,08
Obyčejně se mi uleví, jakmile se pro něco rozhodnu.	3,67	1,00
Nemám rád(a) nejisté situace.	3,15	0,99
Je mi nepříjemné, když nechápu důvod nějaké události, která se mi přihodila.	3,70	0,95
Nemám rád(a), když něčí výrok může znamenat spoustu různých věcí.	3,19	1,18

Využití korelací v konstrukci psychologických testů - reliabilita

- Položky lze sčítat, pokud spolu korelují.
- Položky korelují, existuje-li společný důvod pro určitý způsob odpovídání na ně – měřená charakteristika.

Jak moc spolu musí korelovat?

$$r_{tt} = \frac{k r_M}{1 + (k-1) r_M} \qquad r_{tt} = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} s_i^2}{s_t^2}\right)$$

r_{tt} je vnitřní konzistence, r_M je průměrná korelace mezi položkami, k je počet položek

- při 10 položkách stačí průměrná korelace 0,2

Vnitřní konzistence – **Cronbachovo** α – horní mez reliability

- minimálně 0,7 pro výzkum, 0,9 pro diagnostiku

zpět k *NfS*

	p1	p2	p3	p4	p5	p6	p7	p8	p9
p2	0,73								
p3	0,73	0,81							
p4	0,25	0,41	0,46						
p5	0,42	0,47	0,35	0,35					
p6	0,38	0,51	0,63	0,51	0,47				
p7	0,26	0,31	0,32	0,20	0,16	0,22			
p8	0,32	0,39	0,38	0,64	0,25	0,38	0,17		
p9	0,40	0,32	0,29	-0,02	0,11	0,10	0,42	0,17	
p10	0,47	0,20	0,37	0,06	0,26	0,27	0,22	-0,02	0,39

Průměrná korelace r_M= 0,34

Cronbachova alfa r_{tt} = 0,84

Jaké statistiky už známe

Četnosti

Popisné statistiky jedné proměnné

- momentové: *M*, *SD*, s^2
- pořadové: *min*, *max*, *Md*, Q_1, Q_3, *IQR*, percentily
- kategorické: *Mo*

Ukazatele vztahu mezi dvěma proměnnými

- momentové: Pearsonova *r*, *b*
- pořadové: Kendallovo τ, Spearmanova r_s
- kategorické: r_ϕ, Cramerovo *V*

S jakými předpoklady je spojeno použití těchto statistik?

Co se stane, když nejsou tyto předpoklady splněny?

Předpoklady statistik

Jejich splnění podmiňuje

- matematickou **smysluplnost** výpočtu
  - typicky úroveň měření
- **přesnost**, výpovědní schopnost vypočítané hodnoty
  - typicky tvar rozložení

Při splnění všech předpokladů nese vypočítaná statistika tu informaci, kterou od nich v souladu se statistickou teorií očekáváme.

Statistiky, jejichž smysl není porušením předpokladů příliš ovlivněný, jsou **ROBUSTNÍ**.

- používáme i pro statistiky s minimálními či žádnými předpoklady.

AJ: assumptions, robust

Co ještě omezuje výpovědní schopnost statistik?

- Odlehlé, extrémní hodnoty
  - Není-li statistika příliš ovlivněna výskytem extrémních hodnot, je **REZISTENTNÍ**
  - Resistenci momentových statistik někdy zvyšujeme ořezáváním extrémů, např. trimmed mean
- Efekt podlahy a stropu
  - snižuje ukazatele variability
  - posunuje ukazatele centrální tendence
  - snižuje korelaci
  - … a nic moc s tím nenaděláme, to je věc metodologie