Slide 1

A tárgyalt statisztikai eljárások alkalmazási feltételei

(összefoglalás)

Slide 2

A statisztikai eljárások tárgyalásának fontos része az alkalmazás feltételeinek felsorolása.

Amikor szignifikancia határokat számolunk, a statisztikai modell valószínűségszámítási tulajdonságaira támaszkodunk.

Az előadásban megkülönböztettük az "teljesen korrekt" és a "közelítően korrekt" alkalmazási feltételeket.

Slide 3

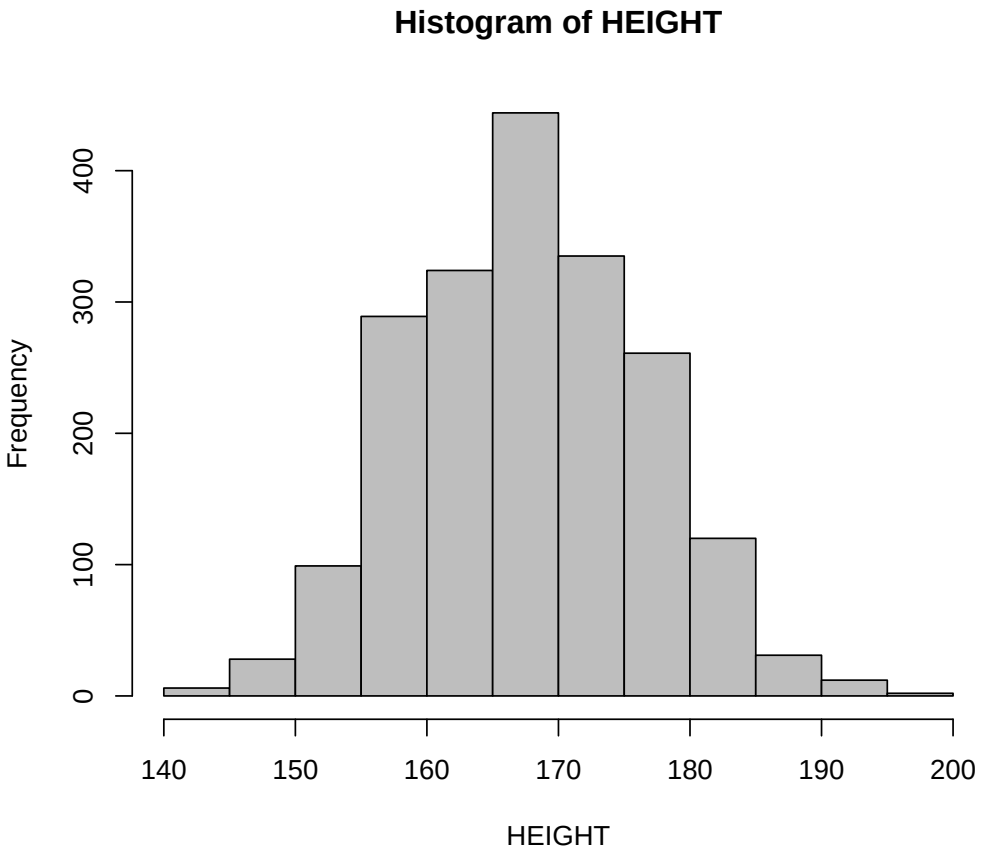
A $Y = \ell(x) + \varepsilon$ lineáris regresszió teljesen korrekt alkalmazásának feltételei:

