

Többváltozós statisztika, regressziós eljárások (BMNPS07600M)

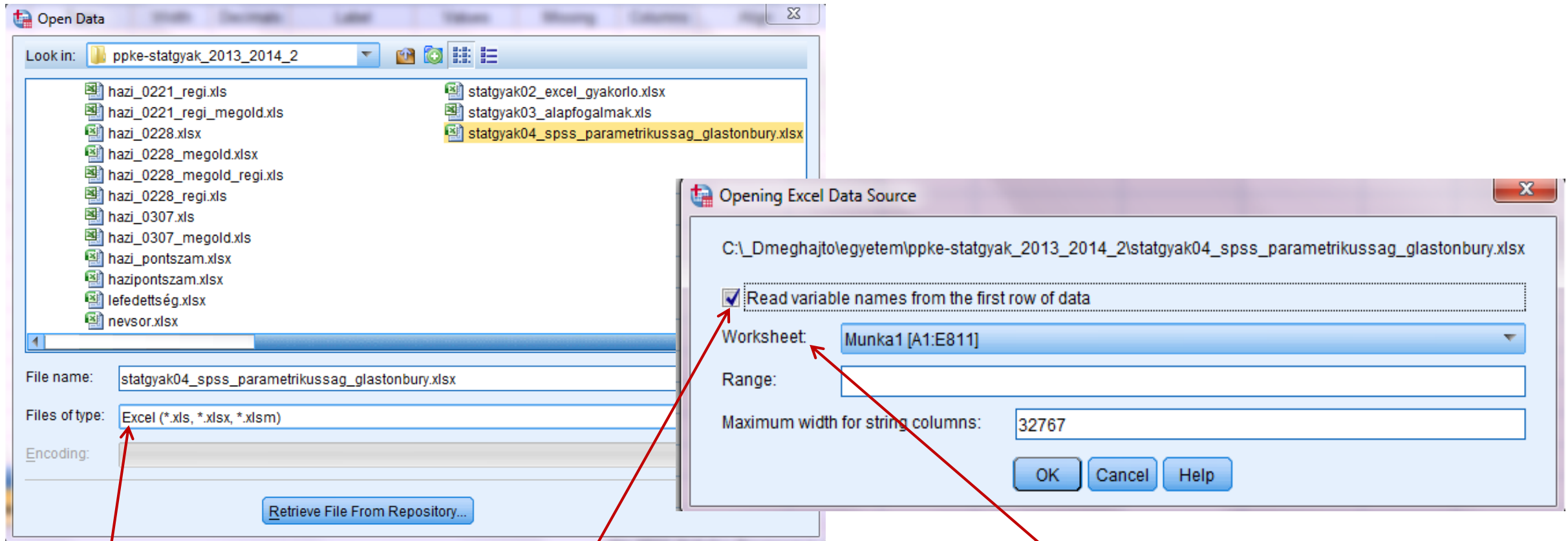
Készítette: Soltész-Várhelyi Klára

Adattisztítás

Parametrikusság & egyéb feltételek

Excel fájl beolvasása

Példa: tobbval03_adatfeldolgozas_glastonbury.xlsx (Andy Field módosított példája)



Be kell állítani, hogy az Excel
Fájlok látszódjanak

Ha az Excelben az első sor a változóneveket
tartalmazza, be kell pipálni

Meg lehet adni, melyik munkafüzet
melyik cellái tartalmazzák az adatokat

Adattisztítás, adatok felkészítése

1. adatok áttekintése

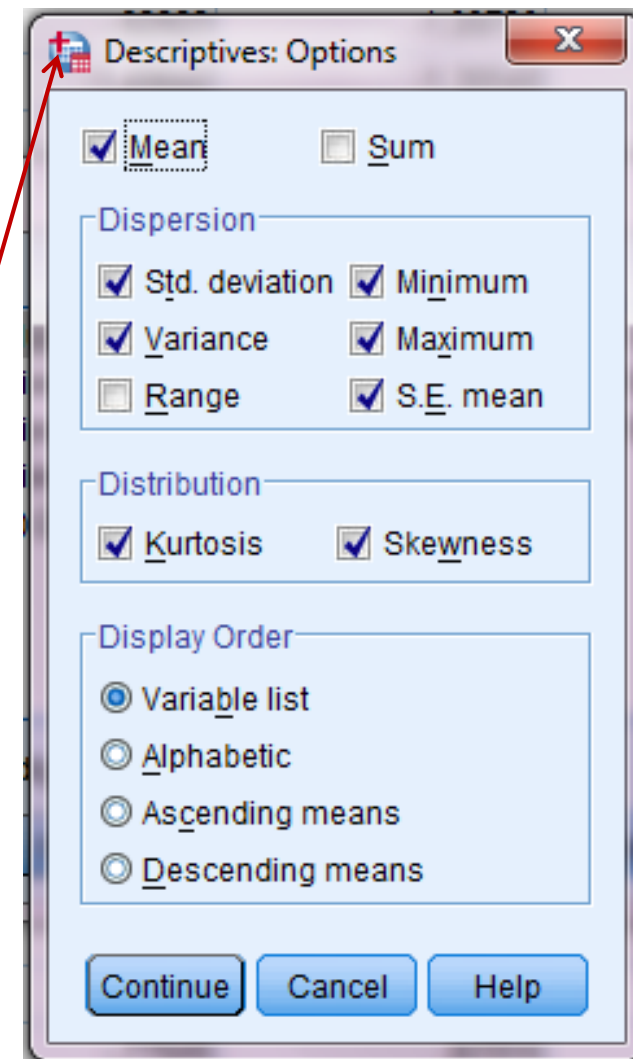
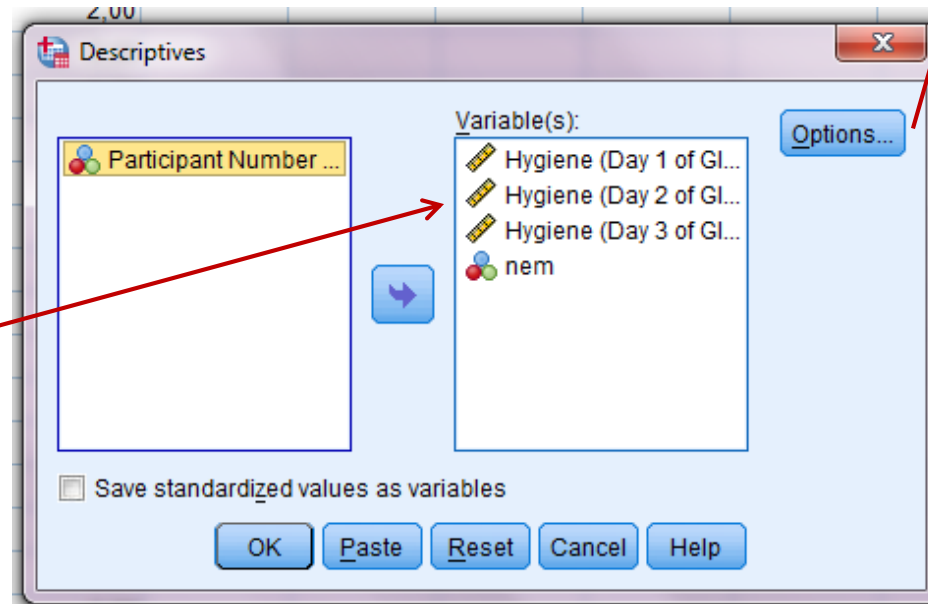
Adatok áttekintése táblázatosan

- **Adatbázis áttekintése**

- Változók: azonosító, higiénia az 1. 2. és 3. napon (0-tól 5-ig folytonos), nem
- Általános kép az adatainkról:
 - Analyze / Descriptive Statistics / Descriptives
DESCRIPTIVES VARIABLES=day1 day2 day3 nem
/STATISTICS=**MEAN STDDEV MIN MAX SEMEAN KURTOSIS SKEWNESS**.
- Lépéseket összefoglaló folyamatra: adatfeldolgozas_flowchart.pdf

-  Olvasd el a hiányzó értékekről szóló részt is!

Minden számunkra
érdekes változót
beteszünk



Adatok áttekintése táblázatosan

Descriptive Statistics

	N	Minimum	Maximum	Mean		Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Std. Error	Statistic	Std. Error
Hygiene (Day 1 of Glastonbury Festival)	810	,02	20,02	1,7983	,03357	,95534	8,648	,086	162,942	,172
Hygiene (Day 2 of Glastonbury Festival)	264	,00	3,44	,9609	,04436	,72078	1,095	,150	,822	,299
Hygiene (Day 3 of Glastonbury Festival)	123	,02	3,41	,9765	,06404	,71028	1,033	,218	,732	,433
Nem	810	1,00	2,00	1,5049	,01758	,50028	-,020	,086	-2,005	,172
Valid N (listwise)	123									

Elemszámok

Átlagok, és Standard error
Leellenőrizhetjük az RSE-t

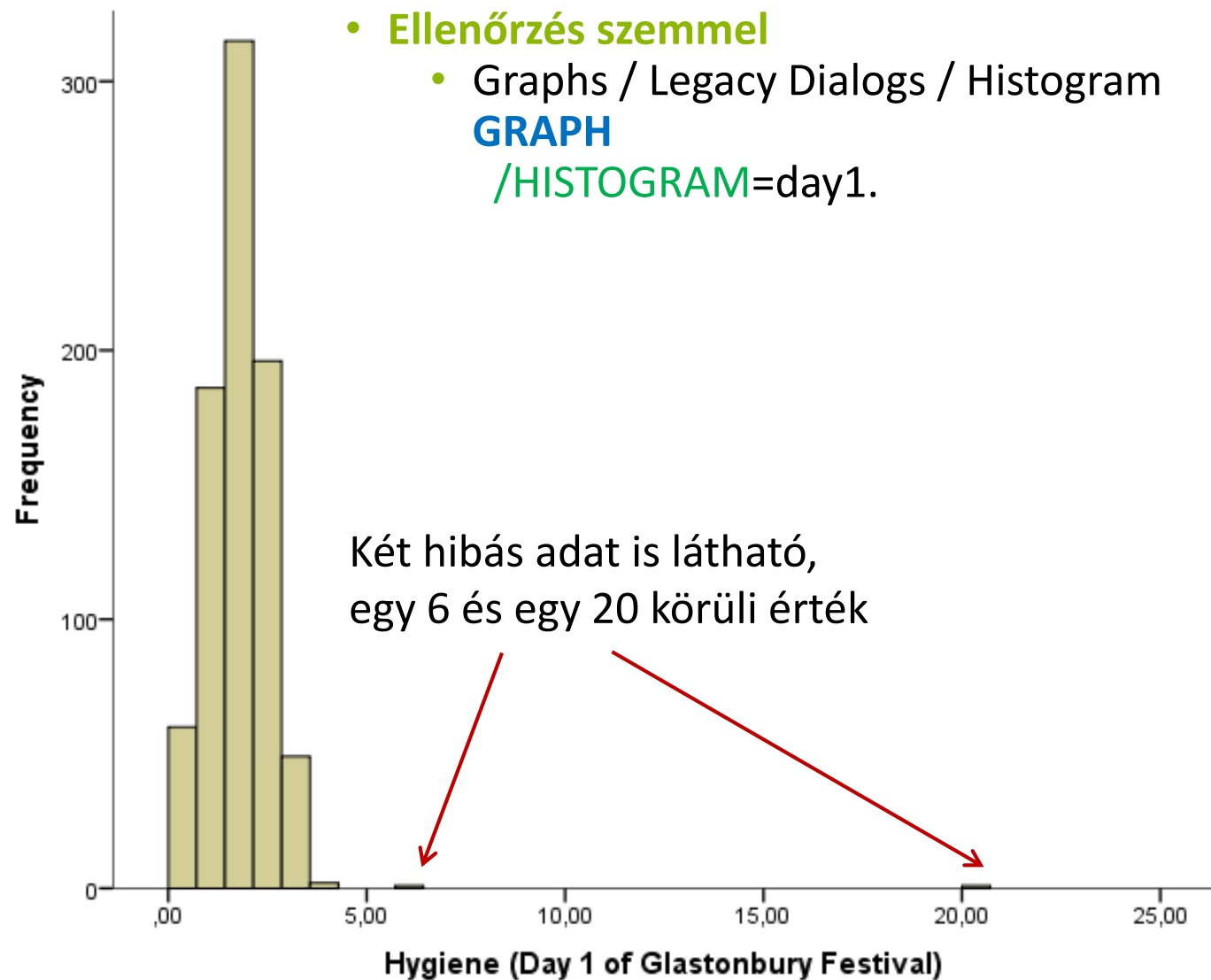
Minimum/maximum

Feltűnik, hogy az 0-tól 5-ig terjedő skála maximuma 20.02!
Természetesen a hiba megtalálása után újra kell kérni a
deskriptív statisztikákat is!!!!

