Correlation

Adatfájlok

- CourseEvaluation
- pontokPrediktorok
- GSSS2018happiness (laborhoz)
- Catterplot

Mi a korreláció

- Két változó értékeinek az együttjárása
- Együttjárás nem jelent ok-okozati összefüggést!
- Skála vagy ordinális változókkal használható
  - Minél több lehetséges értéke van a változóknak, annál jobb
  - Minél több adatunk van, annál jobb

Növekszik-e a várható élettartam az egészségügyi ráfordítással?

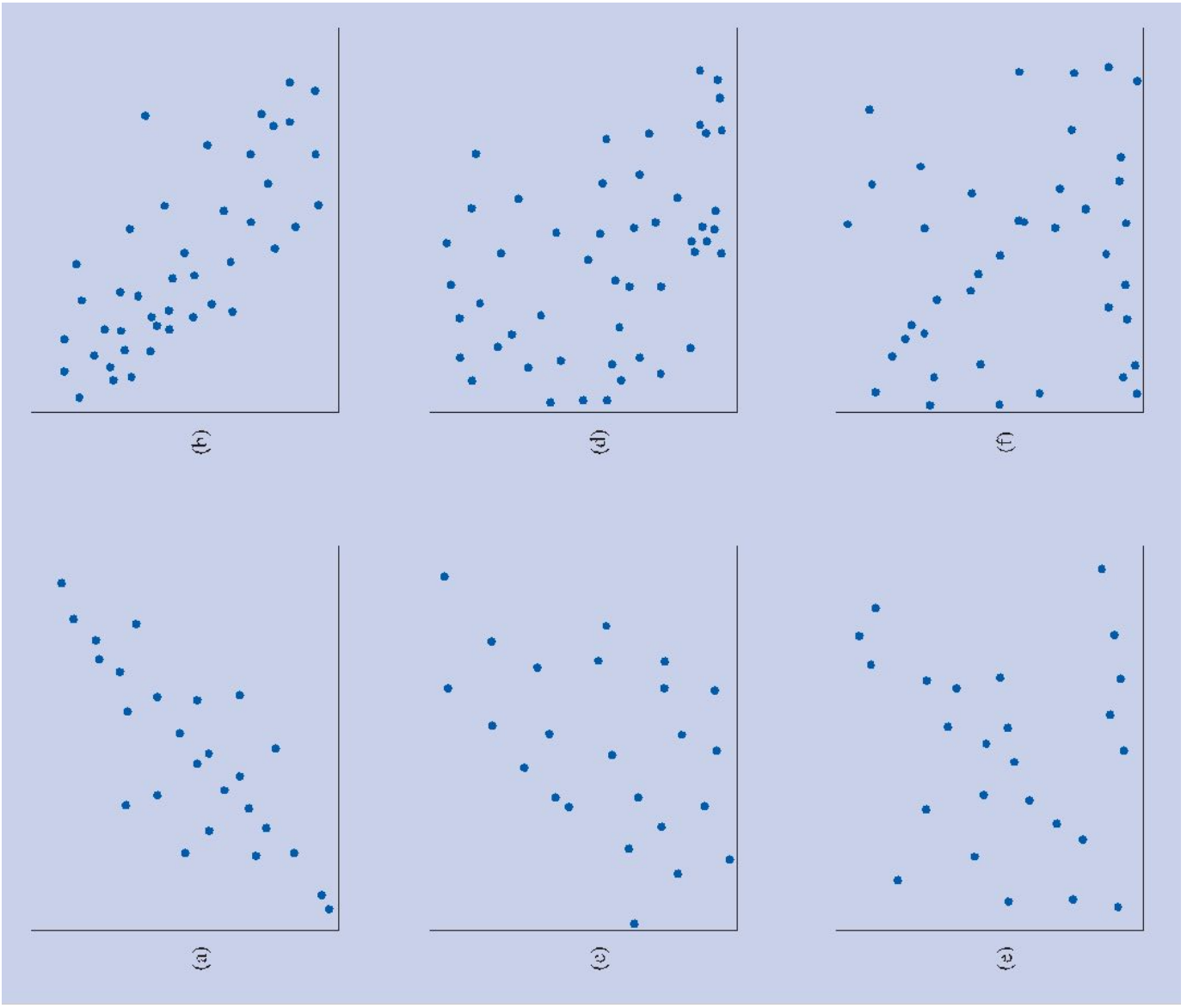
Csökken-e a reakcióidő az életkor növekedésével?