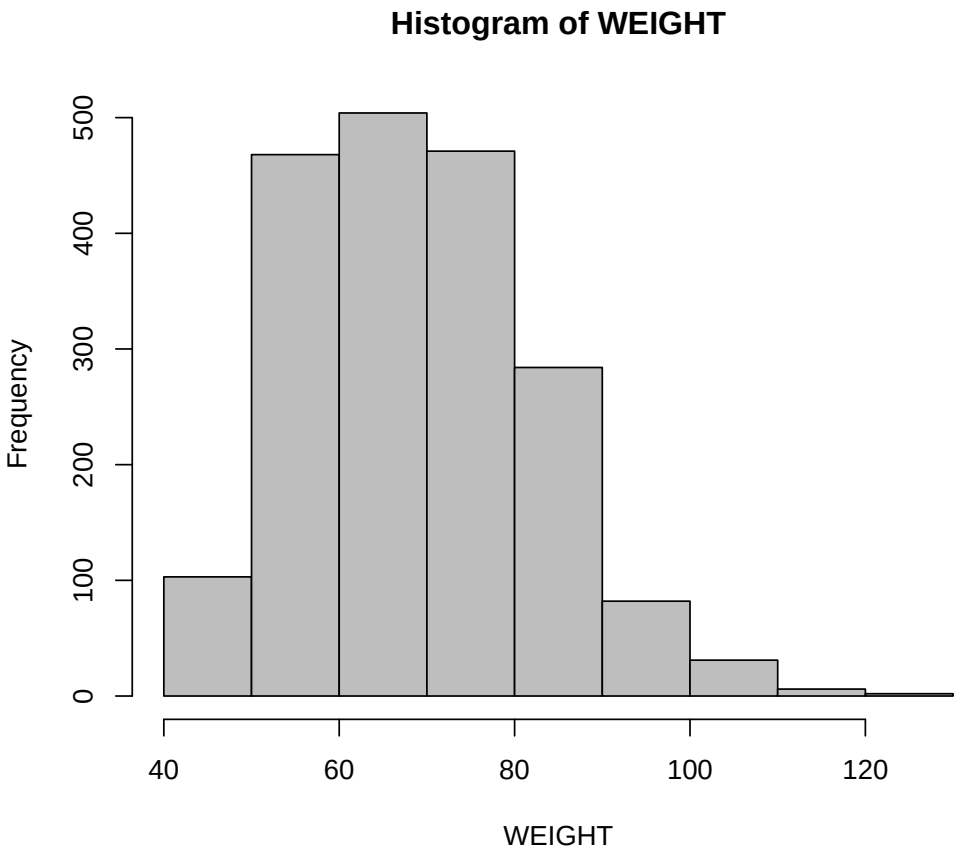
- $Y \mid X = x$ normális eloszlású,
- $\mathbb{E}(Y \mid X = x)$ lineáris x -ben,
- $Var(Y \mid X = x)$ állandó x-ben,
- az Y -ra vett megfigyelések függetlenek.