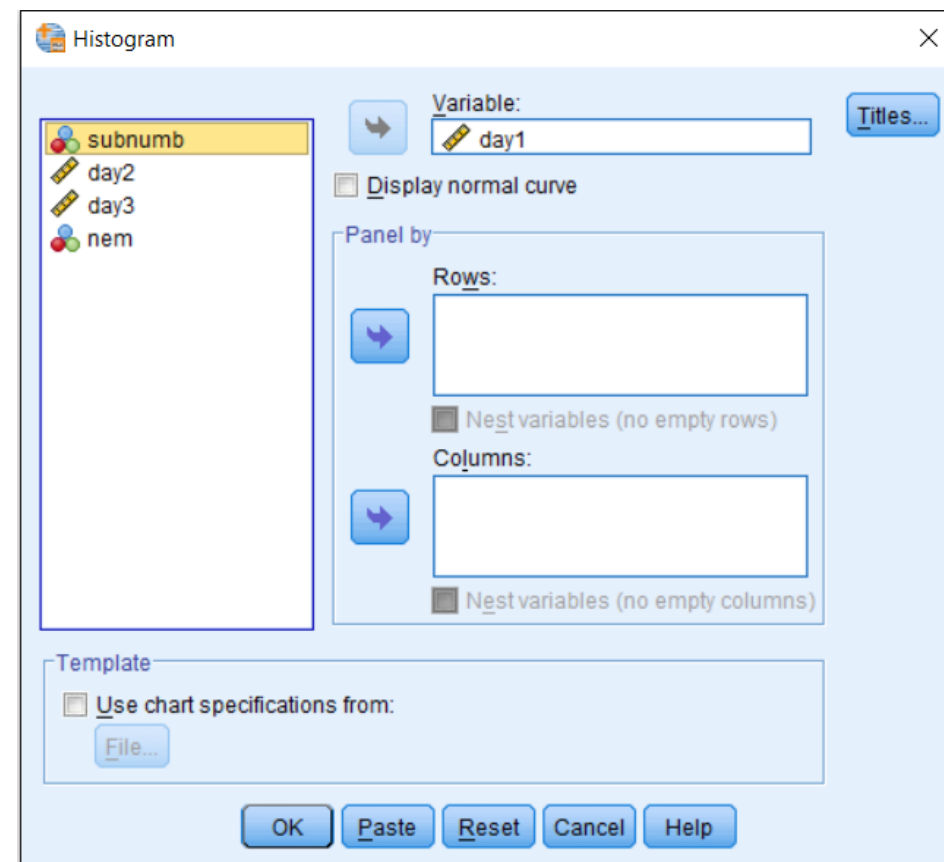
Szórás

Ferdeség/csúcsosság
Később visszatérünk rá

Adatok áttekintése grafikusan: Hisztogram



- Ellenőrzés szemmel
 - Graphs / Legacy Dialogs / Histogram
- GRAPH**
/HISTOGRAM=day1.



Adattisztítás, adatok felkészítése

2. Hibák

3. Outlierek

Hibás értékek megtalálása

- **Hibák megtalálása:**

- Elgépelések, téves adatok, komolytalan kitöltők
- 20-as és 6-os érték egy 0-tól 5-ig terjedő skálán? Gépelési hiba.

- Hogy találjuk meg?

- Analyze / Descriptive Statistics / Explore

EXAMINE **VARIABLES**=day1

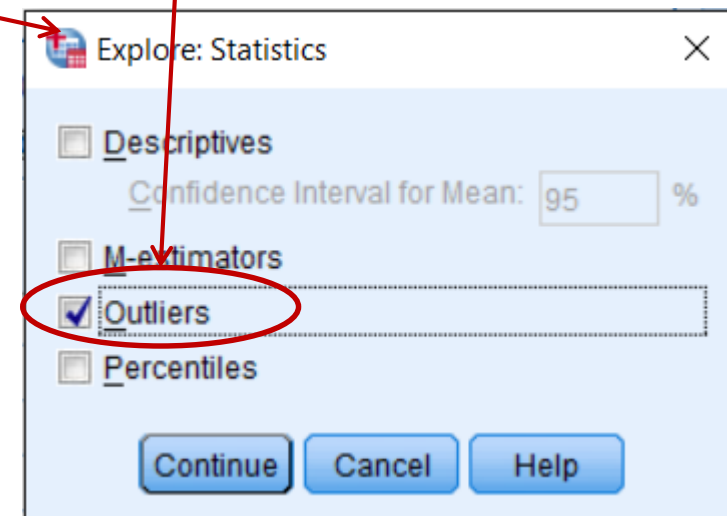
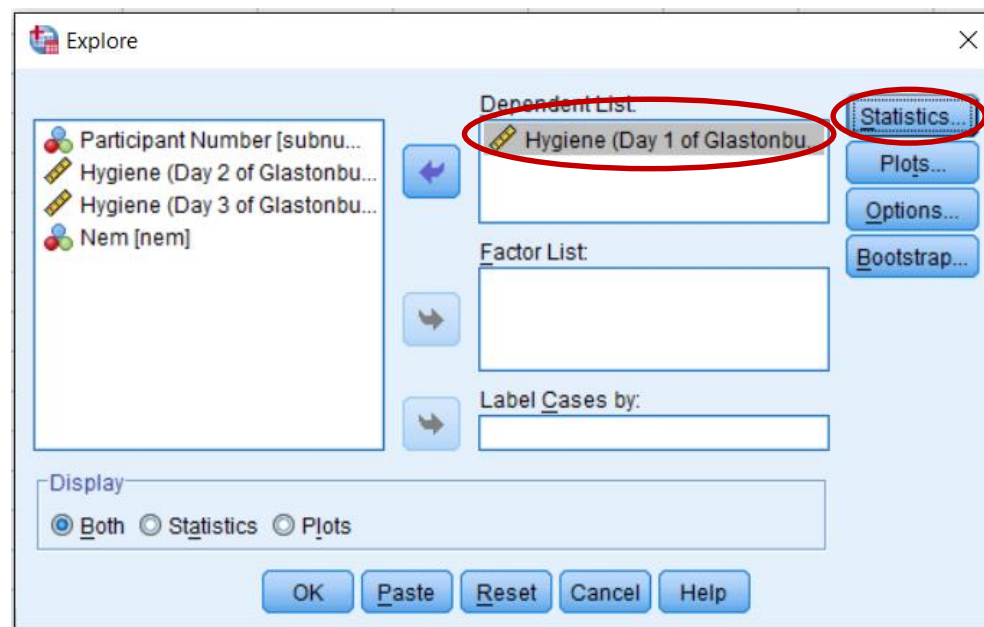
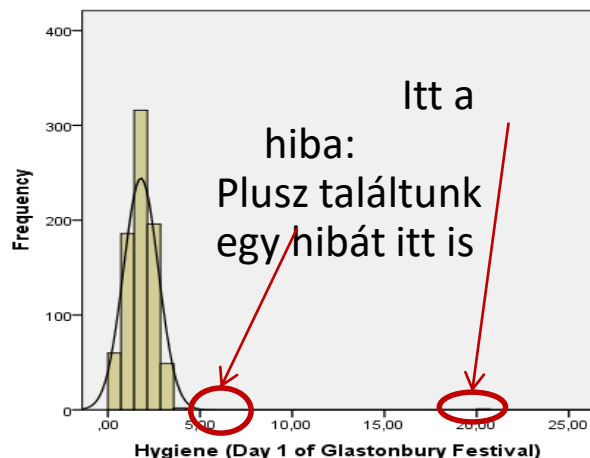
/PLOT NONE

/STATISTICS EXTREME ←

/MISSING PAIRWISE

/NOTOTAL.

Az Outliers elnevezés nagyon félrevezető. A menü **NEM az outliereket** (később megtanuljuk mi is az) listázza, hanem az **öt legmagasabb és öt legalacsonyabb értéket**, mely segít a hibák megtalálásában.



Hibás értékek megtalálása

Extreme Values

			Case Number	Value
Hygiene (Day 1 of Glastonbury Festival)	Highest	1	611	20,02
		2	49	5,88
		3	574	3,69
		4	303	3,58
		5	657	3,44
	Lowest	1	595	,02
		2	366	,05
		3	177	,11
		4	41	,11
		5	569	,23 ^a

a. Only a partial list of cases with the value ,23 are shown in the table of lower extremes.

609	4156	1,44
610	4157	1,64
611	4158	20,02
612	4159	1,20
613	4160	,38
614	4161	1,58
615	4162	1,67
616	4163	1,00

Az Extreme Values táblázatban az öt legmagasabb és öt legalacsonyabb értéket látjuk, és annak a sornak számát, melyben az érték található.

A táblázatból látjuk, hogy a keresett két hiba a 611. és a 49. sorban van.

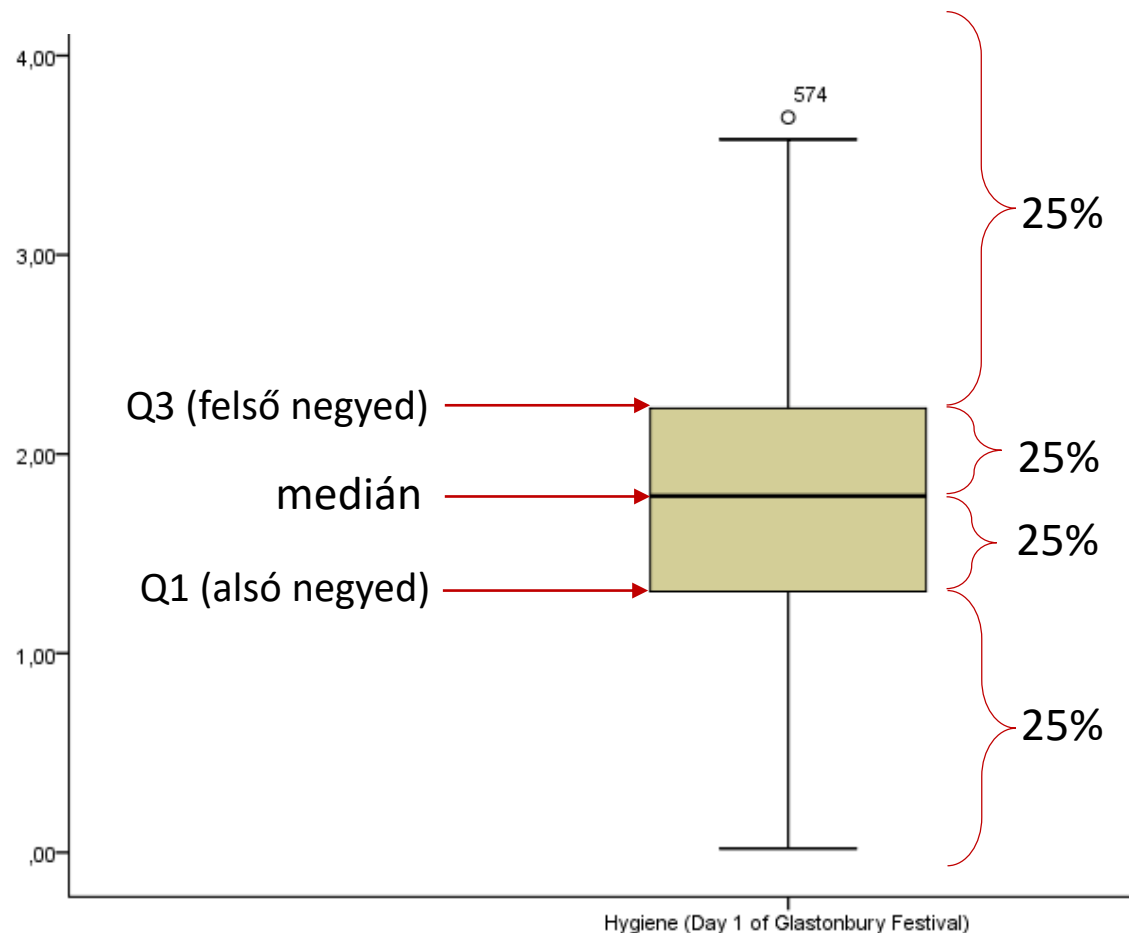
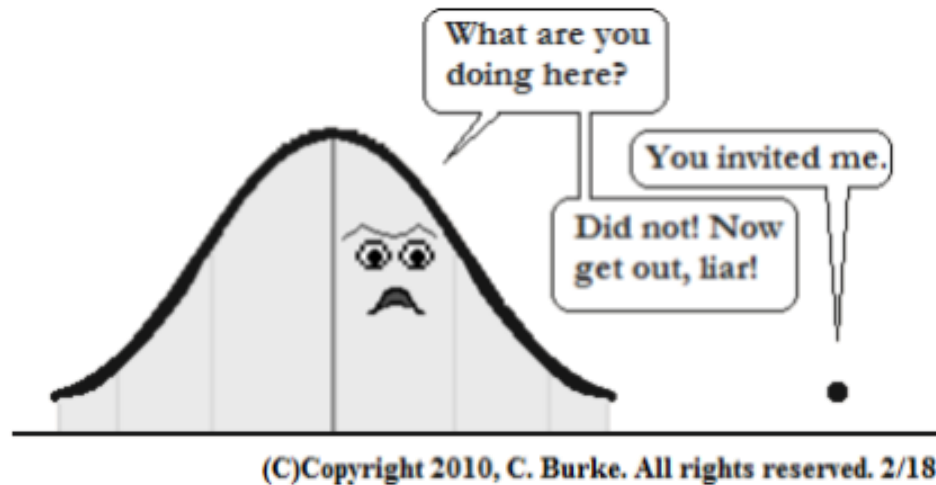
Ha vissza tudjuk keresni a helyes adatot (pl. papírról), akkor javíthatjuk, minden más esetben **TÖRÖLNI KELL**.

Ha kezeltük a hibákat, ellenőrizzük újra a leíró statisztikákat!