Összefügg-e a jövedelem az iskolai teljesítménnyel?

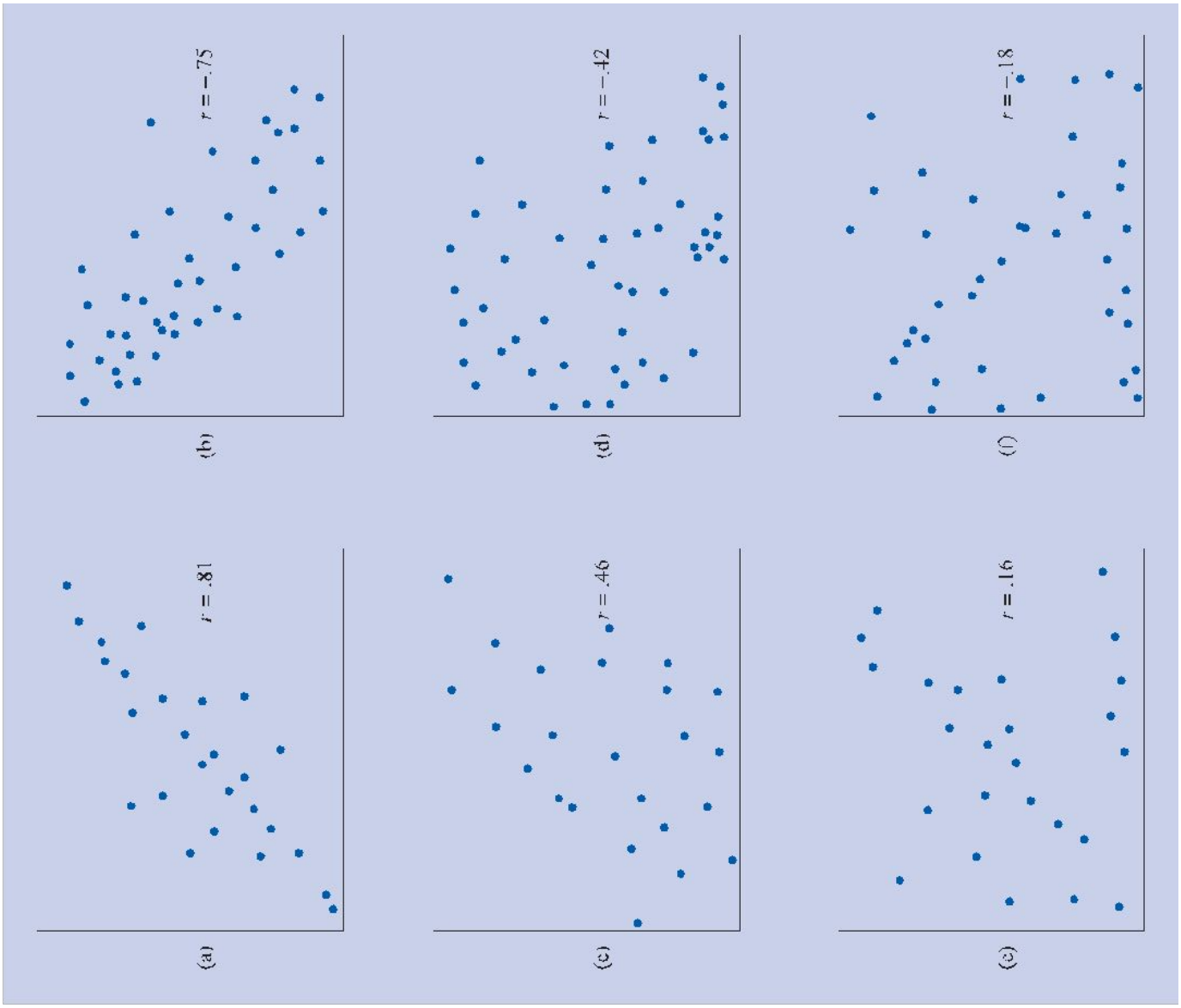
Lineáris és kurvilineáris összefüggés

- **Lineáris**: a kapcsolatot legjobban egy egyenes modellálja. Lehet
  - Pozitív (magasság és cipőméret)
  - Negatív (szorongás és alvás)
- **Kurvilineáris**: a kapcsolatot legjobban egy görbe vagy több, változó irányú egyenes írja le
  - kvadratikus (másodfokú): iskolai teljesítmény és jövedelem
  - kubikus (harmadfokú):