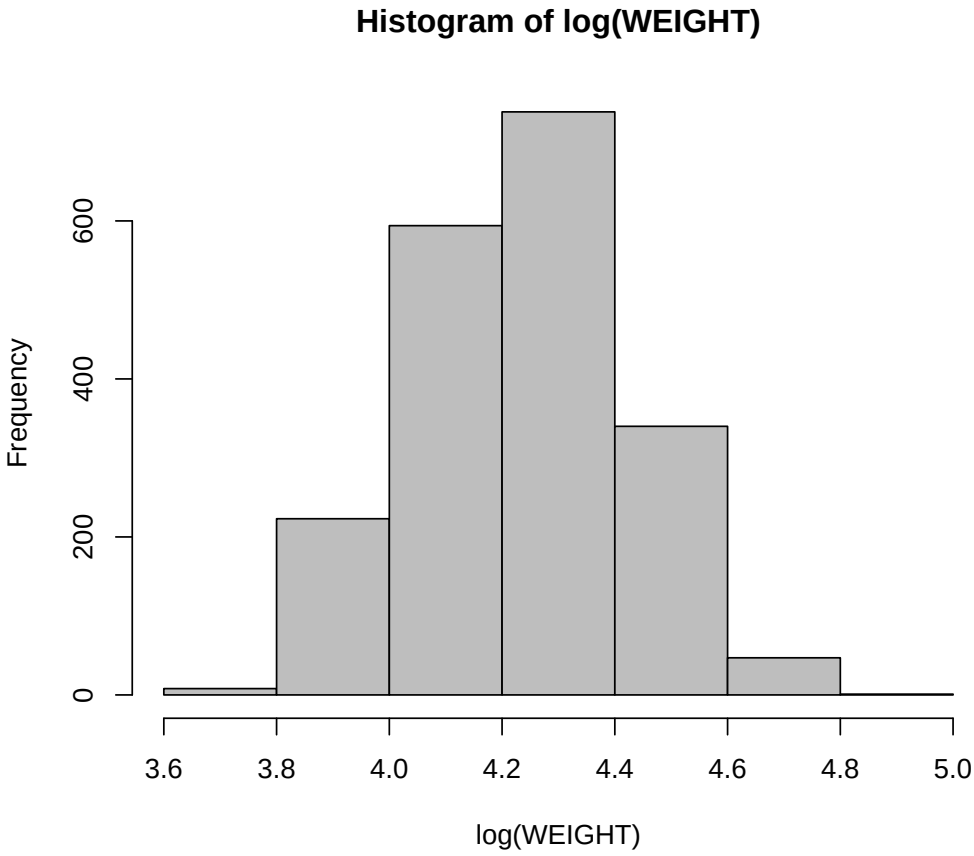
Slide 4

Az első feltétel a normalitás, melyről azt tanultuk, hogy a regresszió robusztus a normalitásra.

A robusztusság a matematikai statisztikában körültekintően (de bonyolultan) definiált fogalom, mi szemléltető példákon kersztül tárgyaltuk.




Histogram of WEIGHT
Frequency
0
100
200
300
400
500
40
60
80
100
120
WEIGHT


Histogram of log(WEIGHT)
Frequency
0
200
400
600
3.6
3.8
4.0
4.2
4.4
4.6
4.8
5.0
log(WEIGHT)

Slide 5

A testmagasság változó hisztogramján nem látjuk semmi jelét annak, hogy ez a változó ne lenne jó közelítéssel normális eloszlású.

Slide 6

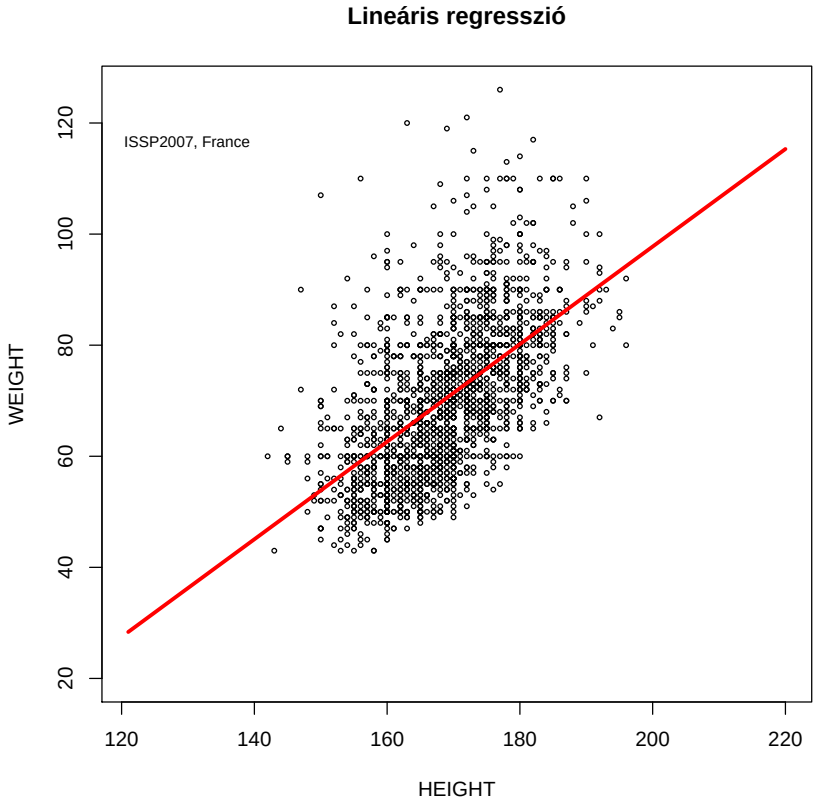
Ámde a testsúly változó hisztogramja jellegzetesen balra ferde eloszlást mutat, a minimális érték sokkal közelebb van a mediánhoz, mint a maximális.

