

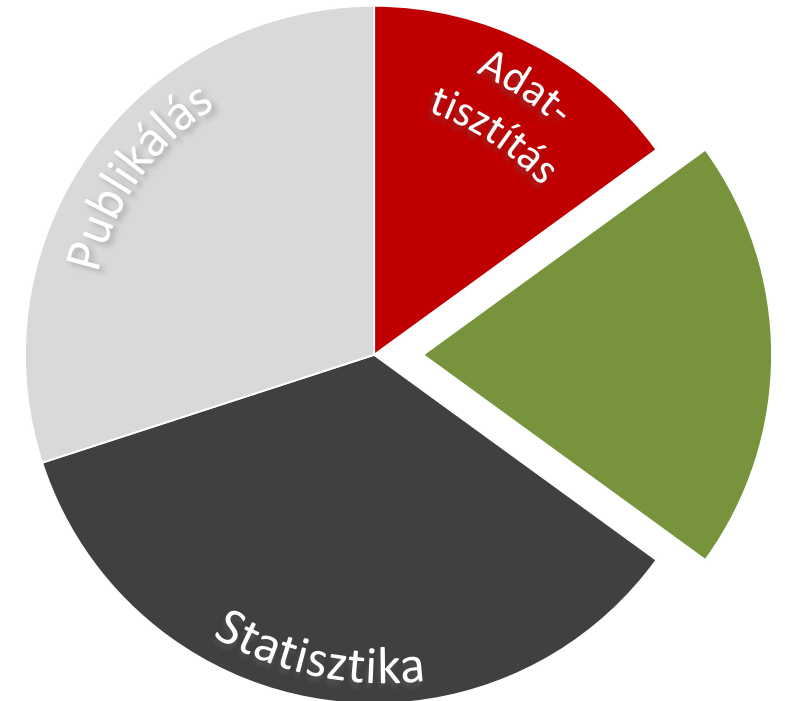
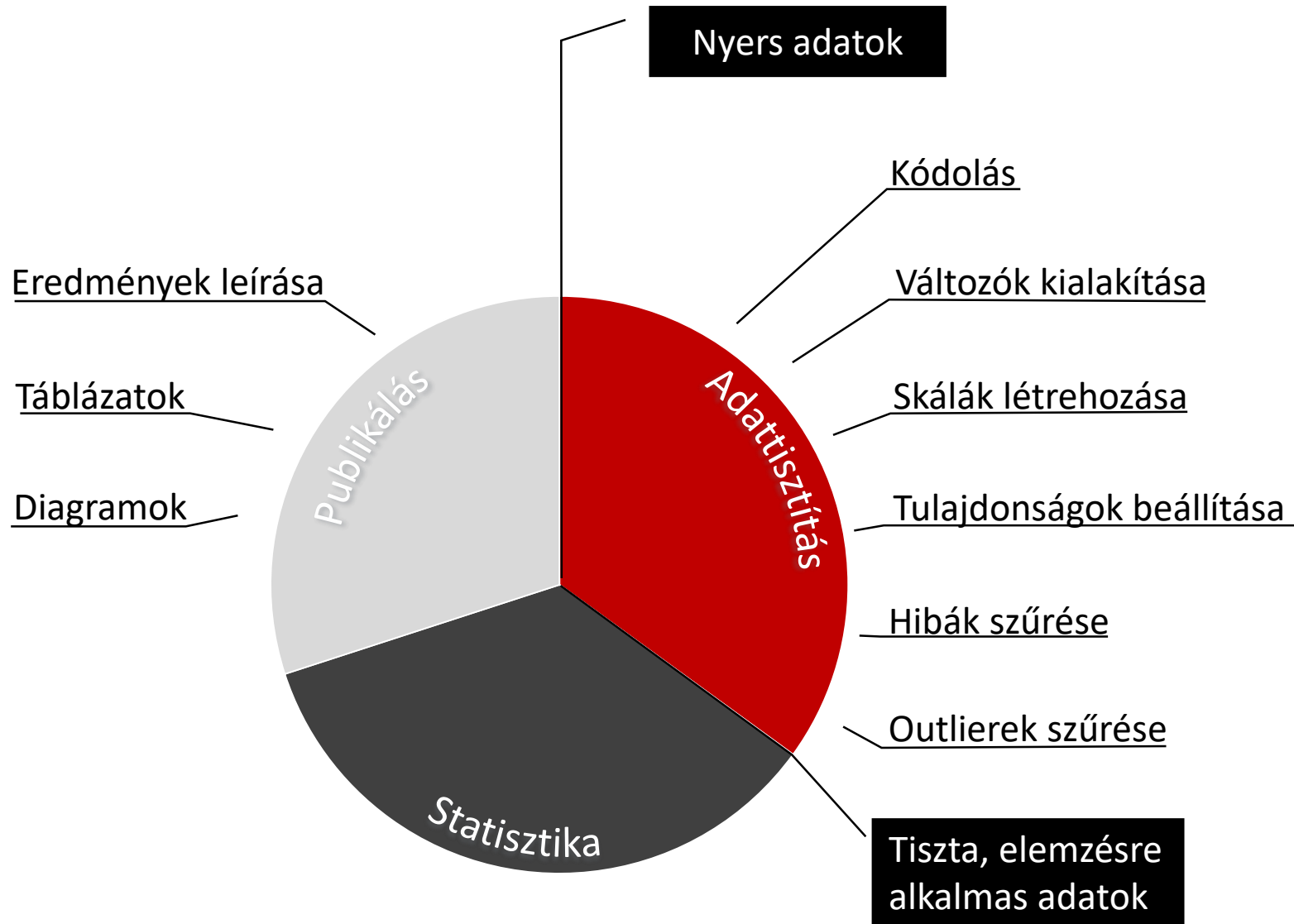
Többváltozós statisztika elméleti alapjai (BMNPS07500M)

Készítette: Soltész-Várhelyi Klára

Adatfeldolgozás

1. Hibák és torzítások

Adattisztítás



Egyéni tapasztalataim alapján az adatok Google formról való letöltése és az első hipotézis tesztelése közötti munkák teszik ki általában egy statisztikai elemzés egyharmadát. Jól átgondolt kísérlettervezéssel ez az idő rövidíthető le.

Hiba

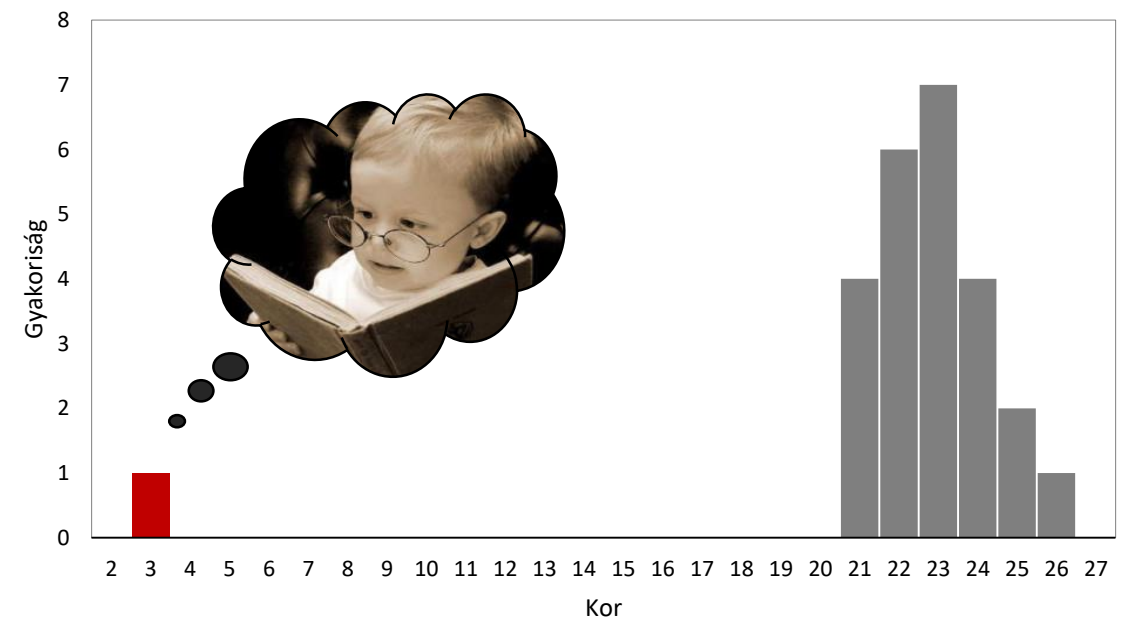


Fig. 1. A statisztika kurzus hallgatóinak életkori eloszlása

- A hiba szó több értelemben is előfordul a kurzus során, ne keverd össze őket! Kettő legfontosabb:
 - **A predikció hibája**
 - vagy más néven a modell pontatlansága / az függő változó varianciájának azon része, amit a prediktorokkal nem tudunk magyarázni / a zaj / a statisztikai érték számolásakor a nevező
 - Ide tartozó fogalmak lesznek majd a SS_R (sum of squared residuals) és MS_R (mean squared residuals)
 - **Mivel többtényezős rendszereket vizsgálunk, és az összes tényezőt sosem vehetjük figyelembe, a modellnek egy jól megtervezett vizsgálatban is mindig lesz hibája** (ettől statisztika, és nem matematika)
 - Pl. a magasságot három tucat gén meg egy rák környezeti faktor befolyásolja, de mi csak a nem és a proteinfogyasztás hatását vizsgáljuk.
 - **Adathiba**
 - Téves adatok
 - Legtöbbször félregépelésekből, mérési problémákból, esetleg nem őszinte válaszadókból adódnak.
 - Pl. ha valaki 165 cm magas, de az adatok bevitelekor véletlenül 175-öt ütöttél be, az hiba. De az is hiba, ha meg volt nyúlva a mérőszalag, és mindenkit 2 centivel rövidebbnek mértél.
 - **Az előzővel szemben törekedni kell arra, hogy adathiba ne legyen az elemzésben.**

Szisztematikus és nemszisztematikus hiba

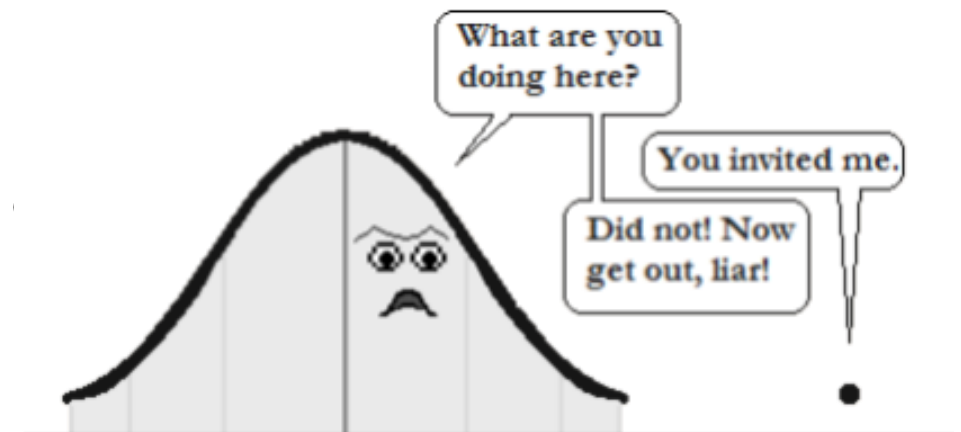
- Módszertanilag a **predikció hibáját két dolog okozhatja**:
 - **Nemszisztematikus hiba**
 - Mivel többtényezős rendszereket vizsgálunk, és az összes tényezőt sosem vehetjük figyelembe, a modellnek egy jól megtervezett vizsgálatban is mindig lesz hibája, **de ezek a figyelembe nem vett tényezők egyforma arányban fordulnak elő minden kísérleti mintánkban.**
 - Pl. a magasságot három tucat gén meg egy rákat környezeti faktor befolyásolja...
 - Pl. egy reakcióidő vizsgálatban egyformán lesznek éberem és fáradtan érkezők minden kísérleti csoportunkban.
 - **A nemszisztematikus hiba természetes része a mérésnek**, de módszertanilag lehet (érdeemes) törekedni a csökkentésére, mert minél kisebb a zaj, annál kisebb hatások is kimutathatóak lesznek
 - **Szisztematikus hiba** (ez egyben egyfajta mérési / adathiba is)
 - Olyan hiba, ami **nem véletlen irányokban befolyásolja a mintáinkat**, így torzítja a mérést.
 - Pl. férfiak és nők memóriateljesítménye közötti különbség & a szexi kísérletvezető nő
 - **Törekedni kell arra, hogy szisztematikus hiba ne legyen az elemzésben.**
 - A vizsgálat megtervezésekor, minta kiválasztásakor & mérések lebonyolításakor ne emeljünk be sz. hibát
 - Adatok áttekintésekor, tisztításakor figyelni kell a sz. hibára utaló jelekre
 - Statisztikák feltételeinek ellenőrzésekor több módszer is van, melyből következtetni lehet sz. hiba jelenlétére

- **Adathiba** sok forrásból származhat:
 - Félregépelések adatbevitelkor
 - Félregépelés, tizedesvessző rossz helyen
 - (Számítási) hibák kiértékeléskor
 - Elszámolások „fejben” számoláskor
 - Rosszul megírt kiértékelő
 - Rossz megoldó kulcs
 - Rosszindulatú / komolytalan kitöltők
 - Végére akar érni, és csupa 1-est nyom
 - Beleunt, és mintázat szerint tölti ki
 - Visszajelzést adsz, ezért megpróbálja egy „alternatív személyisége” szerint is kitölteni
 - Stb. ...

