1838
УНИВЕРЗИТЕТ У БЕОГРАДУ
ФИЛОЗОФСКИ ФАКУЛТЕТ

Статистика у психологији 1

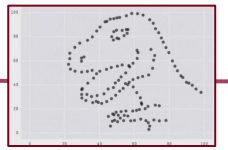
Статистика у истраживању образовања

доц. др Анђела Шошкић, 12.11.2024.

Статистички опис и приказ биваријационих података

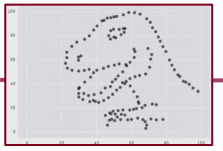
Подсетник пред почетак

- Да ли сте радили из методологије:
  - Како се зову нацрти у којима су све варијабле нумеричке?
  - Све категоричке?
  - А независне категоричке, зависне нумеричке?

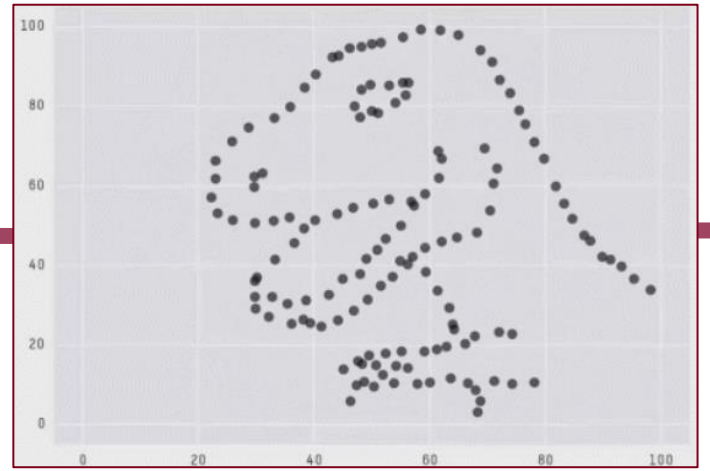


Теме за данас

1. Биваријациони подаци
2. Статистичка независност и повезаност две варијабле
3. Однос две категоричке варијабле
4. Однос две нумеричке варијабле
5. Однос једне бинарне категоричке и једне нумеричке варијабле
6. Тумачење биваријационих односа – опрез!
7. Резиме

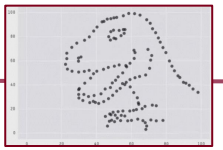


1. Биваријациони подаци



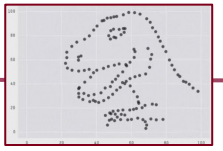
Биваријациони подаци

- **Универијациони** подаци: подаци о једној варијабли
- **Мултиваријациони**: подаци о две или више варијабли прикупљени код истих испитаника (и других ентитета)
  - **Биваријациони** подаци: две варијабле
- Да се не збуните:
  - Негде и уни/би/мулти**варијатни**
  - Исти називи некад се користе и да се означи број *зависних* варијабли, а не свих



Биваријациона расподела

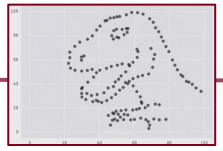
- **Униваријациона** расподела – расподела једне варијабле
- **Биваријациона (заједничка)** расподела – истовремено распоређивање по обема варијаблама
  - Парови вредности, а не појединачне вредности
  - Свака варијабла има и своју униваријациону расподелу, која се у овом контексту зове још и **маргинална**



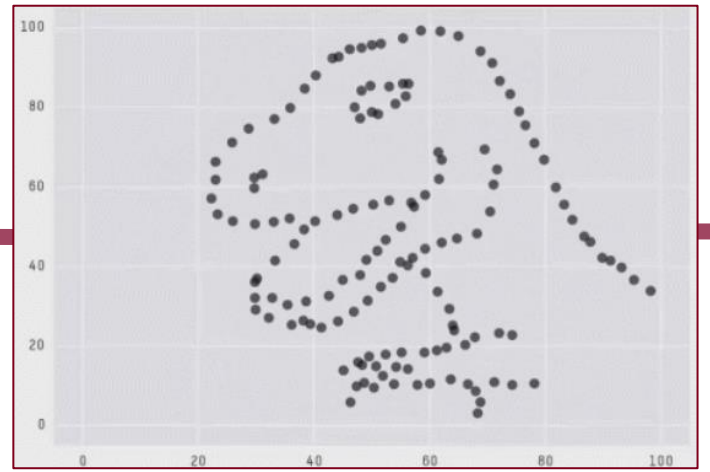
Табеларни приказ биваријационе расподеле

X→ Y ↓	1	2	3	4	5	6	7	8	9	10	f_Y
1		2									2
2	1										1
3			2								2
4		1	1		2						4
5					2	2	1				5
6					1						1
7					1						1
8										1	1
9								2			2
10									1		1
f_X	1	3	3	0	6	2	1	2	1	1	n = 20

- Заједничке фреквенце у ћелијама
- Маргиналне фреквенце у последњем реду и колони



2. Статистичка независност и повезаност две варијабле

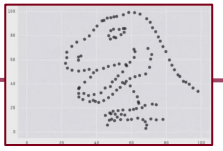


Статистичка повезаност

- **Постоји правилност** у начину на који се упарују вредности две (или више) варијабли
- За нумеричке варијабле: **корелација**
  - Нпр. већа вредност на једној варијабли чешће се јавља са већим вредностима на другој варијабли (или мањим!)
  - У времену: нпр. повећање једне вредности доводи до повећања/смањења друге
- За категоричке варијабле: **асоцијација**
  - Припадност некој групи на једној варијабли чешће је праћена припадношћу неким категоријама друге варијабле
  - У времену: промена припадности групи на једној варијабли чешће ће довести до промене групне припадности на другој варијабли

Статистичка независност

- **Не постоји правилност** у начину на који се упарују вредности две (или више) варијабли
  - За сваки ниво или вредност једне варијабле, иста је расподела вредности друге варијабле
    - На сваком нивоу једне варијабле добија се маргинална, униваријациона расподела друге варијабле



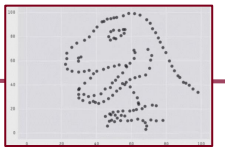
Математички гледано

X→ Y ↓	1	2	3	4	5	6	7	8	9	10	f_Y
1		2									2
2	1										1
3			2								2
4		1	1		2						4
5					2	2	1				5
6					1						1
7					1						1
8										1	1
9								2			2
10									1		1
f_X	1	3	3	0	6	2	1	2	1	1	n = 20

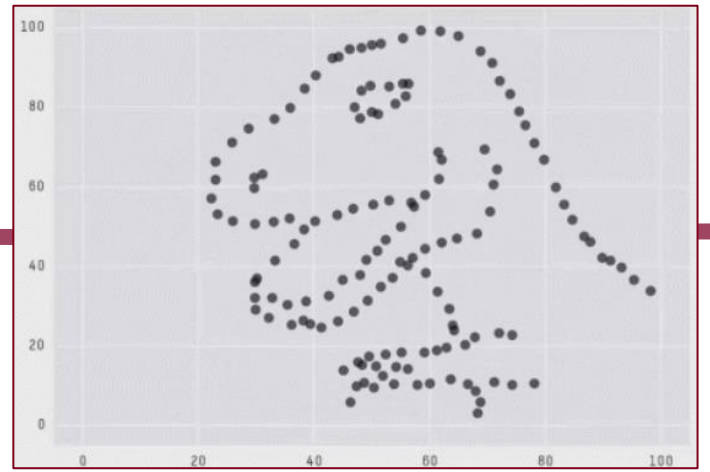
- **p(x, y) = p(x)p(y)** за сваку ћелију

Статистичка повезаност

- Постојање повезаности **не значи** да нешто важи за већину испитаника из неке групе или за већину испитаника са изнад/исподпросечним вредностима на нумеричкој варијабли!
  - Пример: избор играчака код дечака и девојчица



3. Однос две категоричке варијабле



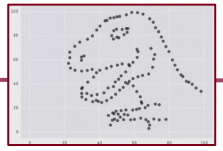
Статистичка повезаност и независност

	мушко	женско	**укупно**
несреће	30	20	**50**
без несреће	42	28	**70**
укупно	**72**	**48**	**120**

	мушко	женско	**укупно**
несреће	42%	42%	**42%**
без несреће	58%	58%	**58%**
укупно	**100%**	**100%**	**100%**

	мушко	женско	**укупно**
несреће	23	27	**50**
без несреће	49	21	**70**
укупно	**72**	**48**	**120**

	мушко	женско	**укупно**
несреће	32%	56%	**42%**
без несреће	68%	44%	**58%**
укупно	**100%**	**100%**	**100%**