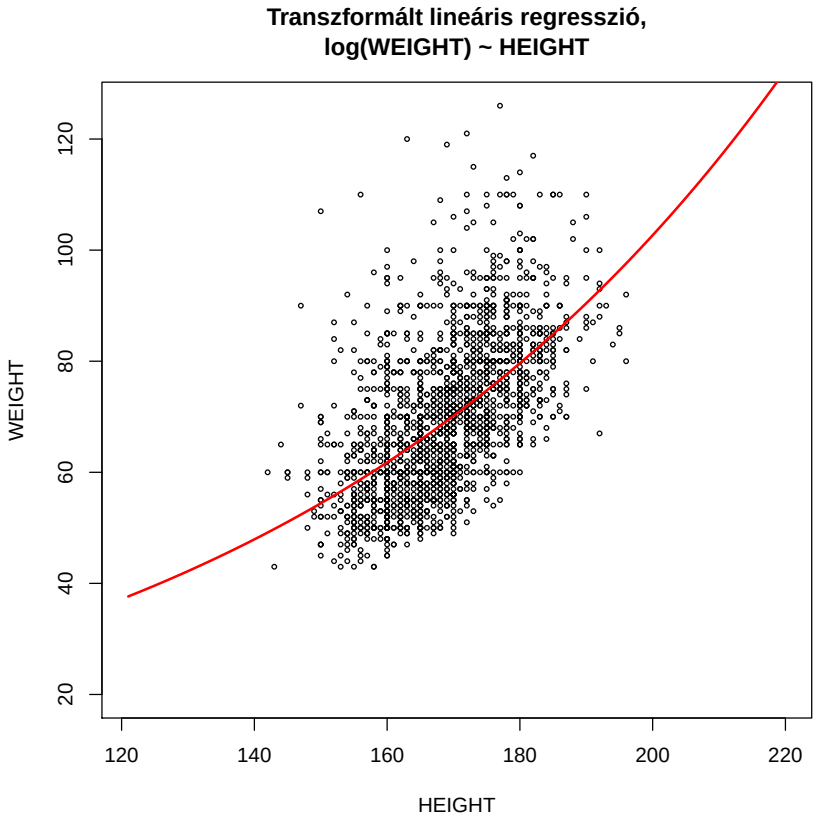
Ezért e változónak az (e -alapú)logaritmusát vettük, e transzformáció után az eloszlás láthatólag közelebb került a normálishoz.

Slide 7

Ezután a testsúly változónak a testmagasság változóra vonatkozó lineáris regresszóját, majd ennek transzformált változatát nézzük meg.

Megbeszéltük, hogy az eredeti és a transzformált modell közül nem lehet mindkettő teljesen korrekt.


Lineáris regresszió
ISSP2007, France
WEIGHT
HEIGHT
20
40
60
80
100
120
120
140
160
180
200
220


Transzformált lineáris regresszió,
log(WEIGHT) ~ HEIGHT
WEIGHT
HEIGHT
20
40
60
80
100
120
120
140
160
180
200
220

