STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

S 102 (M 40)

Toetsen met betrekking tot tweedimensionale normale verdelingen

1953

MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM
STATISTISCHE AFDELING

Leiding: Prof. Dr D. van Dantzig
Chef van de Statistische Consultatie: Prof. Dr J. Hemelrijk

S 102 (M 40)

Toetsen met betrekking tot tweedimensionale normale verdelingen [1].

Inhoud:

1953.

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

1 Tweedimensionale normale verdeling in het algemeen.

De algemene gedaante van de verdelingsdichtheid van een tweedimensionale normale verdeling is:

$$(1)\quad f(x_1,x_2)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\,e^{-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2-2\rho\frac{x_1-\mu_1}{\sigma_1}\cdot\frac{x_2-\mu_2}{\sigma_2}+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right\}}$$

Hierin komen de volgende vijf parameters voor:

$$(2)\quad \begin{cases} \mu_1=\mathcal{E}\,\underline{X}_1 = \text{mathematische verwachting }^{1}) \text{ van } \underline{X}_1, \\ \sigma_1=\sqrt{\mathcal{E}(\underline{X}_1-\mu_1)^2} = \text{spreiding van } \underline{X}_1, \\ \left.\begin{matrix}\mu_2\\ \sigma_2\end{matrix}\right\} \text{ overeenkomstige grootheden voor } \underline{X}_2, \\ \rho=\dfrac{\mathcal{E}(\underline{X}_1-\mu_1)(\underline{X}_2-\mu_2)}{\sigma_1\sigma_2} = \text{correlatiecoëfficiënt van } \underline{X}_1 \text{ en } \underline{X}_2. \end{cases}$$

Het punt (μ_1,μ_2) noemen wij het middelpunt van de verdeling.

De parameter ρ is een maat voor de afhankelijkheid van de variabelen $\underline{X}_1$ en $\underline{X}_2$. Als $\rho=0$ zijn de twee variabelen stochastisch onafhankelijk, als $|\rho|=1$ zijn de variabelen lineair afhankelijk, d.w.z. dat er een rechte lijn in het (X_1,X_2)-vlak is, waar het stochastische punt $(\underline{X}_1,\underline{X}_2)$ met wh 1 op ligt.

Indien wij beschikken over een steekproef van onafhankelijke waarnemingen (X_{1i},X_{2i}) $(i=1,\ldots,n)$ uit de verdeling, kunnen wij de parameters op de volgende wijze schatten:

$$(3)\quad \begin{cases} \mu_1 \text{ door } \bar{X}_1=\frac{1}{n}\sum_{i=1}^{n}X_{1i} \text{ (gemiddelde van de waarnemingen van } \underline{X}_1), \\ \mu_2 \text{ door } \bar{X}_2=\frac{1}{n}\sum_{i=1}^{n}X_{2i} \text{ (gemiddelde van de waarnemingen van } \underline{X}_2), \\ \sigma_1 \text{ door } s_1=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_{1i}-\bar{X}_1)^2} \text{ (spreiding van de waarnemingen van } \underline{X}_1), \\ \sigma_2 \text{ door } s_2=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)^2} \text{ (spreiding van de waarnemingen van } \underline{X}_2), \\ \rho \text{ door } r=\frac{1}{s_1s_2}\cdot\frac{1}{n-1}\sum_{i=1}^{n}(X_{1i}-\bar{X}_1)(X_{2i}-\bar{X}_2) \\ \qquad \text{(correlatiecoëfficiënt van de waarnemingen van } \underline{X}_1 \text{ en } \underline{X}_2). \end{cases}$$

---

[1]) De mathematische verwachting van een grootheid wordt ook wel zijn (theoretische) gemiddelde genoemd.

2 Toets van HOTELLING voor het middelpunt van een tweedimensionale normale verdeling.

Deze toets is een generalisatie van de toets van STUDENT voor het gemiddelde van een normale verdeling (zie Memorandum S 47 (M 8)).

Gegeven zij een steekproef $(x_{1i}, x_{2i})(i = 1, \ldots, n)$ van onafhankelijke waarnemingen uit een tweedimensionale normale verdeling.

De te toetsen hypothese H_0 is, dat een gegeven punt (μ_1, μ_2) het middelpunt van de verdeling is.