- Nem található meg minden hiba, de néhány módszer létezik
 - Félregépelések adatbevitelkor
 - Leíró statisztika minimumra és maximumra – ha 175 helyett 275 cm-t írtál, észre fogod venni
 - Ha tartományon belül van, nem igazán megtalálható, de a félregépelés viszonylag ritka, és elszórt, ezért a nagyon szélsőséges értékeket kivéve nem okoz nagy torzítást
 - (Számítási) hibák kiértékeléskor
 - Ne számolj semmit fejben
 - Ha te írod a kiértékelést, ellenőrizd a skála minimumát és maximumát, esetleg hozz létre tesztkitöltőket
 - A hibás megoldó kulcs gyakoribb, mint gondolnád, fúsd át az itemeket
 - Rosszindulatú / komolytalan kitöltők
 - Az „alternatív személyiség” ellen lehet védekezni, ha a kérdőíved végén megkérdezed, hogy ezek valós válaszok-e
 - Komolytalan kitöltőkre utalhat, ha a válaszainak kicsi a szórása – ezt Excelben érdemes leellenőrizni – főleg internetes merítés esetén

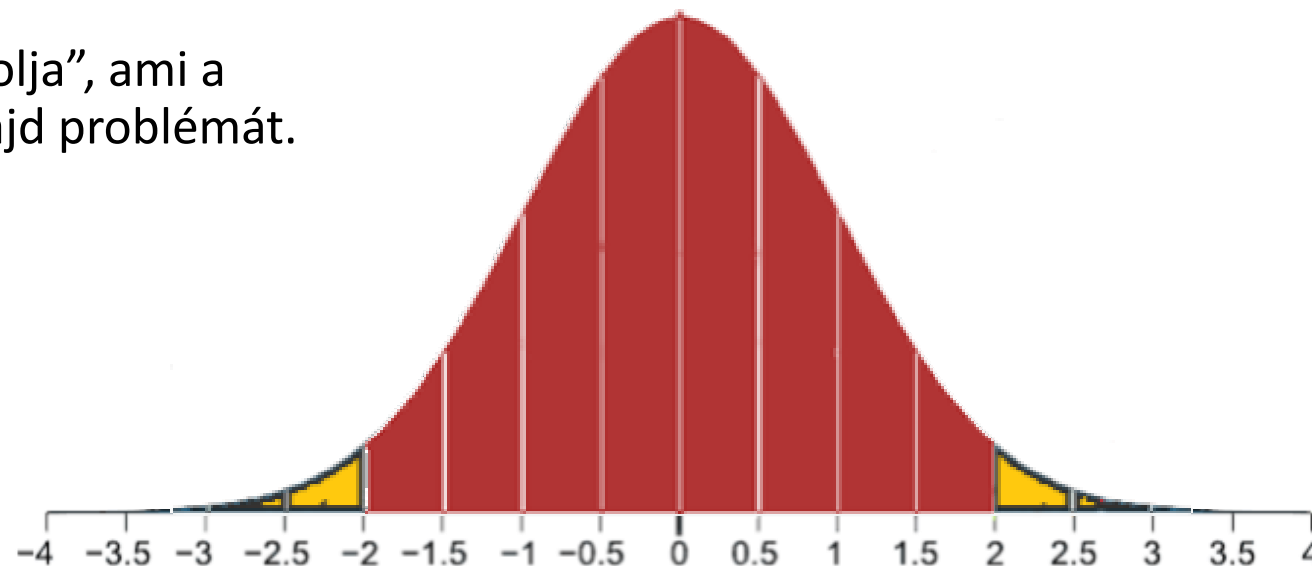
- NEP-PI-R – rossz automatikus kiértékelő – Hogyan jöttünk rá? A skála minimuma kevesebb volt, mint kellene
- EDI - rossz megoldó kulcs – Alacsony reliabilitás mutató
- Interakciós vizsgálat – nem tiszta kategóriák
- Utánkövetéses vizsgálat Szomáliában – összekevert azonosítók – eltérő életkorok, nevek
- Anorexia vizsgálat – hagyományos kérdőívek használata (használatatlansága), ha torz az önkép – gyanúsan nem kijövő eredmények után grafikus megjelenítés azonosítókkal
- Vizuális kísérlet – Program rosszul mentette ki az eredményeket – gyanúsan kicsi szórások
- Arcok kódolása – komolytalan kitöltők – először outlierként, aztán az adatokat átnézése a válaszok szórására
- Demográfiai adatok – szöveges infók a korban – gyanúsan sok hiányzó érték

Outlier



- Általános definíció: outlier az,
 - „ami valid érték, de nagyon eltér a többi adatunktól”
 - Miért probléma:
 - Bár valid érték, de nem jól reprezentálja a populációt – ez főleg kis minták esetén probléma (egy 15 fős mintában van egy 2m magas nő – a valóságban nincs a nők közel 7%-a 2m körül)
 - Torzítja a statisztikai mutatókat: elhúzza az átlagot, megnöveli a szórást.
- Hogy találjuk meg őket:
 - ~~Átlag $\pm$ 2SD szabály~~ (elavult)
 - **Outlier labelling rule**

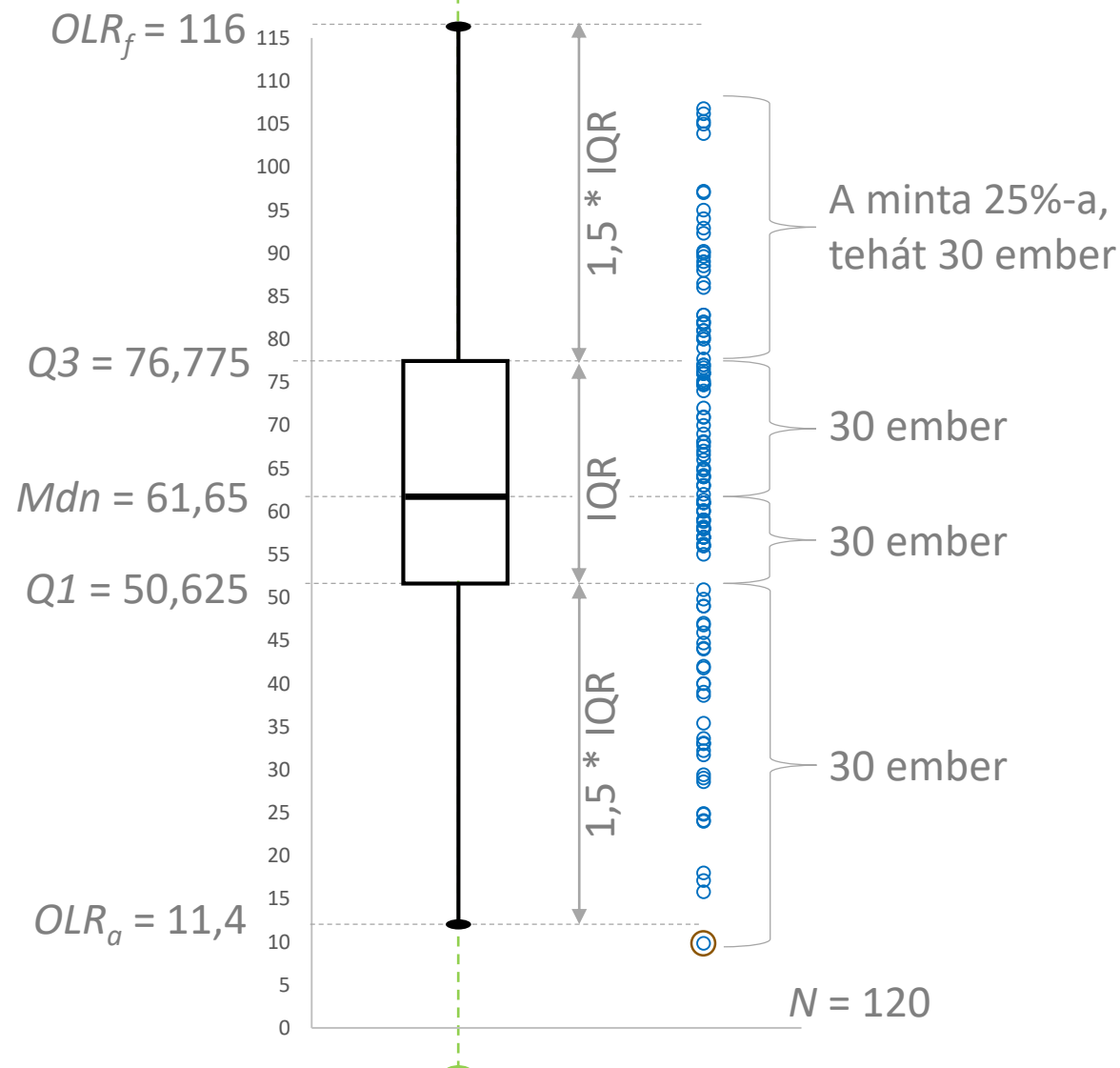
- Mit jelent az, hogy „ami nagyon eltér a többi adatunktól”
 - **Átlag $\pm$ 2 vagy 3 SD szabály**
 - **ELAVULT, NE HASZNÁLD!!**
 - Tekintsük outliernek azt az elemet, aki az átlagtól 2 (vagy 3) szórányira van!
 - Probléma:
 - A szabály leginkább csak normáeloszlás esetén értelmezhető
 - A ± 2 -es z-értékhez (a 2SD-hez) tartozó valószínűség 2,275% - tehát definíció szerint a minta majdnem 5%-a 2SD kívül fog elhelyezkedni.
 - Ördögi kör – ha kiveszem a szélsőséges értékeket, csökken a szórás, akkor újabb értékek kerülnek a 2SD-n kívülre...
 - Az eloszlás görbét úgymond „csonkolja”, ami a normalitás szempontjából jelent majd problémát.





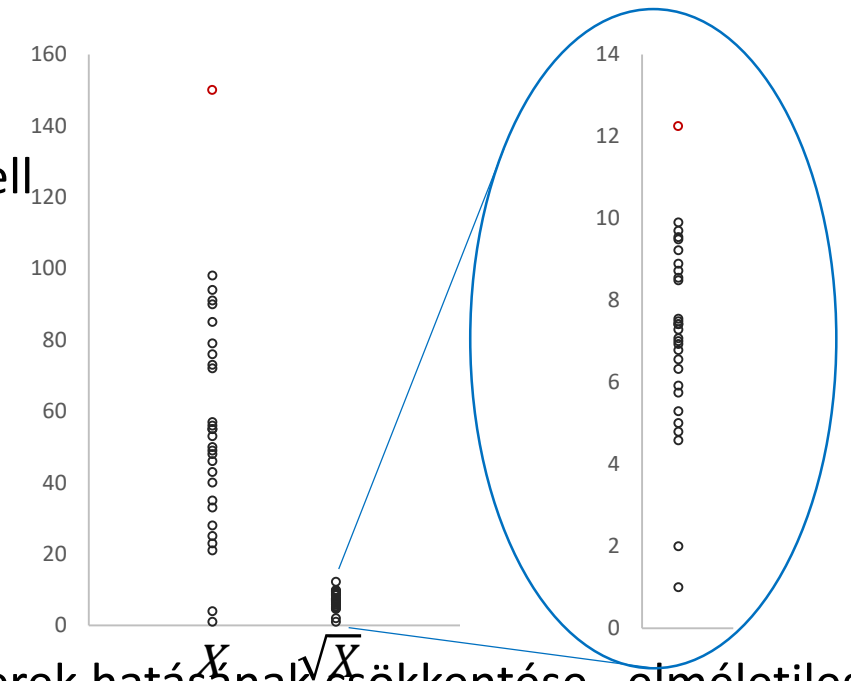
• Outlier labelling rule

- ami távolabb az alsó/felső negyedétől, mint a középső 50% nagyságának másfélszerese
- Határok:
 - $Q1 - 1,5 * IQR$ és $Q3 + 1,5 * IQR$
- Módosított határok
 - $Q1 - 2,2 * IQR$ és $Q3 + 2,2 * IQR$
- Tukey, J.W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley. – először veti fel az ötletet.
- Hoaglin, D.C., Iglewicz, B., and Tukey, J.W. (1986). Performance of some resistant rules for outlier labeling, *Journal of American Statistical Association*, 81, 991-999.
- Hoaglin, D. C., and Iglewicz, B. (1987), Fine tuning some resistant rules for outlier labeling, *Journal of American Statistical Association*, 82, 1147-1149. – felveti a 2,2-es szorzó használatát