Adatot törölni szintaxszal körülményes, így a kurzuson nem tanuljuk, de kommentként írjuk oda, hogy manuális változtatás történt & hol, mit, mire!

Outlierek

- Általános definíció: outlier az,
 - „ami nagyon eltér a többi adatunktól”
 - Outlier labelling rule - ami távolabb az alsó/felső negyedétől, mint a középső 50% nagyságának másfélszerese
- Miért baj, ha van? Eltorzítja a statisztikákat.



Outlierek

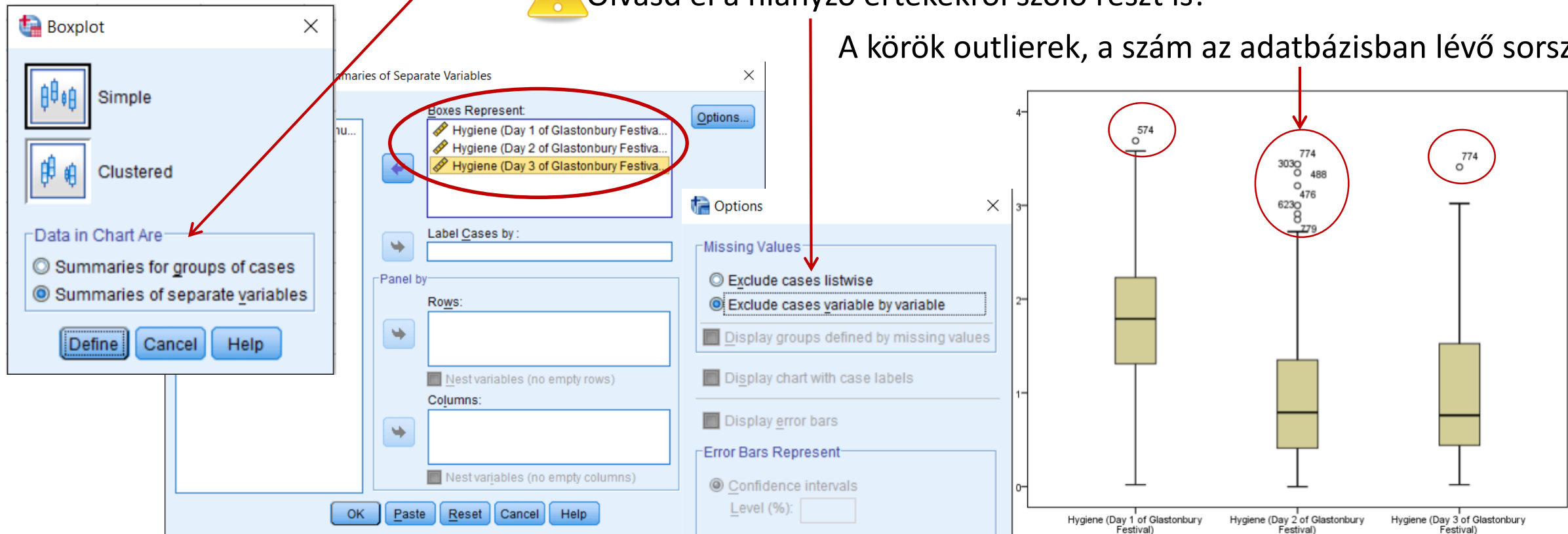
- Graphs / Legacy Dialogs / Boxplot
EXAMINE VARIABLES=day1 day2 day3
/COMPARE VARIABLE
/PLOT=BOXPLOT
/STATISTICS=NONE
/NOTOTAL
/MISSING=PAIRWISE.

Groups of cases, ha több csoportot akarsz megjeleníteni
Separate variables, ha több változót akarsz megjeleníteni



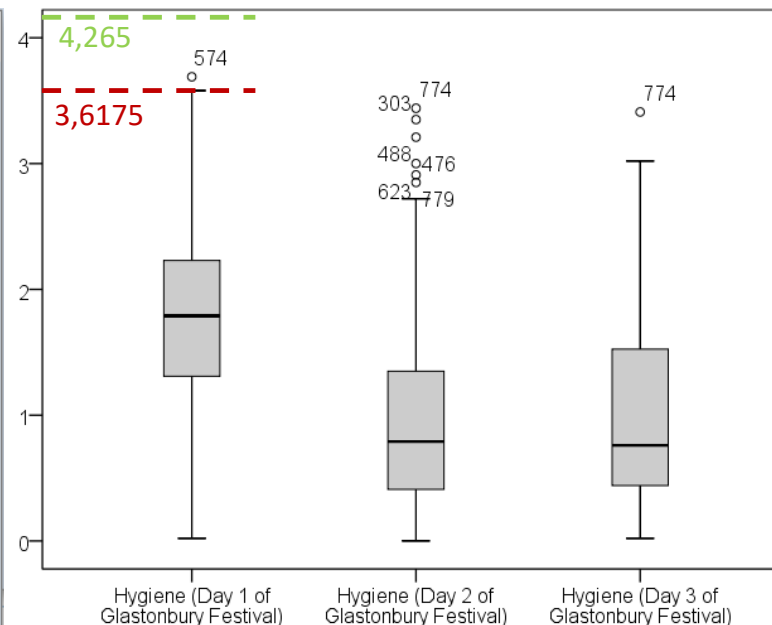
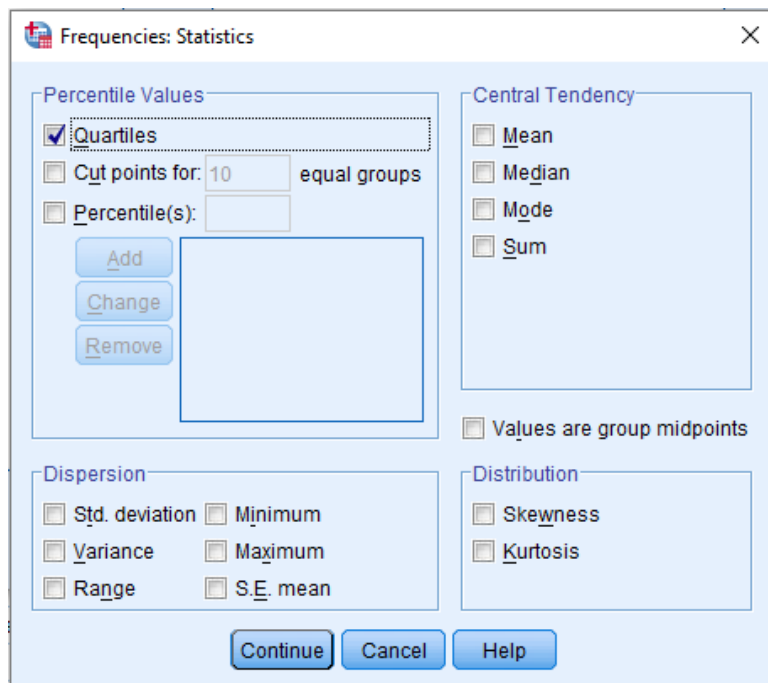
Olvasd el a hiányzó értékekről szóló részt is!

A körök outlierek, a szám az adatbázisban lévő sorszám



Outlierek

- A 1,5-es szorzó helyett sokkal jobban működik a 2,2-es szorzó, de az SPSS-ben nem lehet ezt kérni, ezért kézzel kell számolni a határt.
- Analyze / Descriptive / Frequencies / menüben Statistics almenü, és a Quartiles-t kattintjuk be.
 - (a főoldalon a Display Frequency Tables mellől kiszedhető a pipa, mert hosszúak lennének a táblázatok és nincs is rájuk szükség)



Q1 az alsó negyedelő
Q3 pedig a felső negyedelő

		Statistics		
		Hygiene (Day 1 of Glastonbury Festival)	Hygiene (Day 2 of Glastonbury Festival)	Hygiene (Day 3 of Glastonbury Festival)
N	Valid	810	264	123
	Missing	0	546	687
	Percentiles			
	25	1,3050	,4100	,4400
	50	1,7900	,7900	,7600
	75	2,2300	1,3500	1,5500

- Az alsó határ: $Q1 - (Q3 - Q1) * 2,2 = 1,305 - (2,23 - 1,305) * 2,2 = -0,73$
- A felső határ: $Q3 + (Q3 - Q1) * 2,2 = 2,23 + (2,23 - 1,305) * 2,2 = 3,6175$

Q1 és Q3 közötti rész az interquartilis terjedelem (IQT=Q3-Q1), mely a középső 50% értéket tartalmazza

- 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**
 - PRO: torzító hatás megszűnik KON: információt veszünk
 - Leggyakrabban használt, kis minta esetén bajos
 - Gondoljuk át, miből jöhet egy szélsőséges érték
 - **2. Transzformáljuk az adatokat**
 - PRO: nincs információvesztés KON: nehéz értelmezni
 - Transzformációkkal 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**
 - PRO: torzító hatás megszűnik kevés információvesztéssel KON: -
 - 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!

Boxplot csillagok és körök

- **Körök:** ami távolabb az alsó/felső negyedről, mint a középső 50% nagyságának **másfélszerese**
- **Csillagok:** ami távolabb az alsó/felső negyedről, mint a középső 50% nagyságának **háromszorosa**
- Nagyon gyakran a csillagok valamilyen elírásból származnak, míg a körök outlierok a mintában, de ez nem feltétlenül igaz:
 - Példa1: nemi adatoknál a 0 (ami egyértelműen elírás) nem jelenik meg csillagként, de még körként sem.
 - Példa2: 5 fokú skálán a 6 ésték (ami elírás), lehet, hogy nem jelenik meg csillagként, csak körként
- Megjelenhetnek a skálán elképzelhető értékek is csillagként (például sikerült beválogatnod egy NBA játékost). Ezek bár elképzelhető értékek, de annyira szélsőségesek, hogy mindenképp kezdeni kell velük valamit.

Lehetséges publikálás

A adatokban a hagyományos outlier labelling rule (Tukey, 1977) alapján több szélsőséges értéket is találtunk (az első napi higiénia értékekben egy esetet 3.69-es értékkel, a második napi értékekben 3.21 és 3.44 között hat darabot, és harmadik napi higiénia értékben 3.41-es értékkel egy darabot).

Az így megtalált outlierok közül a felülvizsgált, 2.2-es szorzót használó outlier labelling szabály (Hoaglin és Iglewicz, 1987) alapján csak egy minősülnek szélsőséges értéknek, a második napi higiénia értékekben található 3.44-es érték.

A megtalált szélsőséges értéket a winzorizálás szabályainak megfelelően átírtuk a rangsorban következő legnagyobb már nem szélsőséges értéknél egy egységgel nagyobbra, azaz 3,35-re.

- Hivatkozások:
- Tukey, J. W. (1977). Exploratory data analysis.
- Hoaglin, D. C., & Iglewicz, B. (1987). Fine-tuning some resistant rules for outlier labeling. Journal of the American Statistical Association, 82(400), 1147-1149.

független megfigyelések
függetlenség
független minták
független hibák
skála típusú adatok
monotonitás
dichotomitás
ferdeség
ordinális adatok
korreláltság
korrelálatlanság
folytonosság
linearitás
normalitás
kollinearitás
multikollinearitás
unimodalitás
szfericitás
kurtózis
szóráshomogenitás
összefüggő minták
random mintavételezés
kovarianciák homogenitása

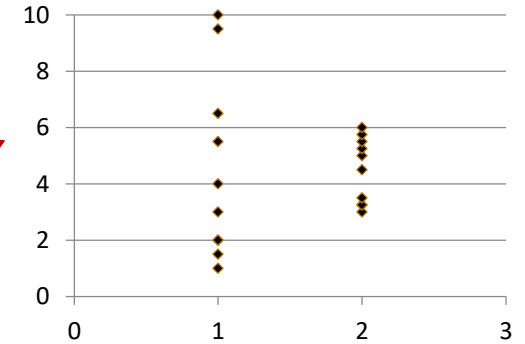
4. Feltételek ellenőrzése

Parametrikus adatok

- Minden tesztnek, amit végzünk vannak feltételei. A feltételek közül négy nagyon sok tesztnél előfordul, ezért ezeket külön vesszük. **Parametrikus tesztek** megkívánják, hogy az adatok parametrikusak legyenek, tehát az alábbi négy feltételt kielégítsék.

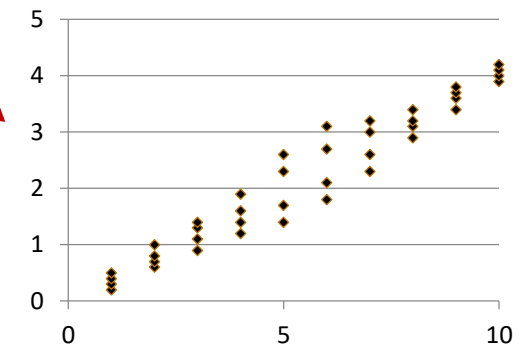
Ellenőrizni kell

- Normál eloszlás**
- Varianciák homogenitása / Szóráshomogenitás**
 - A varianciáknak egyformának kell lenniük az egész mintában
 - Ha csoportokat vizsgálunk, a különböző csoportok varianciájának kell azonosnak lennie
 - Ha korrelációt, akkor a változónk varianciájának stabilnak kell lennie a másik változó minden szintjén



Kialakítani kell

- Legalább intervallumskála típusú adatok**
 - Az adatok legalább intervallumskála szintűek
- Függetlenség**
 - a kísérleti személyek válasza független egymástól
 - (nem feltétlenül az adatok, csak a válaszadók, lásd repeated measures)



Milyen mintán ellenőrizzük a feltételeket?