Статистичка повезаност и независност

- **Асоцијација: постоји правилност** у начину на који се упарују категорије две варијабле
  - Припадност некој групи на једној варијабли чешће је праћена припадношћу неким категоријама друге варијабле *него што је то случај са осталим групама прве варијабле*
- **Независност:**
  - За сваки ниво једне варијабле, иста је расподела по категоријама за другу варијаблу
  - Припадност свакој групи једне варијабле праћена је истом учестаношћу по категоријама друге варијабле

	мушко	женско	**укупно**
несреће	32%	56%	**42%**
без несреће	68%	44%	**58%**
укупно	**100%**	**100%**	**100%**

	мушко	женско	**укупно**
несреће	42%	42%	**42%**
без несреће	58%	58%	**58%**
укупно	**100%**	**100%**	**100%**

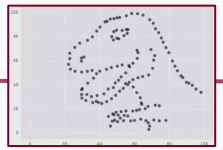
Статистичка независност

	мушко	женско	**укупно**
несреће	30	20	**50**
без несреће	42	28	**70**
укупно	**72**	**48**	**120**

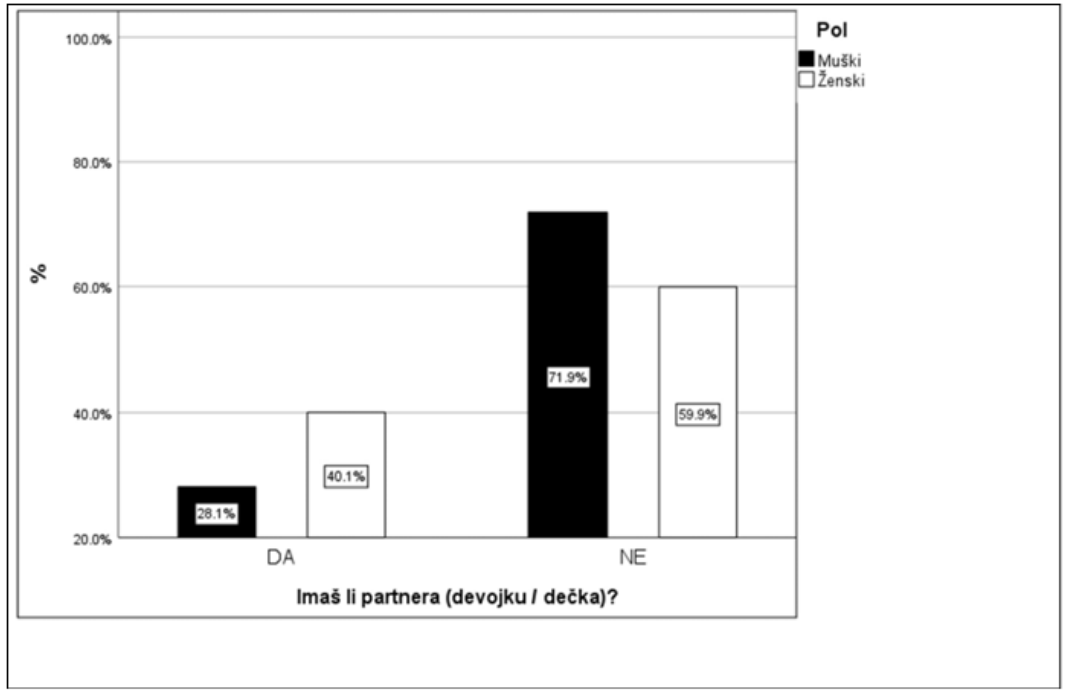
	мушко	женско	**укупно**
несреће	42%	42%	**42%**
без несреће	58%	58%	**58%**
укупно	**100%**	**100%**	**100%**

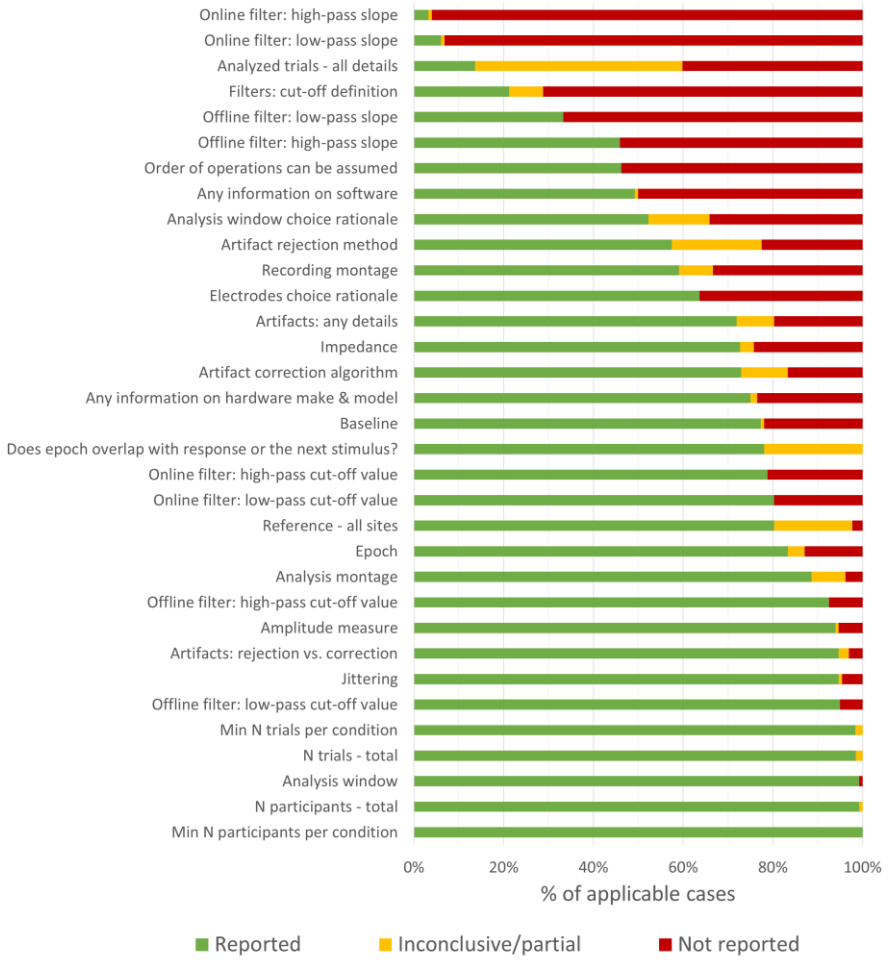
$$30 = \frac{50}{120} * \frac{72}{120} = \frac{50 * 120}{120} = \frac{3600}{120}$$

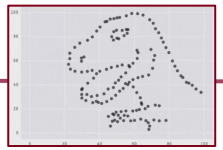
Рекли смо малочас: **p(x, y) = p(x)p(y)** за сваку ћелију



Графички приказ асоцијације две категоричке варијабле







Мере асоцијације две категоричке варијабле

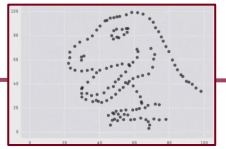
- Количник шанси

$$OR = \frac{\frac{f_{11}}{f_{12}}}{\frac{f_{21}}{f_{22}}} = \frac{f_{11} * f_{22}}{f_{12} * f_{21}}$$

- Односи заједничких фреквенци
- Две дихотомне варијабле
  - Ако је више категорија – односи шанси за појединачне категорије и једну референтну категорију, има их више
- Други назив: количник унакрсних производа

- Тумачење: поређење са 1
- Мора се логаритмовати да би могле да се пореде асоцијације различитог смера

	мушко	женско	**укупно**
несреће	30, f11	20, f12	**50**
без несрећа	42, f21	28, f22	**70**
укупно	**72**	**48**	**120**

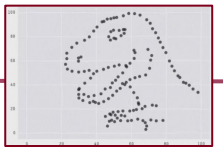


Мере асоцијације две категоричке варијабле

- Фи-коефицијент

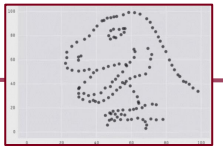
$$\Phi = \frac{f_{11}f_{22} - f_{12}f_{21}}{\sqrt{(f_{11} + f_{12})(f_{11} + f_{21})(f_{12} + f_{22})(f_{21} + f_{22})}}$$

  - Две дихотомне варијабле
  - Распон 0-1
  - Слична логика као количник шанси, али не гледамо *колико пута* се разликују шансе, него *за колико* (одузимање, а не множење)