Outlierek

- Mit kezdünk velük?
 - **-1. Hiba vagy outlier?**
 - Ha hiba, akkor nincs helye a mérlegelésnek, mennie kell
 - **0. Elemezzük a helyzetet**
 - Pl. nem normális eloszlásnál mit jelentenek?
 - **1. Töröljük őket**
 - Ésszel!!
 - Leggyakrabban használt, kis minta esetén bajos
 - Gondoljuk át, miből jöhet egy szélsőséges érték
 - **2. Transzformáljuk az adatokat**
 - Lineáris transzformációkkal (pl. gyökvonással) az outlierek hatásának csökkentése – elméletileg legjobb megoldás, gyakorlatilag minél jobban transzformáljuk az adatokat, annál nehezebb pszichológiailag releváns következtetéseket levonni majd
 - **3. Winzorizálás - Átírjuk az értéket valami olyanra, ami már nem outlier**
 - Csalás lenne? De ha benne hagyjuk, az is torzítja az adatokat, a két rossz közül ez a kisebb
 - Mire?
 - A következő legnagyobb -már nem outlier- érték plusz 1 egység (ill. a következő legkisebb -1 egység)
 - (Átlag + 2 SD)



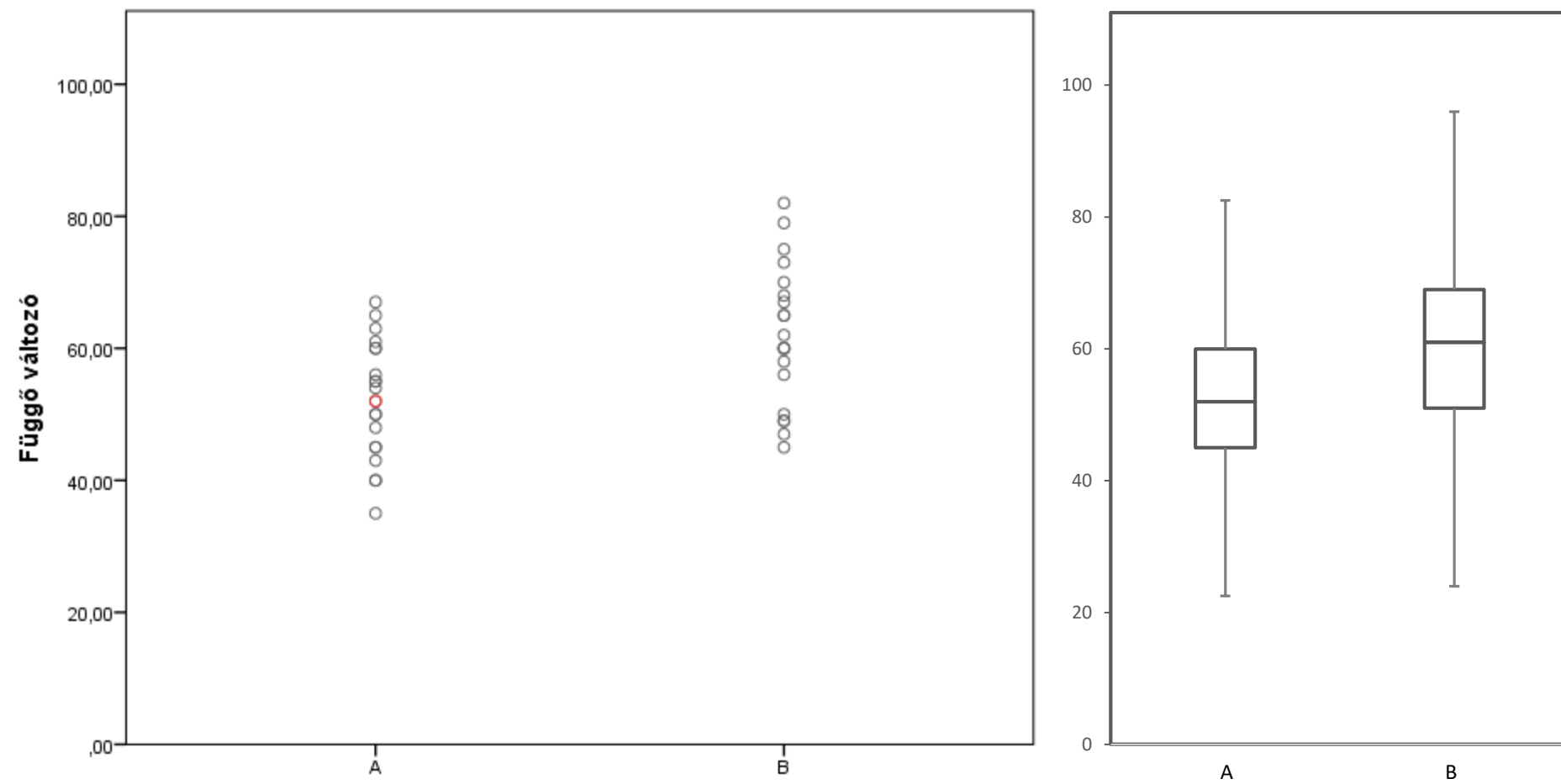
Hiba vagy outlier?

- **Hiba**

- Mi hiba?
 - Olyan érték, mely az adott skálán nem képzelhető el, az érték valamilyen elgépelésből származik, ezért nincs információ értéke. Például nemi adatoknál 3 vagy 0-5-ig terjedő skálán 7 vagy magasságban adatokban 257 (oké, kivéve, ha a Guinness rekordok könyvéből veszed az adatokat 😊)
- Honnan tudom, hogy vannak-e hibás értékek, és ha igen, hol?
 - A leíró statisztikában minimum és maximum nem megfelelő
 - A Extreme Value-kat érdemes megnézni
 - A boxploton ÁLTALÁBAN a csillagos értékek valamilyen elírásból származnak
- Mit kell velük kezdeni?
 - Hibás értékeknél nincs kérdés, ha tudod, javítod őket, ha nem tudod javítani, törlöd

- **Outlier**

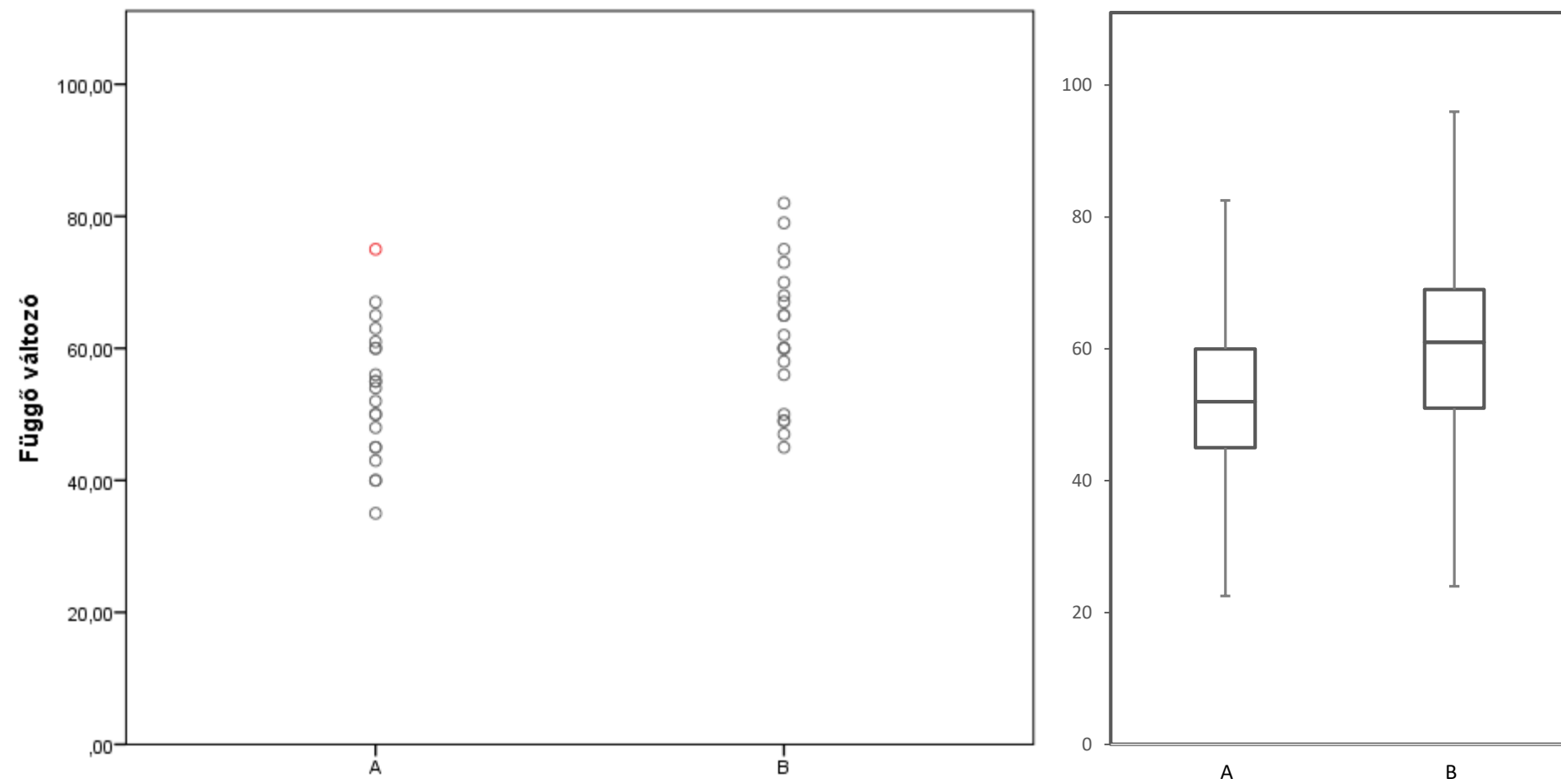
- Mi outlier?
 - Olyan érték, mely elképzelhető az adott skálán, de annyira szélsőséges, hogy torzíthatja a statisztikákat
- Honnan tudom, hogy vannak-e outlierok, és ha igen, hol?
 - A bloxploton ÁLTALÁBAN a körökkel jelzett értékek outlierok
- Mit kell velük kezdeni?
 - Eldöntheted, lásd az előző diát!



Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	985,738 ^a	1	985,738	10,325	,003	,209
Intercept	133575,006	1	133575,006	1399,165	,000	,973
group	985,738	1	985,738	10,325	,003	,209
Error	3723,238	39	95,468			
Total	137804,000	41				
Corrected Total	4708,976	40				

a. R Squared = ,209 (Adjusted R Squared = ,189)