- A feltétel-ellenőrzést nem l'art pour l'art végezzük, hanem mindig egy adott statisztikai teszthez kötötten – így fontos, hogy azokon az adatokon végezzük el a feltételtesztelést, amin majd a teszt is dolgozik.

Különbözik-e a három nap higiénia pontszáma?

Teszt: Repeated Measures ANOVA

subnumb	day1	day2	day3	nem
2338	,84	.	.	nő
2384	3,03	.	.	férfi
2401	,88	,08	.	nő
2405	,85	.	.	nő
2432	,97	1,41	,29	nő
2467	1,56	.	.	nő
2478	3,02	.	.	férfi
2490	2,29	.	.	nő
2504	1,11	,44	,55	nő
2509	2,17	.	.	nő
2510	,82	,20	,47	férfi
2511	2,64	1,35	1,61	nő
2514	1,41	.	.	férfi
2515	1,76	1,64	1,58	nő
2520	1,38	,02	.	férfi
2521	2,79	.	.	férfi
2529	1,50	.	.	nő
2533	1,91	2,05	.	férfi
2535	2,32	.	.	férfi
2538	2,05	.	.	nő
2549	2,17	,70	,76	nő
2551	2,05	.	.	nő
2558	1,61	.	.	nő

Különbözik-e az első napi higiénia a hármas értéktől?

Teszt: Egymintás t-próba

subnumb	day1	day2	day3	nem
2338	,84	.	.	nő
2384	3,03	.	.	férfi
2401	,88	,08	.	nő
2405	,85	.	.	nő
2432	,97	1,41	,29	nő
2467	1,56	.	.	nő
2478	3,02	.	.	férfi
2490	2,29	.	.	nő
2504	1,11	,44	,55	nő
2509	2,17	.	.	nő
2510	,82	,20	,47	férfi
2511	2,64	1,35	1,61	nő
2514	1,41	.	.	férfi
2515	1,76	1,64	1,58	nő
2520	1,38	,02	.	férfi
2521	2,79	.	.	férfi
2529	1,50	.	.	nő
2533	1,91	2,05	.	férfi
2535	2,32	.	.	férfi
2538	2,05	.	.	nő
2549	2,17	,70	,76	nő
2551	2,05	.	.	nő
2558	1,61	.	.	nő

Korrelál-e az első és második napi higiénia?

Teszt: Pearson korreláció

subnumb	day1	day2	day3	nem
2338	,84	.	.	nő
2384	3,03	.	.	férfi
2401	,88	,08	.	nő
2405	,85	.	.	nő
2432	,97	1,41	,29	nő
2467	1,56	.	.	nő
2478	3,02	.	.	férfi
2490	2,29	.	.	nő
2504	1,11	,44	,55	nő
2509	2,17	.	.	nő
2510	,82	,20	,47	férfi
2511	2,64	1,35	1,61	nő
2514	1,41	.	.	férfi
2515	1,76	1,64	1,58	nő
2520	1,38	,02	.	férfi
2521	2,79	.	.	férfi
2529	1,50	.	.	nő
2533	1,91	2,05	.	férfi
2535	2,32	.	.	férfi
2538	2,05	.	.	nő
2549	2,17	,70	,76	nő
2551	2,05	.	.	nő
2558	1,61	.	.	nő

- 0. lépés: grafikusan
 - Átfogó kép az adatokról
 - **Hisztogram**
- 1. statisztikailag
 - három elégséges feltétele van a normalitásnak, **BÁRMELYIK** teljesül a három közül, a mintát **normál eloszlásúnak tekinthetem**
 - A) **Central-limit theorem**
 - B) **Shapiro-Wilk** vagy **Kolmogorov-Smirnov** teszt
 - C) **z-tesztek**

Parametrikus adatok / normál

- **A) Central limit theorem** vagy központi határeloszlás tétel
 - Minden minta-eloszlás normálisnak tekinthető, ha elég nagy a minta
 - Mit jelent az, hogy elég nagy? Legalább 40 fő (ha több csoport van, akkor minden csoportban)
A CLT működésének demonstrációja: [a weblapon central_limit_theorem.xlsx](#) néven
- Publikálva:
 - A minták elemszámaiból következően (az elemszám minden mintában meghaladja a 40 főt) a central limit theorem törvényszerűségei miatt (Lumley et al., 2002) feltételezhető a mintákon a normalitás feltételének teljesülése.
- Hivatkozások:
 - Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. Annual review of public health, 23(1), 151-169.

Parametrikus adatok / normál

- **B) Shapiro-Wilk vagy Kolmogorov-Smirnov teszt**

- K-S teszt kevésbé jó, de ismertebb, és lassan „megy ki a divatból”
- Megadja, hogy a változónk eloszlása szignifikánsan eltér-e a normálistól
- Analyze / Descriptive Statistics / Explore

EXAMINE **VARIABLES**=day1 day2 day3

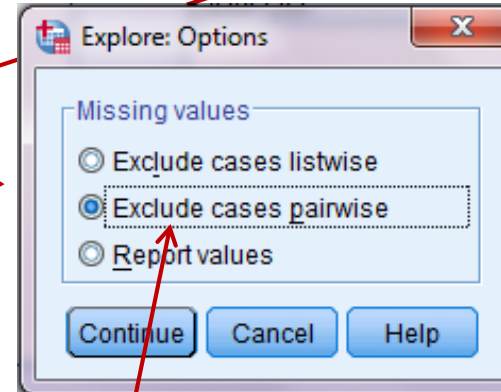
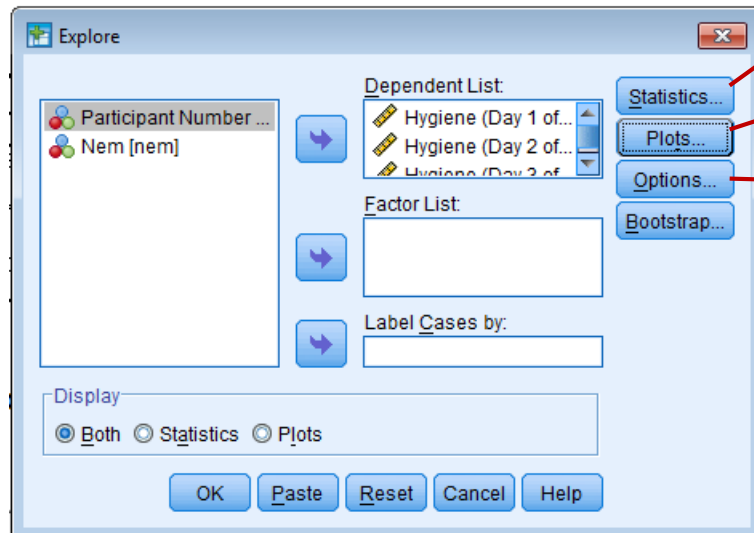
/PLOT **HISTOGRAM** **NPLOT**

/STATISTICS **DESCRIPTIVES**

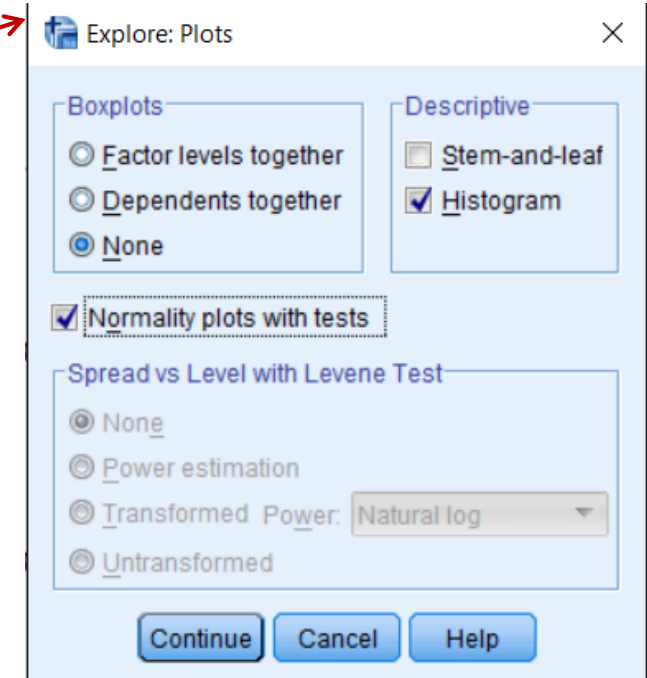
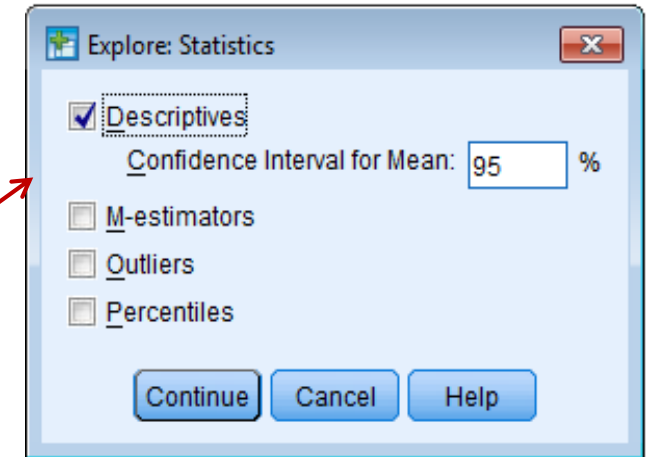
/CINTERVAL 95

/MISSING **PAIRWISE**

/NOTOTAL.



Olvasd el a hiányzó értékekről szóló részt is!

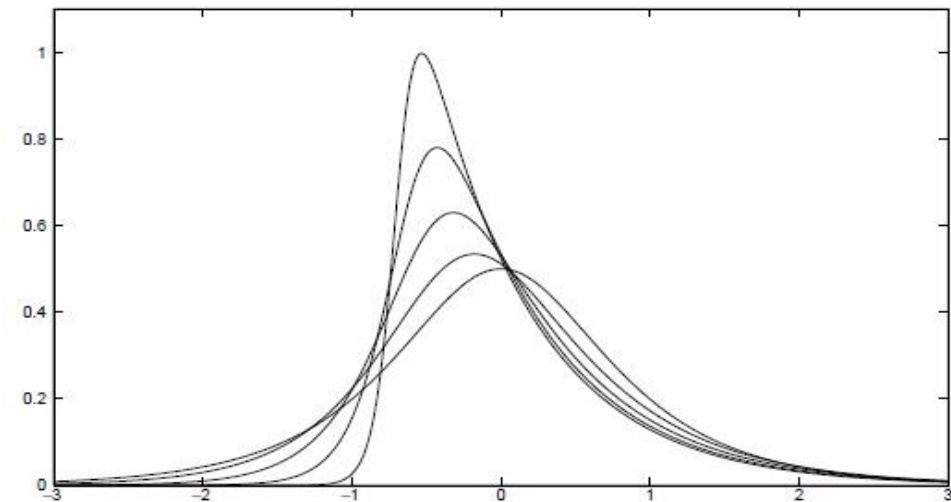
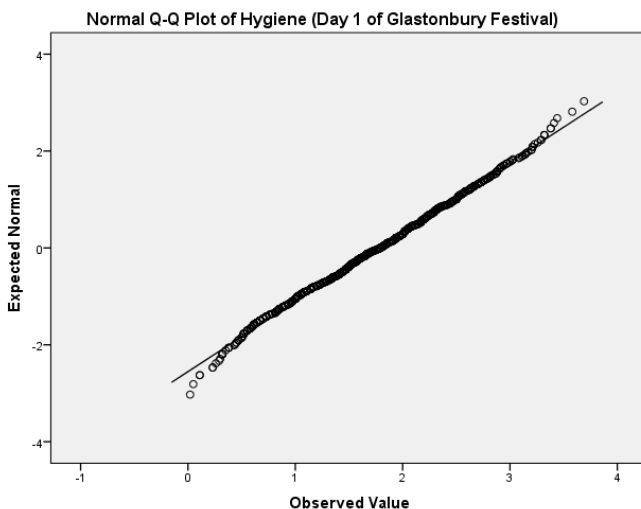


Parametrikus adatok / normál

A S-W és K-S próba azt teszteli, hogy van-e szignifikáns különbség a mi eloszlásunk és a normáeloszlás között, így ha szignifikáns, az azt jelenti, hogy az adataink eloszlása szignifikánsan eltér a normál eloszlástól, nem feltételezhetjük a normál eloszlást

Tests of Normality						
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Hygiene (Day 1 of Glastonbury Festival)	,029	810	,097	,996	810	,032
Hygiene (Day 2 of Glastonbury Festival)	,121	264	,000	,908	264	,000
Hygiene (Day 3 of Glastonbury Festival)	,140	123	,000	,908	123	,000

a. Lilliefors Significance Correction



Q-Q plot:

A standardizált mértértékek és a normál eloszlás estén várt értékeket megjelenítő diagram.