Мере асоцијације две категоричке варијабле

- Крамеров V коефицијент
  - „Проширење“ фи-коефицијента за случајеве када једна или више варијабли нису дихотомне
  - Формула се изводи из хи-квадрат теста за закључивање о асоцијацији у популацији на основу узорка (имате је у тексту за студенте на Мудлу)
  - Згодан јер:
    - Распон 0-1
    - Упоредив када су табеле различитих димензија
  - Незгодан за: закључивање о популацијским вредностима (пристрасан)
    - Више речи о томе касније



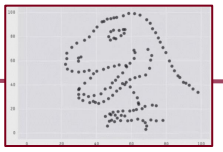
Мере асоцијације две категоричке варијабле

- Коефицијент контигенције

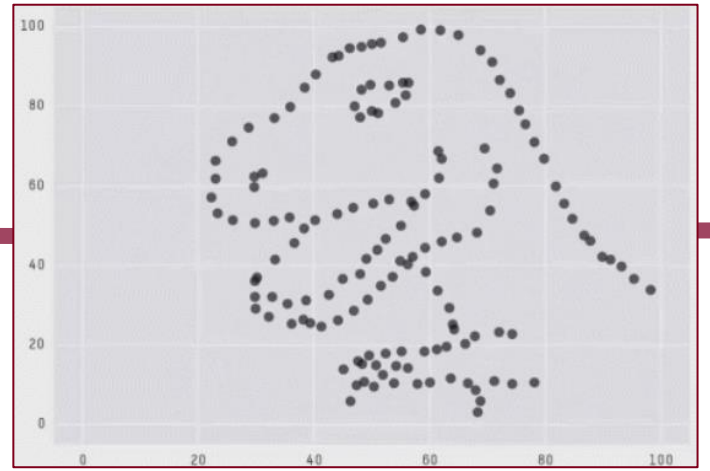
$$C_{xy} = \sqrt{\frac{(\sum_{i=1}^{n} \frac{f_o^2}{f_t}) - n}{\sum_{i=1}^{n} \frac{f_o^2}{f_t}}}$$

- n – број испитаника
- f_o – опажене фреквенце
- f_t – теоријске фреквенце (кад не би било асоцијације)

- 0 до максималне вредности која зависи од димензија табеле, не већи од 1
  - Не могу се поредити C вредности за табеле различитих димензија



4. Однос две нумеричке варијабле

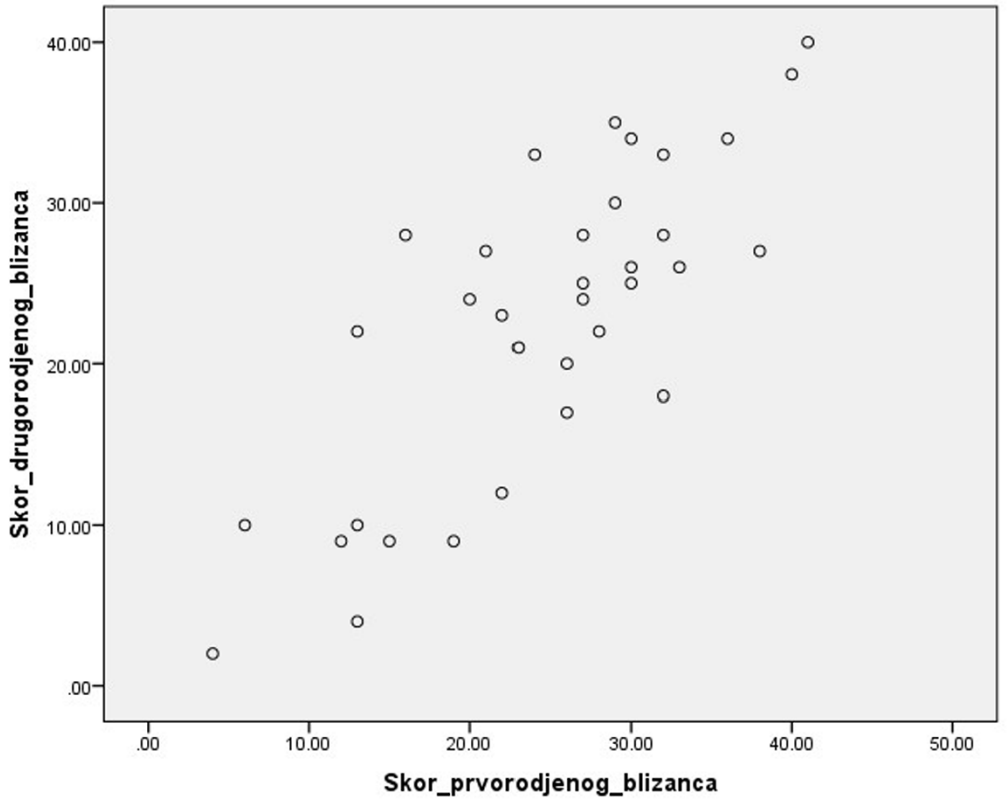


Шта рекосмо да је корелација?

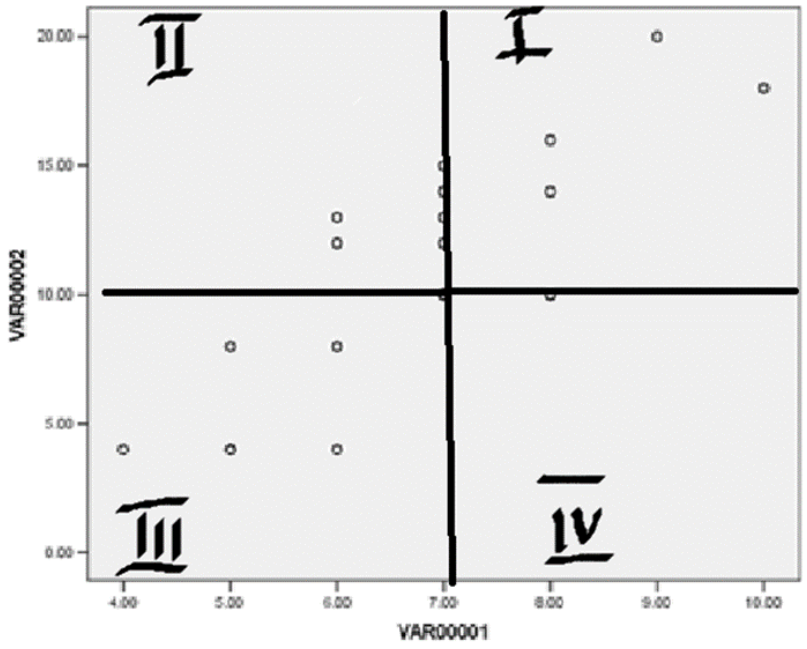
- **Бинарна корелација:** Постоји правилност у начину на који се упарују вредности две нумеричке варијабле
  - Најједноставнији случај: већа вредност на једној варијабли чешће се јавља са већим вредностима на другој варијабли (или мањим!)

Корелација у табели фреквенција и на дијаграму распршења

X→ Y ↓	1	2	3	4	5	6	7	8	9	10	f_Y
1		2									2
2	1										1
3			2								2
4		1	1		2						4
5					2	2	1				5
6					1						1
7					1						1
8										1	1
9								2			2
10									1		1
f_X	1	3	3	0	6	2	1	2	1	1	n = 20

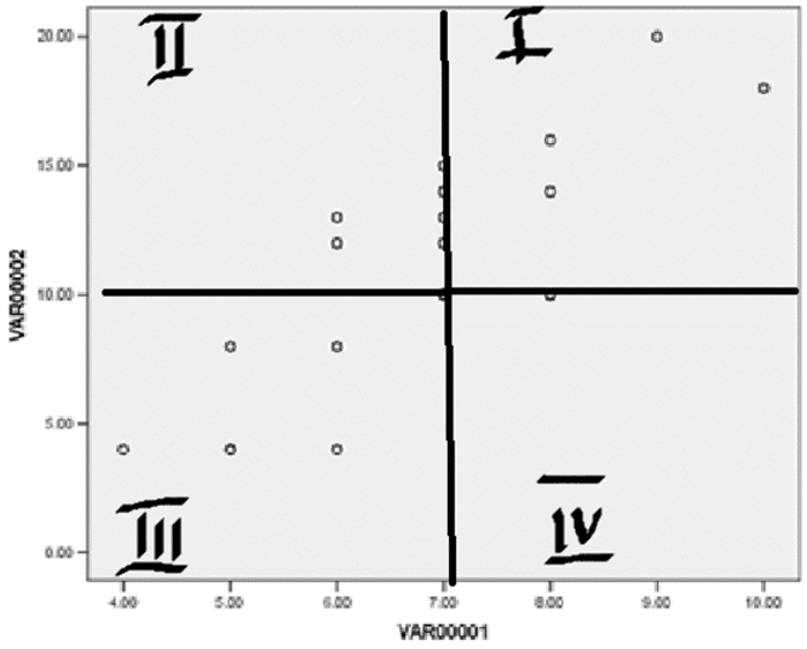


Ковариjанса



- варијанса $SD_x = \frac{\sum_{i=1}^{n}(x_i - M_x)^2}{n}$
- **ко**варијанса $C_{xy} = \frac{\sum_{i=1}^{n}(x_i - M_x)(y_i - M_y)}{n}$