Pontdiagram



A korrelációs együttható mint hatásméret



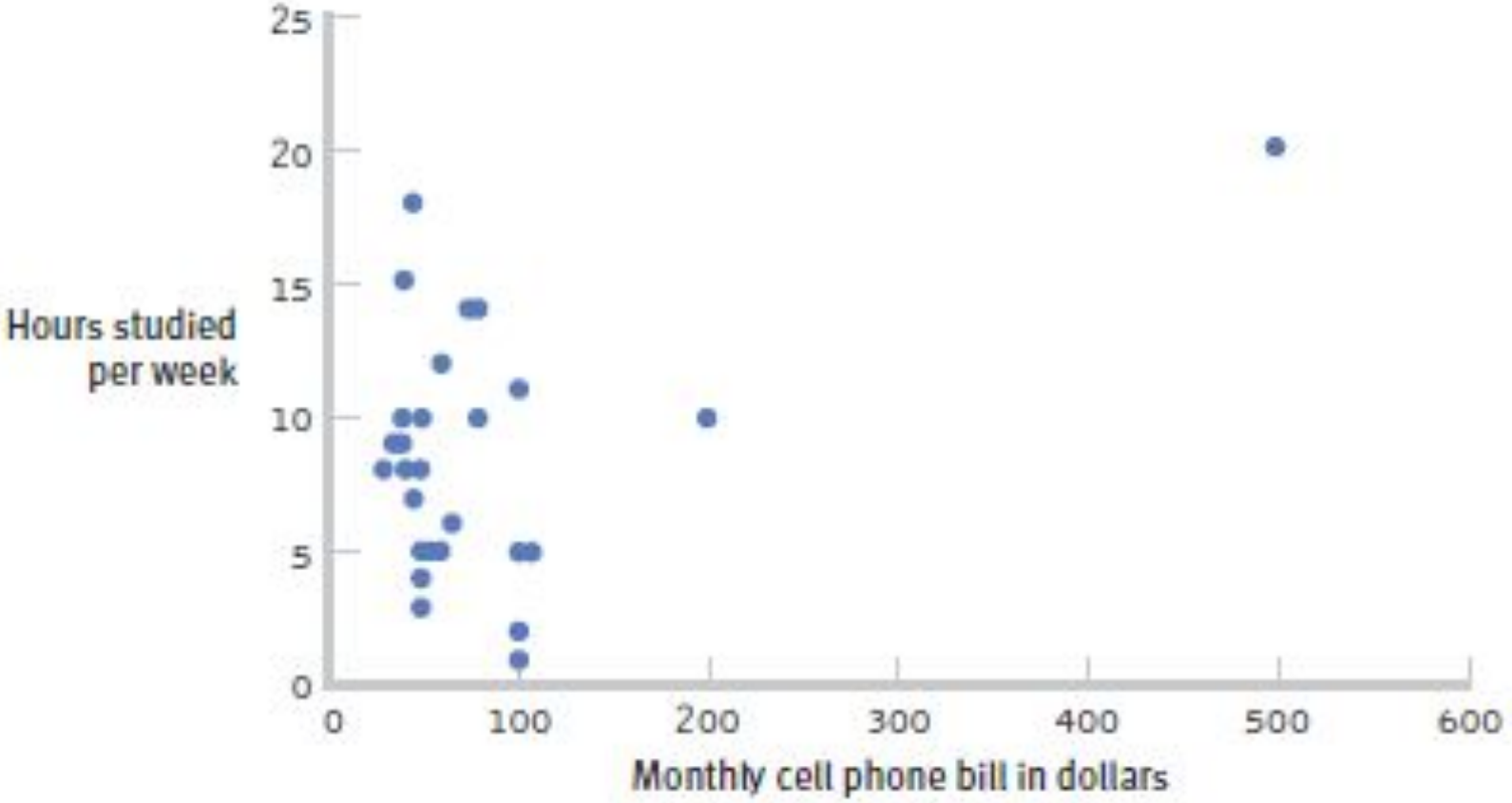
A korrelációs együttható

- Értéke -1.00 és +1.00 között lehet
  - Előjel jelzi az összefüggés irányát
  - A nullától való távolság az összefüggés erejét (hatásméret)
- Leggyakoribb statisztikák
  - Pearson product-moment correlation (r, r_p): normál eloszlású skála adatok
  - Spearman rank-order correlation (r_s, ρ, rho, ró) : ordinális adatok, vagy nem-normál eloszlású skála adatok