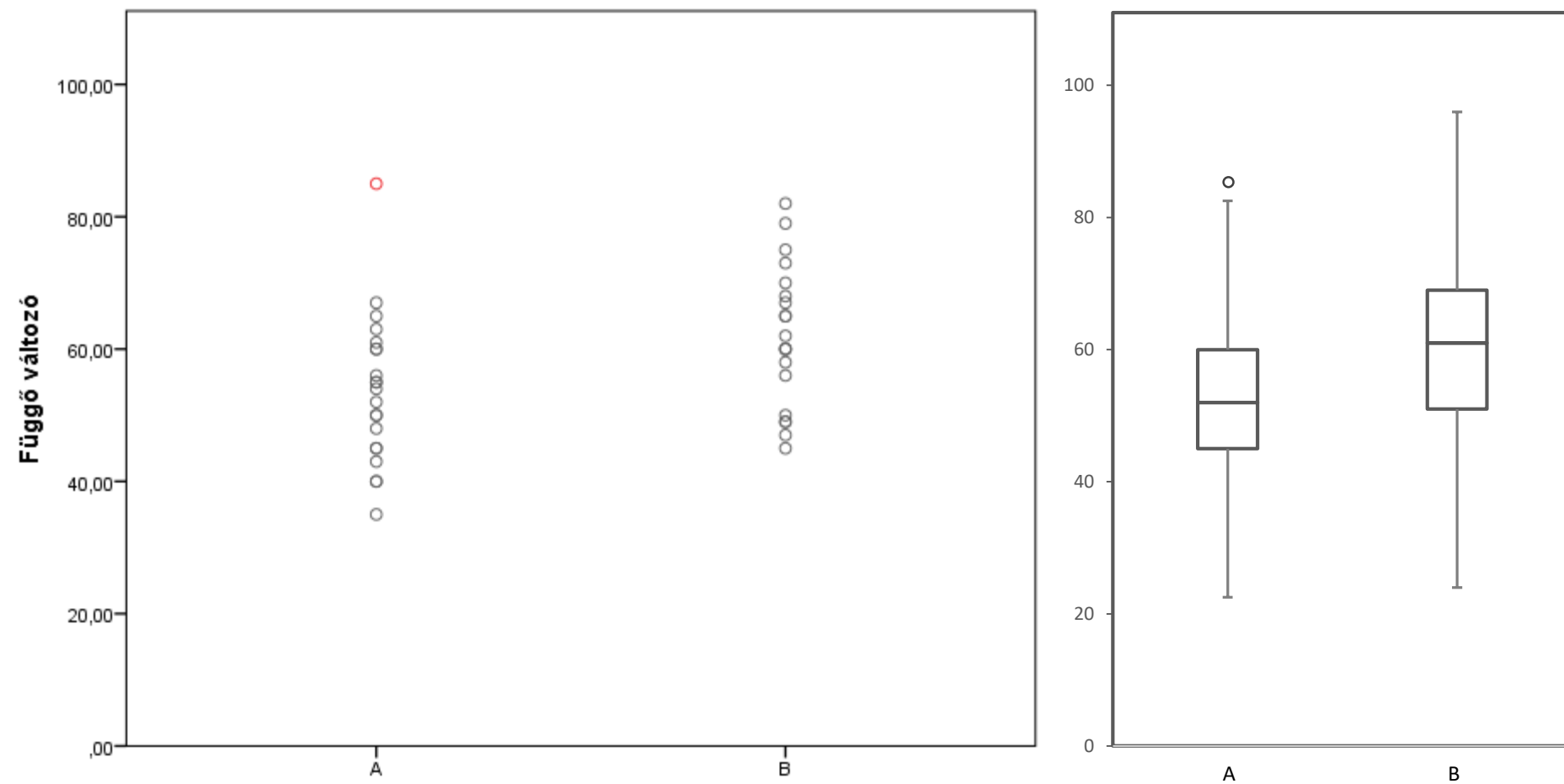
csoport	Mean	Std. Deviation	N
A	52,1905	8,77849	21
B	62,0000	10,71644	20
Total	56,9756	10,85009	41



Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	777,909 ^a	1	777,909	7,192	,011	,156
Intercept	136149,617	1	136149,617	1258,766	,000	,970
group	777,909	1	777,909	7,192	,011	,156
Error	4218,286	39	108,161			
Total	140725,000	41				
Corrected Total	4996,195	40				

a. R Squared = ,156 (Adjusted R Squared = ,134)

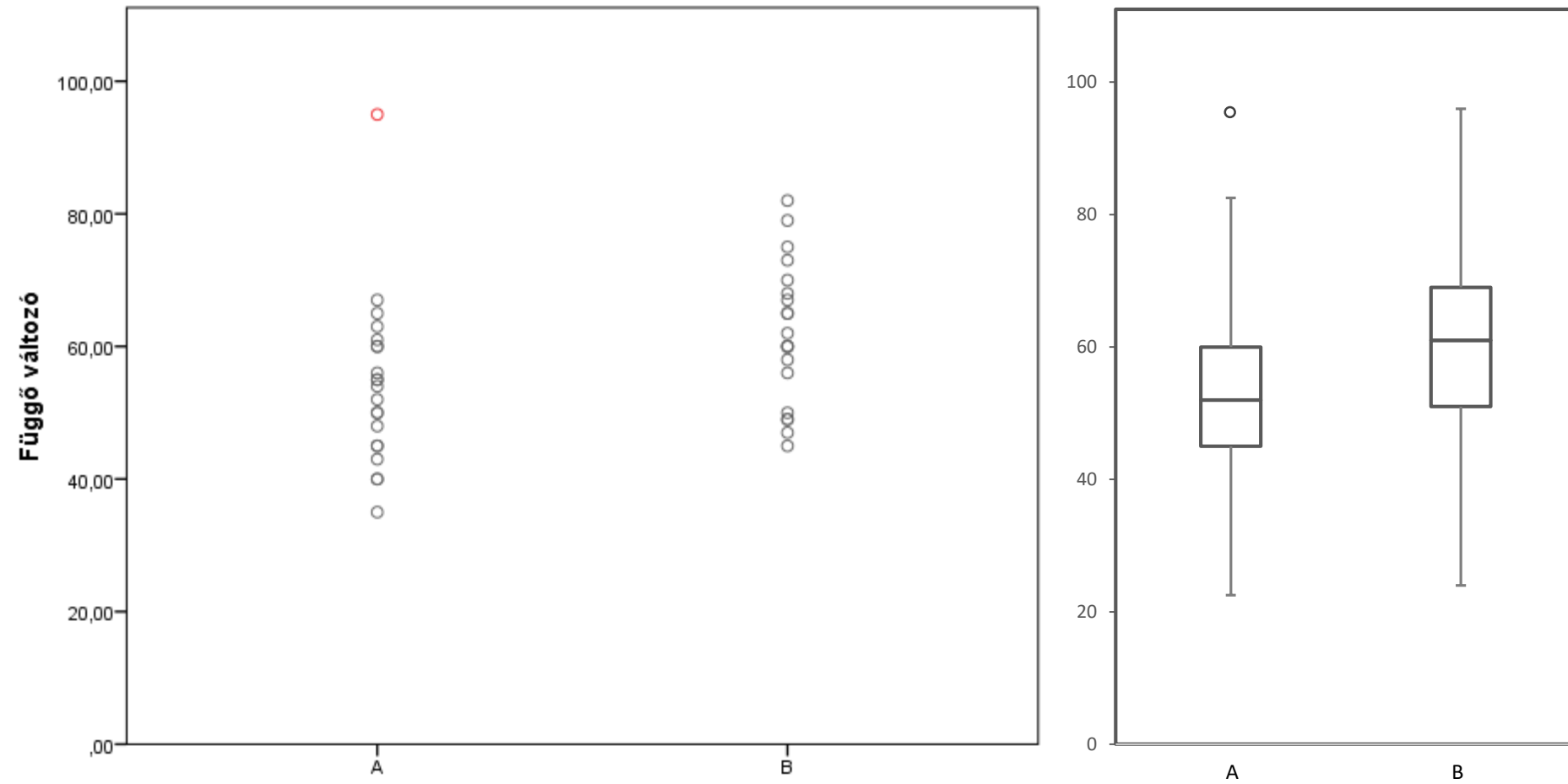
csoport	Mean	Std. Deviation	N
A	53,2857	10,09031	21
B	62,0000	10,71644	20
Total	57,5366	11,17609	41



Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	695,215 ^a	1	695,215	5,711	,022	,128
Intercept	137276,678	1	137276,678	1127,634	,000	,967
group	695,215	1	695,215	5,711	,022	,128
Error	4747,810	39	121,739			
Total	142325,000	41				
Corrected Total	5443,024	40				

a. R Squared = ,128 (Adjusted R Squared = ,105)

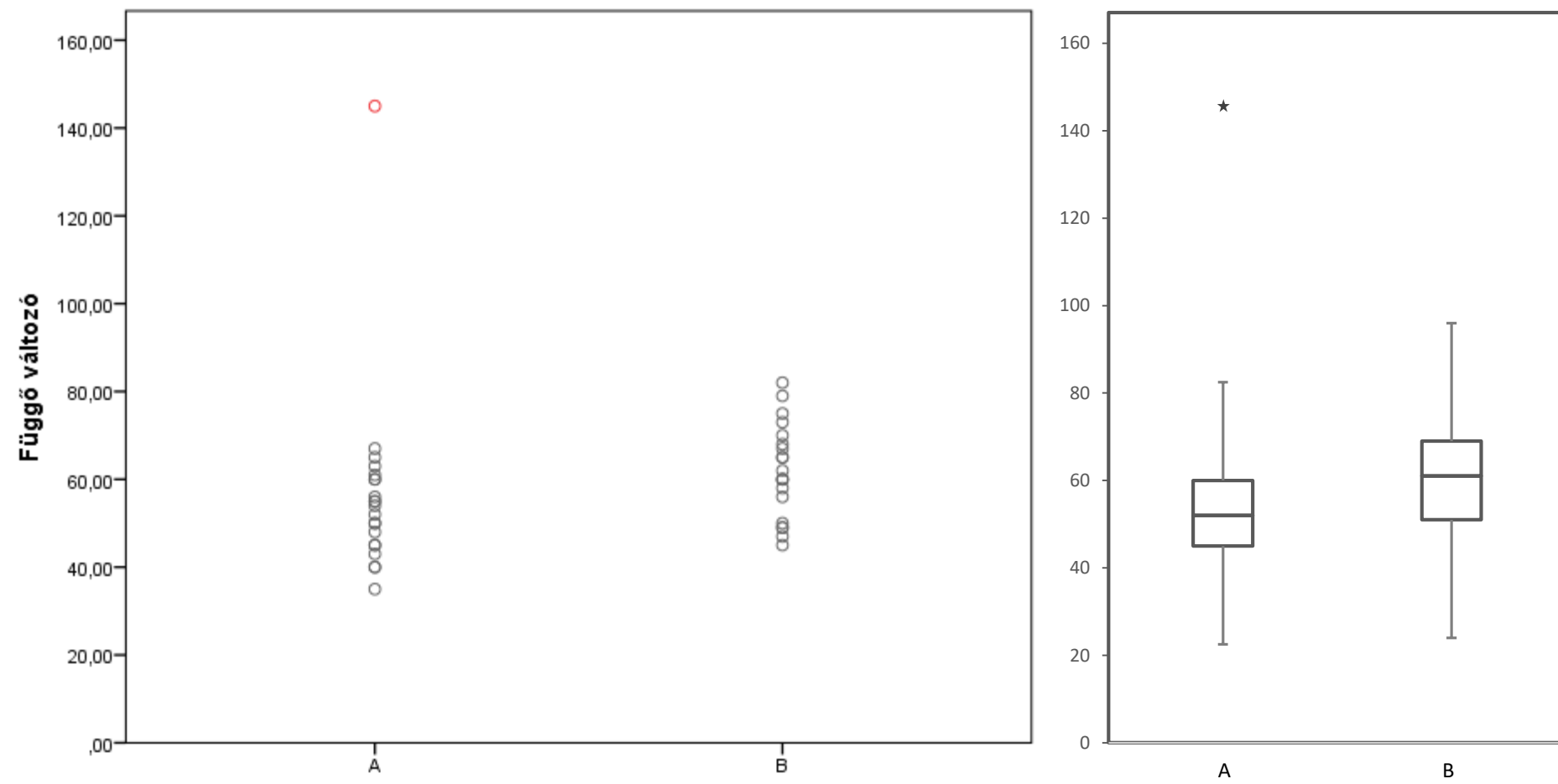
csoport	Mean	Std. Deviation	N
A	53,7619	11,32654	21
B	62,0000	10,71644	20
Total	57,7805	11,66515	41



Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	617,166 ^a	1	617,166	4,402	,042	,101
Intercept	138408,386	1	138408,386	987,219	,000	,962
group	617,166	1	617,166	4,402	,042	,101
Error	5467,810	39	140,200			
Total	144125,000	41				
Corrected Total	6084,976	40				

a. R Squared = ,101 (Adjusted R Squared = ,078)

