PARAMÉTERES PRÓBÁK

Szórásra, varianciára vonatkozó próbák

Egy minta esetén: ***X^2-próba varianciára***

Két minta esetén: **F-próba**

Több minta esetén: **Levene-próba** *VAGY* **Bartlett-próba**

Várható értékre vonatkozó próbák

Egy minta esetén: ***egymintás t-próba***

Két független minta esetén: **Welch-próba**

Két (páronként) összefüggő minta esetén: **páros t-próba**

Több független minta esetén: **Varianciaanalízis (ANOVA)**

egytényezős, teljes véletlen elrendezésű ANOVA

Eloszlásra vonatkozó NEMPARAMÉTERES PRÓBÁK

X^2-próba illeszkedésvizsgálatra

H_0: az alapsokaság eloszlása megegyezik a hipotetikus eloszlással
H_1: az alapsokaság eloszlása nem egyezik meg a hipotetikus eloszlással

Feltétel: a megfigyelt és a várt gyakoriság kategóriánként legalább 5 legyen.

X^2-próba függetlenségvizsgálatra

H_0: a két változó független egymástól
H_1: a két változó nem független egymástól *(…!)*

Feltétel: a megfigyelt és a várt gyakoriság kategóriánként legalább 5 legyen.

Alternatív lehetőség függetlenségvizsgálatra: a ***Fisher-féle egzakt próba***

X^2-próba homogenitásvizsgálatra

H_0: a minták azonos eloszlású alapsokaságokból származnak
H_1: a minták nem azonos eloszlású alapsokaságokból származnak

Hol tartunk?

A statisztika feladatai
– adatgyűjtés
az adatok bemutatása, szemléltetése (statisztikák, ábrák) ✓

Megismertünk néhány statisztikai mutatót (várható érték, variancia, szórás, medián, kvartilisek), és az adatok szemléltetésére használt ábrák közül párat (hisztogram, boxplot, …).

– becslések készítése az adatok, tehát a minta alapján (pontbecslés, intervallumbecslés), ✓

Pontbecslések, konfidenciaintervallum várható értékre (és varianciára).

– hipotézisvizsgálat
tipikus esetei:
→ megcáfoljuk, hogy két (vagy több) mennyiség átlaga (…) egyenlő ✓
→ megcáfoljuk, hogy két mennyiség független ✓

Megismertük a hipotézisvizsgálat alapjait, felépítését, a döntési algoritmusokat, és a számtalan hipotézisvizsgálat közül néhány, gyakran előforduló módszerről részletesebben is beszéltünk. Még néhányat megemlítettünk, sok százról nem esett szó.

– asszociáció, korreláció, regresszió (változók közötti kapcsolat kimutatására, jellemzésére)

Na, ezekről még egyáltalán nem esett szó!

KORRELÁCIÓ- ÉS REGRESSZIÓANALÍZIS

Két változó közötti összefüggéseket vizsgálunk. (Ha van összefüggés…)

Asszociáció: VAN-E kapcsolat a két – akár nominális – változó között?

Korreláció: VAN-E (lineáris!) kapcsolat a két – folytonos – változó között?

Regresszió: MILYEN kapcsolat van a két (folytonos) változó között?

Asszociáció – *VAN-E kapcsolat a két – akár nominális – változó között? Mennyire erős a kapcsolat…?*

- nominális változókra is (színek, vegyszerek, …)
- kontingenciatáblázat alapján
- általában 0 és 1 közötti az értéke, 1: ha a *kérdéses jellegű* kapcsolat áll fenn a két változó között
- Cramer-féle V: $V = \sqrt{\frac{\mathrm{X}^2_{számított}}{n \cdot \min(oszlopok\ száma - 1, sorok\ száma - 1)}}$

  Pearson-féle C: $C = \sqrt{\frac{\mathrm{X}^2_{számított}}{\mathrm{X}^2_{számított} + n}}$

  Goodman-Kruskal-féle λ:

  $\lambda = \frac{P(Y\text{-re vonatkozó tévedés, ha } X \text{ értékét nem ismerjük}) - P(Y\text{-re vonatkozó tévedés, ha } X \text{ értékét ismerjük})}{P(Y\text{-re vonatkozó tévedés, ha } X \text{ értékét nem ismerjük})}$

  nem szimmetrikus!