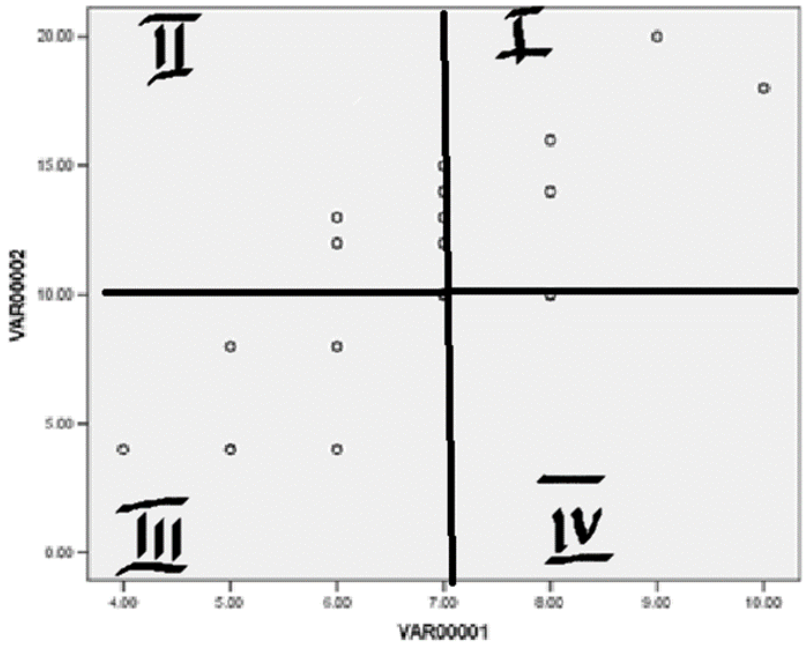
Линеарна корелација



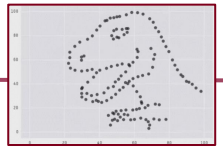
- ковариjанса $C_{xy} = \frac{\sum_{i=1}^{n}(x_i - M_x)(y_i - M_y)}{n}$

x_i	$x_i - X$	y_i	$y_i - Y$	$(x_i - X)(y_i - Y)$
2	-3.17	10	5.00	-15.85
8	2.83	11	6.00	16.98
9	3.83	7	2.00	7.66
10	4.83	1	-4.00	-19.32
1	-4.17	0	-5.00	20.85
1	-4.17	1	-4.00	16.68
Mx = 5.17		My = 5		$\sum (x_i - X)(y_i - Y) =$ **27.00**

Коваријанса



- коваријанса
  - последица формуле: мери линеарни однос
  - распон: -∞ до +∞
  - 0 = нема линеарне повезаности
  - величина зависи од мерне јединице



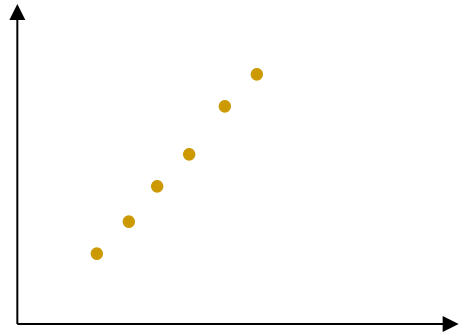
(Браве)-Пирсонов коефицијент линеарне корелације

- Стандардизована коваријанса

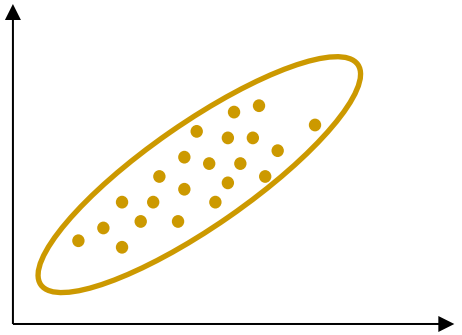
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - M_x)(y_i - M_y)}{nSD_xSD_y} = \frac{\sum_{i=1}^{n} z_x z_y}{n}$$

- -1 до 1, 0 = нема корелације
- Исти предзнак као C

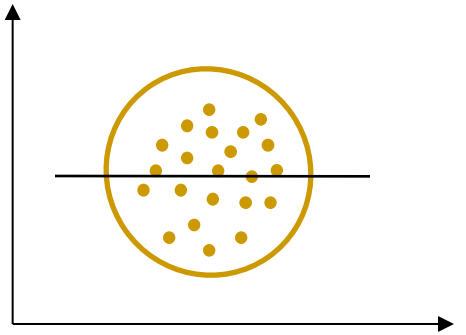
Интензитет и смер корелације и Браве-Пирсонов коефицијент



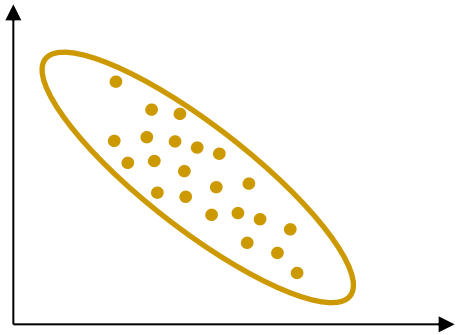
r=1



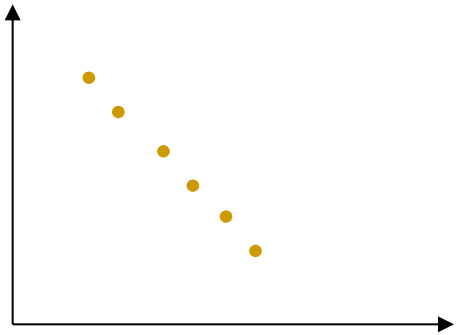
r>0



r=0



r<0

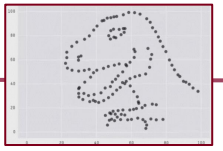


r=-1

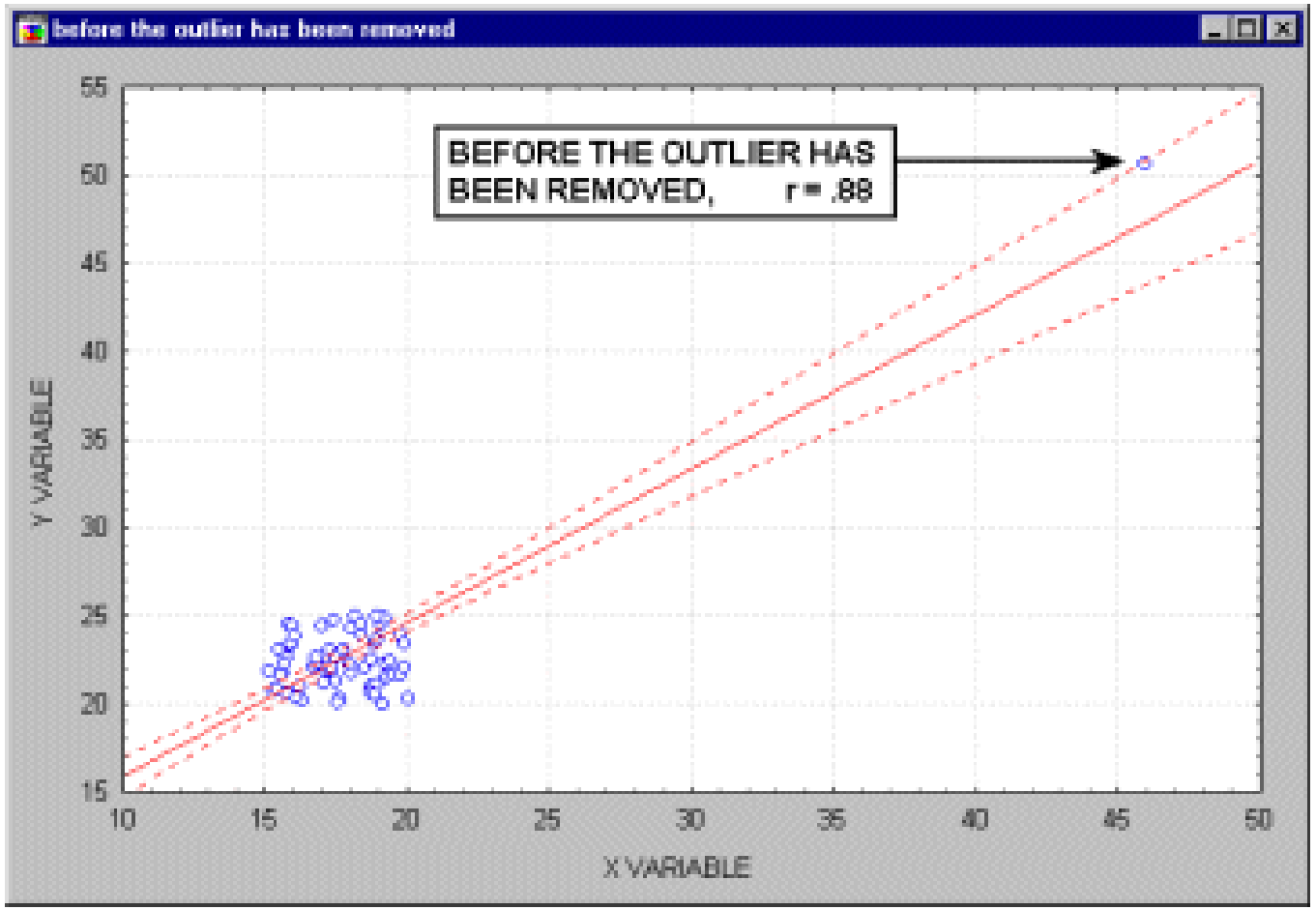
- Позитивна корелација
- Негативна корелација
- Нема корелације