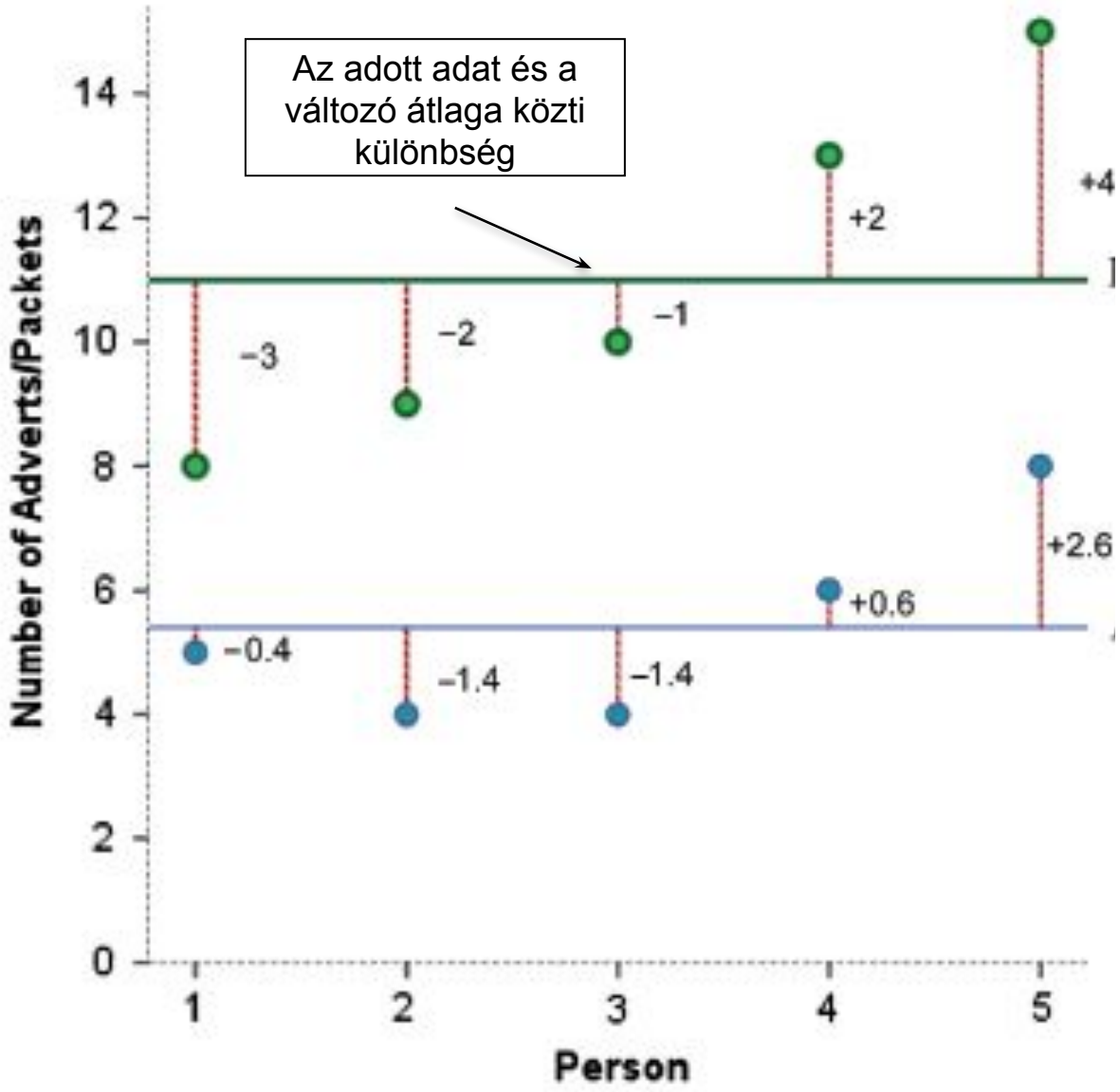
A korrelációelemzés

- A korrelációs együttható: hatásméret (leíróstatisztika)
- inferenciális statisztika:
  - mi a null-hipotézis valószínűsége
  - azaz mi annak a valószínűsége, hogy nincs összefüggés a két változó között
  - azaz mi annak a valószínűsége, hogy a kapott korrelációs együttható a véletlen műve: **p**

Korreláció és kiugró értékek



FORMÁLISAN



Packs of candy purchased

Ads watched

A kovariancia

- Minden egyénre kiszámoljuk az egyén adata és a változó átlaga közötti különbséget az egyik majd a másik változóra
- Ezeket a különbségeket összeszorozzuk: ha mindkét különbség pozitív vagy mindkét különbség negatív, pozitív lesz a szorzat. Ha az egyik különbség pozitív, a másik negatív (azaz ellenkező irányba mozognak az adatok az átlaghoz képest), a szorzat negatív