Slide 8

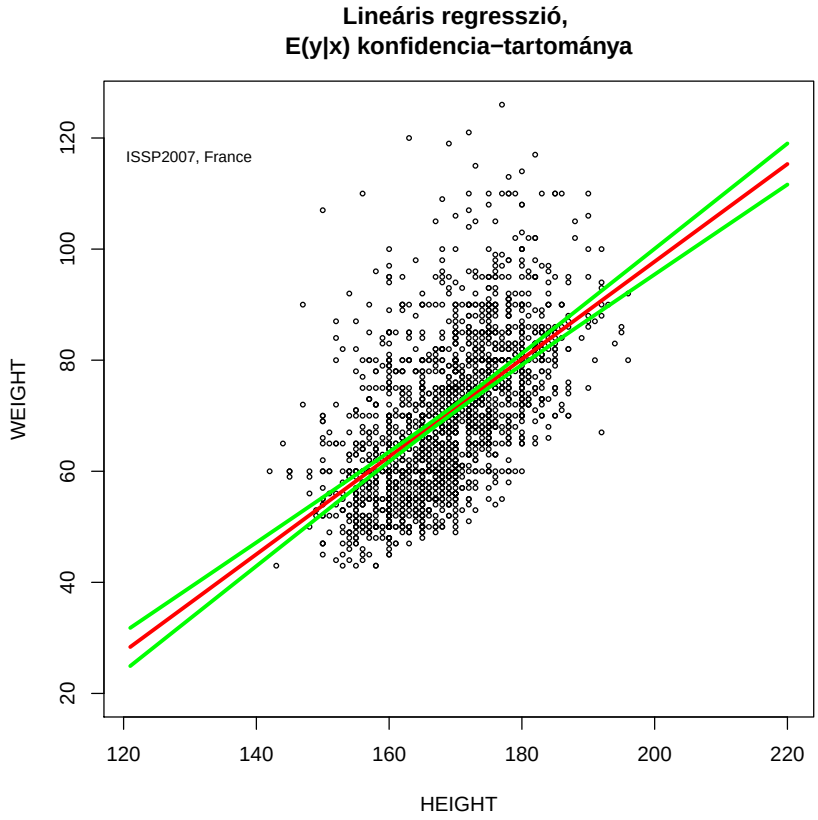
Megbeszéltük, hogy lineáris regressziót alkalmazunk a $log(Y)$ változóra, majd a becsült értékeket a $log()$ transzformáció inverzével, az $exp()$ transzformációval az Y eredeti értékeivel azonos skálán ábrázoljuk, és az R^2 illeszkedési mutatót is ebből számoljuk.

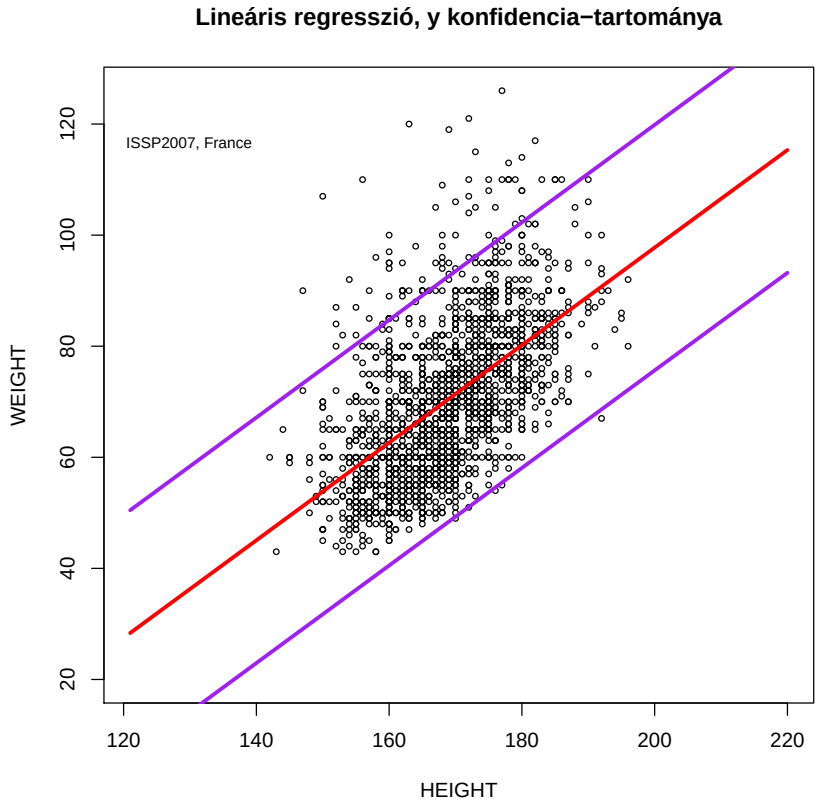
Slide 9

Az ISSP példában a transzformáció során nem sikerült jelentős R^2 javulást elérni, viszont javult (=csökkent) a regresszió reziduálisainak x -től való függése.

Slide 10

A becslések konfidencia-intervallumát kétféle ábrán közöljük:
- a feltételes várható értékre adott becslés konfidencia-határai (a szélesség függ x-től),
- az Y ra adott becslés konfidencia-határai (a szélesség nem függ x-től).