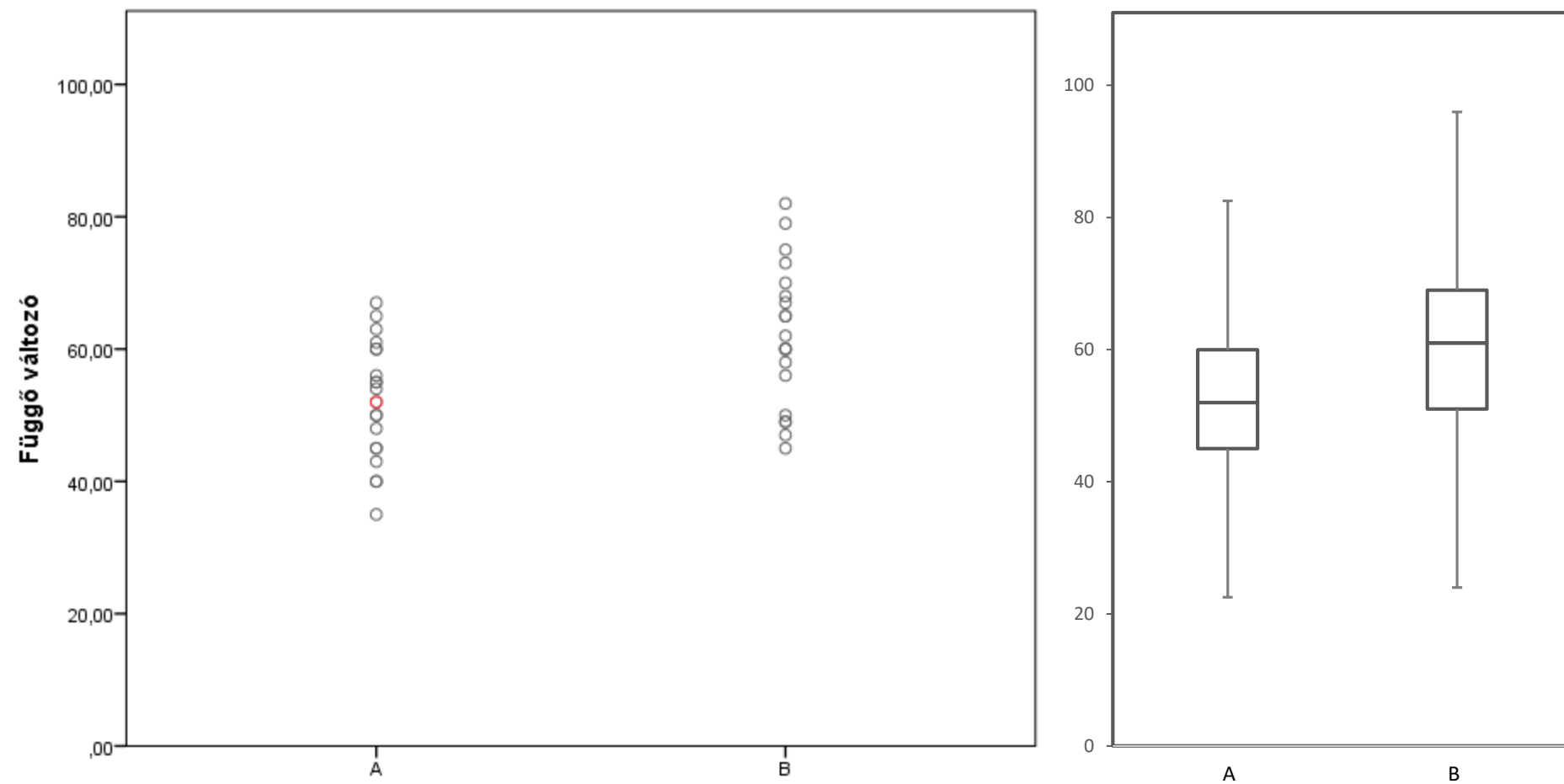
csoport	Mean	Std. Deviation	N
A	54,2381	12,81758	21
B	62,0000	10,71644	20
Total	58,0244	12,33387	41



Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	296,609 ^a	1	296,609	,970	,331	,024
Intercept	144136,609	1	144136,609	471,392	,000	,924
group	296,609	1	296,609	,970	,331	,024
Error	11924,952	39	305,768			
Total	156125,000	41				
Corrected Total	12221,561	40				

a. R Squared = ,024 (Adjusted R Squared = -,001)

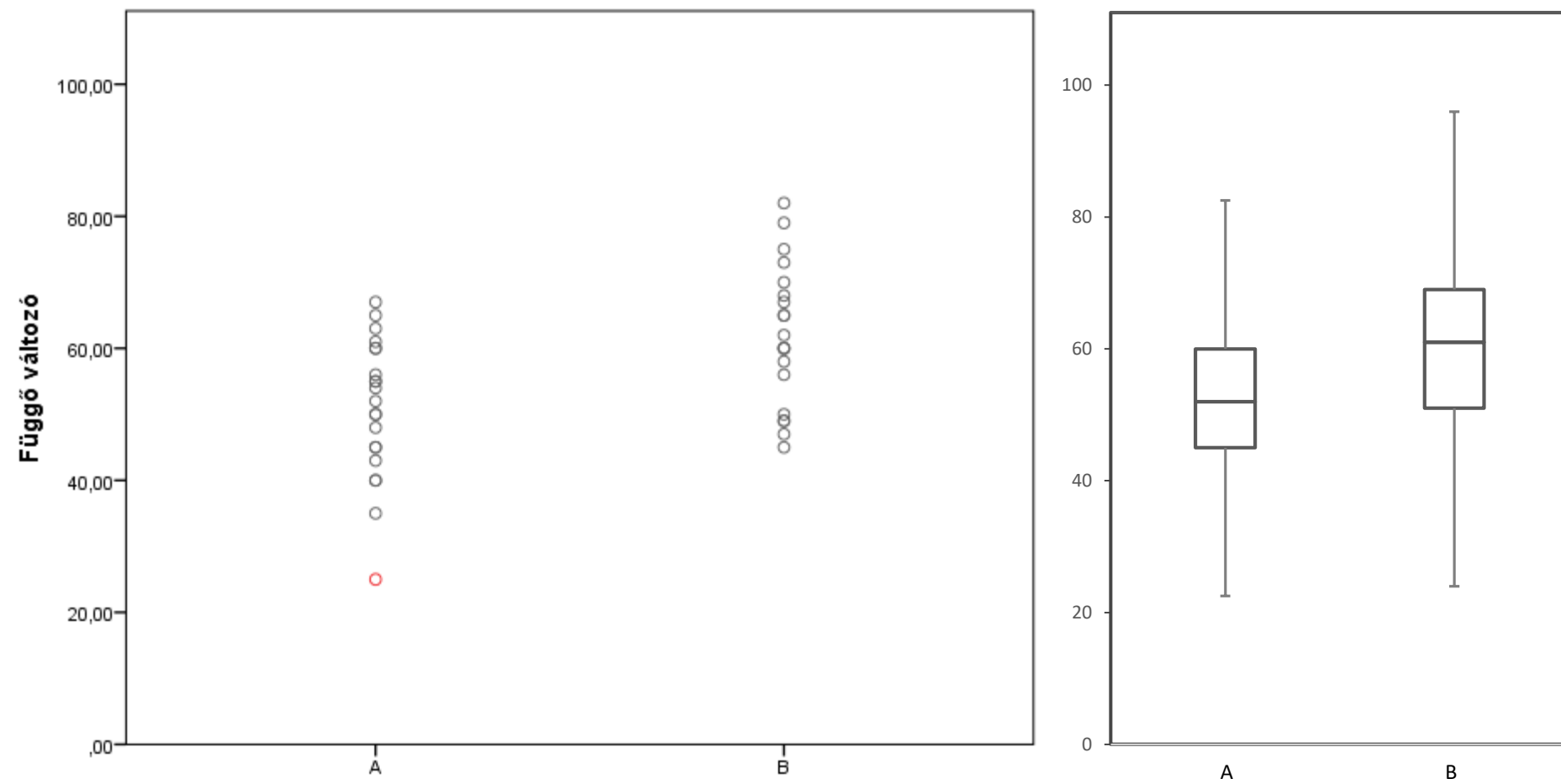
csoport	Mean	Std. Deviation	N
A	56,6190	22,07142	21
B	62,0000	10,71644	20
Total	59,2439	17,47967	41



Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	985,738 ^a	1	985,738	10,325	,003	,209
Intercept	133575,006	1	133575,006	1399,165	,000	,973
group	985,738	1	985,738	10,325	,003	,209
Error	3723,238	39	95,468			
Total	137804,000	41				
Corrected Total	4708,976	40				

a. R Squared = ,209 (Adjusted R Squared = ,189)

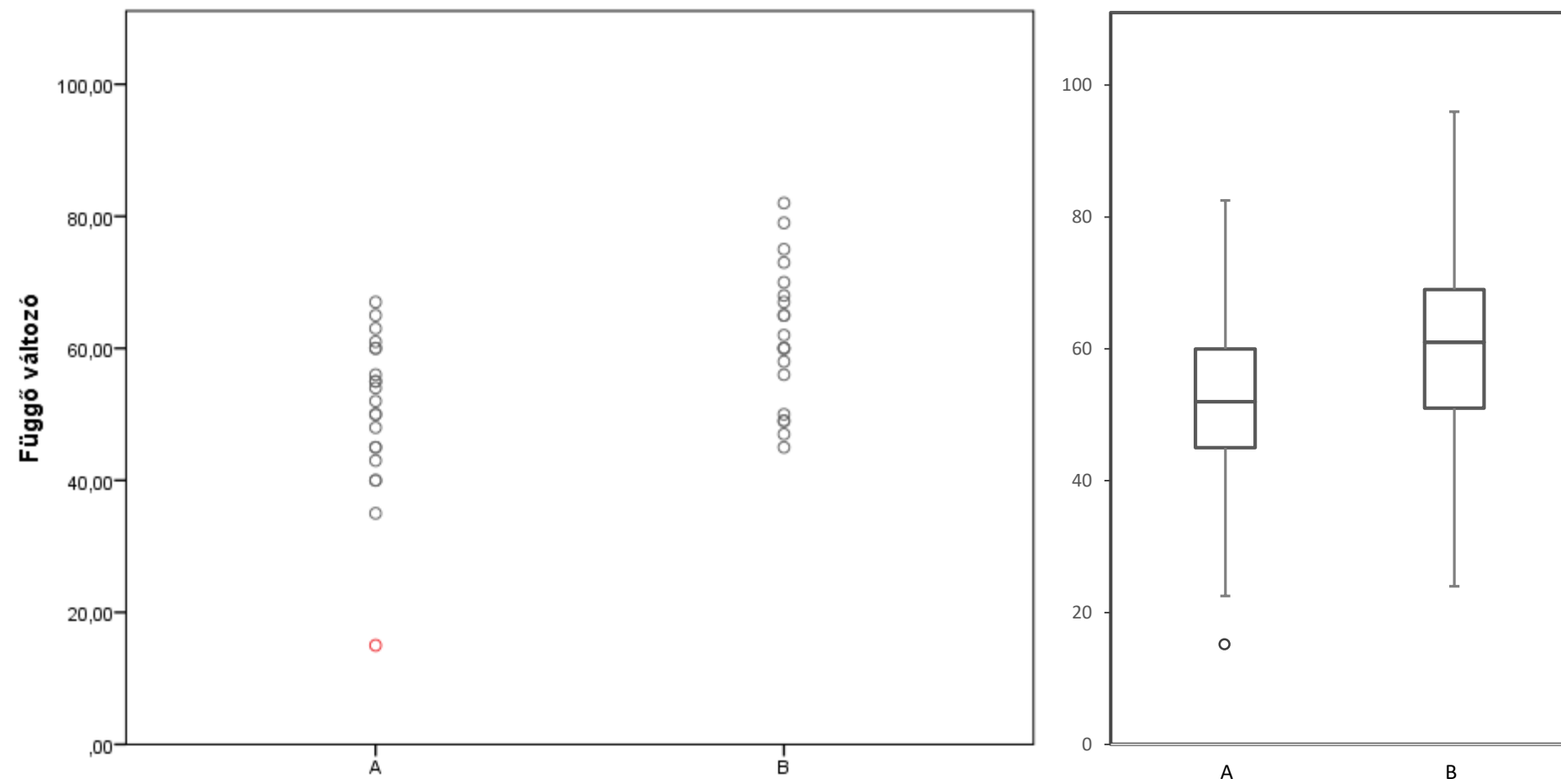
csoport	Mean	Std. Deviation	N
A	52,1905	8,77849	21
B	62,0000	10,71644	20
Total	56,9756	10,85009	41



Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	1261,069 ^a	1	1261,069	11,107	,002	,222
Intercept	130583,995	1	130583,995	1150,180	,000	,967
group	1261,069	1	1261,069	11,107	,002	,222
Error	4427,810	39	113,534			
Total	135725,000	41				
Corrected Total	5688,878	40				

a. R Squared = ,222 (Adjusted R Squared = ,202)