- Összeadjuk a szorzatokat
- Elosztjuk az eredményt N - 1-gyel.

Kovariancia:

$$\text{cov}(x,y) = \frac{\Sigma(X_i - M_x)(Y_i - M_y)}{N - 1}$$

$$= \frac{(-0.4)(-3) + (-1.4)(-2) + (-1.4)(-1) + (-0.6)(2) + (2.6)(4)}{4}$$

$$= \frac{(1.2) + (2.8) + (1.4) + (1.2) + (10.4)}{4} = \frac{17}{4} = 4.25$$

A Pearson Korrelációs Együttható (r)

Elosztjuk a kovarianciát a két változó szórásának szorzatával

$$r = \frac{cov_{xy}}{SD_x SD_y} = \frac{\Sigma\left(x_i - Mx\right)\left(y_i - M_y\right)}{(N\text{-}1)SD_x SD_y}$$

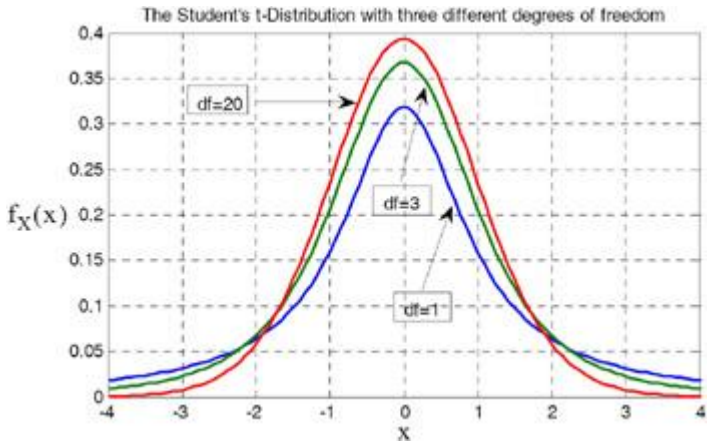
Hipotézis-tesztelés

- Keresünk egy eloszlást, aminek ismerjük a paramétereit (ez a Student t lesz)
- Az r értéket konvertáljuk t-re