Lineáris regresszió,
E(y|x) konfidencia−tartománya
ISSP2007, France
WEIGHT
HEIGHT
20
40
60
80
100
120
120
140
160
180
200
220


Lineáris regresszió, y konfidencia−tartománya
ISSP2007, France
WEIGHT
HEIGHT
20
40
60
80
100
120
120
140
160
180
200
220

Slide 11

A transzformált lineáris regresszió előnye különösen jól látszik az utóbbi ábrázolásnál, mert az X (testmagasság) nagyobb értékeinél az Y (testsúly) láthatóan nagyobb szórásának megfelelően a konfidencia-határok széttartóak.

Slide 12

Az elmondottak a regressziónak a normalitás követelményében való robusztusságát illusztrálják: a $log()$ transzformációval közelebb jutottunk a kimeneti változó normális eloszlásához, és kedvezőbb illeszkedést kaptunk.

Ez a részletesebb kifejtése annak, hogy a lineáris regresszió "közelítően korrekt módon" alkalmazható, ha a kimeneti változóra vonatkozó normalitás közelítően teljesül.

Slide 13

A lineáris regresszió teljesen korrekt alkalmazásának feltételei közül a normalitás, a feltételes várható érték linearitása és a feltételes szórás állandósága olyan követelmények, melyekre nézve az eljárás robusztus.

Ámde a mintaelemek függetlenségének feltétele nem ilyen: ennek "kis" megsértése is eredményezhet "nagyon" rosszul illeszkedő modellt.

Slide 14

Összefoglalóan azt mondhatjuk, hogy ahol a z- és t-próba, a regresszió és az ANOVA esetén a teljesen korrekt alkalmazási feltételben szereplő normalitásra nézve minden eddig tárgyalt eljárás robusztus.

Slide 15

Ámde nincs robusztusság a mintavételi függetlenség tekintetében: a kétmintás z- és t-próbáknál, az illeszkedésvizsgálatnál a csoportokban és a csoportok között is teljesülnie kell a mintavételi függetlenségnek.

Slide 16

Az összetartozó mintás eljárásoknál (páros t-próba, függetlenségvizsgálat, regresszió-számítás) a megfigyelési egységek között kell teljesülnie a függetlenségnek, de ugyanazon megfigyelési egységen belül nem.

Slide 17

Két azonos elsőfajú hibavalószínűségű próba közül hatásosabb (erősebb) az, amelyiknek kisebb a másodfajú hibavalószínűsége.

A próbákat természetesen azonos mintanagyság mellett kell összehasonlítani, ugyanazon az alapsokaságon.

Nem garantálható, hogy különböző alapsokaságok és mintanagyságok mellett mindig ugyanaz a próba bizonyul erősebbnek.

Slide 18

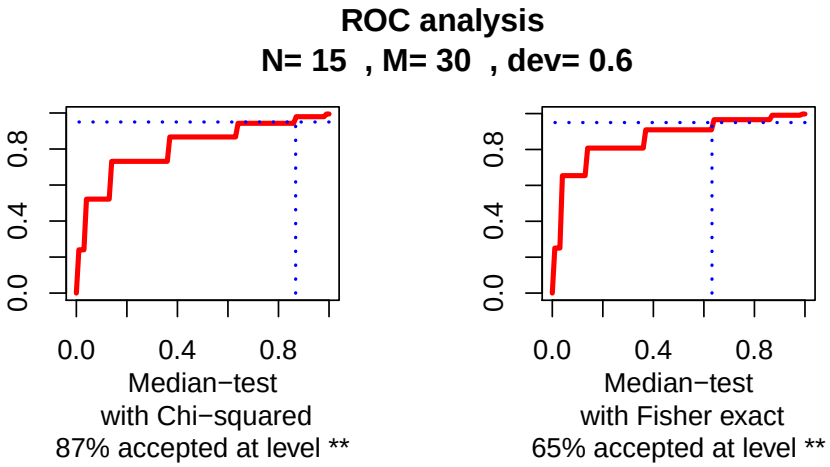
Például ha a z-próba alkalmazási feltételei teljesülnek, akkor z-próba hatásosabb (erősebb), mint a t-próba (mint tanultuk, a két eljárás között annyi a különbség, hogy a t-próba nem veszi figyelembe az alapsokasági szórásra vonatkozó információt).

