Többváltozós statisztika elméleti alapjai (BMNPS07500M)

Készítette: Soltész-Várhelyi Klára

Adatfeldolgozás

2. Gyakori feltételek

Feltételek ellenőrzése

Parametrikus adatok

- Minden tesztnek, amit végzünk vannak feltételei. A feltételek közül öt nagyon sok tesztnél előfordul, ezért ezeket külön vesszük.

- **Parametrikusság négy feltétele**
  - Parametrikus tesztek megkívánják, hogy az adatok parametrikusak legyenek, tehát az alábbi négy feltételt kielégítsék.
    - **Függetlenség**
    - **Legalább intervallumskála típusú adatok**
    - **Normál eloszlás**
    - **Szóráshomogenitás**
    - **Linearitás**

Kialakítani kell

Ellenőrizni kell

- Összefoglaló flowchart: adatfeldolgozas_flowchart.pdf

Függetlenség

- Minek a **függetlenség**e a parametrikusság feltétele?
  - Csoportok? Minták? Válaszadók? Válaszok? Változók?
  - **Válaszadók (mérések, azaz sorok) függetlenségét feltételezzük**.
    - A nagy számok törvénye és a central limit theorem független megfigyelések esetén működik.
  - Példa a sérülésére:
    - IQ és tanulmányi eredmény összefüggését vizsgálva több iskolából hozott adat. Az egy iskolába járók hasonló tanulmányi környezetben vannak – nem függetlenek egymástól.
    - Hamburger- és üdítőméret összefüggését vizsgálva megfigyelések reggeli és esti órákból is.
    - Reakcióidő mérésben egy ember válaszainak összevonása helyett az összes mérés egymás alá téve. Az egy személytől származó adatok nem függetlenek egymástól.
    - Két részvény mozgásának összefüggés-vizsgálatához a tőzsdei árfolyam monitorozása egy éven keresztül. Egy napi árfolyam nem független a megelőző napoktól
      - Ez az ún. szeriális korreláció – kiemelten figyelni kell az elkerülésére
  - Ne zavarjon össze, későbbiekben más függetlenségi feltételek is lesznek (pl. ANOVA csoportok függetlensége, regresszió hibatagok függetlensége stb.)

Skála típusú adatok

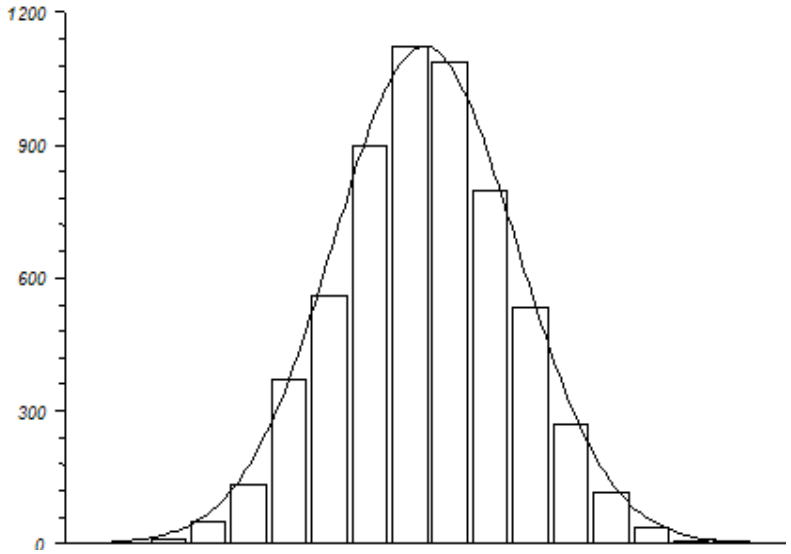
- A parametrikus tesztek **legalább intervallum skála típusú (függő) változó(k**o)n működnek
  - **Intervallum skála típusú** - az elemek sorba rendezhetőek, az elemek közötti különbség is kifejezhető, lehet arányokat számítani, de nincs természetes nulla pont.
  - **Arány skála típusú** – az értékek nagyság szerint sorba rendezhetőek, az elemek közötti különbség kifejezhető, lehet arányokat számítani, és van természetes nulla pont.
  - **Skála típusú** – A legtöbb statisztikai elemzés során nem használjuk ki a természetes nulla pont előnyeit, ezért az az intervallum és arány skálát sokszor egységesen, skála típusként kezeljük
    - Együtt kezeljük őket, de azért okozhatnak különbséget a statisztikáinkban, például a b_0 regressziós együttható egész érdekesen viselkedik, ha intervallum skála típusú adatokkal dolgozunk)
    - **Legalább intervallum szintű** – a skála típusú változókra szoktunk így hivatkozni, legalább intervallum szintű, tehát vagy intervallum skála típusú vagy arányskála típusú.
  - **Folytonos** – tehát minden ponton értelmezhető (legalább egy intervallumon belül), például az 1 és 2 cm között van 0,5cm, de 0,25 vagy 0,22453cm is, a skála tetszőleges finomsággal felbontható. Bár ez így rendkívül pontatlan, és a két fogalom nem azonos, sokszor a skála típusú változókra utalunk a folytonos változó elnevezéssel.