$$t_r = \frac{r}{\sqrt{\frac{1 - r^2}{N - 2}}} \qquad t_r = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

Hipotézis-tesztelés

- A t eloszlásból leolvasható a kapott t (és így az r) valószínűsége
- Ez annak a valószínűsége, hogy az r értékünk pusztán a véletlen műve



Spearman együttható

- Ordinális vagy nem-normál eloszlású skála adatokra
- A logika ugyanez, de a nyers adatok helyett a pontszámok rangsorával számol

Eredeti pontszám	**Spearmanhoz használt pontszám**
12	1
45	2
129	3

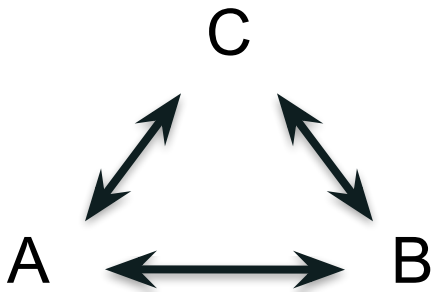
Parciális korreláció

Bivariate correlation:

Used to assess the relationship between two variables.

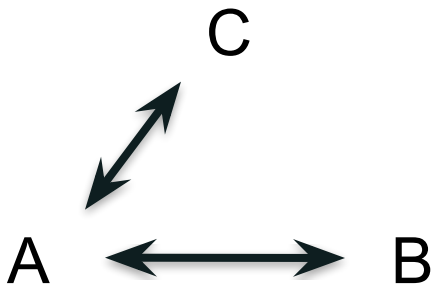
Partial correlation:

When we do a partial correlation between two variables, we control for the effects of a third variable. Specifically, the effect that the third variable has on ***both*** variables in the correlation is controlled.



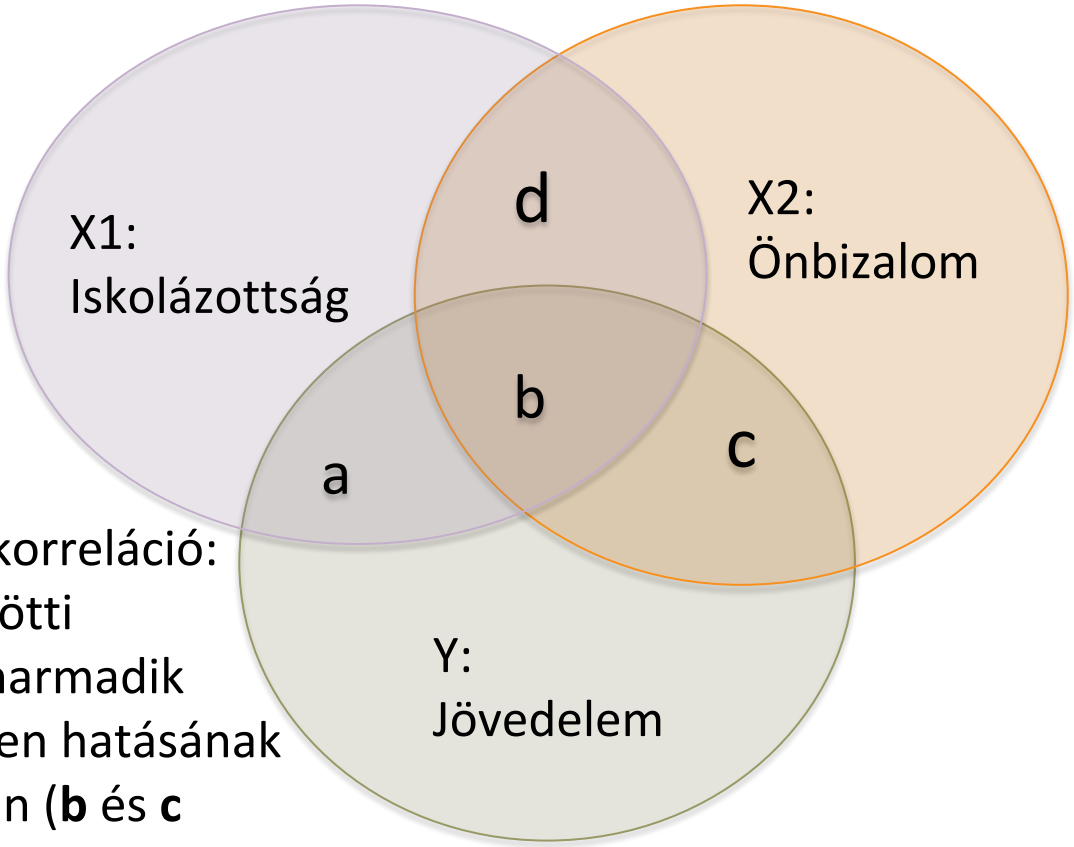
Semi-partial (or part) correlation:

