STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

S 73 (M 34)

Toets voor de hypothese dat een regressiecoëfficiënt nul is, wanneer de afwijkingen van de regressielijn normaal verdeeld zijn met gelijke spreidingen.

MATHEMATISCH CENTRUM,
2de Boerhaavestr. 49,
A M S T E R D A M - O.

Statistische Afdeling
S 73 (M 34)

Toets voor de hypothese dat een regressiecoëfficiënt nul is, wanneer de afwijkingen van de regressielijn normaal verdeeld zijn met gelijke spreidingen[1].

Ondersteld wordt:

1° De grootheid ξ kan zonder fouten worden waargenomen, terwijl de fouten[2] in de waarnemingen voor de grootheid η onderling onafhankelijk normaal[3] verdeeld zijn met gemiddelde 0 en onbekende, maar steeds dezelfde, spreiding σ.

2° Tussen de grootheden η en ξ bestaat een lineair verband:

$$\eta = \beta\xi + \alpha,$$

waarin de grootheden β en α onbekende parameters zijn.

3° Bij iedere waarneming x_i $(i = 1, \ldots, h)$ van ξ wordt een groepje, onderling onafhankelijke, waarnemingen $y_{i1}, \ldots, y_{in_i}$ voor η verricht.

Gevraagd wordt op grond van het waarnemingsmateriaal:

$$\begin{array}{l} (x_1,y_{11}),\ (x_1,y_{12}),\ \ldots,\ (x_1,y_{1n_1}) \\ \vdots \\ (x_h,y_{h1}),\ (x_h,y_{h2}),\ \ldots,\ (x_h,y_{hn_h}) \end{array}$$

de hypothese te toetsen dat de regressiecoëfficiënt β gelijk aan nul is:

$$H_0:\ \beta = 0.$$

Om de hypothese H_0 te toetsen, worden schattingen b, a en s^2 voor β, α en σ^2 berekend met behulp van de methode der kleinste kwadraten, welke in dit geval overeenkomt met de methode

---

1) Dit memorandum is slechts bedoeld ter oriëntatie en streeft niet naar volledigheid of volledige exactheid.

2) Onder fouten vallen in dit geval, behalve meetfouten, ook toevallige afwijkingen van andere aard, b.v. physiologische afwijkingen.

3) De stochastische grootheid $\underline{x}$ bezit een normale verdeling met gemiddelde μ en spreiding σ, indien voor iedere a geldt:

$$P[\underline{x} \leqq a] = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{1}{2}\frac{(u-\mu)^2}{\sigma^2}}\, du.$$

der meest aannemelijke schattingen (Eng: method of maximum likelihood).

Dit geeft:

$$b = \frac{\sum_{i=1}^{h} n_i(x_i - \overline{x})y_i}{\sum_{i=1}^{h} n_i(x_i - \overline{x})^2},$$

$$a = \overline{y} - b\overline{x},$$

$$s^2 = \frac{\sum_{i=1}^{h} \sum_{j=1}^{n_i} (y_{ij} - bx_i - a)^2}{N - 2} =$$

$$= \frac{\sum_{i=1}^{h} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2 + \sum_{i=1}^{h} n_i(y_i - bx_i - a)^2}{N - 2},$$

waarin:

$$N = \sum_{i=1}^{h} n_i; \qquad \overline{x} = \frac{\sum_{i=1}^{h} n_i x_i}{N};$$

$$\overline{y}_i = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i} \quad \text{en} \quad \overline{y} = \frac{\sum_{i=1}^{h} n_i y_i}{N}.$$

Beschouwen we niet alleen de waarnemingen y_{ij} ($j = 1, \ldots, n_i$; $i = 1, \ldots, h$), maar de verzameling van alle mogelijke waarden, die waargenomen kunnen worden, dan bezitten zij op deze verzameling een waarschijnlijkheidsverdeling, welke afhankelijk is van de (exact waargenomen) waarden $x_1, \ldots, x_h$. De schattingen b, a en s^2 zijn dan eveneens stochastisch, hetgeen door onderstreping van de letters wordt aangegeven.

De grootheden $\underline{b}$ en $\underline{s}$ zijn onderling onafhankelijk verdeeld, hetgeen b.v. door Mood [1] , pp. 292 - 294, bewezen wordt. De grootheid $\underline{b}$ bezit een normale verdeling met gemiddelde β en spreiding:

$$\frac{\sigma}{\sqrt{\sum_{i=1}^{h} n_i(x_i - \overline{x})^2}},$$

terwijl de grootheid

$$\frac{(N-2)\underline{s}^2}{\sigma^2}$$

verdeeld is volgens een χ^2-verdeling met (N-2) vrijheidsgraden. Hieruit volgt dat de stochastische grootheid

$$\underline{t} = \frac{(\underline{b} - \beta)\sqrt{\sum_{i=1}^{h} n_i(x_i - \overline{x})^2}}{\underline{s}}$$

verdeeld is volgens de verdeling van Student met (N-2) vrijheidsgraden.

Onder de hypothese H_o: $\beta = 0$ gaat deze grootheid over in

$$\underline{t} = \frac{\underline{b}\sqrt{\sum_{i=1}^{h} n_i(x_i - \overline{x})^2}}{\underline{s}},$$

die dus onder de hypothese H_o weer bovengenoemde Studentverdeling bezit.

Is $\beta \neq 0$, dus de hypothese niet juist, dan zullen verder van nul gelegen waarden van $\underline{t}$ grotere waarschijnlijkheid bezitten, dan dan wanneer $\beta = 0$. Voor toetsing van de hypothese gebruikt men daarom in het twee-zijdige geval (als zowel $\beta > 0$ als $\beta < 0$ kan optreden) een twee-zijdige kritieke zône:

$$|\underline{t}| \geqq t_o,$$

waarbij t_o, behorende bij een bepaalde onbetrouwbaarheid α, opgezocht kan worden in tabellen, vermeld in onderstaande litteratuurlijst.

Indien slechts alternatieve mogelijkheden van de vorm $\beta > 0$ of $\beta < 0$ worden toegelaten, gebruikt men een eenzijdige kritieke zône:

$$\underline{t} > t_1 \text{ resp. } \underline{t} < -t_1,$$

waarin t_1 in dezelfde tabellen kan worden opgezocht als t_o, door te zoeken in een tabel voor de tweezijdige toets met onbetrouwbaarheidsdrempel 2α.

Het bij de tabellen vermelde aantal vrijheidsgraden (vaak aangegeven door ν of n-1 of n') is in dit geval gelijk aan N-2.

Opmerking: Er kan bewezen worden (zie [5]), dat de hier beschreven toets voldoet aan het, door J. Neyman ingevoerde, λ-principe (zie [6]).

Litteratuur:

[1] A.M. Mood, Introduction to the theory of statistics, Mc Graw-Hill, New York, Toronto, London, 1950, hoofdstuk 13 (regressietheorie) en p. 425 (tabel).

[2] H. Cramér, Mathematical methods of statistics, Princeton Univ. Press, Princeton, 1946.

[3] M.G. Kendall, The advanced theory of statistics I, Griffin and Co., London, 1947, pp. 440 - 441 (tabel).

[4] C.W. Emmens, Principles of biological assay, Chapman and Hall Ltd, London, 1948, pp. 18 - 19 (tabel).

[5] H.B. Mann, Analysis and design of experiments, Dover Publications, New York, 1949, pp. 41 - 42.

[6] J. Neyman, First course in probability and statistics, Henry Holt and Co., New York, 1950, pp. 338 - 343.