Skála típusú adatok

- Néhány kivétel:
  - **Kvázi intervallum típusú** – a változó szigorú értelemben véve ordinálisnak tekintendő, de bizonyos tulajdonságai miatt olyan próbákban is használható, melyek feltétele a legalább intevallum szintű változó.
    - **Intelligencia** – az IQ számolásából adódóan csak ordinális változónak tekinthető, de az elemzések során kvázi intervallum skála típusúnak tekintjük.
    - **Likert-skála** – Statisztikailag a Likert-skála egy iteme ordinálisnak tekintendő, de sok esetben kvázi intervallum típusúnak tekintjük, és az itemekből számolt skálákkal már, mint skála típusú változóval szoktunk számolni.
  - **Dummy változó** – megoldás arra, hogy miként lehetne kategoriális változókat betenni olyan elemzésbe, melybe eredetileg csak skála típusúak kerülhettek.
    - Tegyük fel, hogy van egy nemzetiség változónk angol, német, francia kategóriákkal, és szeretnénk ezt a regresszióba betenni. Készítsünk egy darab nemzetiség változóból két dummy változót: (1) angolság, ahol 1, ha angol, és 0, ha nem angol, vagyis német vagy francia (2) németség, ahol 1, ha német, és 0, ha nem német, vagyis angol vagy francia. A harmadikra dummyra, a franciaságra nincs szükség, hiszen egyértelműen adódik, ha a németség és angolság is nulla, akkor franciáról van szó. Az így kapott két dummy változó már dichotóm, tehát betehető a regressziós elemzésbe.

Emlékeztető a normál eloszlásról

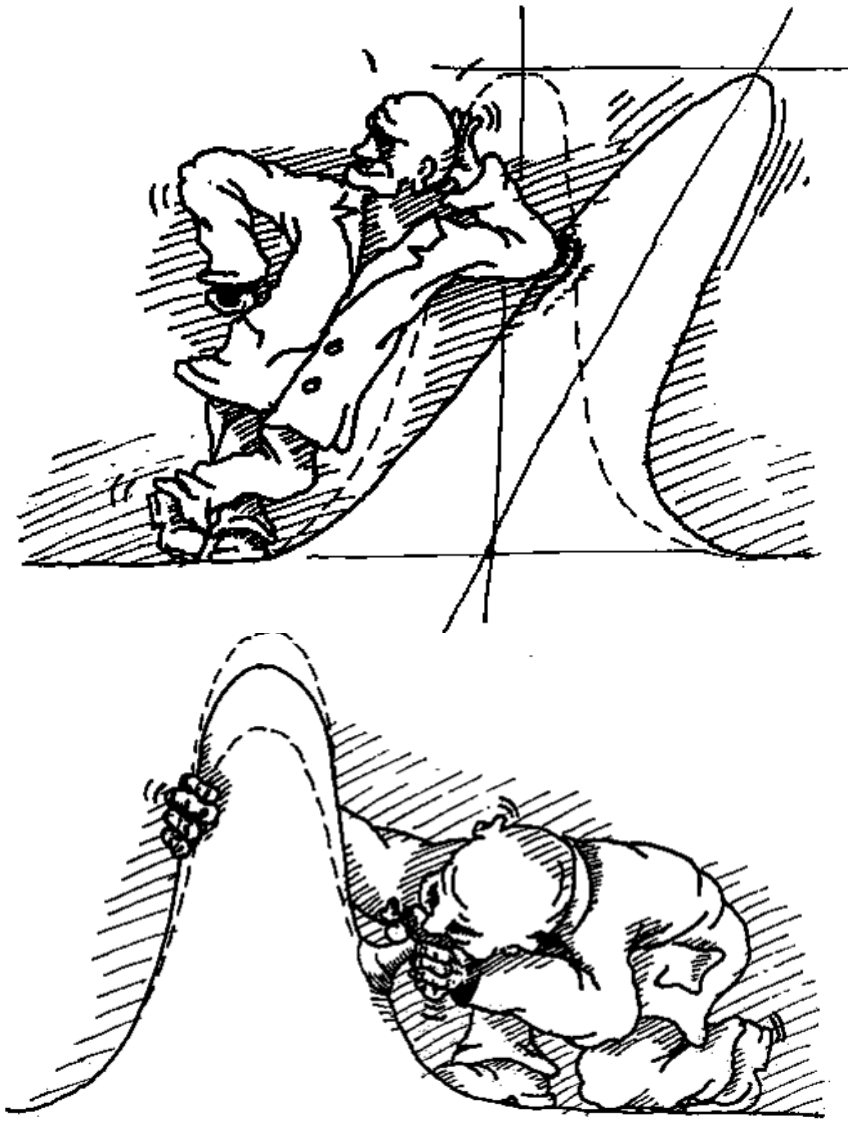
- **Normál** vagy **Z** vagy **Gauss-eloszlás**
  - **Folytonos** változók eloszlásának leírására alkalmas, **Haranggörbe alakú**, **Unimodális, Szimmetrikus** az átlag körül, és bár értéke az egész számegyenesen nézve soha nem csökken nullára, három szórás távolságra gyakorlatilag annak tekintető
  - A társadalomtudományokban vizsgált tulajdonságok nagy részének eloszlása jól leírható vele főként a central limit theorem miatt, ezért a statisztikai próbák nagy része valamilyen szempontból feltételezi a normál eloszlást



- **Central limit theorem** vagyis **központi határeloszlás tétel**
  - Adott feltételek teljesülése esetén **kellően nagy számú egymástól független**, meghatározható átlaggal és szórással rendelkező **változó számtani átlaga a populációban normál eloszláshoz közelít**, **függetlenül a változók eloszlásától**.
  - Emberi nyelven: ha sok, elég nagy mintát veszünk egy akármilyen eloszlású populációból, a mintaátlagok eloszlása normális lesz.