Az egyenes vonal jelöli, hogy milyenek lennének az értékek, ha normális lenne az eloszlás

A pontok jelölik, hogy milyenek ténylegesen az értékeink

Parametrikus adatok / normál

Publikálása:

Shapiro-Wilk: $W([\text{szabadságfok}]) = [\text{W-érték}] \ p = [\text{p-érték}]$

Kolmogorov-Smirnov: $D([\text{szabadságfok}]) = [\text{D-érték}] \ p = [\text{p-érték}]$

A normalitás feltételének teszteléséhez Ghasemi és Zahedi-Asl (2012) útmutatását követve az ismertebb Lilliefors korrigált Kolmogorov-Smirnov teszt helyett Shapiro-Wilk normalitás tesztet használtam. A teszt eredményei alapján a higiénia eloszlásgörbéje mindhárom nap szignifikánsan eltér a normál eloszlástól. Első nap: $W(810) = .996 \ p = .032$ második nap: $W(264) = .908 \ p < .001$ és harmadik nap: $W(123) = .908 \ p < .001$

Tests of Normality						
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Hygiene (Day 1 of Glastonbury Festival)	,029	810	,097	,996	810	,032
Hygiene (Day 2 of Glastonbury Festival)	,121	264	,000	,908	264	,000
Hygiene (Day 3 of Glastonbury Festival)	,140	123	,000	,908	123	,000

a. Lilliefors Significance Correction

Hivatkozások:

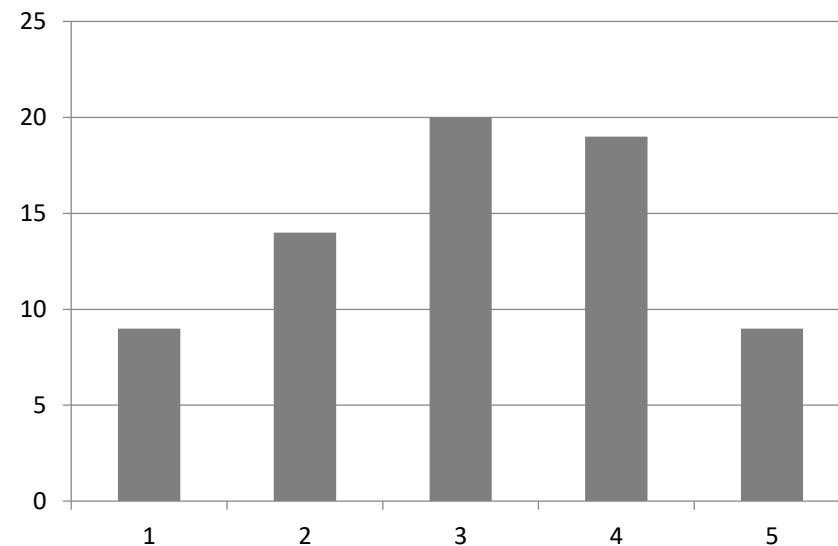
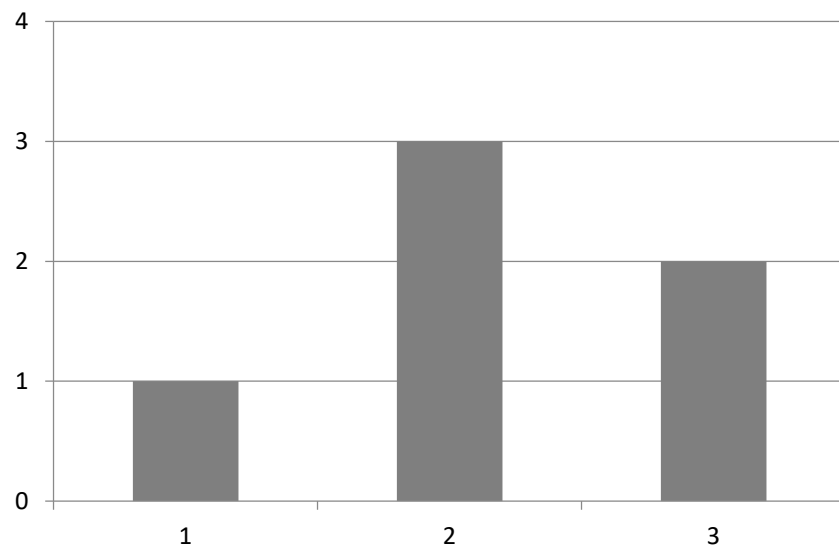
Ghasemi, A., & Zahedi-Asl, S. (2012). Normality tests for statistical analysis: a guide for non-statisticians. International journal of endocrinology and metabolism, 10(2), 486.

Thode, H. C. (2002). Testing for normality (Vol. 164). CRC press.

Steinskog, D. J., Tjøstheim, D. B., & Kvamstø, N. G. (2007). A cautionary note on the use of the Kolmogorov-Smirnov test for normality. Monthly Weather Review, 135(3), 1151-1157.

Parametrikus adatok / normál

- Miért szignifikáns a Shapiro-Wilk próba, amikor látjuk, hogy haranggörbe alakú?
- Minél nagyobb az elemszám, annál kisebb normalitástól való eltérés is szignifikáns.



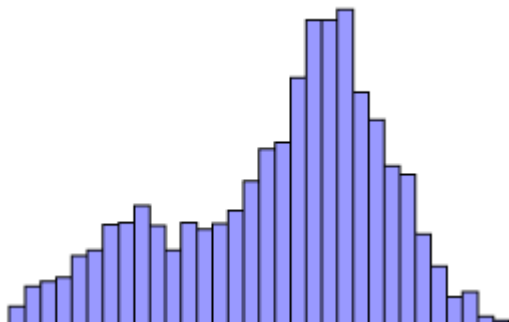
- Melyik esetben lehetünk biztosabbak, hogy az eloszlás nem normális, nem pedig csak szerencsétlenül vettük fel az adatokat?

Parametrikus adatok / normál

• C) Z-tesztek

- Ha 15 adatnál több van, az eloszlásgörbe legfeljebb csak enyhén ferde, enyhén csúcsos, unimodális (egy csúcsa van), outlierok nélküli, akkor feltételezhetjük a normalitást.
- Ferdeség, csúcsosság ellenőrzése:
 - A normálgörbe ferdesége és csúcsossága 0.
 - Z-tesztekkel ellenőrizhetjük, hogy a mintánk ferdesége és csúcsossága szignifikánsan eltér-e a 0-tól.
 - Ha $|z_{skew}| > 1,96$, akkor a mintaeloszlás ferdesége szignifikánsan eltér a normálistól
 - Ha $|z_{kurt}| > 1,96$, akkor a mintaeloszlás csúcsossága szignifikánsan eltér a normálistól

$$z_{skew} = \frac{S}{SE_{skew}} \quad z_{kurt} = \frac{K}{SE_{kurt}}$$

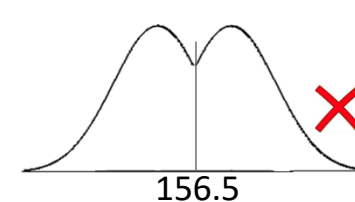
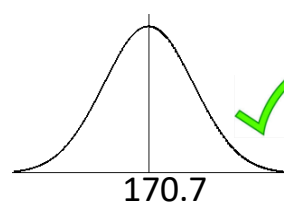
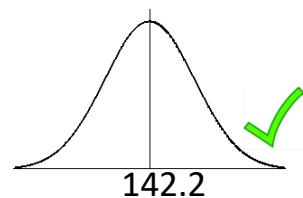


Descriptive Statistics

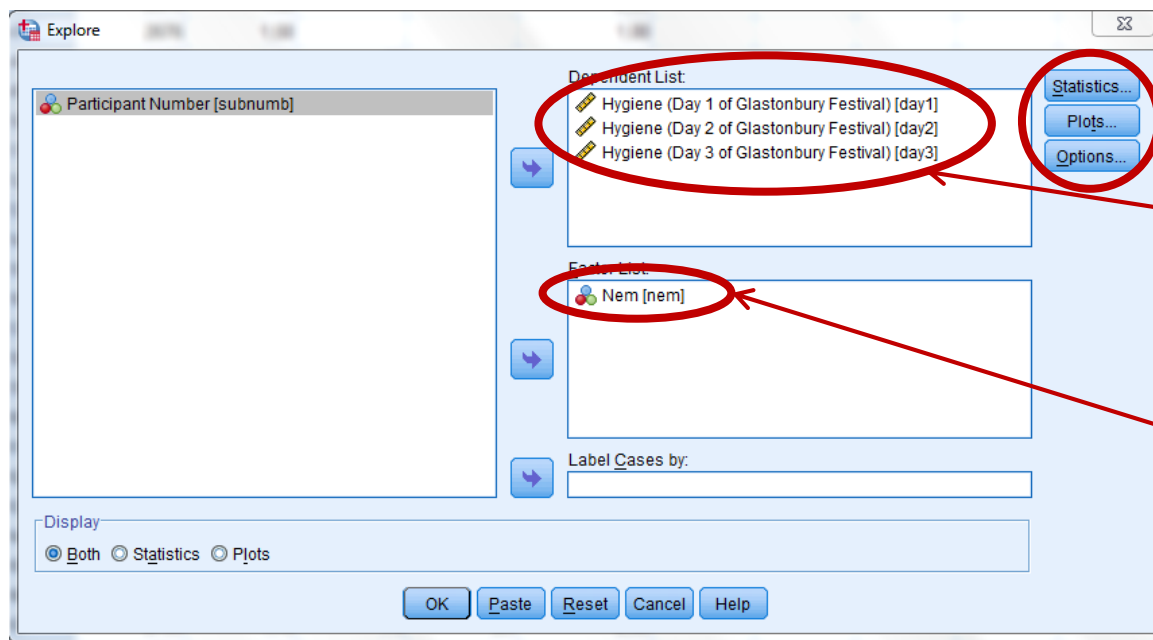
	N	Skewness		Kurtosis	
		Statistic	Std. Error	Statistic	Std. Error
Hygiene (Day 1 of Glastonbury Festival)	810	-,004	,086	-,410	,172
Hygiene (Day 2 of Glastonbury Festival)	264	1,095	,150	,822	,299
Hygiene (Day 3 of Glastonbury Festival)	123	1,033	,218	,732	,433
Valid N (listwise)	123				

Parametrikus adatok / normál

- Parametrikusság tesztelése mintákként külön
 - miért? Azt várjuk, hogy a minták különböznek egymástól -> eltér az átlaguk -> ha közösen nézzük őket, nem lesz unimodális az eloszlás
 - női magasság Bolíviában és Hollandiában valamint, ha együtt néznénk



- Analyze / Descriptive Statistics/ Explore



ugyanazzal a beállítással,
mint az előbb

csoportosító változó

Parametrikus adatok / normál

- Ugyanazokat a táblázatokat és grafikonokat kapjuk meg, mint az előbb, de most a férfi és női mintára külön

Case Processing Summary

		Cases					
		Valid		Missing		Total	
		N	Percent	N	Percent	N	Percent
Hygiene (Day 1 of Glastonbury Festival)	férfi	401	100,0%	0	0,0%	401	100,0%
	nő	409	100,0%	0	0,0%	409	100,0%
Hygiene (Day 2 of Glastonbury Festival)	férfi	133	33,2%	268	66,8%	401	100,0%
	nő	131	32,0%	278	68,0%	409	100,0%
Hygiene (Day 3 of Glastonbury Festival)	férfi	62	15,5%	339	84,5%	401	100,0%
	nő	61	14,9%	348	85,1%	409	100,0%

Tests of Normality

		Kolmogorov-Smirnov ^a			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Hygiene (Day 1 of Glastonbury Festival)	férfi	,025	401	,200*	,997	401	,619
	nő	,041	409	,089	,991	409	,011
Hygiene (Day 2 of Glastonbury Festival)	férfi	,130	133	,000	,891	133	,000
	nő	,117	131	,000	,921	131	,000
Hygiene (Day 3 of Glastonbury Festival)	férfi	,149	62	,002	,921	62	,001
	nő	,143	61	,003	,901	61	,000

*. This is a lower bound of the true significance.