When we do a semi-partial correlation, we control for the effect that the third variable has on only one of the variables in the correlation.



Parciális korreláció

- Két változó közötti összefüggés egy harmadik változó hatásának eltávolítása után: Az iskolázottság hatása a jövedelemre az önbizalom hatásának eltávolítása után (**a** marad, **b**, **c**, és **d** kizárva)



Szemiparciális korreláció: Két változó közötti összefüggés a harmadik változó közvetlen hatásának eltávolítása után (**b** és **c** kizárva, de **d** nem)

Kendall’s tau

Kendall’s tau (τ):

Used Ordinális vagy nem-normál eloszlású skála adatokra, ha viszonylag kicsi a minta és/vagy sok az azonos pontszám.

Regressziós egyenes és Konfidencia Intervallumok

- Regressziós egyenes, “legkisebb négyzetek” módszer:
  - az egyes pontok távolsága az egyenestől a lehető legkisebb legyen
- Konfidencia Intervallum: az intervallum, amibe az r értéke **a populációra vonatkoztatva** 95% bizonyossággal esik

1. Példa: courseEvaluation
JASP: Regression > Correlation Matrix

Van-e összefüggés a kurzusra kapott jegy és a kurzus hallgatói értékelése között. Adatok: kurzusra kapott jegyek átlaga és hallgatói értékelés átlaga 5-pontos skálán mérve.

1. Eloszlás (hisztogram, boxplot, stb)
2. Kiugró értékek (pontdiagram) - mi legyen?
3. Regression > Correlation Matrix
4. Pearson, Spearman vagy Kendall's tau?
5. Egyszélű vagy kétszélű a hipotézis?
6. Konfidencia Intervallumok
7. Döntés

(Ha valakit érdekel a Vovk-Sellke maximum p-ratio: https://jasp-stats.org/2017/06/12/mysterious-vs-mpr/)

Jelentés

hipotézis

adatgyűjtés

It is often thought by course instructors that the way in which students evaluate a course will be related, in part, to the grades that are given in that course. In an attempt to test this hypothesis we collected data on 15 courses in a large university, asking students to rate the overall quality of the course (on a five-point scale) and report their anticipated grade in that course.

problémák

One extreme outlier was identified and removed from further analyses. The remaining 14 courses showed a very weak (non-significant) positive correlation between expected grade and course evaluation (r = .12[-.44, .61], p = .68, see Figure 1)

We therefore have no conclusive evidence for the hypothesis that students' course ratings are affected by their grades.