Emlékeztető a normál eloszlásról

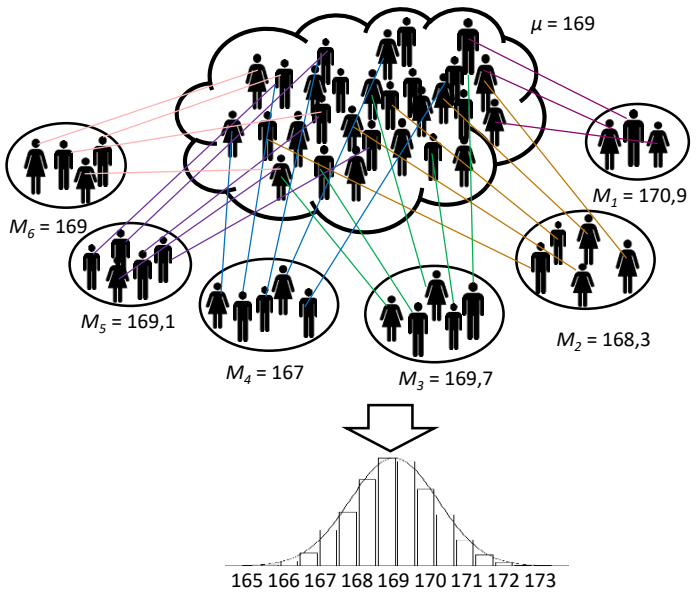
- Két fontos eltérés
  - Ferdeség (skewness) és csúcsosság (kurtosis)
- **Ferdeség**:
  - A normál eloszlás ferdesége 0 (szimmetrikus)
  - Pozitív ferdeség: az az eloszlásgörbe pozitív irányba nyúlik el
  - Negatív ferdeség: az eloszlásgörbe negatív irányba nyúlik el
  - Értéke az egész számegyenesen értelmezett, de ±3-at ritkán haladja meg
- **Csúcsosság**:
  - A normál eloszlás csúcsossága 0
  - Pozitív csúcsosság: az az eloszlásgörbe csúcsos
  - Negatív csúcsosság: az eloszlásgörbe lapos
  - Értéke az egész számegyenesen értelmezett, de ±3-at ritkán haladja meg

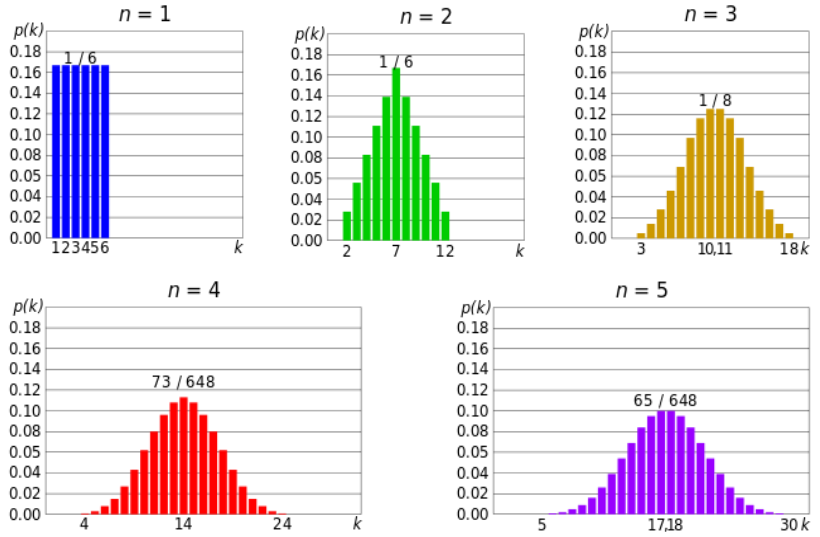


1998 G. Meixner

Normalitás feltétele

- Minek kell normál eloszlást követni?
  - A függő változónak? A populációnak? A mintának? A mintaátlagoknak? A becslés hibájának?
- 1. A **mintából becsült érték**nek: a mi esetünkben ez mindig a **mintaátlagok normál eloszlását** jelenti
  - Tehát a normalitás feltétele az, hogy ha veszek több mintát, kiszámolom a mintaátlagokat, akkor a kapott átlagok normál eloszlást kell kövessenek
  - Hogyan biztosítható ez, hiszen a mintaátlagok eloszlása nem ismert?
    - **Central limit theorem** kimondja, ha sok, elég nagy mintát veszünk egy akármilyen eloszlású populációból, a mintaátlagok eloszlása mindig normális lesz.
    - Tehát, ha elég nagy a mintám, akkor feltételezhetem, hogy a normalitás feltétele teljesül.
  - Mi az elég nagy minta?
    - Pontosan nem tudjuk, de még az olyan populációnál is, mely eloszlása normálistól a lehető legjobban eltér (exponenciális) **30-40 fő** felett eléri a mintaátlagok eloszlása a normálgörbe-formát