Тумачење Браве-Пирсоновог коефицијента

- Предложени различити критеријуми
- Један пример:
  - 0.00 do ±0.30 ⇨ ниска повезаност
  - ±0.30 do ±0.50 ⇨ средња
  - ±0.50 do ±1.00 ⇨ висока
- Поћи од тога колике корелације се иначе срећу у тој области и шта други истраживачи узимају као критеријум

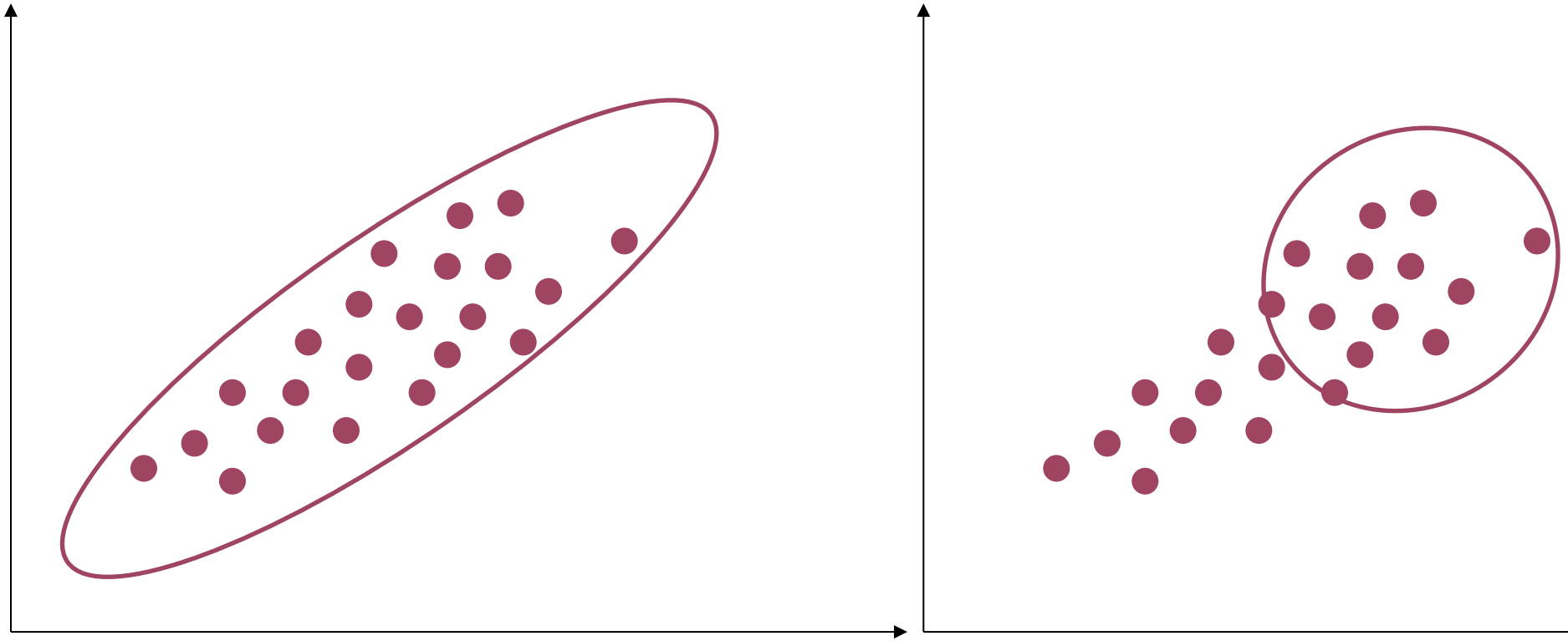


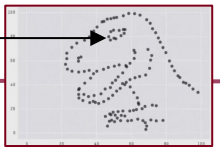
Браве-Пирсонов коефицијент је веома осетљив на изнимке!



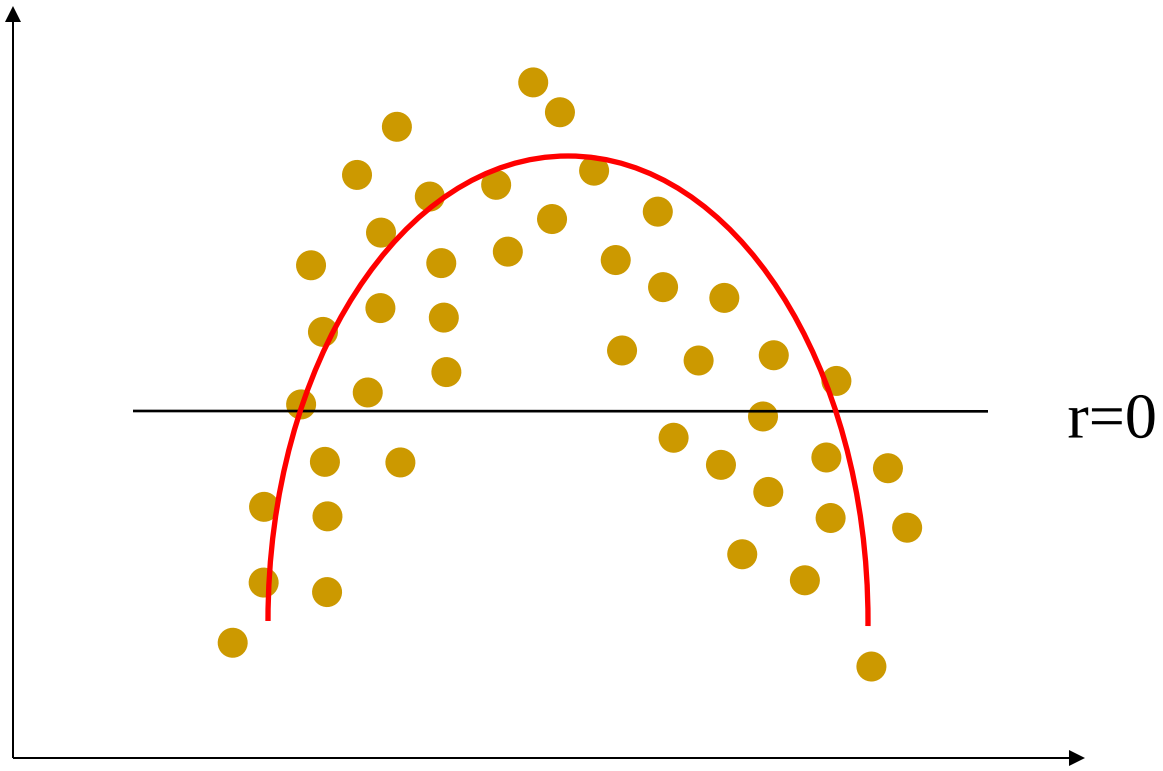
Линеарна корелација и рестрикција распона

- r је смањен у случају рестрикције распона (нпр. ако постоји председелекција на некој или обе варијабле