De toetsingsgrootheid is de $\underline{T}^2$ van HOTELLING, gedefinieerd als volgt:

$$(4)\quad \underline{T}^2 = \frac{n}{1-\underline{r}^2}\left\{\left(\frac{\underline{\bar{x}}_1 - \mu_1}{\underline{s}_1}\right)^2 - 2\underline{r}\,\frac{(\underline{\bar{x}}_1 - \mu_1)(\underline{\bar{x}}_2 - \mu_2)}{\underline{s}_1\,\underline{s}_2} + \left(\frac{\underline{\bar{x}}_2 - \mu_2}{\underline{s}_2}\right)^2\right\}.$$

Als de nulhypothese juist is, zal

$$(5)\quad \underline{F} = \frac{1}{2}\cdot\frac{n-2}{n-1}\,\underline{T}^2,$$

verdeeld zijn als de $\underline{F}$ van FISHER-SNEDECOR met $\nu_1 = 2$ en $\nu_2 = n-2$ vrijheidsgraden (zie Memorandum S 53 (M 24)).

De kritieke zone is van het type $\underline{T}^2 \geqq T_0^2$ en komt dan overeen met een tweezijdige toets van STUDENT.

Litteratuur: A.HALD, Statistical Theory with engineering applications, New York and London, 1952, p. 607.

3 Toets van HOTELLING voor twee steekproeven.

De toets is een generalisatie van de toets van STUDENT voor twee steekproeven (zie Memorandum S 47 (M 9)). Zij wordt gebruikt voor het toetsen van de hypothese H_o dat twee steekproeven $(X_{1i}, X_{2i})(i = 1,\ldots,n)$ en $(Y_{1j}, Y_{2j})(j = 1,\ldots,m)$ beschouwd kunnen worden als steekproeven uit dezelfde tweedimensionale normale verdeling.

Als toetsingsgrootheid wordt hierbij gebruikt de $\underline{T}^2$ van HOTELLING voor twee steekproeven:

$$(5)\quad \underline{T}^2 = \frac{m\,n}{(m+n)(1-\underline{r}^2)}\left\{\left(\frac{\bar{\underline{X}}_1 - \bar{\underline{Y}}_1}{\underline{S}_1}\right)^2 - 2\,\underline{r}\,\frac{(\bar{\underline{X}}_1 - \bar{\underline{Y}}_1)(\bar{\underline{X}}_2 - \bar{\underline{Y}}_2)}{\underline{S}_1\,\underline{S}_2} + \left(\frac{\bar{\underline{X}}_2 - \bar{\underline{Y}}_2}{\underline{S}_2}\right)^2\right\},$$

waarin:

$$\bar{\underline{X}}_1 = \frac{1}{n}\sum_{i=1}^{n}\underline{X}_{1i}\quad,\quad \bar{\underline{X}}_2 = \frac{1}{n}\sum_{i=1}^{n}\underline{X}_{2i}$$

$$\bar{\underline{Y}}_1 = \frac{1}{m}\sum_{j=1}^{m}\underline{Y}_{1j}\quad,\quad \bar{\underline{Y}}_2 = \frac{1}{m}\sum_{j=1}^{m}\underline{Y}_{2j}$$

$$(6)\qquad \underline{S}_1 = \sqrt{\frac{\sum_{i=1}^{n}(\underline{X}_{1i} - \bar{\underline{X}}_1)^2 + \sum_{j=1}^{m}(\underline{Y}_{1j} - \bar{\underline{Y}}_1)^2}{m+n-2}}$$

$$\underline{S}_2 = \sqrt{\frac{\sum_{i=1}^{n}(\underline{X}_{2i} - \bar{\underline{X}}_2)^2 + \sum_{j=1}^{m}(\underline{Y}_{2j} - \bar{\underline{Y}}_2)^2}{m+n-2}}$$

Als de hypothese H_o juist is bezit:

$$(7)\quad \underline{F} = \frac{1}{2}\cdot\frac{m+n-3}{m+n-2}\cdot\underline{T}^2$$

de verdeling van de $\underline{F}$ van FISHER-SNEDECOR met $\nu_1 = 2$ en $\nu_2 = m+n-3$ vrijheidsgraden.

De kritieke zone is van het type $T^2 \geqq T_o^2$ en komt dan overeen met een tweezijdige toets van STUDENT.

Litteratuur: A.HALD, Statistical Theory with engineering applications, New York and London, 1952, p. 616.

4 Toets voor de correlatiecoëfficiënt van een tweedimensionale normale verdeling.

De te toetsen hypothese H_0 is, dat de (theoretische) correlatiecoëfficiënt van een tweedimensionale normale verdeling gelijk is aan een gegeven getal ρ.

Gegeven zij weer een steekproef $(X_{1i}, X_{2i})(i = 1, \ldots, n)$ van onafhankelijke waarnemingen uit de verdeling.

De toetsingsgrootheid is de correlatiecoëfficiënt $\underline{r}$ van deze waarnemingen, gedefinieerd in (3).

De verdeling van $\underline{r}$ is voor de waarden van ρ (0,0(0,1)0,9) en =3(1)25,50,100,200,400 getabelleerd door F.N.DAVID (zie litteratuur).