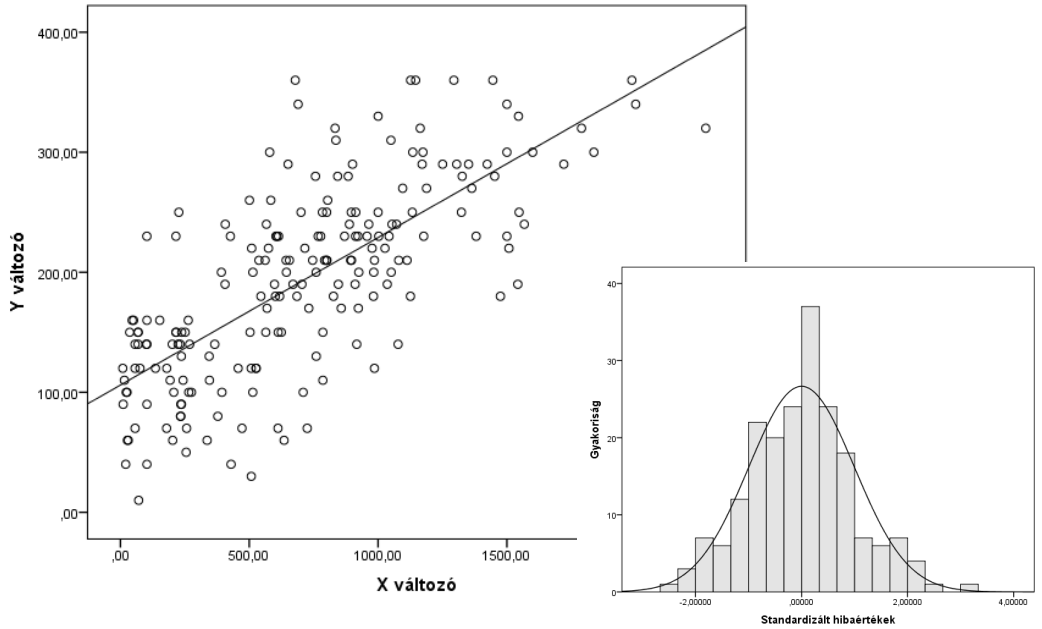


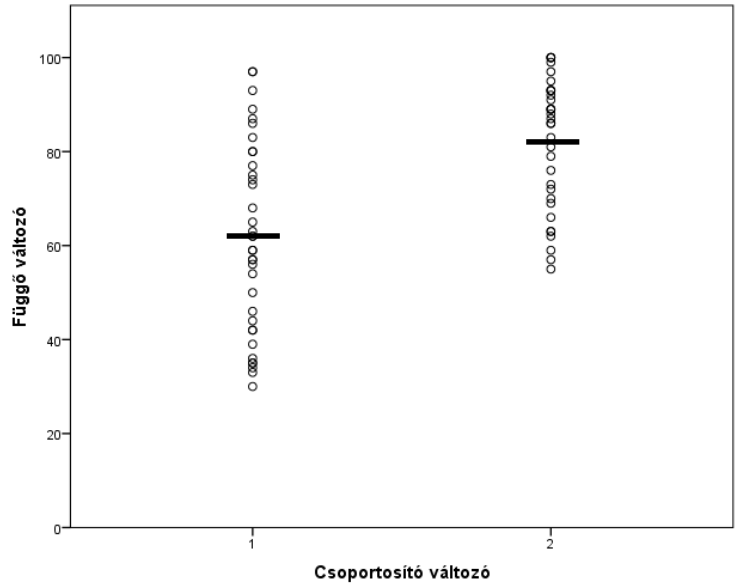


Normalitás feltétele

- 2. a **becslés hibájának normál eloszlását** is feltétezzük
  - A becslés értéke független a becslés hibájától, tehát a hiba a véletlen mintavételezésből adódó zaj, ergo normál eloszlást mutat.
  - Minták összehasonlítása esetén a becsült értékek a mintaátlagok lesznek, a hiba pedig a mintákon belüli változatosság, ezért nem egészen pontos, de elterjedt gyakorlat a **minták** (a függő változó mintánkénti) normál eloszlását ellenőrizni (pl. t-próba feltételeként).

- Végül populáció normál eloszlása
  - Nem ellenőrizhető
  - Nem feltétel
  - De egy normál eloszlást követő populációban könnyebben teljesülnek a feltételek (pl. kisebb minta is elég lenne)
  - Szerencsére a pszichológia és egyéb társadalomtudományok által vizsgált tulajdonságok általában sokténye zősek, ezért eloszlásuk gyakran közel van a normál eloszláshoz (central limit theorem miatt).