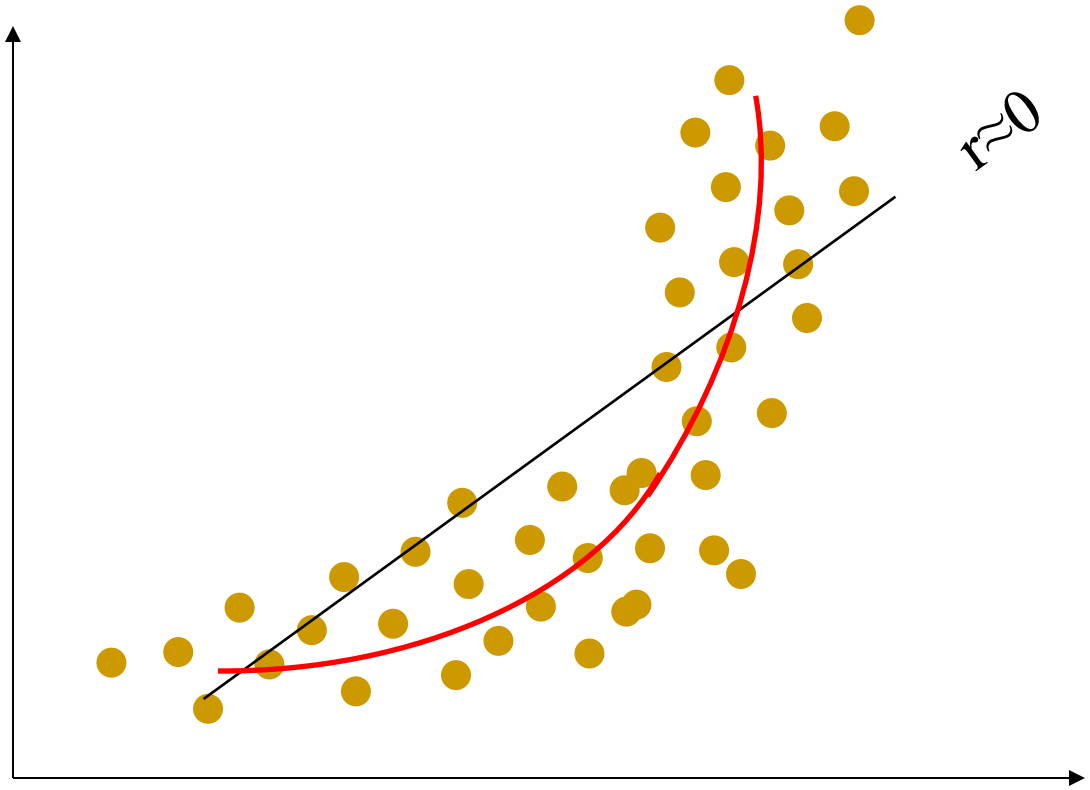




Нелинеарна корелација

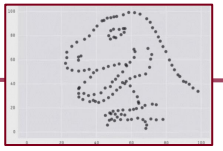


Нелинеарна корелација



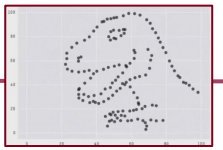
Нелинеарна корелација

- Шта ако имамо нелинеарни однос?
  1. трансформишемо податке на једној или обе варијабле нелинеарним трансформацијама
     - Проблем – то постаје незгодно за интерпретацију
  2. нелинеарни модели код којих се одступање гледа не у односу на праву, него у односу на кривуље различитих облика
  3. изделити цео распон варијабли на неколико сегмената, па сваки гледати посебно

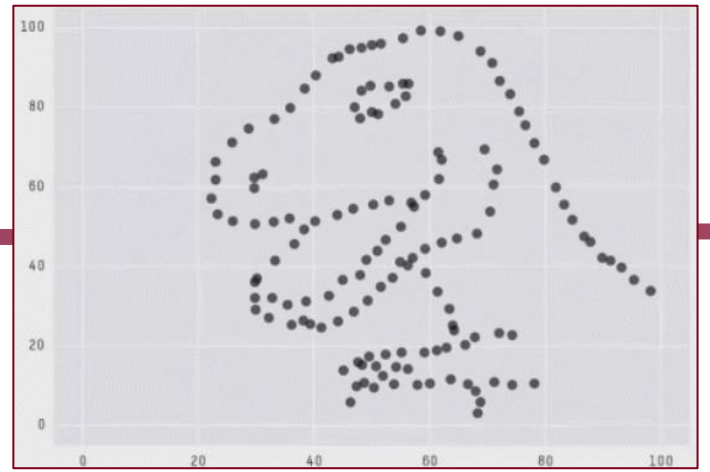


Још неке напомене о Браве-Пирсоновом коефицијенту линеарне корелације

- Не мења се ако линеарно трансформишемо варијабле без промене предзнака
  - множење или дељење позитивним бројем, сабирање и одузимање
  - нпр. промена мерне јединице за већину физичких мера
- **Услови за примену:**
  - Обе варијабле, су, бар теоријски, континуиране
  - Однос је линеаран
  - Парови података су статистички независни (један испитаник/ентитет даје само један пар мера)
  - Заједничка дистрибуција две варијабле је биваријациона нормална расподела
    - На неке врсте одступања од ове расподеле није много осетљив (робустан је), али на друге није
    - Ако то није испуњено? Корелације рангованих података (отом потом)



5. Однос бинарне и нумеричке варијабле



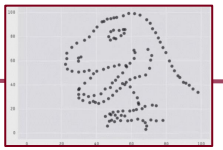
Повезаност нумеричке и бинарне варијабле

- Поинт-биserијска корелација

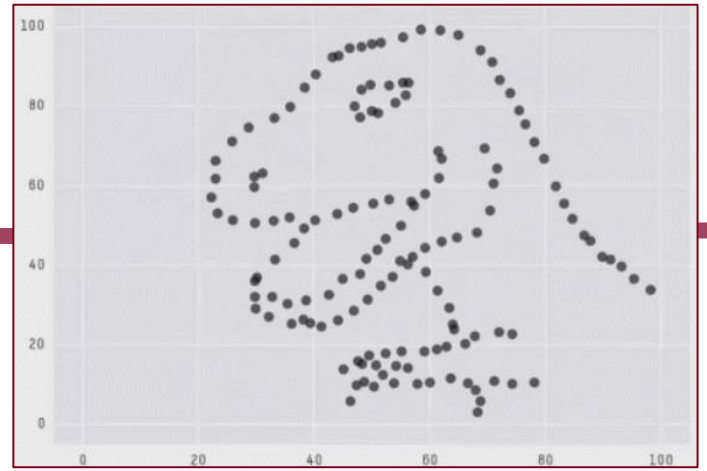
$$r_{pb} = \frac{M_0 - M_1}{s_y}\sqrt{\frac{n_0}{n}\frac{n_1}{n}}$$

  - M_0, M_1 – просеци група
  - Sy – стандардна девијација за све испитанике, обе групе заједно
  - n_0, n_1 – величине група
  - n – број испитаника у обе групе заједно

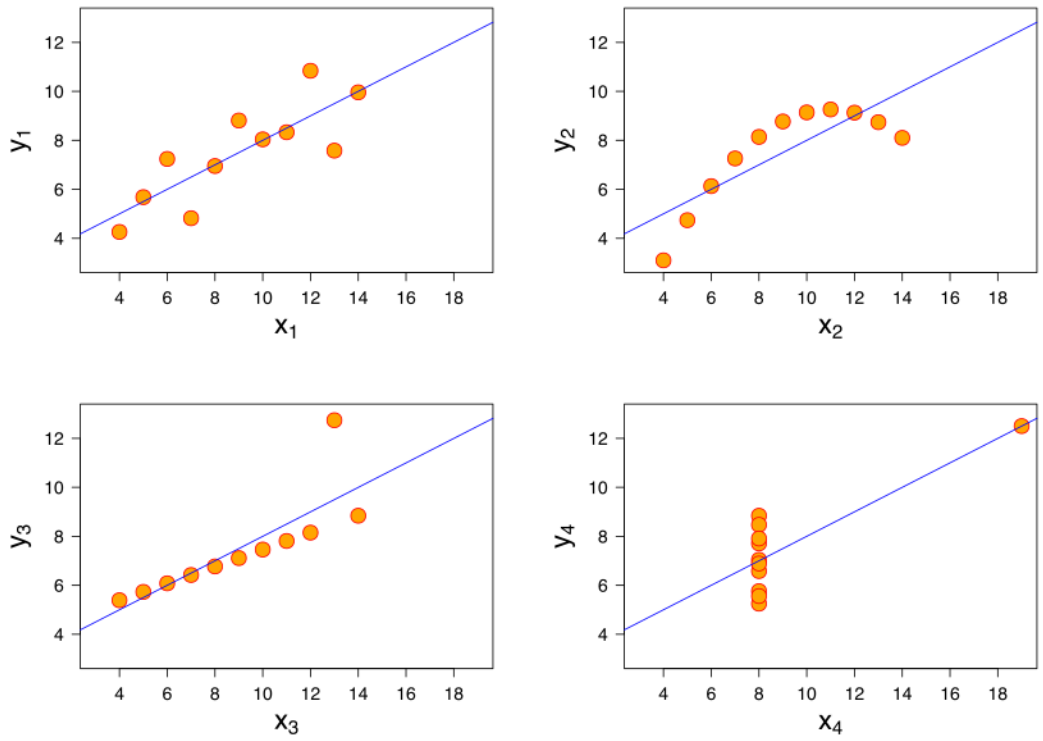
- Посебан случај Пирсонове корелације када је једна варијабла дихотомна
- Од -1 до 1, предзнак говори о томе која група има већи просек на нумеричкој варијабли
- Опрез: важно да величина група буде приближно једнака, у противном максимална вредност коју достиже умањена