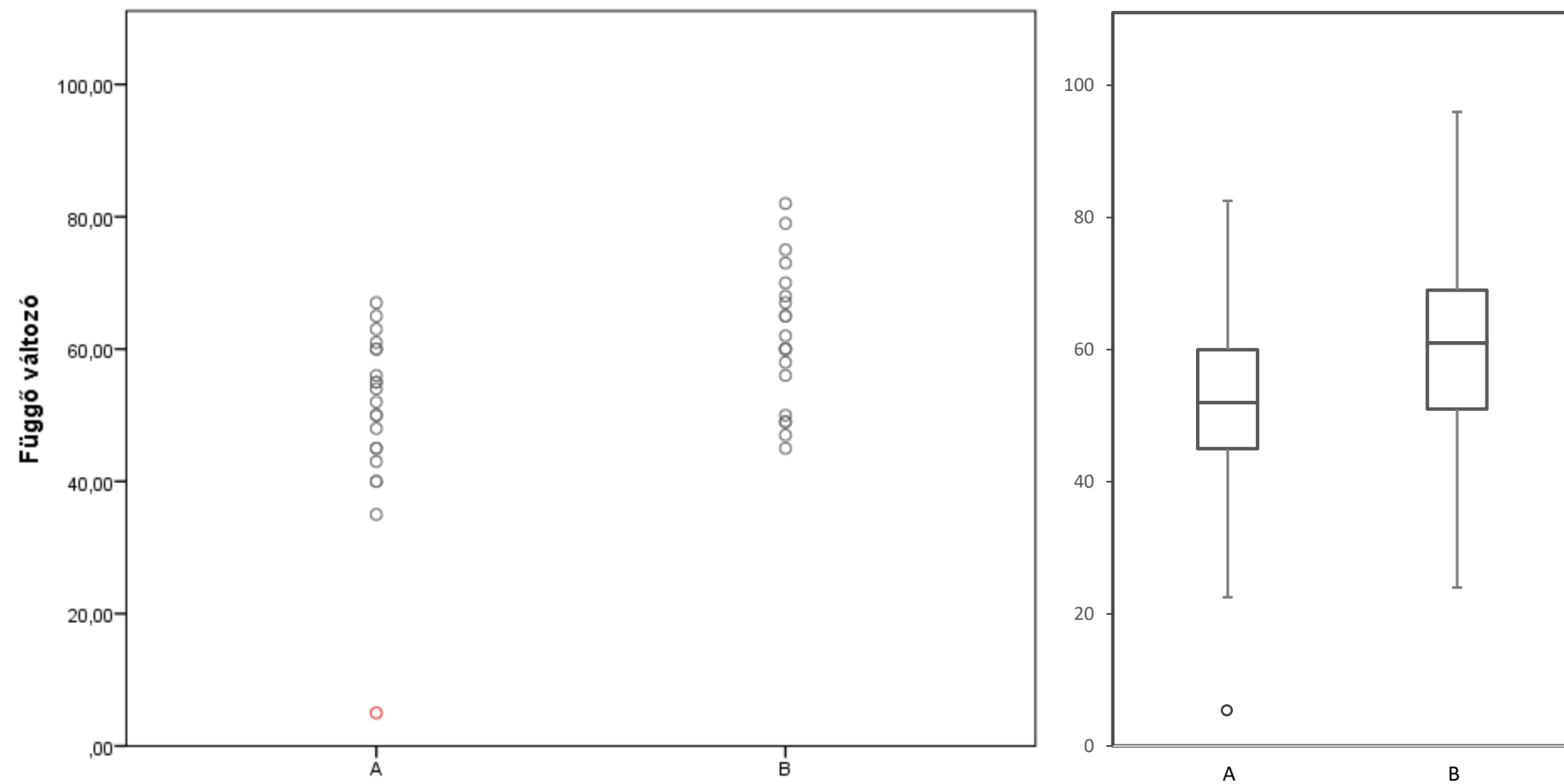
csoport	Mean	Std. Deviation	N
A	50,9048	10,59672	21
B	62,0000	10,71644	20
Total	56,3171	11,92568	41



Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	1371,638 ^a	1	1371,638	10,611	,002	,214
Intercept	129484,808	1	129484,808	1001,739	,000	,963
group	1371,638	1	1371,638	10,611	,002	,214
Error	5041,143	39	129,260			
Total	135325,000	41				
Corrected Total	6412,780	40				

a. R Squared = ,214 (Adjusted R Squared = ,194)

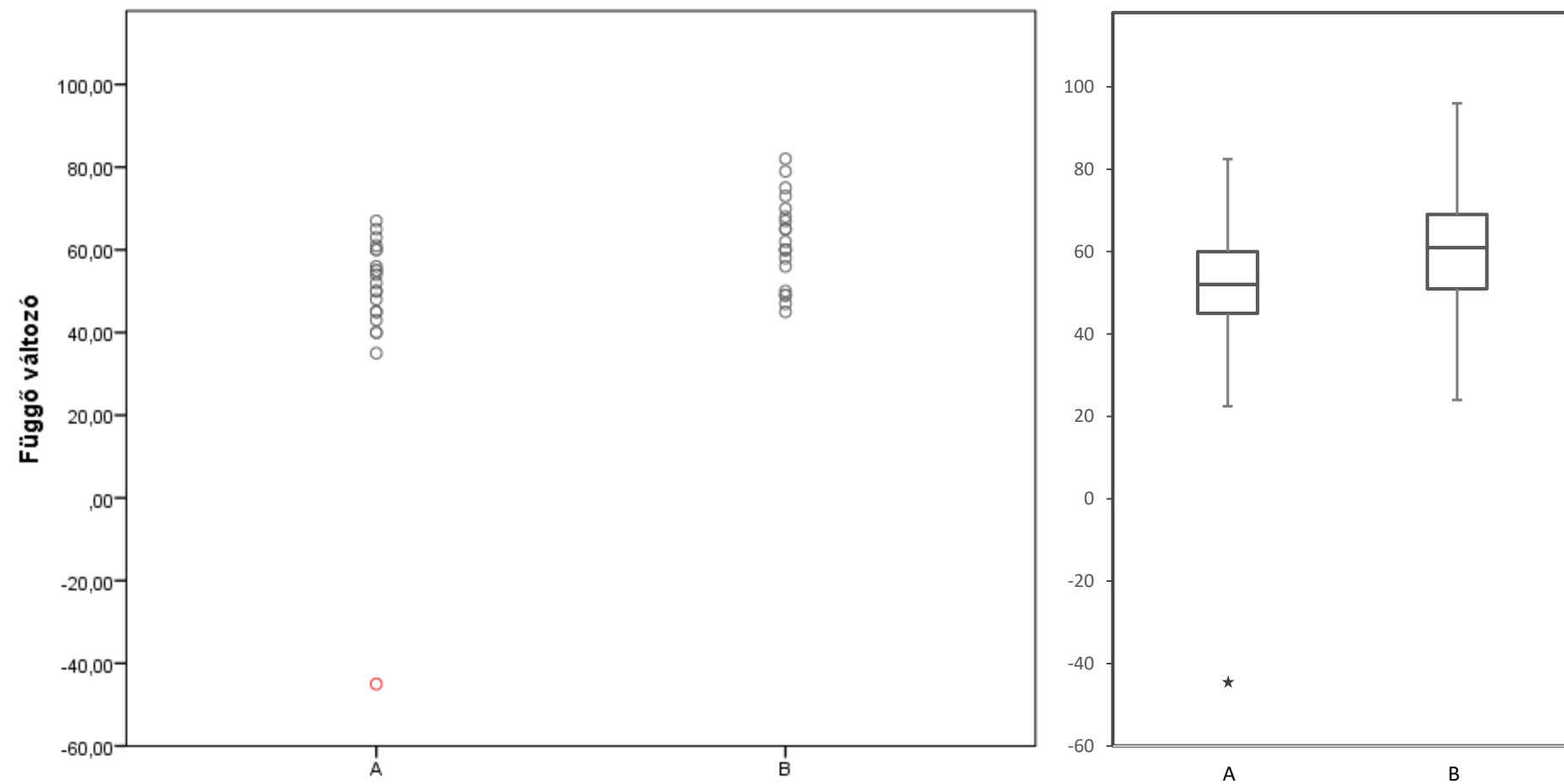
csoport	Mean	Std. Deviation	N
A	50,4286	11,95647	21
B	62,0000	10,71644	20
Total	56,0732	12,66173	41



Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	1486,852 ^a	1	1486,852	9,921	,003	,203
Intercept	128390,267	1	128390,267	856,674	,000	,956
group	1486,852	1	1486,852	9,921	,003	,203
Error	5844,952	39	149,871			
Total	135125,000	41				
Corrected Total	7331,805	40				

a. R Squared = ,203 (Adjusted R Squared = ,182)

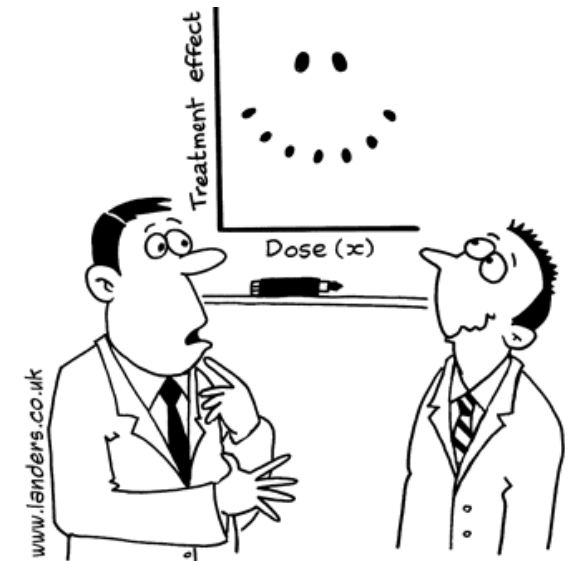
csoport	Mean	Std. Deviation	N
A	49,9524	13,53320	21
B	62,0000	10,71644	20
Total	55,8293	13,53865	41



Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	csoport	Mean	Std. Deviation	N
Corrected Model	2132,613 ^a	1	2132,613	6,538	,015	,144	A	47,5714	22,95555	21
Intercept	122987,247	1	122987,247	377,050	,000	,906	B	62,0000	10,71644	20
group	2132,613	1	2132,613	6,538	,015	,144	Total	54,6098	19,27029	41
Error	12721,143	39	326,183							
Total	137125,000	41								
Corrected Total	14853,756	40								

a. R Squared = ,144 (Adjusted R Squared = ,122)

Többdimenziós outlierek

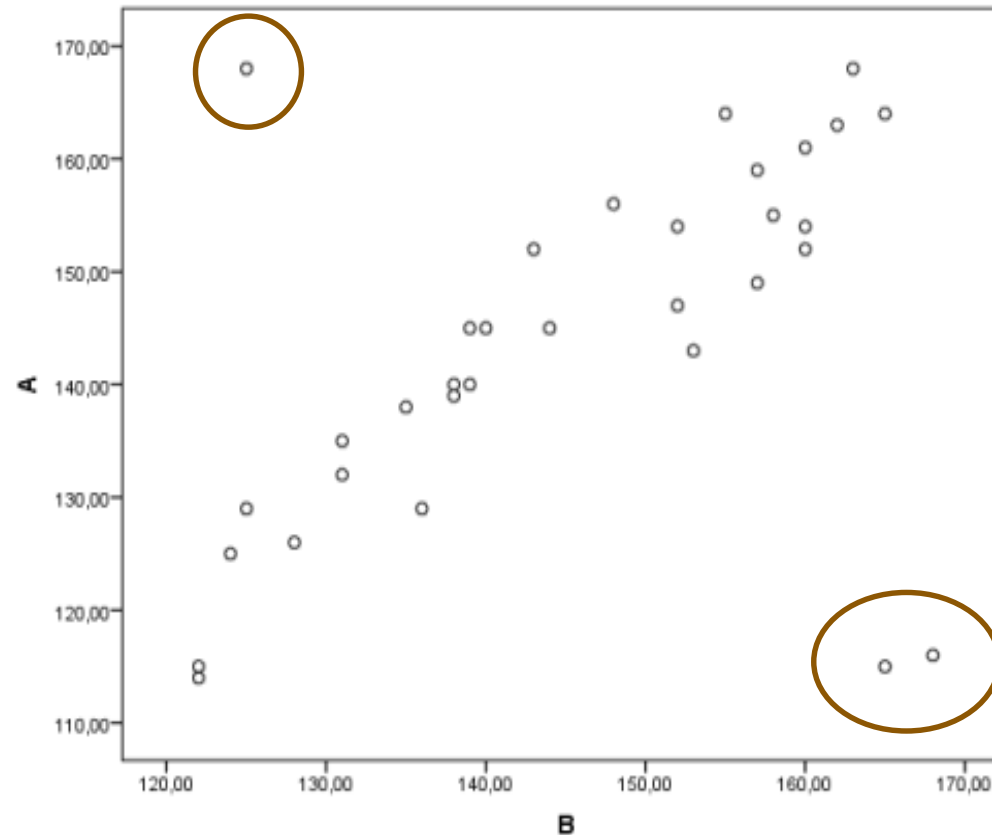


"It's a non-linear pattern with outliers.....but for some reason I'm very happy with the data."