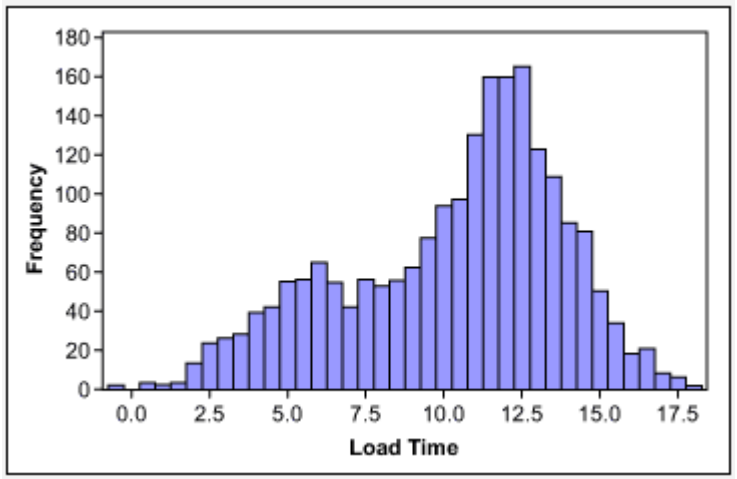


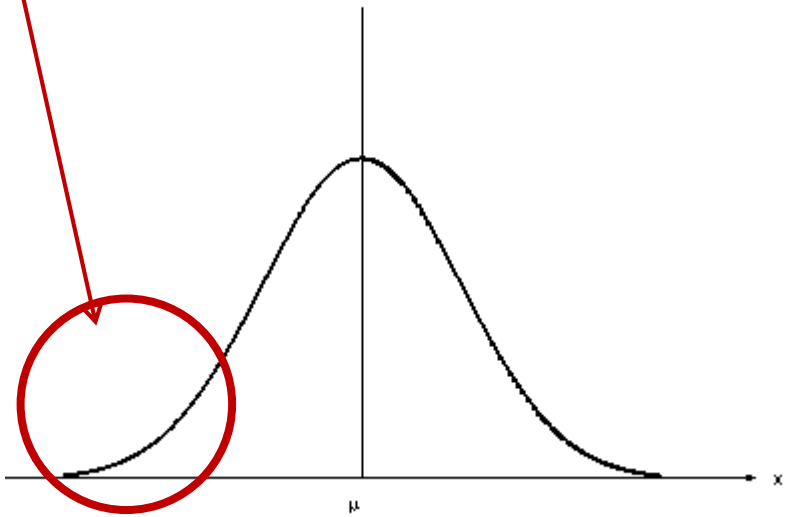
Normalitás feltételének ellenőrzése

- Grafikusan
  - **Hisztogram**
- Statisztikailag
  - három **elégséges** feltétele van a normalitásnak, **BÁRMELYIK teljesül a három közül, a normalitás feltételét teljesültnek tekinthetem**
    - A) **Central-limit theorem**
      - Mintánként 40 fő felett feltételezhetjük, hogy a mintaátlagok eloszlása normál eloszlást követ a central limit theorem miatt, tehát a normalitás feltétele teljesül
    - B) **Shapiro-Wilk** vagy **Kolmogorov-Smirnov teszt**
      - Nem parametrikus tesztek, ellenőrzik, hogy a mintánk eloszlása szignifikánsan különbözik-e egy előre meghatározott eloszlástól (a mi esetünkben a normál eloszlástól)
      - A szignifikáns eredmény jelentése, hogy a minta eloszlása szignifikánsan különbözik a normál eloszlástól, tehát a normalitás feltétele nem teljesül
    - C) **z-tesztek**
      - Mintánként 15 fő felett, ha a minta outlierek nélküli, eloszlása unimodális, ferdesége és csúcsossága nem tér el szignifikánsan a normál eloszlás esetén várhatótól, a feltételezhetjük a normalitás feltételének teljesülését

Ha nincs normalitás?

- Mit csináljak, ha nem-normális?
  - **Kutatói mérlegelés kérdése, melyik mellett döntesz**
  - **Találd meg a választ, miért nem normális!**
    - Outliererek
    - 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!**
    - Outlierek és ferde eloszlás esetén például jól működhet az adatokból való gyökvonás vagy a logaritmizálás
  - **Használj non-parametrikus próbákat!**

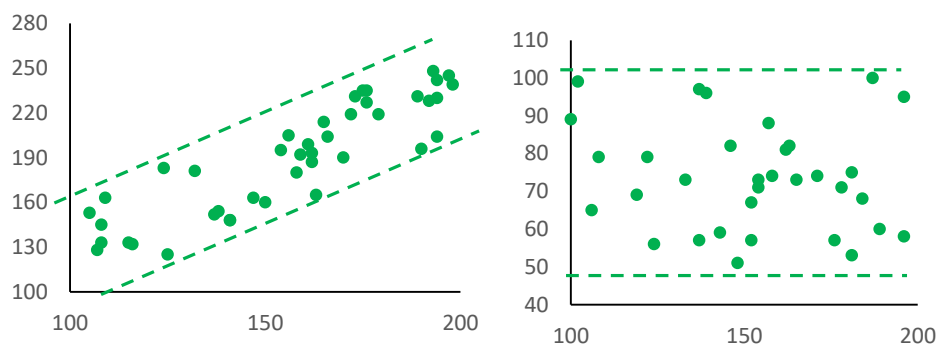


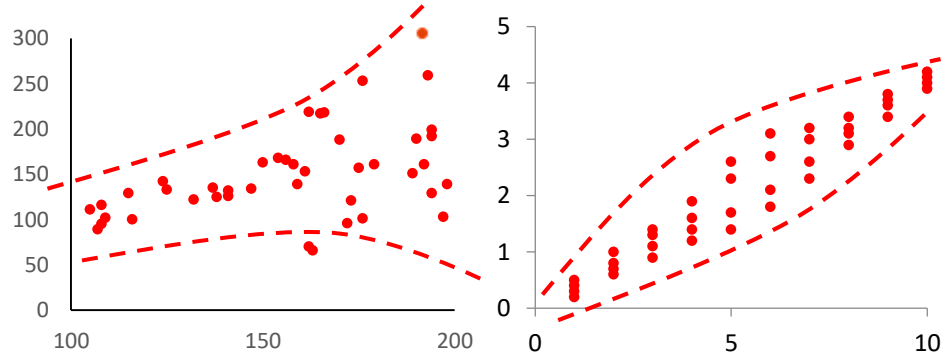


Szóráshomogenitás

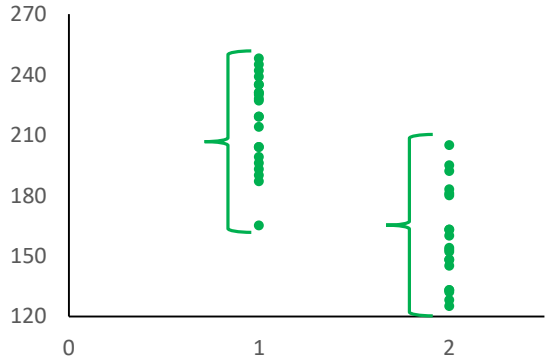
- **Varianciák homogenitása / Szóráshomogenitás**
  - A varianciáknak egyformának kell lenniük az egész mintában
  - Ha korrelációt, akkor a változónk varianciájának stabilnak kell lennie a másik változó minden szintjén
  - Ha csoportokat vizsgálunk, a különböző csoportok varianciájának kell azonosnak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