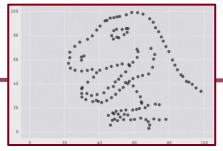
6. Тумачење биваријационих података – опрез!



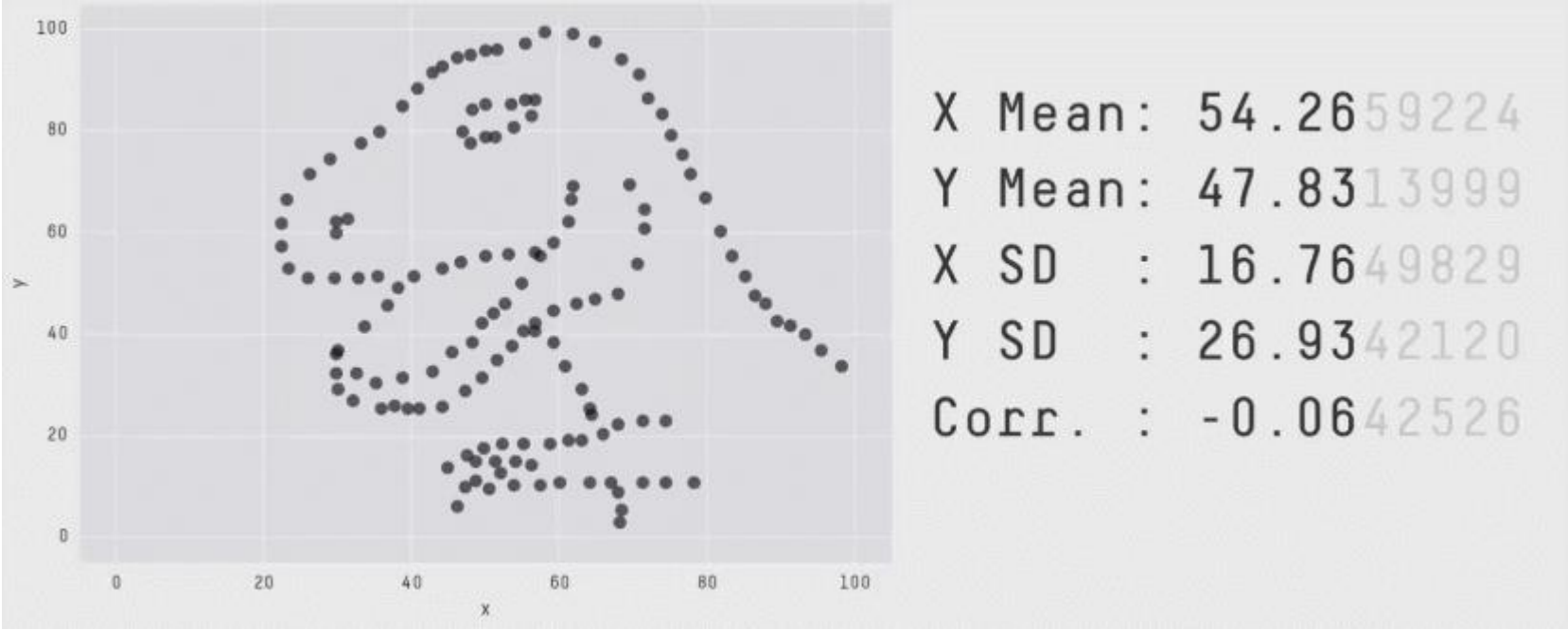
А. Не ослањајте се само на статистике



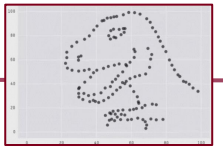
- Илустрација: Енскамов квартет (Енскам, 1973)
- $M_x = 9$, $S^2_x = 11$
- $M_y = 7.5$, $S^2_y = 4.125$
- $r_{xy} = 0.816$
- Линерана корелација и универијациони статистици су исти у свим случајевима