Rangkorreláció – *Mennyire erős a monotonitási kapcsolat a két változó között?*

- rangszám: nagyság szerinti sorszám (legkisebb: 1, legnagyobb: n, esetleg: kapcsolt rangok)
- legalább ordinális változó (az értékeinek létezik természetes sorrendje)
- Spearman-féle rangkorrelációs együttható $r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} {d_i}^2}{n^3 - n}$, ahol d_i az összetartozó rangok különbsége
- értéke -1 és 1 közötti, -1 körüli értékek erős negatív, 1 körüli értékek erős pozitív monotonitást jeleznek

Korreláció: – *VAN-E lineáris kapcsolat a két változó között? Mennyire erős a lineáris kapcsolat…?*

- folytonos változókra
- (összetartozó) adatpárok alapján
- *Pearson-féle korrelációs együttható:*

$$R(X,Y) = \frac{E\left(\left(X - E(X)\right) \cdot \left(Y - E(Y)\right)\right)}{D(X) \cdot D(Y)} = \frac{E(X \cdot Y) - E(X) \cdot E(Y)}{D(X) \cdot D(Y)}$$

- értéke -1 és 1 közötti, -1 körüli értékek erős negatív, 1 körüli értékek erős pozitív korrelációt jeleznek (negatív, illetve pozitív meredekségű egyenes), 0 körüli érték esetében a két változó között nincsen (lineáris) kapcsolat
- *a korreláció az asszociáció speciális esete!*

Regresszió – *MILYEN (függvényszerű) kapcsolat van a két (vagy több) változó között?*

- folytonos változókra, összetartozó adatpárok alapján!
- a ***függő változó*** (y) értékeit szeretnénk minél pontosabban megadni a ***magyarázó változó(k)*** (x vagy x_1, x_2, … x_k) értékének (értékeinek) ismeretében
- a magyarázó változó értéke „nem függ a véletlentől”

Egyváltozós lineáris regresszió

Modellegyenlet: **$Y = b_0 + b_1 \cdot X + \varepsilon$**

ahol Y a függő változó (response variable, dependent variable), X a magyarázó változó (explanatory variable, independent variable), b_0 a konstans tag / y tengellyel való tengelymetszet, b_1 együtthatója, azaz az egyenes meredeksége, és ε a véletlen hibatag / reziduum / maradéktag

ε várható értéke 0, eloszlása normális kell legyen, illetve értéke legyen független x értékétől

Célunk: b_0 és b_1 (*együtthatók, koefficiensek*) értékére minél jobb becslést adni, az adataink felhasználásával.

Ehhez a ***legkisebb négyzetek elve*** szerint *az* $\sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2 = \sum_{i=1}^{n}\left(y_i - (b_0 + b_1 \cdot x_i)\right)^2$ *négyzetösszeget (azaz a reziduumok négyzetösszegét) minimalizáljuk*.

(ábra!)

Ez alapján megkapjuk a b_0 és b_1 értékeket, aminek segítségével felírhatjuk az *illesztőfüggvény* egyenletét. (Például *y = -2 + 5x*, vagy *y = 1 - 1/2·x*.)

Ezután azt vizsgáljuk, hogy a (már kész) modellünk mennyire jó, azaz mennyire jól írja le az Y értékét az X értékének ismeretében.

1. ***Determinációs együttható**: R^2* értéke 0 és 1 közötti (ez az *R korrelációs együttható* négyzete): a modell mekkora részben magyarázza a függő változó (y) ingadozását.

   Adjusztált R^2: több magyarázó változó esetén a változók számával korrigált R^2 érték, minél több változó van a modellben, annál kisebb az értéke.

2. *Modellre vonatkozó F-próba („modellre vonatkozó ANOVA")*

   H_0: A modell nem magyarázza a függő változó ingadozását („a modell rossz").
   H_1: A modell magyarázza a függő változó ingadozását („a modell jó").

   $F_{számított} = \frac{SS_{Modell}}{SS_{Hiba}} \cdot \frac{n-1-k}{k}$, ahol SS_{Modell} a függő változó ingadozásának a modellünk által magyarázott része, SS_{Hiba} a függő változó ingadozásának a modell által nem magyarázott (véletlennek tulajdonított) része, n a mintaelemszám (a modell elkészítéséhez felhasznált adatpárok száma), és k a szabadsági fok: a becsült paraméterek száma mínusz 1 (jelen esetben 1, mert b_0-t és b_1-et becsültük, ami két paraméter).

3. *Együtthatókra vonatkozó t-próbák*

   i) b_0-ra: H_0: $b_0 = 0$
   H_1: $b_0 \neq 0$

   Ha $b_0 = 0$, akkor nincs rá szükség, akár ki is hagyhatjuk a modellből.

   ii) b_1-re: H_0: $b_1 = 0$
   H_1: $b_1 \neq 0$

   Ha $b_1 = 0$, akkor a modellünk használhatatlan. (De még mindig jobb ha ezt legalább tudjuk…)

4. *Feltételellenőrzés* *a) Reziduumok normalitása*

   b) Reziduumok függetlensége x-től

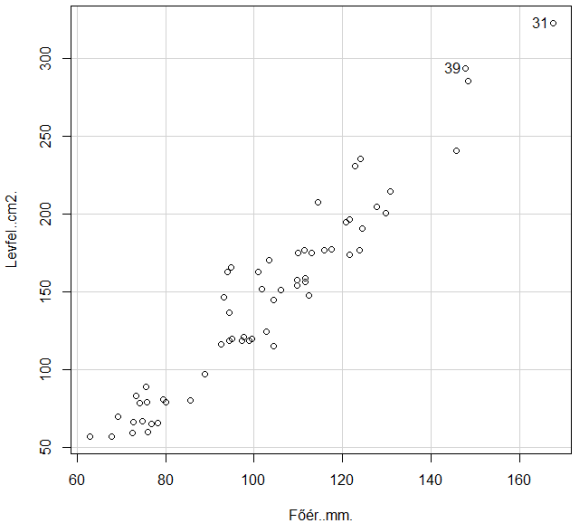
Figyelem! A modellünk csak az X változó azon tartományára érvényes, ahonnan az adataink származnak!