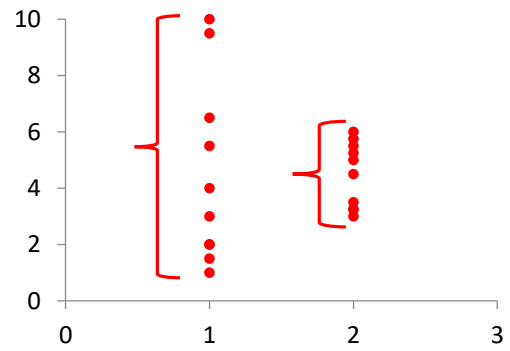
Változók közötti kapcsolat





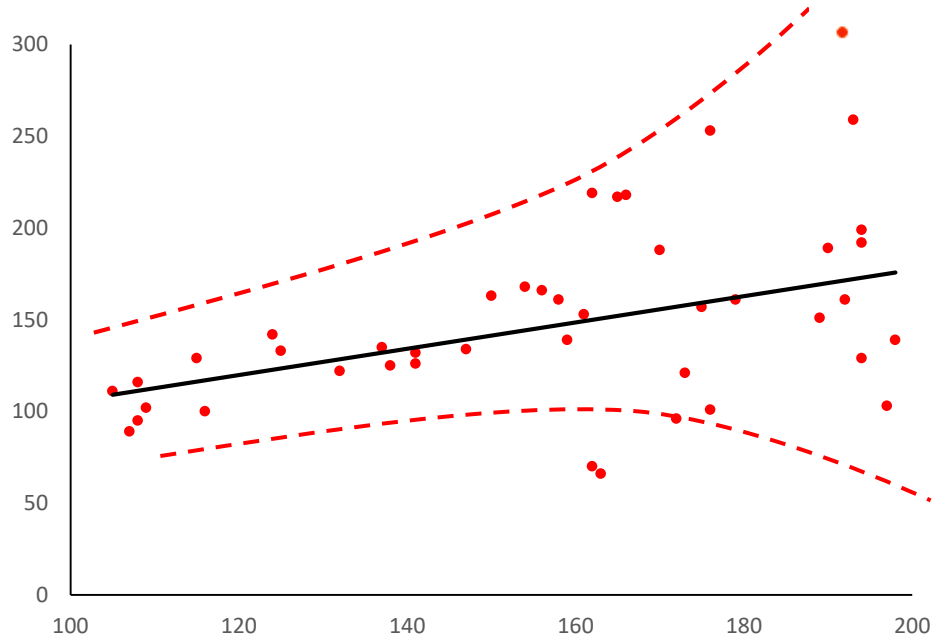
Minták közötti különbség

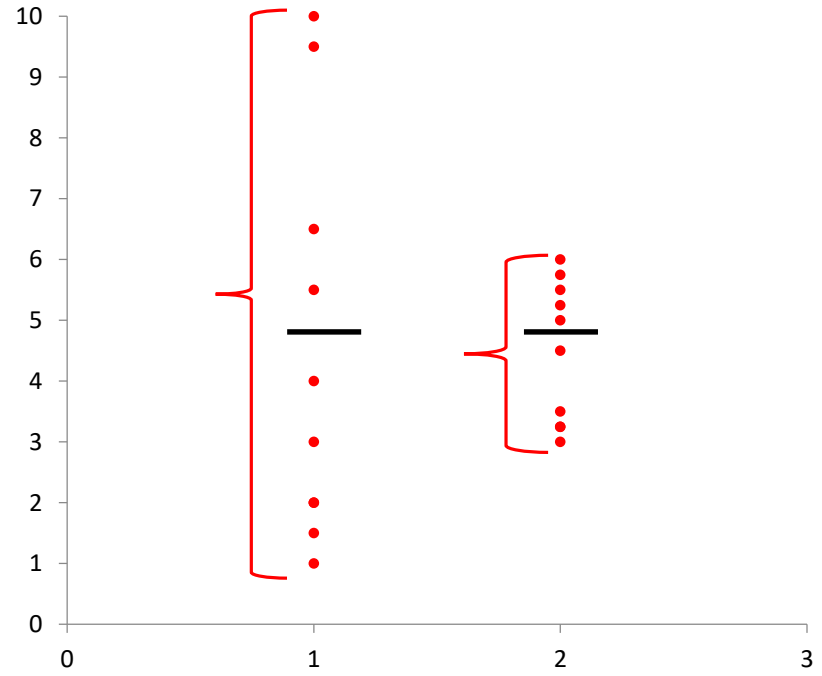




Szóráshomogenitás

- **Miért baj, ha sérül a feltétel?**
  - A modell nem mindenhol lesz egyformán pontos
  - (Pontatlanná válik a becsült variancia-kovarancia mátrix, ezért) pontatlanok lesznek a paraméterekhez tartozó standard error-ok – pontatlanok lesznek a szignifikancia értékek

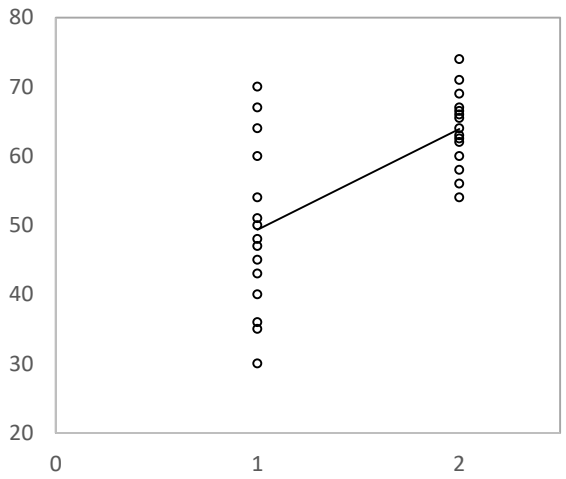




Szóráshomogenitás és elemszámok aránya

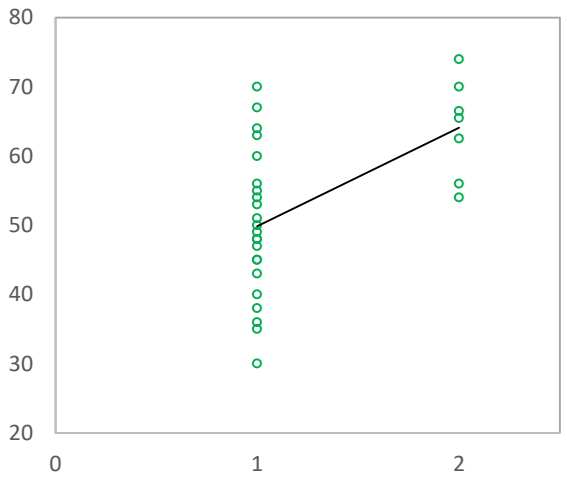
- **Következmények?**
  - A szóráshomogenitás sérülésének hatásai függnek a minta más tulajdonságaitól is.