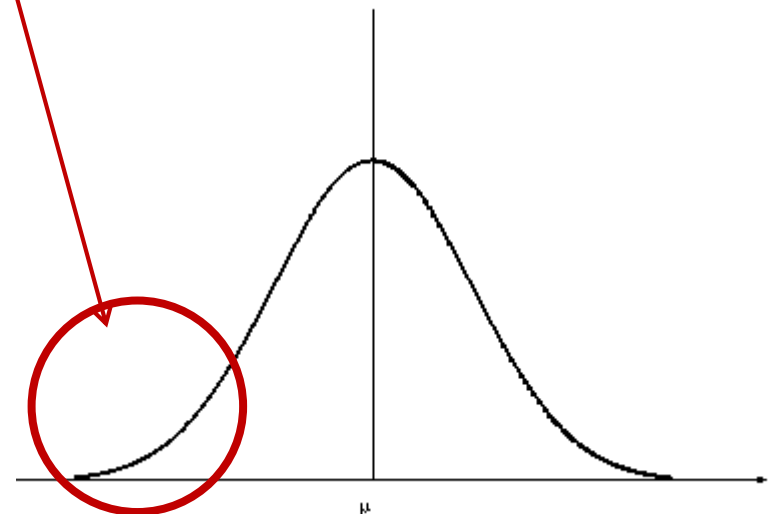
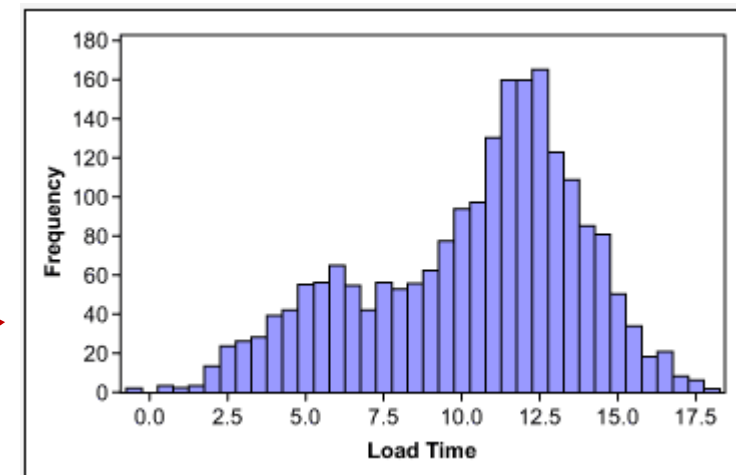
a. Lilliefors Significance Correction

Descriptives

		Statistic	Std. Error
Hygiene (Day 1 of Glastonbury Festival)	Mean	1,7719	,03483
	95% Confidence Interval for Mean	Lower Bound 1,7034	
		Upper Bound 1,8404	
	5% Trimmed Mean	1,7702	
	Median	1,7900	
Hygiene (Day 2 of Glastonbury Festival)	Mean	1,7704	,03414
	95% Confidence Interval for Mean	Lower Bound 1,7033	
		Upper Bound 1,8375	
	5% Trimmed Mean	1,7697	
	Median	1,8200	
	Variance	,477	
Hygiene (Day 3 of Glastonbury Festival)	Std. Deviation	,69042	
	Minimum	1,000	

Parametrikus adatok / normál

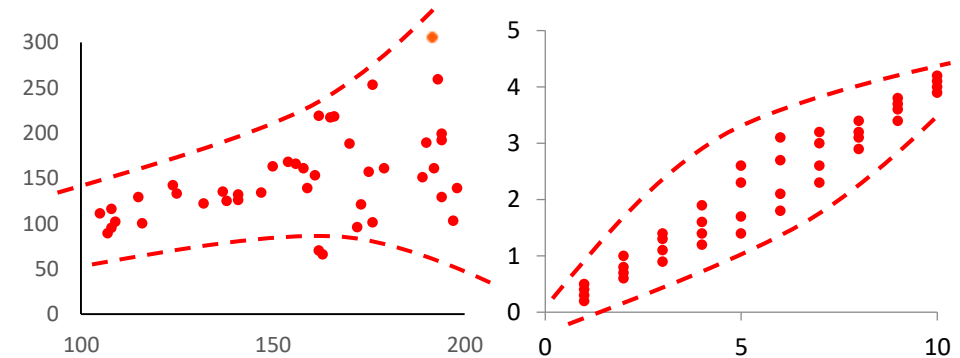
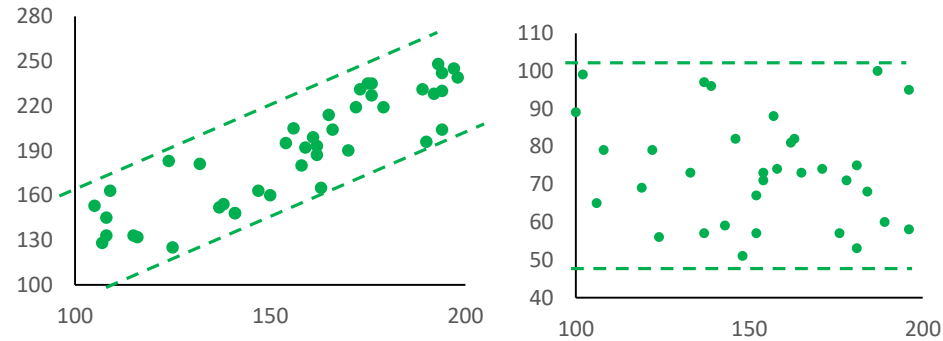
- Mit csináljak, ha nem-normális?
 - **Kutatói mérlegelés kérdése, melyik mellett döntesz**
 - **Találd meg a választ, miért nem normális!**
 - Outlierek
 - Valamiért több dolgot sikerült egyszerre mérnünk
 - Nem elég érzékeny skála
 - Kiválogatott adat (az adatoknak csak egy részét látom)
 - Természetes határ (neuron-tüzelés)
 - Az adat valamilyen más eloszlást követ
 - **Növeld az elemszámot!**
 - Central limit teorem miatt
 - **Traszformáld az adatokat!**
 - **Használj non-parametrikus próbákat!**



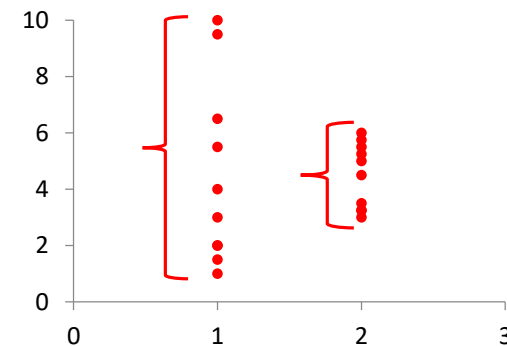
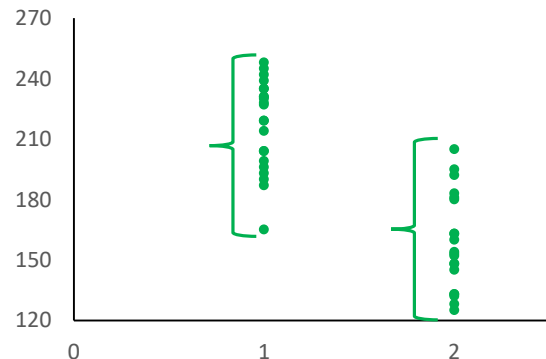
Parametrikus adatok / Szóráshomogenitás

- **Varianciák homogenitása / Szóráshomogenitás**
 - A varianciáknak egyformának kell lenniük az egész mintában
 - Ha korrelációt, akkor a változónk varianciájának stabilnak kell lennie a másik változó minden szintjén
 - Ha csoportokat vizsgálunk, a különböző csoportok varianciájának kell azonosnak lennie

Változók közötti
kapcsolat



Minták közötti
különbség



Parametrikus adatok / Szóráshomogenitás

- Korrelációnál, regressziónál **homoszkedaszticitás**nak nevezzük
 - korrelációnál: diagram vagy legnagyobb és legkisebb variancia különbsége
 - regressziónál: beépített elemzés (részletesebben regressziónál)
- Graphs / Legacy dialogs / ScatterDot /

Simple Scatterplot

Y Axis:
Hygiene (Day 2 of Glastonbury Festival) [day2]

X Axis:
Hygiene (Day 1 of Glastonbury Festival) [day1]

Set Markers by:

Label Cases by:

Panel by

Rows:

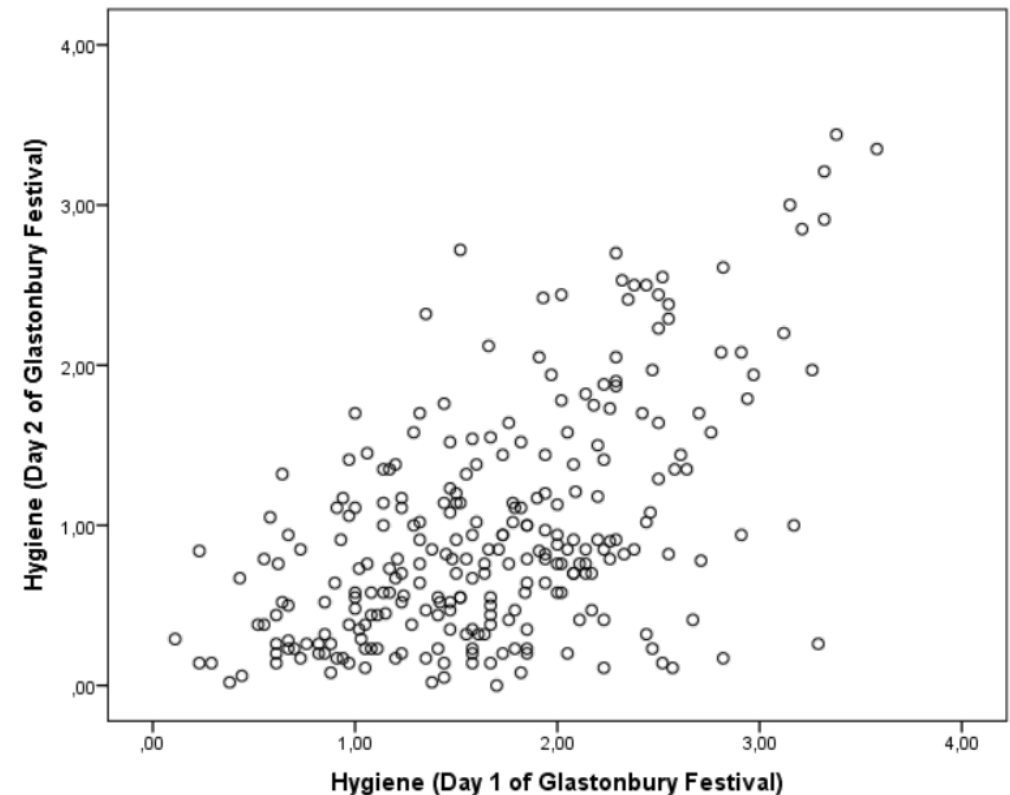
☐ Nest variables (no empty rows)

Columns:

☐ Nest variables (no empty columns)

Titles...

Options...



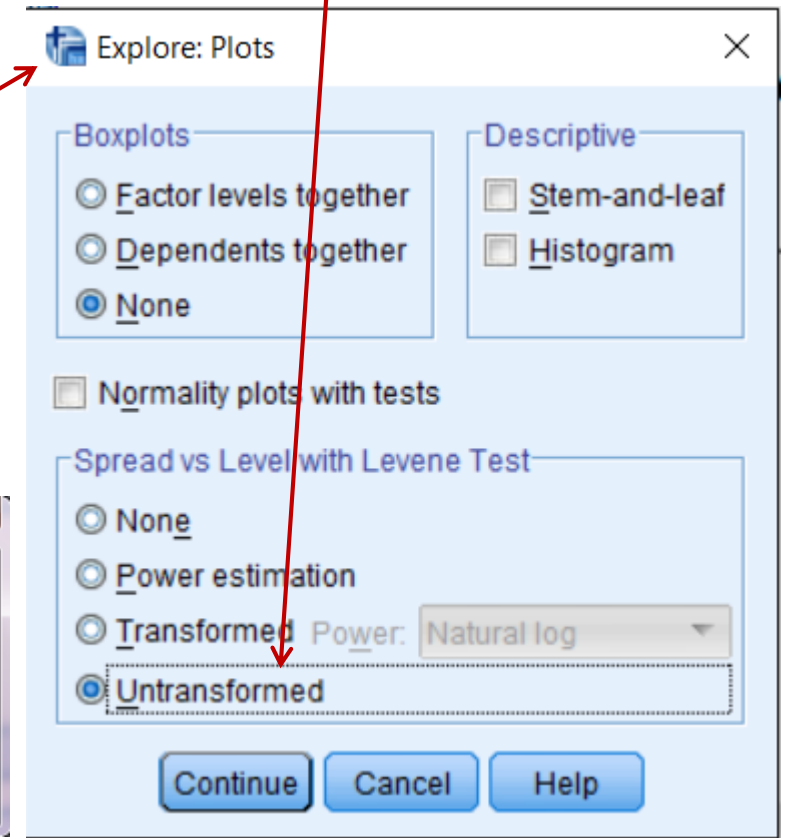
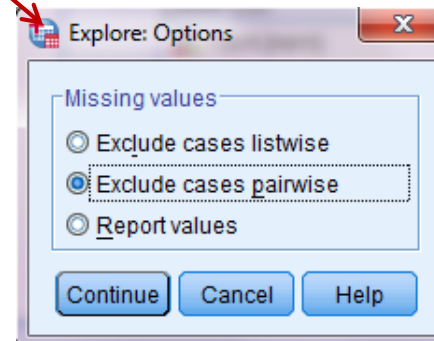
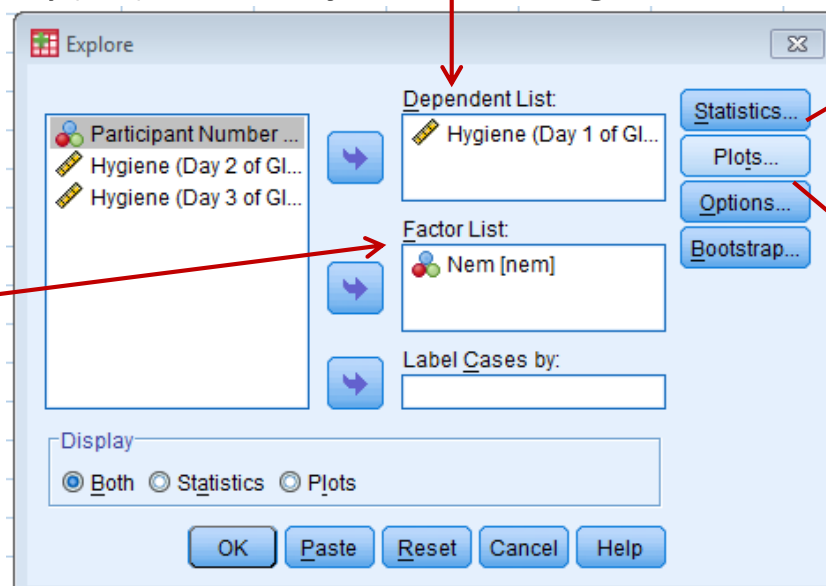
Parametrikus adatok / Szóráshomogenitás

- Csoportok összehasonlításánál **Levene teszttel**
 - Analyze / Descriptive Statistics / Explore
EXAMINE VARIABLES=day1 day2 day3 **BY** nem
/PLOT SPREADLEVEL(1)
/STATISTICS DESCRIPTIVES
/CINTERVAL 95
/MISSING PAIRWISE
/NOTOTAL.