(…)

Példa: főérhossz-levélfelület (1. hét)

Egy 60 elemű szőlőlevélmintában a főerek hosszát és a levélfelületet mérték. Van-e ezek között szoros függvényszerű kapcsolat, melynek segítségével a későbbiekben a főér hosszából a levélfelület becsülhetővé válik?

Itt a függő változó (y) a levélfelület, a magyarázó változó (x) a főérhossz.



Az ábra alapján a lineáris jó közelítés.

Általában: elsősorban szakmai indokok szerint válasszanak modellt, csak végső esetben az ábra alapján!!!

Modellegyenlet: $Y = b_0 + b_1 \cdot X + \varepsilon$

Konkrétan: *levélfelület* $= b_0 + b_1 \cdot$ *főérhossz* $+ \varepsilon$

Call:
lm(formula = Levfel..cm2. ~ Főér..mm., data = foer)

Residuals:

Min	1Q	Median	3Q	Max
-35.318	-12.442	-4.508	10.884	40.641

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-125.1971	10.5785	-11.84	<2e-16 ***
Főér..mm.	2.6370	0.1007	26.18	<2e-16 ***

Signif. codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**

Residual standard error: 17.48 on 58 degrees of freedom
Multiple R-squared: 0.922, Adjusted R-squared: 0.9206
F-statistic: 685.1 on 1 and 58 DF, p-value: < 2.2e-16

Az illesztőfüggvény egyenlete: **Y = -125,2 + 2,6·X**

avagy

levélfelület = -125,2 + 2,6·főérhossz

Regressziós diagnosztika – a modell elemzése, vizsgálata

1. Determinációs együttható: $R^2 = 0{,}922$

Adjusztált $R^2 = 0{,}921$

2. Modellre vonatkozó F-próba („modellre vonatkozó ANOVA”)

H_0: A modell nem magyarázza a függő változó ingadozását („a modell rossz”).

H_1: A modell magyarázza a függő változó ingadozását („a modell jó”).

$p < 2{,}2 \cdot 10^{-16}$

3. Együtthatókra vonatkozó t-próbák

i) b_0-ra: H_0: $b_0 = 0$

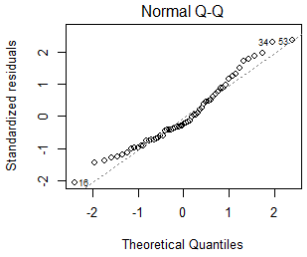
H_1: $b_0 \neq 0$

$p < 2{,}2 \cdot 10^{-16}$

ii) b_1-re: H_0: $b_1 = 0$

H_1: $b_1 \neq 0$

$p < 2{,}2 \cdot 10^{-16}$

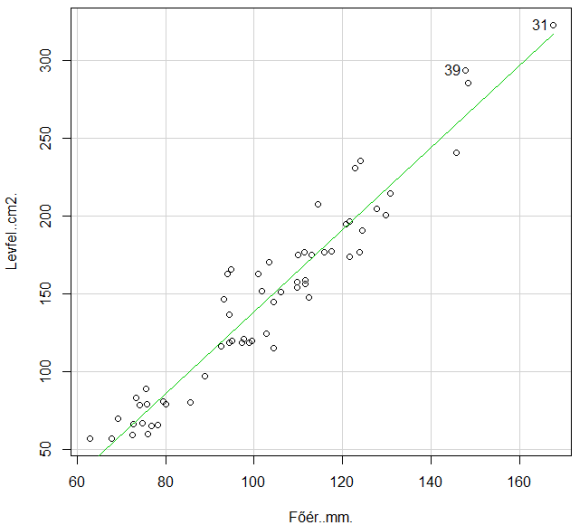


4. Reziduumok normalitása

RegModel.1$residuals

RegModel.1$fitted.values

Ábrán: Options / Least-squares line



Megjegyzés: a reziduum fogalma itt az ANOVA esetében értelmezett reziduum fogalomtól nem független ugyan, de némi különbséggel definiáljuk. ***Reziduum****ok* alatt a mért (y) értékeknek a (illesztett modell alapján) becsült értékektől való eltérését értjük.*

* reziduum, véletlen hibatag, maradéktag