Többdimenziós outlierek

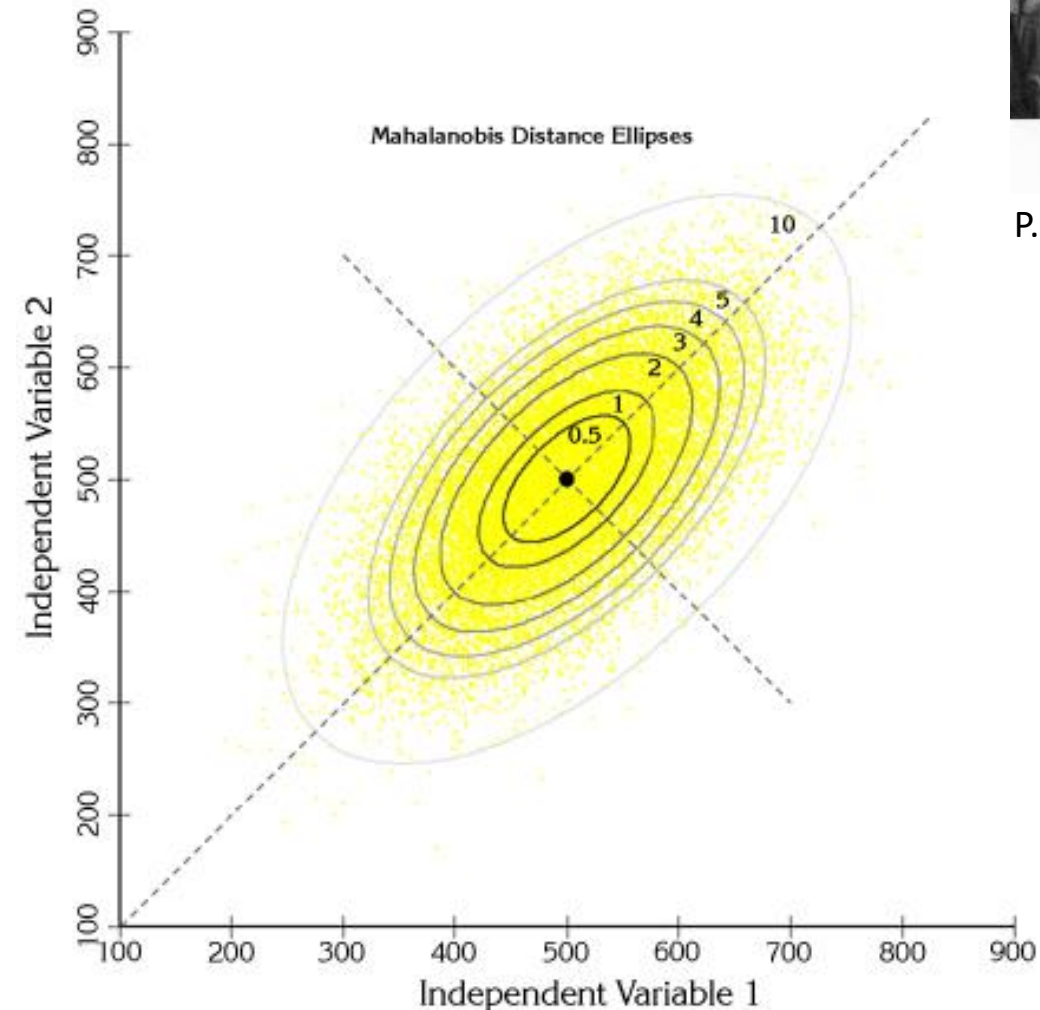
- **Többdimenziós outlierek**

- Többdimenziós outlier külön-külön egyik dimenzió belül sem szélsőséges érték, csak az adott kombináció szélsőséges.
- Például 50kg-os és 190cm magas ember



- **Mahalanobis távolság:**

- Egy adott érték távolsága a pontfelhő közepétől, figyelembe véve a kovarianciát és a varianciák különbözőségét.
- **Megadja, hogy egy adott pont hány szórásnyi távolságra van egy többdimenziós pontfelhő közepétől.** (Minden egyes adatponthoz tartozik egy Mahalanobis távolság érték)
- Mivel a távolságot szórás alapján adja meg, **nincs mértékegysége és nem függ a változók skálázástól.**
- Egy köríven (vagy többdimenziós esetben egy héjon) elhelyezkedő pontok mind azonos szórásnyra vannak a középponttól.
- Főkomponensek mentén számolódik.
- A távolságok négyzete χ^2 eloszlást mutat.



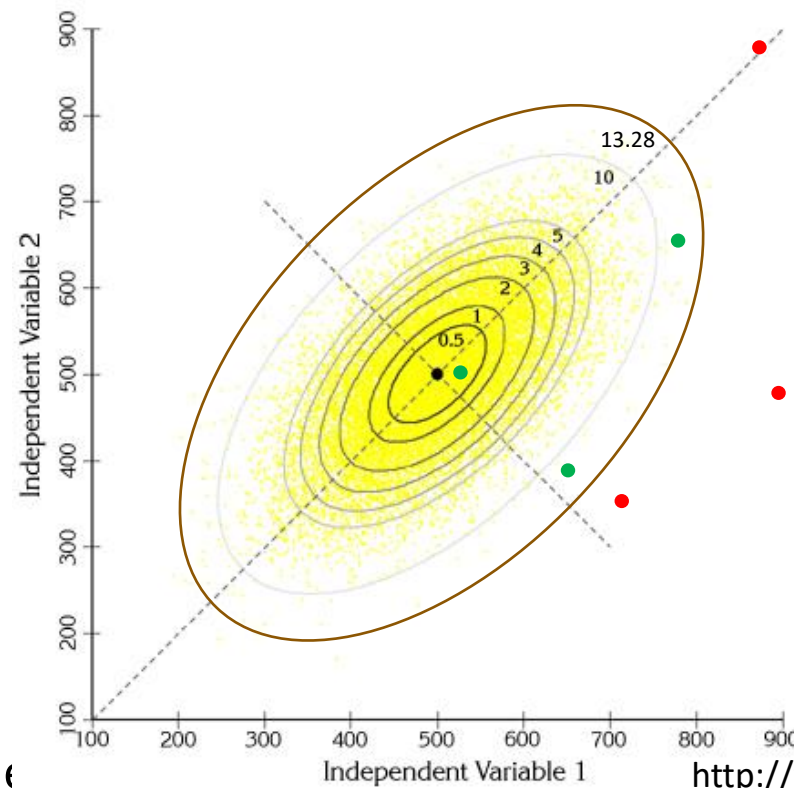
P. C. Mahalanobis

P. C. Mahalanobis

- **Mahalanobis távolság:**

- Outlier egy adat, ha a hozzá tartozó Mahalanobis távolság meghaladja a kijelölt határértéket
- A határ pontos értéke függ az elemszámtól és a változók mennyiségétől, valamint attól, mennyire szigorú alfa kritérium szint mellett dolgozunk.

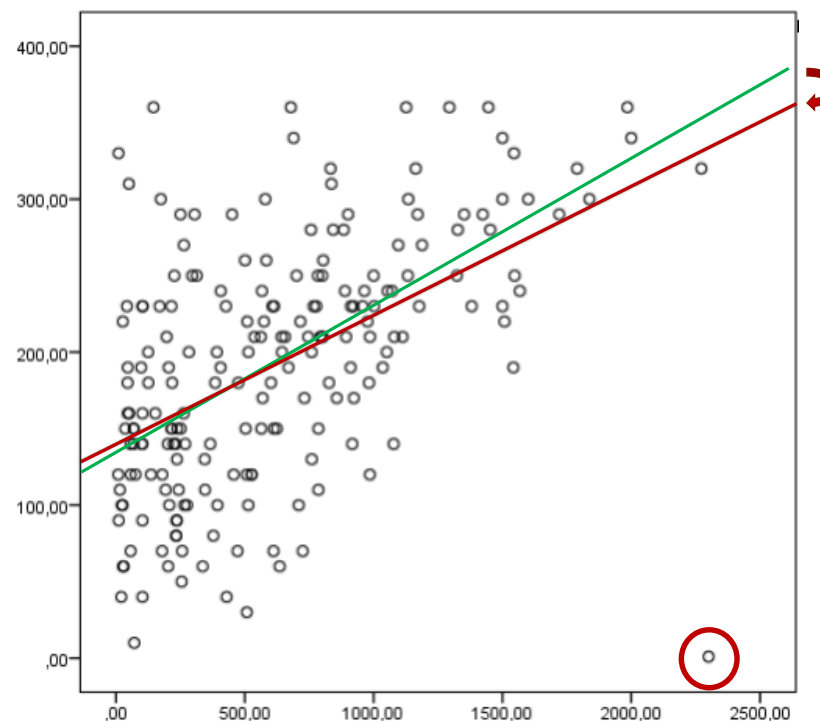
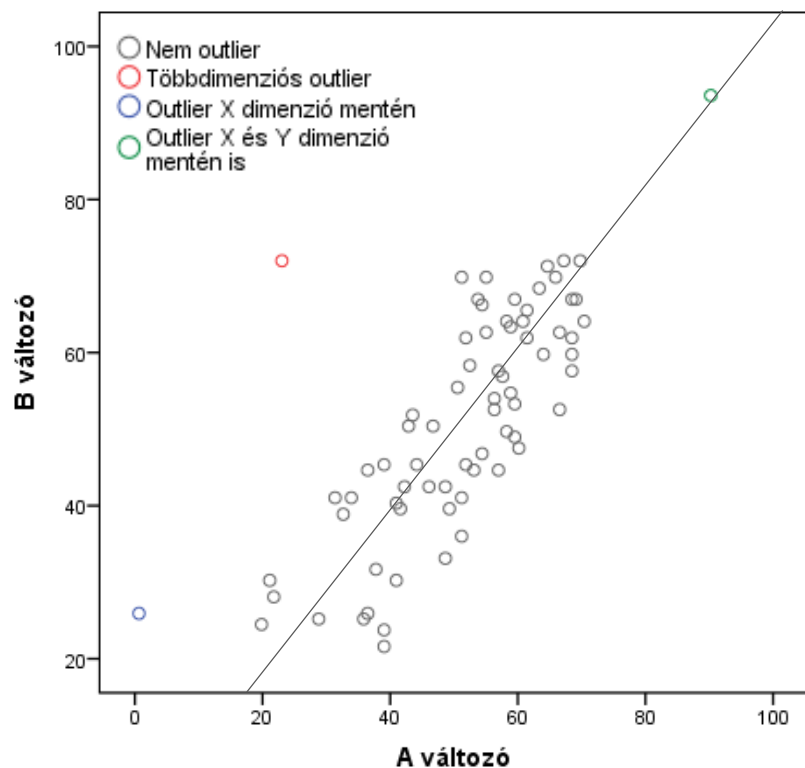
Változók száma	Határérték $p < .001$ szinten
2 (korreláció)	13.28
3	16.27
4	18.47
5	20.52
6	22.46
7	24.32
8	26.13
9	27.88
10	29.59



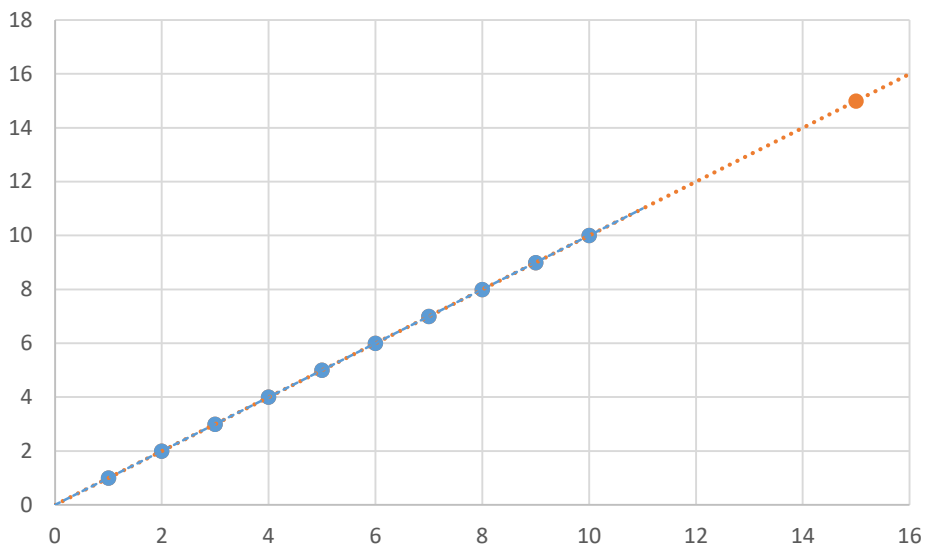
- Vegyük észre, egy a modellünket alátámasztó (pontos pont) is outliernek ítélt.