De kritieke zone (tweezijdig) bij een gegeven onbetrouwbaarheid α, bestaat uit de intervallen $\underline{r} \leqq r_1$ respectievelijk $\underline{r} \geqq r_2$, bepaald door:

$$P[r \leq r_1 \mid \rho] = \tfrac{1}{2}\alpha \qquad \text{resp.} \quad P[r \geq r_2 \mid \rho] = \tfrac{1}{2}\alpha$$

Bij eenzijdige toetsing is de kritieke zone van het type $r \geq r_0$ (rechtszijdig), r_0 bepaald door:

$$P[r \geq r_0 \mid \rho] = \alpha$$

Voor grote waarden van n maakt men gebruik van het feit, dat

$$(8) \qquad \underline{z} = \tfrac{1}{2} \ln \frac{1+\underline{r}}{1-\underline{r}} = 1{,}1513 \log \frac{1+\underline{r}}{1-\underline{r}}$$

bij benadering normaal verdeeld is met gemiddelde:

$$(9) \qquad \zeta = \tfrac{1}{2} \ln \frac{1+\rho}{1-\rho}$$

en spreiding

$$(10) \qquad \sigma_{\underline{z}} = \frac{1}{\sqrt{n-3}}$$

Litteratuur:

A.HALD, Statistical Theory with engineering applications, New York and London, 1952, p. 608.

F.N.DAVID, Tables of the Ordinates and Probability integral of the distribution of the correlation coefficient in small samples, London, 1938.

6 Toets om na te gaan, of een tweedimensionale verdeling normaal is.

Gegeven is een steekproef $(x_{1i}, x_{2i})(i = 1, \ldots, n)$. Wij wensen de hypothese H_0 te toetsen, dat dit een steekproef is uit een tweedimensionale normale verdeling. Hiertoe berekenen we de grootheden $\bar{x}_1$, $\bar{x}_2$, s_1, s_2 en r volgens (3).

Beschouw nu bij een λ_ε gegeven door

$$(13) \qquad 1 - e^{-\frac{1}{2} \cdot \frac{\lambda_\varepsilon^2}{1-r^2}} = \varepsilon$$

de ellips:

$$(14) \left(\frac{x_1 - \bar{x}_1}{s_1}\right)^2 - 2r\frac{(x_1 - \bar{x}_1)(x_2 - \bar{x}_2)}{s_1 s_2} + \left(\frac{x_2 - \bar{x}_2}{s_2}\right)^2 = \lambda_\varepsilon^2$$

De lange as gaat door het punt $(\bar{x}_1, \bar{x}_2)$ en heeft een helling gegeven door:

$$(15) \qquad h = \frac{2r}{\frac{s_1}{s_2} - \frac{s_2}{s_1} + \sqrt{\left(\frac{s_1}{s_2} - \frac{s_2}{s_1}\right)^2 + 4r^2}}$$

De normale verdeling met $\mu_1 = m_1$, $\mu_2 = m_2$, $\sigma_1 = s_1$, $\sigma_2 = s_2$ en $\rho = r$ noemt men de aangepaste normale verdeling. Voor deze verdeling geldt, dat de kans op een waarnemingen binnen de door (13) en (14) gegeven ellips gelijk aan ε is. Verder is deze verdeling symmetrisch t.o.v. de assen van die ellips.

Wij construeren nu de ellips gegeven door (13) en (14) voor een aantal waarden van ε b.v.: $\varepsilon = \frac{1}{6}, \frac{2}{6}, \cdots, \frac{5}{6}$. De ellipsen met hun gemeenschappelijke assen verdelen dan het platte vlak in 24 vakken. De verwachtingen volgens de aangepaste normale verdeling van de aantallen punten binnen deze vakken zijn dan $\frac{1}{24}n$. Wij tellen nu de aantallen waargenomen punten die in ieder der vakken liggen, aan te duiden met $n_1, n_2, \ldots, n_{24}$ en berekenen:

$$\chi^2 = \sum_{j=1}^{24} \frac{\left(n_j - \frac{1}{24}n\right)^2}{\frac{1}{24}n}$$

Als H_0 juist is, zal deze grootheid voor grote n bij benadering een χ^2-verdeling hebben met 24-6=18 vrijheidsgraden (er zijn 5 parameters aangepast ; men vergelijke hiervoor het memorandum S 47 (M 17) over de χ^2-toets voor aanpassing). Wij zullen de hypothese H_0 verwerpen als χ^2 groter is dan χ_0^2 gedefinieerd door

$$P[\chi^2 \geqq \chi_0^2] = \alpha$$

als α de gewenste onbetrouwbaarheidsdrempel is.

Litteratuur: A.HALD, Statistical Theory with engineering applications, New York and London, 1952, p. 602.