SD1> SD2 és N1 = N2



$$\frac{hatás}{hiba}$$

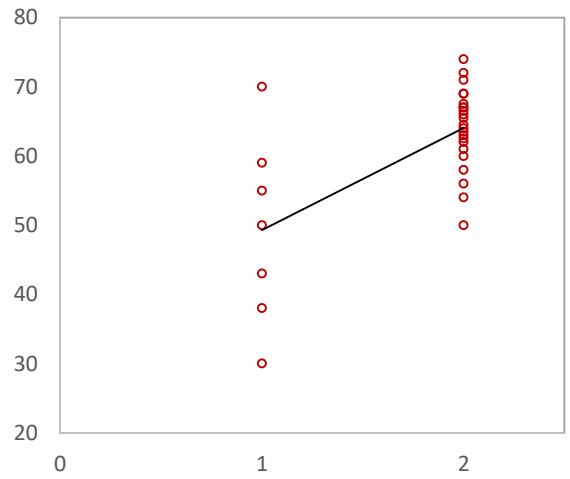
SD1> SD2 és N1 > N2



$$\frac{hatás}{hiba}$$ ⇨ Stat. érték csökken, p-érték nő

A modell kevésbé lesz szignifikáns, tehát a próba szigorodott

SD1> SD2 és N1 < N2



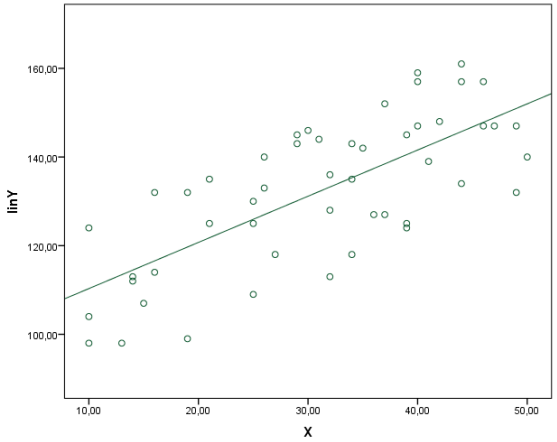
$$\frac{hatás}{hiba}$$ ⇨ Stat. érték nő, p-érték csökken

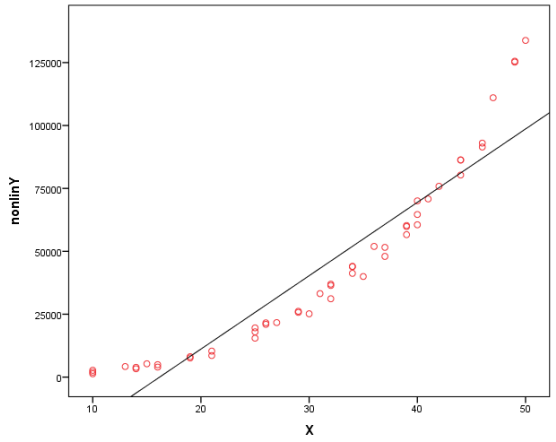
A modell szignifikánsabb lett, tehát a próba megengedőbbé vált, az elsőfajú hiba valószínűsége megnőtt

Linearitás

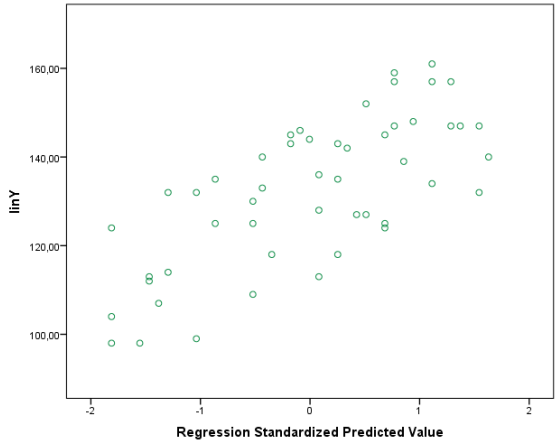
- **Linearitás**
  - A legtöbb statisztika, amit használni fogunk lineáris modellt használ.
  - Ha nem teljesül, pontatlanok lesznek a felállított modell általi becslések – ez kiemelten igaz extrapoláció esetén, plusz alul fogod becsülni a magyarázóerőket.