Különbség az outlierek és a befolyásos értékek között

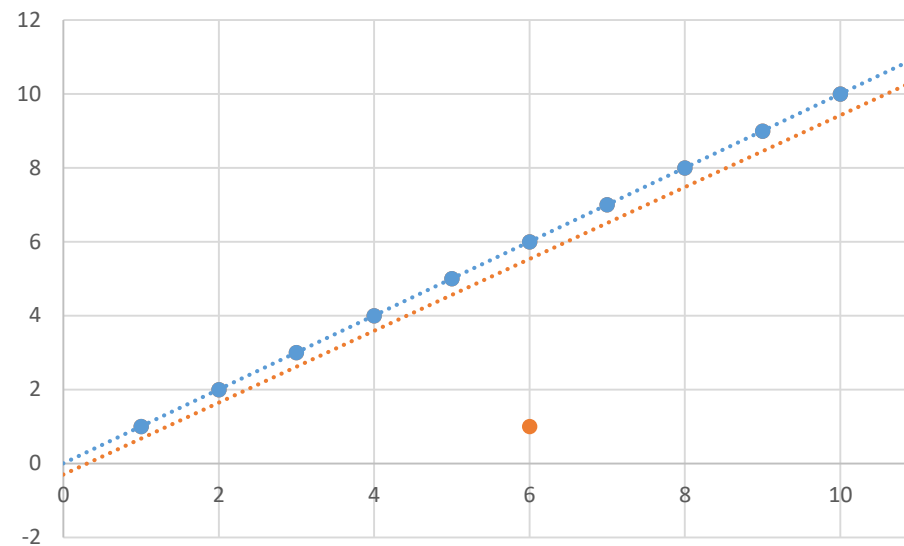
- **Outlier** olyan érték, ami nagyon eltér a többitől
- **Befolyásos érték** olyan érték, aki nagyon megváltoztatja a modellt (bővebben regresszió kapcsán)
- Az outlierek (akár egy vagy több dimenziósak) nem mindig vannak jelentős hatással a modellre, ekkor akár az elemzésben benn is hagyhatóak (a hibát azért ekkor is növelik – így rontják a magyarázóerőt). De ha a modellben is hagyjuk őket, akkor is érdemes megvizsgálni, miért térnek el a többiektől.



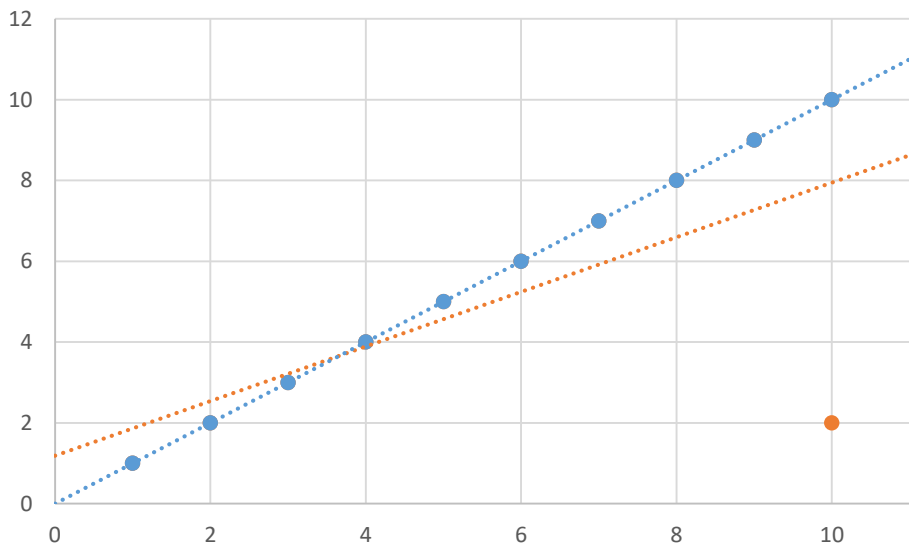
Outlierek, befolyásos értékek & bejósolható értékek



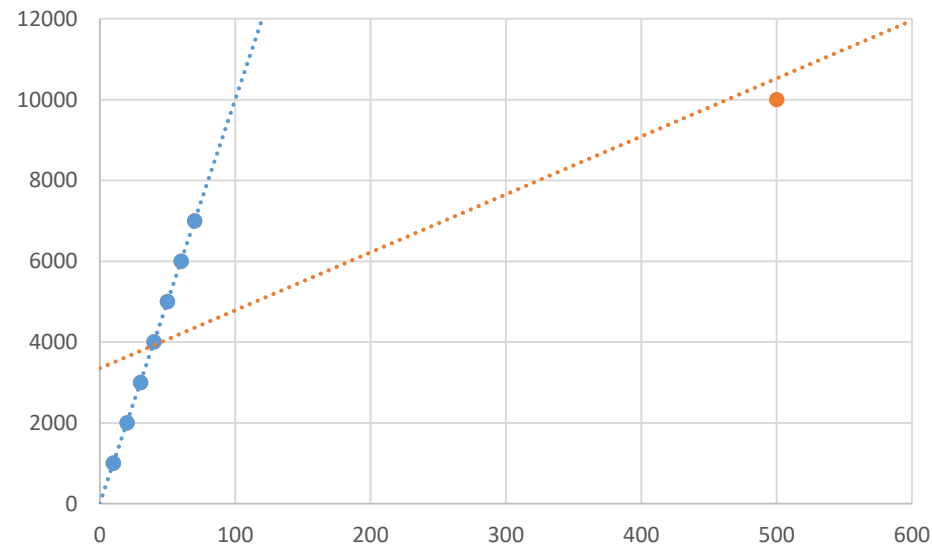
Outlier mind az X, mind az Y tengely mentén, de nem több-dimenziós outlier. Nem befolyásolja a modellt, és jól bejósolható.



Több dimenziós outlier, de nem befolyásolja a modellt, és nem jósolható jól be.

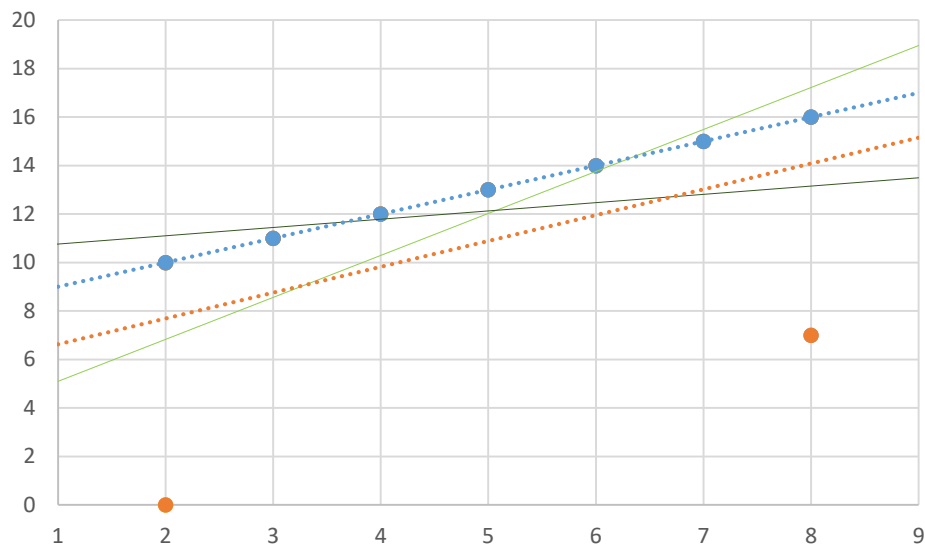


Csak több dimenzió mentén outlier, befolyásolja a modellt, és nem jósolható jól be.

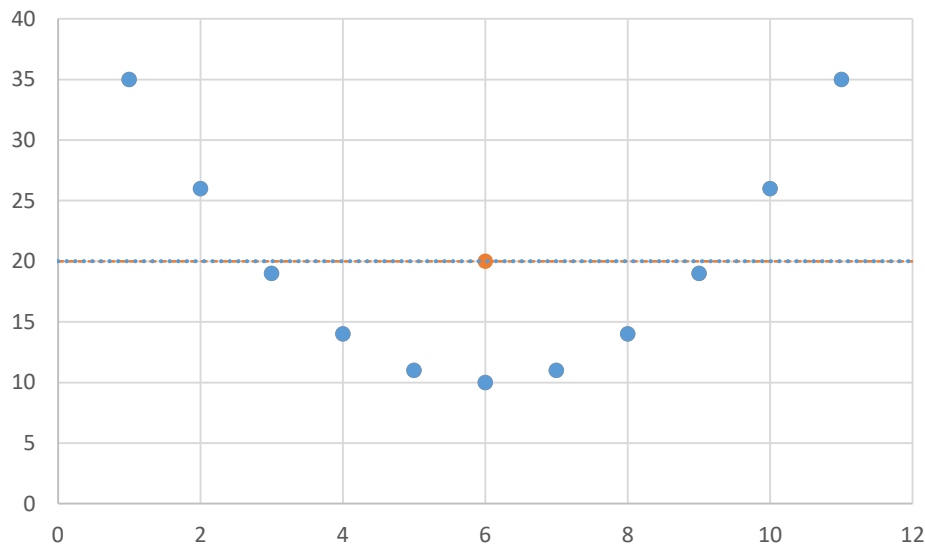


Outlier mind az X, mind az Y tengely és többdimenzió mentén is. Erősen befolyásolja a modellt, önmaga jól bejósolható, de a többi érték bejósolhatóságát nagyban csökkenti. Példa: londoni pubok és bűnözés kapcsolata - Dr. Richard Roberts példája via Andy Field.

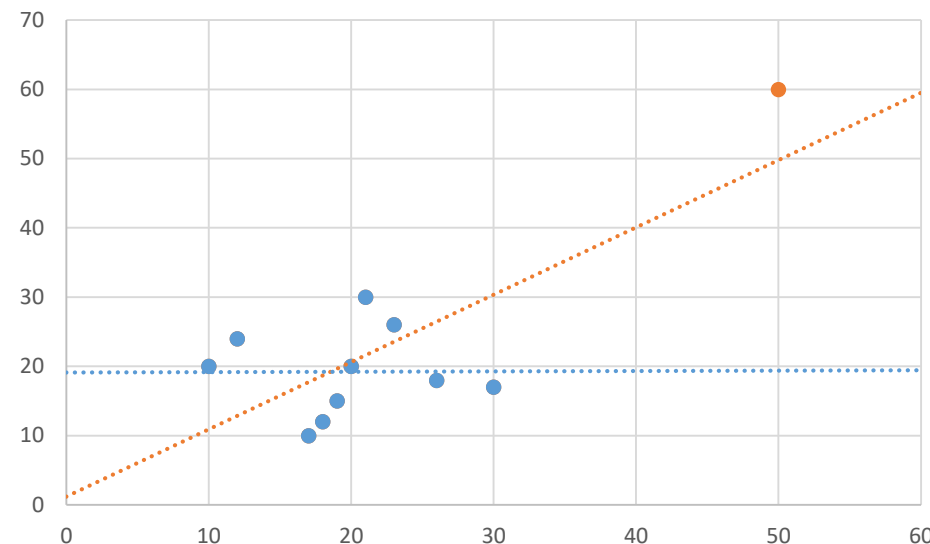
Outlierek, befolyásos értékek & bejósolható értékek



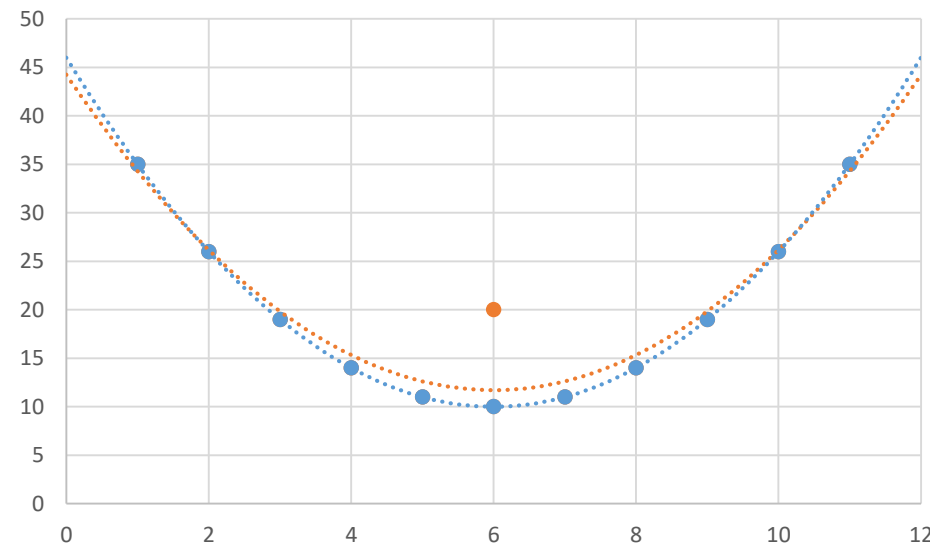
Mindkettő outlier Y mentén, a modellt közösen mégsem befolyásolják annyira jelentősen – egyenként annál inkább. Ki kell venni őket, mert így túl nagy a véletlen mintavételezés szerepe.



Félretéve, hogy láthatóan nagyon rossz modellt választottunk, nem outlier sem X, sem Y mentén, de több dimenzióban igen. A modellt nem befolyásolja és jól bejósolható.



Outlier X és Y mentén is. Így lehet egy nem létező kapcsolatot létezőnek látni.



A helyes modell ugyanazon problémára. A modellt most sem befolyásolja jelentősen, de nem jól bejósolható.