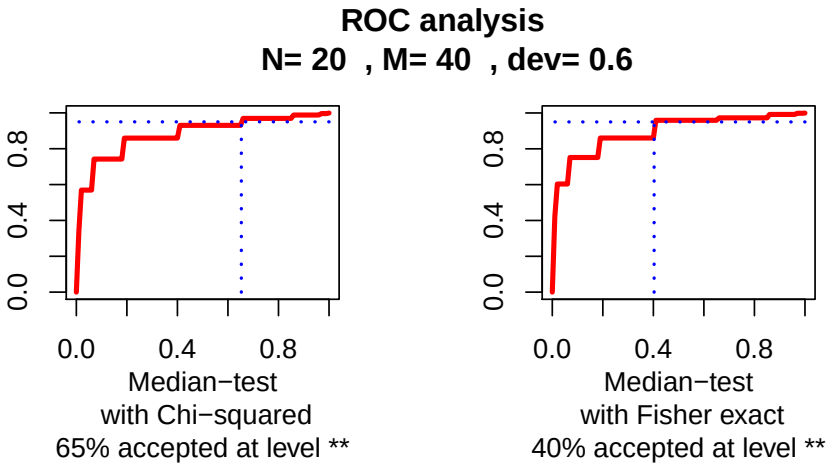
Slide 19

Ha az alapsokasági szórásra vonatkozó információ hibás, és ezt az adatot a z-próbánál használjuk, akkor az Első- ás másodfajú hibavalószínűségek előadás 3. mintafeladatánál látott módon alalkulnak a hibavalószínűségek.

Slide 20

Az általunk tanult minden próbára igaz, hogy azonos alapsokaság és elsőfajú hibavalószínűség mellett nagyobb mintanagyságnál kisebb a másodfajú hiba.





Slide 21

Az ROC elemzések az alábbi mintára vonatkoznak:

$\{X_1, ...X_N\}$ EVM $\mathcal{N}(+0.6, 1)$-re,
$\{Y_1, ...X_M\}$ EVM $\mathcal{N}(-0.6, 1)$-re,
a két mintavétel egymástól független

Legyen
$MED =$ az $N + M$ elemű egyesített minta mediánja.

Slide 22

Megszámoljuk, melyik mintában hányan vannak MED-nél nagyobbak, hányan MED-nél kisebbek.

N_1 olyan X van, ami $\leq MED$
N_2 olyan X van, ami $> MED$
M_1 olyan Y van, ami $\leq MED$
M_2 olyan Y van, ami $> MED$

Slide 23

Táblázatban:

	X	Y
$\leq MED$	N_1	M_1
$> MED$	N_2	M_2

Slide 24

Az összefüggésvizsgálat előadáson megbeszéltük, hogy ebben az esetben a próba a hipotézis elfogadásával másodfajú hibát vét, ezért a bemutatott szimulációs eredmények szerint a medián próba erősebb a (Pearson-féle) khi-négyzet próbánál.

Slide 25

Hatásosabb két torzítatlan becslés közül az, amelyiknek kisebb az SH-ja.

Megbeszéltük például (Mintavétel öf. 6. oldal), hogy az EVM és a rétegzett mintavétel egyaránt alkalmas az alapsokasági átlag torzítatlan becslésére, de az utóbbi általában hatásosabb.

Slide 26

A tárgyalt becslési eljárások mind rendelkeznek azzal a tulajdonsággal, hogy nagyobb mintanagyság mellett kisebb az SH.

Ezért azt is mondhatjuk, hogy az a hatásosabb becslés, amely ugyanazt az SH értéket (torzított becslés esetén SE értéket) kisebb mintanagyság mellett éri el.

Slide 27

Megbeszéltük a háromváltozós összefüggésvizsgálat egy speciális kérdését, a feltételes függetlenséget.

Példákkal bizonyítottuk, hogy két változó függetlenségéből nem következik a feltételes függetlenségük, és a feltételes függetlenségből sem a függetlenségük.