Az adatok nem transzformáltak
(de megnézheted, milyen lenne a
szóráshomogenitás, ha transzformálnád őket)

A változó(k), amely(ek) variáciájának homogenitását vizsgálom

Csoportosító
változó



Parametrikus adatok / Szóráshomogenitás

Test of Homogeneity of Variance

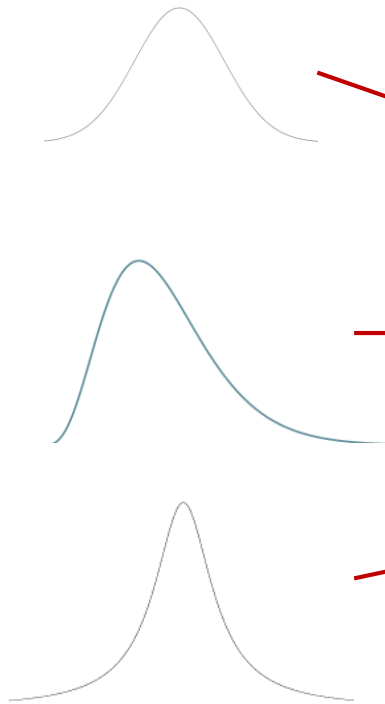
		Levene Statistic	df1	df2	Sig.
Hygiene (Day 1 of Glastonbury Festival)	Based on Mean	,234	1	808	,629
	Based on Median	,222	1	808	,638
	Based on Median and with adjusted df	,222	1	804,111	,638
	Based on trimmed mean	,233	1	808	,629

← Nem szignifikáns

- Nem tudtunk kimutatni szignifikáns különbséget a varianciában, tehát teljesül a szóráshomogenitás feltétele
- Publikálása: $F([df1], [df2]) = [\text{Levene teszt értéke}] \text{ } p = [\text{szignifikancia}]$
 $F(1, 808) = .234 \text{ } p = .629$
- Mikor használható a Levene-teszt?
 - A Levene-teszt is érzékeny a normál eloszlás hiányára.
 - Ha nem teljesül a normál eloszlás a Levene-teszt helyett nemparametrikus szóráshomogenitás vizsgálatot kell végezni

Szóráshomogenitás tesztelése normál eloszlás hiányában

- A Levene-teszt is érzékeny a normál eloszlás hiányára.
- **I. Megoldás:** az átlag helyett a mediánt vagy trimmelt átlagot használjuk (jobb megoldás, mint az átlag, de még ez sem elég robosztus)



Test of Homogeneity of Variance

	Levene Statistic	df1	df2	Sig.
Hygiene (Day 1 of Glastonbury Festival) → Based on Mean	,234	1	808	,629
→ Based on Median	,222	1	808	,638
Based on Median and with adjusted df	,222	1	804,111	,638
→ Based on trimmed mean	,233	1	808	,629



Szóráshomogenitás tesztelése normál eloszlás hiányában

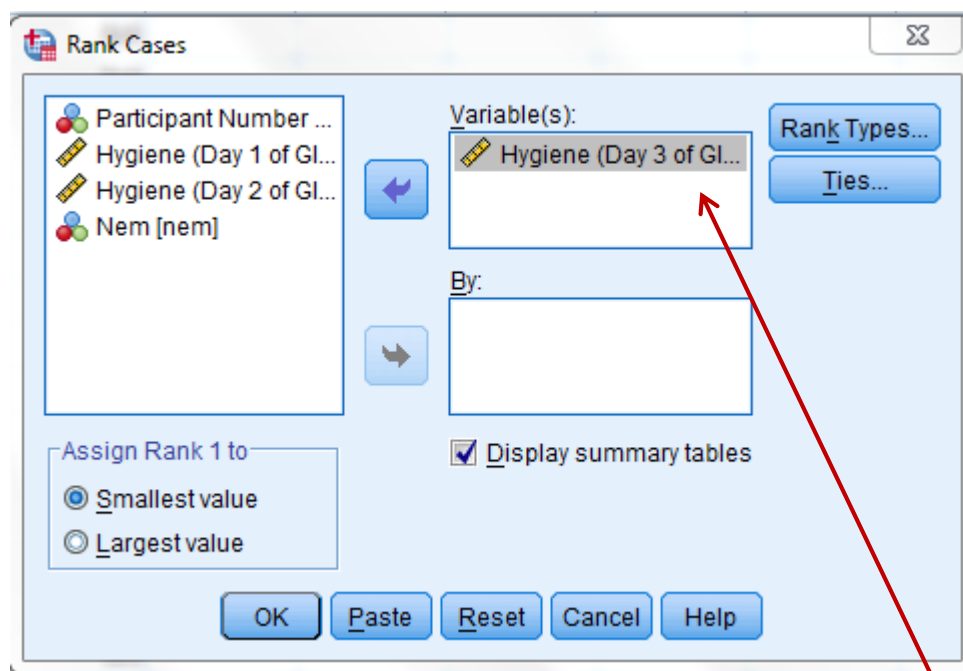
- **II. Megoldás:** rangsorolás kézzel, majd a rangok szórásának összehasonlítása
 - Ehhez ki kell számolnunk három új változót:
 - 1. Rang: az elemezni kívánt változó elemeit rangsorolni kell
 - 2. Rangátlag: ki kell számolni csoportonként a rangsor átlagát
 - 3. Eltérés: ki kell számolni az elemek rangjának eltérését a csoportonkénti rangsor átlagától
 - majd az eltérés változón ANOVA tesztet végzünk



Szóráshomogenitás tesztelése normál eloszlás hiányában

- Számoljunk most csak a day3 értékeivel!
- 1. rang kiszámolása
 - Transform / Rank Cases

automatikusan létrejött egy új változó, ami a day3 változó értékeit növekvő sorrendben rangsorolja



rangsorolni kívánt változó

day1	day2	day3	nem	Rday3
3,03	.	.	férfi	.
3,02	.	.	férfi	.
,82	,20	,47	férfi	38,000
1,41	.	.	férfi	.
1,38	,02	.	férfi	.
2,79	.	.	férfi	.
1,91	2,05	.	férfi	.
2,32	.	.	férfi	.
1,66	,85	.	férfi	.
2,76	.	.	férfi	.
3,23	.	.	férfi	.
,97	,38	,76	férfi	61,500
2,57	,11	,02	férfi	1,500
,26	.	.	férfi	.
1,73	.	.	férfi	.
1,94	,82	1,67	férfi	99,000
1 01			férfi	

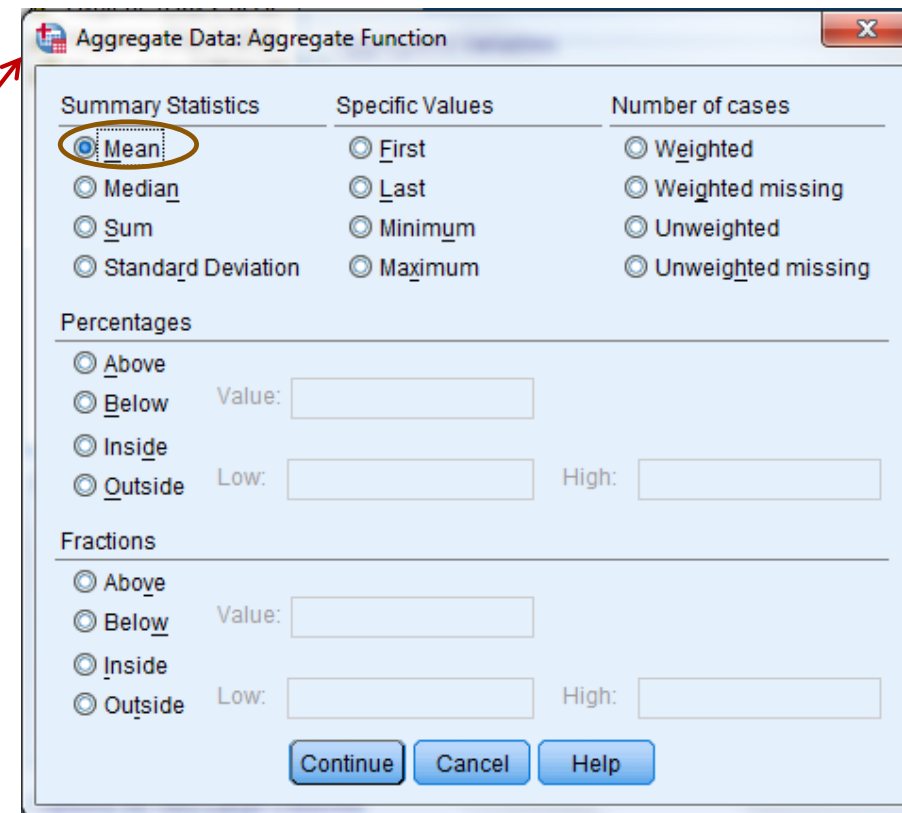
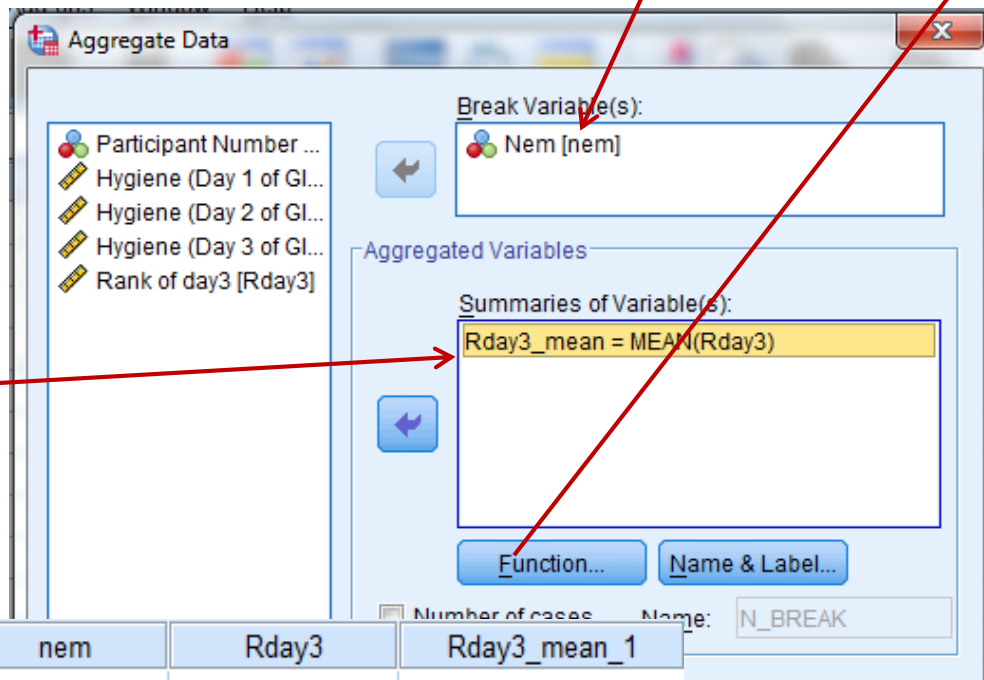


Szóráshomogenitás tesztelése normál eloszlás hiányában

- 2. csoportonként a rang átlaga
 - Data / Aggregate

csoportosító változó

átlagolandó változó



y1	day2	day3	nem	Rday3	Rday3_mean_1
2,11	.	.	férfi	.	65,74
,64	.	.	férfi	.	65,74
2,61	1,44	.	férfi	.	65,74
1,28	.	.	férfi	.	65,74
1,26	.	.	férfi	.	65,74
,84	.	.	nő	.	58,20
,88	,08	.	nő	.	58,20
,85	.	.	nő	.	58,20

létrejött új változó:

külön a rangok átlaga férfiaknál és nőknél
- egy-egy csoport minden tagja ugyanazt az értéket (a csoport rangátlagát) kapja