konklúzió

utalás ábrára

teszt

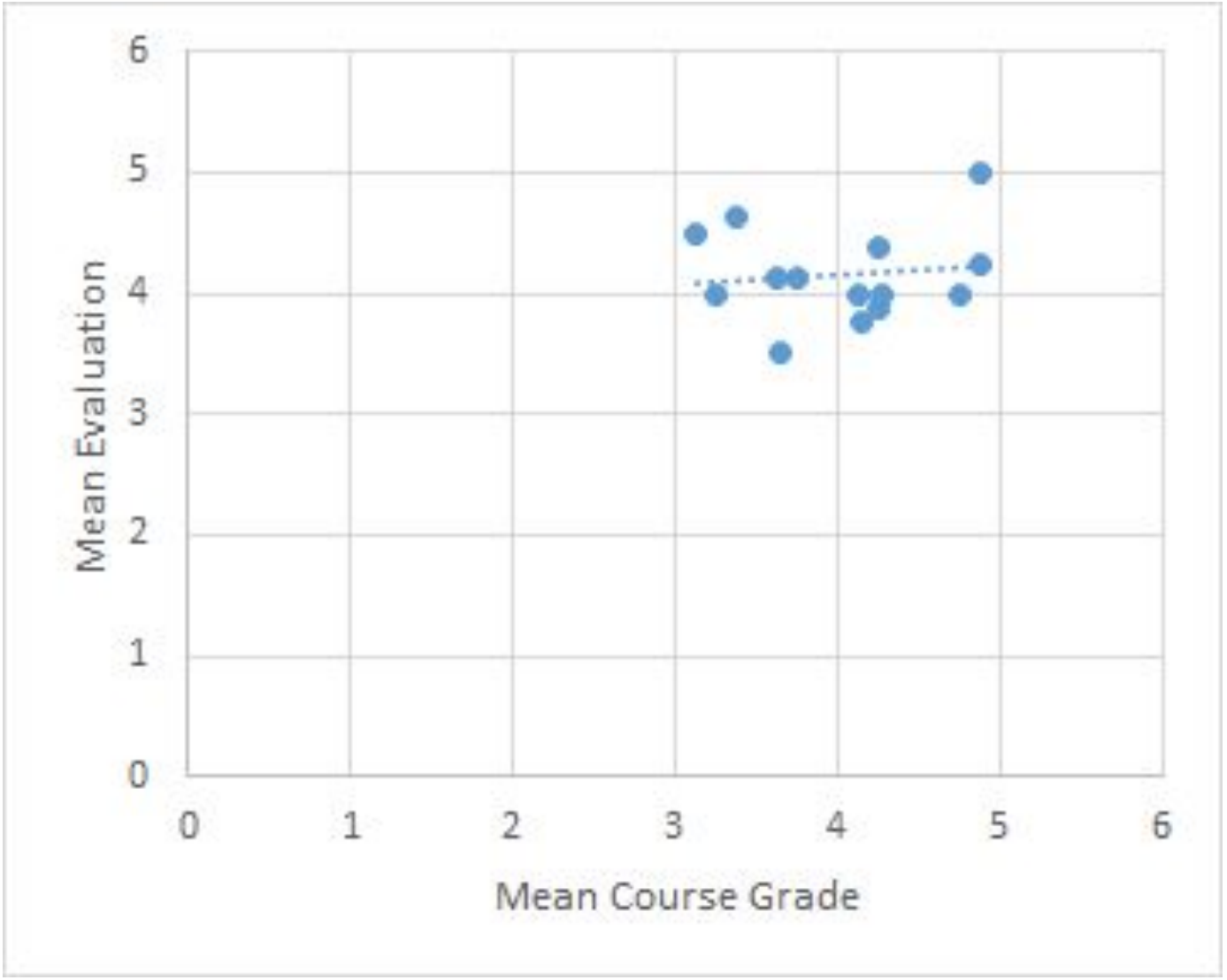


Figure 1. Correlation between Course Grade and Student Rating.

2. példa: Statisztika gyakorlatok és zh eredmények

- Előző években házi feladatok átlaga és zh-n elért pontszám