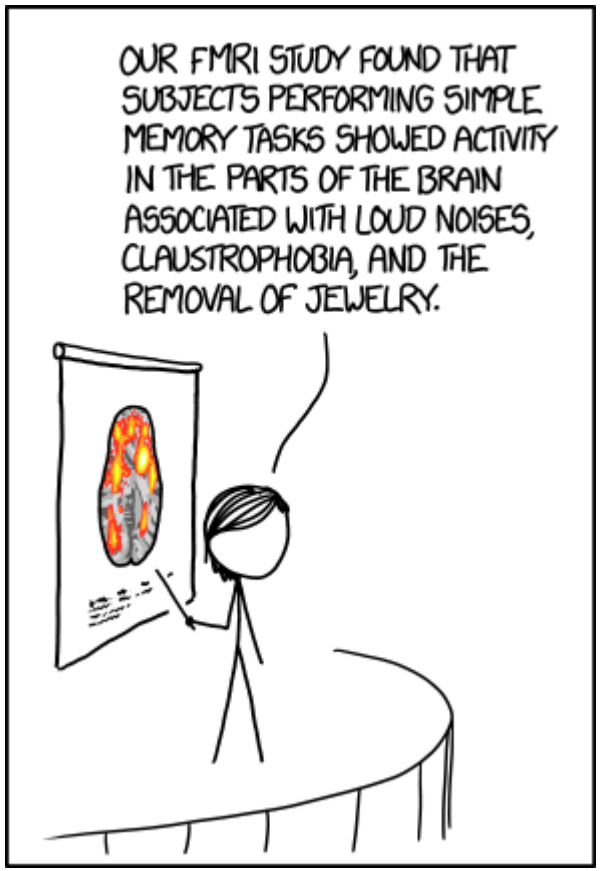
А. Не ослањајте се само на статистике



Илустрација: Datasaurus Dozen

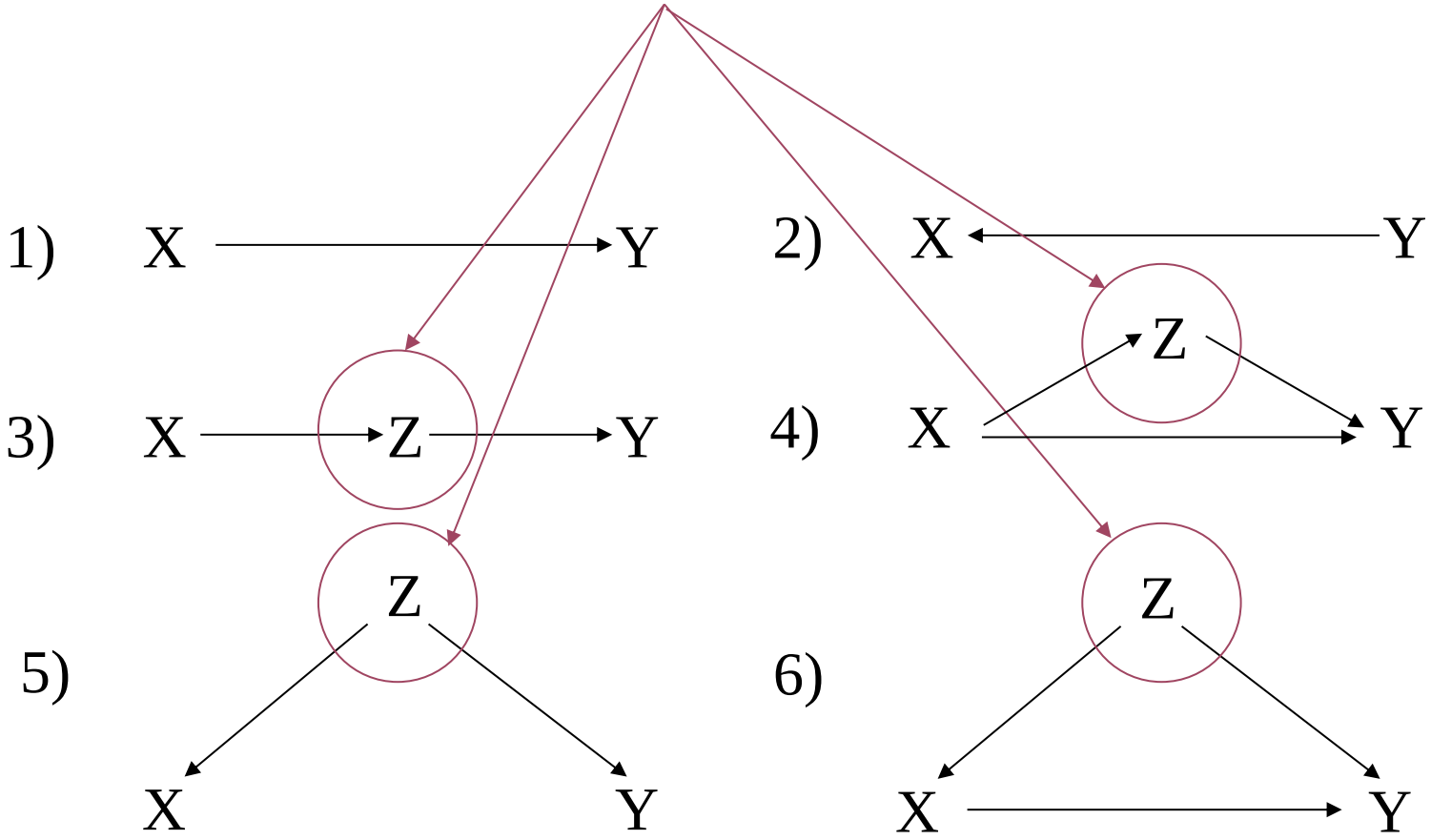


Б. Повезаност није исто што и узрочност



Б. Повезаност није исто што и узрочност

конфундирајућа варијабла

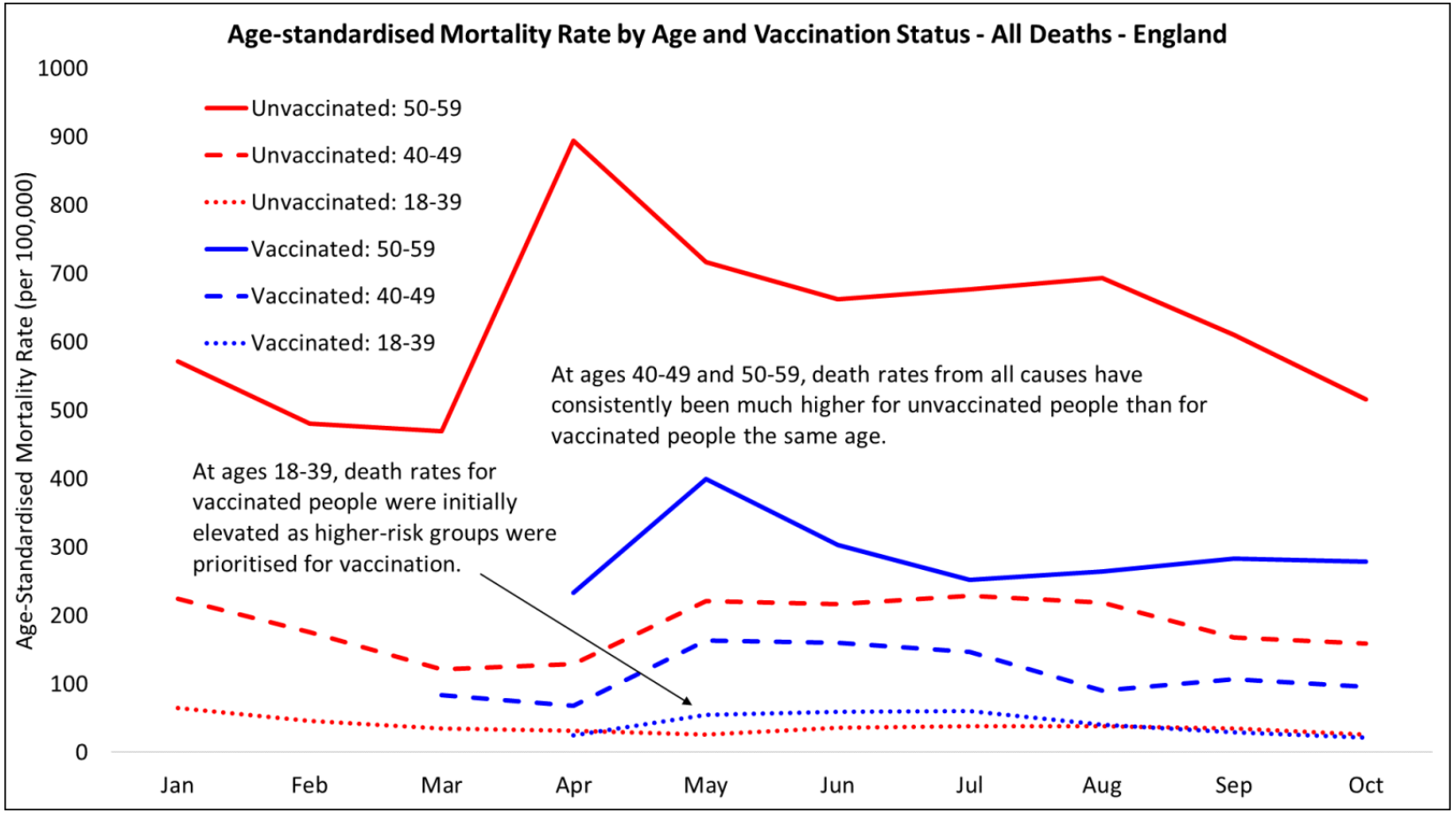


7) случајност!

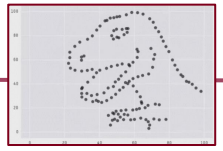
Пример из праксе

- Која је ово корелација?
  - Између којих варијабли?
- АЛИ, на почетку вакцинације:
  - Вакцинисани су прво стари
  - Међу млађима, вакцинисани су прво они из ризичних група
  - Конфундирајуће варијабле

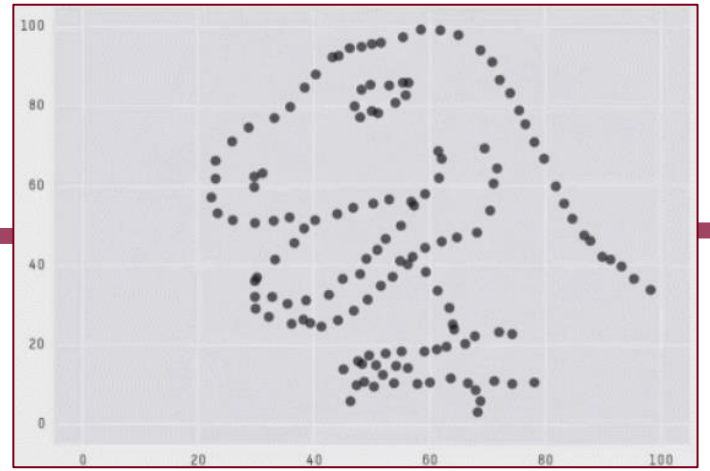
Пример из праксе



- Кад се изделе у категорије по угрожености и узрасту, корелација нестаје!
  - Симпсонов парадокс

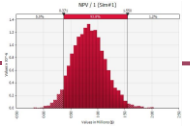


7. Да сумирамо...



Да сумирамо...

- Шта значи да постоји статистичка повезаност две варијабле?
- А независност?
- Како се зове повезаност две нумеричке?
- А две категоричке?
- Које формуле имамо за асоцијацију дихотомних варијабли?
- А ако варијабле не морају бити дихотомне?
- Шта користимо најчешће као меру корелације?
- Када је адекватно користити Браве-Пирсонов r?
- Ако Браве-Пирсонов коефицијент корелације применимо на случај када је једна варијабла бинарна, који коефицијент добијамо?
- Зашто корелација не мора да значи и узрочност?



Крај, за данас ☺