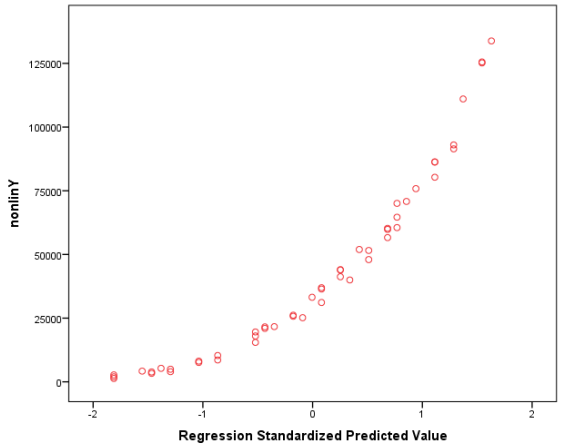
Prediktor és kimeneti érték kapcsolata



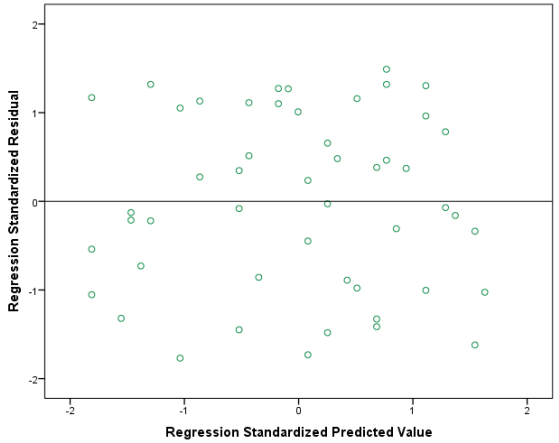


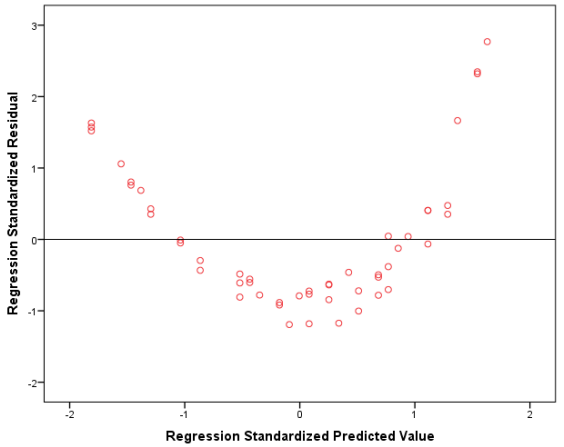
Predikált és kimeneti érték kapcsolata





Predikált érték és hiba kapcsolata



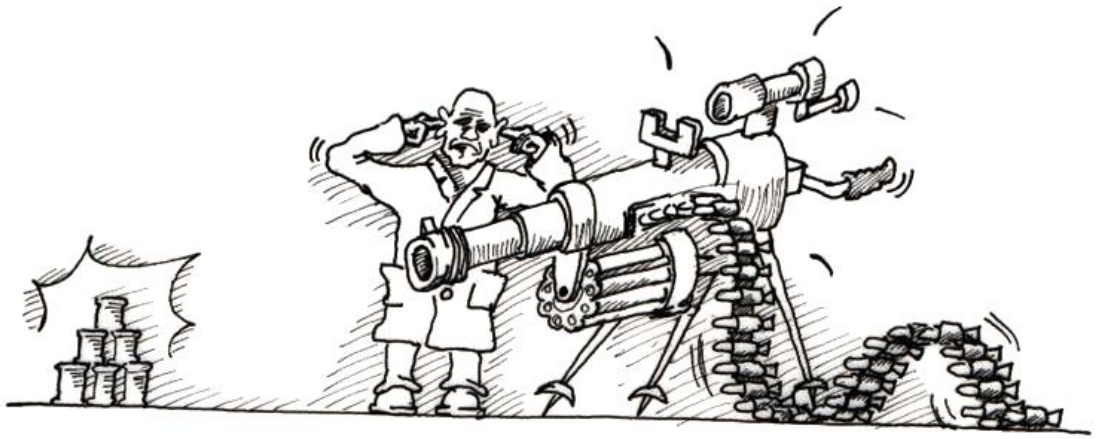


Feltételek – mennyire fontosak?

- **Ha egy próba robosztus egy feltétel megszegésére egy adott mintán, akkor a feltétel sérülése ellenére is értelmezhetőek az eredmények**
  - Azt látjuk, hogy a próbák feltételeinek robosztussága hat egymásra, például:
    - ANOVA robosztus a szóráshomogenitás bármilyen mértékű sérülésére, ha a minták elemszáma hasonló, a normalitás teljesül, és N>15
    - ANOVA robosztus a szóráshomogenitás kis mértékű sérülésére, ha a minták elemszáma hasonló, a normalitás teljesül és a legkisebb és legnagyobb variancia aránya < 3.5
    - ANOVA NEM robosztus a szóráshomogenitás sérülésére, bármekkora is az elemszám, ha a minták elemszáma különböző, és a legkisebb és legnagyobb variancia aránya > 8
    - ANOVA NEM robosztus a szóráshomogenitás sérülésére, ha változók eloszlása egymással ellentétes irányban ferde
    - ANOVA robosztus a normalitás sérülésére, ha a szóráshomogenitás teljesül és az elemszám N > 15

Feltételek – mennyire fontosak?

- **Robosztusság ellenőrzése Monte-Carlo szimulációval**
  - 1. Létrehozunk egy populációt amiben nincs jelen a hatás! (pl. X tulajdonság átlaga, szórása, eloszlása azonos A és B csoportban)
  - 2. A létrehozott populációból random mintavételezéssel veszünk nagyon sok mintát, és elvégezzük rajtuk a vizsgálni kívánt próbát (pl. t-próbát ezer vagy 10ezer alkalommal)
    - Hány esetben várom, hogy szignifikáns eredményt kapjak?
      - A statisztikák 5%-nál várunk szignifikáns eredményt (ennyi lesz az első fajú hibát)
  - 3. Ezt követően kiválasztok mintákat, ahol a vizsgálni kívánt feltétel valamilyen mértékben sérül (pl. az egyik mintának kétszer akkora a szórása mint a másiknak), plusz egyéb követelményeknek is megfelelnek (pl. adott eloszlásúak, elemszámúak stb.)
    - **Egy próba robosztus egy feltétel megszegésére, ha a szignifikáns eredmények továbbra is 4-6% között maradnak**



1998. G. Meixner