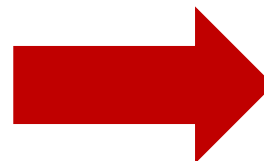
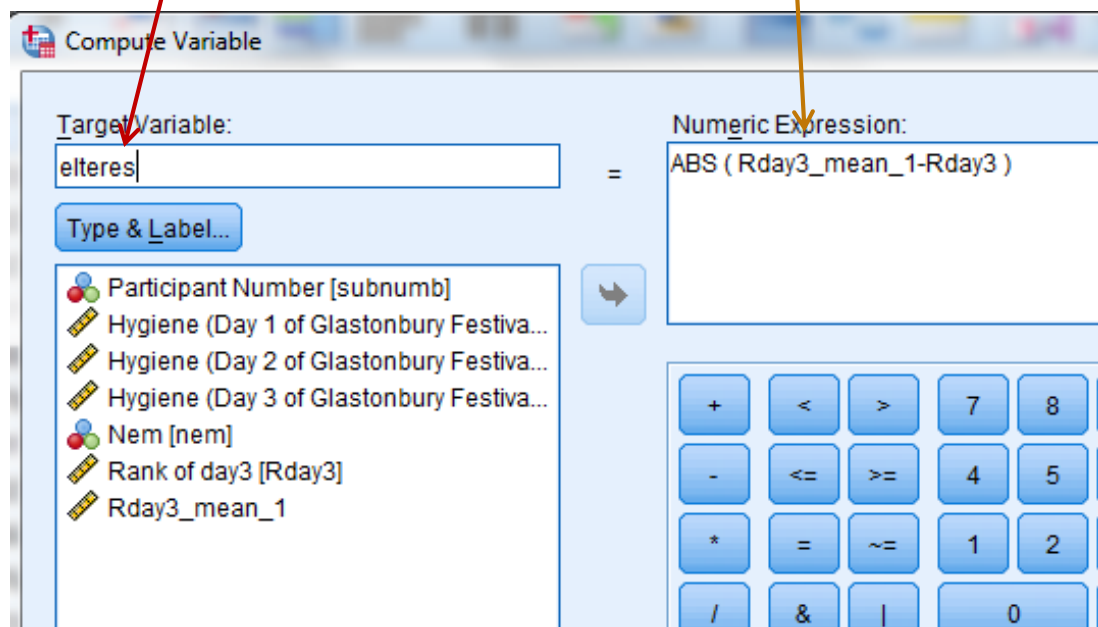
Szóráshomogenitás tesztelése normál eloszlás hiányában

- 3. személyek rangjának a csoportjuk rangátlagától való eltérésének abszolút értéke
 - Transform / Compute Variable

hogyan számoljuk az új változót ki?
eltérés = abszolútérték (rangátlag – rang)

új változó neve: elteres

ABS (Rday3_mean_1 – Rday3)



létrejön az új változó

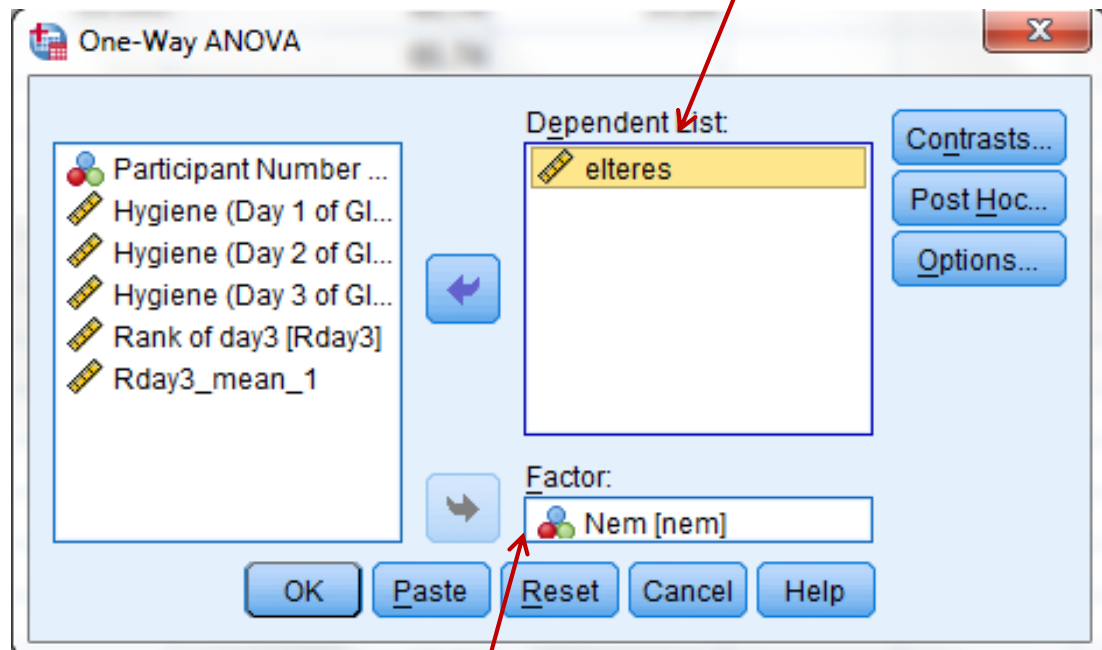
ay3	Rday3_mean_1	elteres
1,500	65,74	64,24
-	65,74	-
-	65,74	-
99,000	65,74	33,26
-	65,74	-
-	65,74	-
-	65,74	-
-	65,74	-
-	65,74	-
-	65,74	-
5,500	65,74	60,24
-	65,74	-
-	65,74	-
-	65,74	-
43,500	65,74	22,24
-	65,74	-
-	65,74	-



Szóráshomogenitás tesztelése normál eloszlás hiányában

- 4. az új eltérés változónkon One-way ANOVA végzése
 - Analyze / Compare Means / One-way ANOVA

rangok és a rangátlag közötti eltérés



csoportosító változó

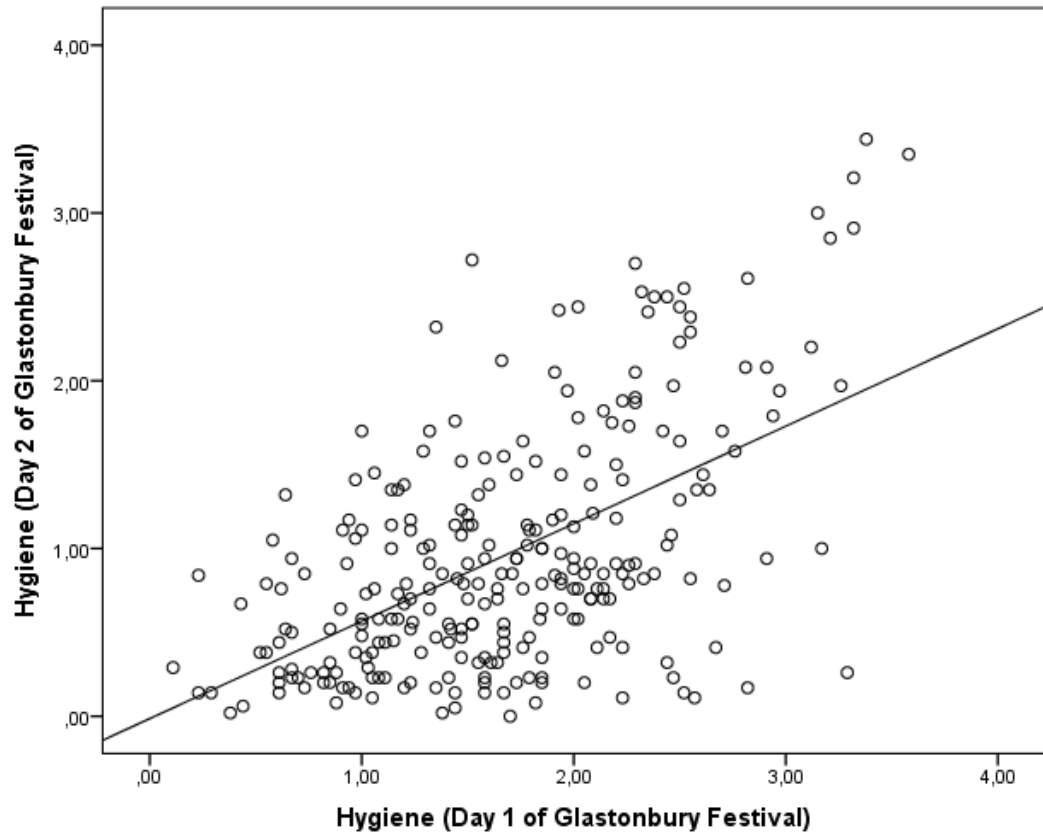
ANOVA

ANOVA					
eltérés	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1793,465	1	1793,465	5,947	,016
Within Groups	36490,262	121	301,572		
Total	38283,728	122			

A null-hipotézis az, hogy a csoportok között nincs különbség a rangok rangátlagtól való eltérésében (a rangok egyformán szóródnak mindkét csoportban)
Ha az ANOVA szignifikáns lett, akkor el kell vetnünk a nullhipotézist: különbség van a csoportok között a rangok szóródásában -> nem teljesül a szóráshomogenitás

- **Linearitás**

- Ha nem teljesül, akkor a lineáris modellel alul fogod becsülni a hatást
- Leginkább szem alapján döntünk
- Scatter/dot – a szerkesztőben Element / Fit Line at Totallal felhelyezhető a lineáris modellnek megfelelő egyenes



Properties

Chart Size Lines **Fit Line** Variables

☐ Display Spikes ☐ Suppress intercept

Fit Method

☐ Mean of Y ☐ Quadratic

☒ Linear ☐ Cubic

☐ Loess

% of points to fit: 50

Kernel: Epanechnikov

Confidence Intervals

☒ None ☐ Mean ☐ Individual

?: 95

☐ Attach label to line

Apply Cancel Help

Amikor a parametrikus tesztek kritériumai nem teljesülnek

- **Nem parametrikus tesztek**

- Ha nem teljesül a parametrikus tesztek kritériumából a folytonosság vagy a normál eloszlás kritériuma

- **Rangsor**

- Megkeressük a változó legkisebb értékét, hozzárendelünk 1-et, majd a következő értékhez 2-őt és így tovább
- a legtöbb nempara teszt a rangsorolást követően a rangokon a parametrikus próbákhoz hasonló elemzést végez

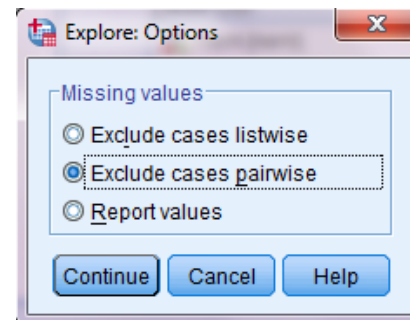
- **Kevésbé hatásos statisztikák-e, mint parametrikus társaik?**

- A rangsorolással elvesztjük a különbségek nagyságáról szerzett információ egy részét
- Ezért ha parametrikus adatokon hasonlítjuk össze, a nem-parametrikus próbák kevésbé tudják a hatást megtalálni – de miért használnánk parametrikus adatokon őket?
- Nem-parametrikus adatokon viszont a parametrikus próbák megbízhatatlanok (az elsőfajú hiba 5% csak normál eloszlás esetén jön ki)

Hiányzó értékek

- **A hiányzó értékek kezelése**

- Hogy listwise vagy pairwise vizsgáljuk az adatok parametrikusságát, kérünk leíró statisztikákat, stb., függ a statisztikai próbától amit végezni fogunk az adatokon.
- **Mindig tudni kell, hogy a statisztikai próba, amelyet végzünk hogyan kezeli a hiányzó értékeket (mi a minta, melyek dolgozni fog), és annak megfelelően kérni ki a leíró statisztikákat**
- Példa a mi esetünkben:
 - Repeated Measure mindenképp listwise fogja elemezni az adatokat
 - A változók normalitását is listwise kell vizsgálni (hiba az első nap a 810 fős minta normalitását ellenőrizni, ha a vizsgálatba csak az a 123 kerül, akinek mindhárom nap van adata)
 - Hasonlóképp hiba a leíró statisztikákat 810 főre kérni. A Descriptive Statistics / Descriptive-en belül nem kérhetünk listwise hibakezelést. Ekkor a Descriptive Statistics / Explore-t kell használni.
 - Ha Repeated Measure helyett Mixed Models használunk majd, ott beállíthatunk különböző hiányzó érték-kezelési módokat
 - Ha korrelációt végzünk, ott is beállíthatjuk, miként kezeje a hiányzó értékeket
- **Ha több statisztikai próbát végzünk, és különböző hiányzóérték-kezelési módokat használunk, vagy bizonyos próbáknál szűrt adatokon dolgozunk, akkor minden próbához külön le kell kérni a hozzá tartozó leíró statisztikákat!**



Feltételek – mennyire fontosak?

- **Robosztusság ellenőrzése Monte-Carlo szimulációval**
 - Előállítanak mintákat, ahol a nullhipotézis teljesül – ekkor a statisztikák 5%-nál várunk szignifikáns eredményt (ennyi lesz az első fajú hibát)
 - Elvégzem a próbát sok (pl. ezer vagy 10ezer) alkalommal
 - Egy próba robusztus egy feltétel megszegésére, ha a szignifikáns eredmények 4-6% között maradnak
- **Ha egy próba robusztus egy feltétel megszegésére egy adott mintán, akkor a feltétel sérülése ellenére is értelmezhetőek az eredmények**
 - Azt látjuk, hogy a próbák feltételeinek robusztussága hat egymásra, például:
 - ANOVA robusztus a szóráshomogenitás bármilyen mértékű sérülésére, ha a minták elemszáma hasonló, a normalitás teljesül, és $N > 15$
 - ANOVA robusztus a szóráshomogenitás kis mértékű sérülésére, ha a minták elemszáma hasonló, a normalitás teljesül és a legkisebb és legnagyobb variancia aránya < 3.5
 - ANOVA NEM robusztus a szóráshomogenitás sérülésére, bármekkora is az elemszám, ha a minták elemszáma különböző, és a legkisebb és legnagyobb variancia aránya > 8
 - ANOVA NEM robusztus a szóráshomogenitás sérülésére, ha változók eloszlása egymással ellentétes irányban ferde
 - ANOVA robusztus a normalitás sérülésére, ha a szóráshomogenitás teljesül és az elemszám $N > 15$