Mitschrift zu
KV Statistik 2
bei DI Petz

F l o r i a n K ö n i g
0255220/521
florian.koenig@oeh.jku.at

29. März 2004

1 Klassenbildung

$\sqrt{N}$ ist eine gute Zahl von Klassen bei kleinen N, ansonsten besser $^2\log N$.

2 Lagemaßzahlen

Ganz allgemein kann man zur Verbesserung der Ausreißerempfindlichkeit das Datenmaterial *trimmen*. 5 % oben und unten weg sind normal, seltener schon 10 %, maximal jedoch 50 %.

2.1 Minimum, Maximum

Sind am ausreißerempfindlichsten.

$M_i = \min\limits_{j=1}^{n} x_j \qquad M_a = \max\limits_{j=1}^{n} x_j$

2.2 Arithmetisches Mittel

Ist sehr ausreißerempfindlich.

$\bar{x} = \frac{1}{N} \sum\limits_{i=1}^{N} x_i \qquad \bar{x} = \frac{1}{N} \sum\limits_{i=1}^{N} f_i \cdot x_i$

N Klassenzahl oder Anzahl der Urlistenelemente

$N = \sum\limits_{i=1}^{k} f_i$ mit k als Klassenzahl bzw. Anzahl der verschiedenen Elemente.

2.3 Dezile

Dezile sind mitunter sehr ausreißerempfindlich.

2.4 Median

Ist wenig ausreißerempfindlich, so wie Quartile ganz allgemein.

$\tilde{x} = \tilde{x}_{0,5} = Q_2 = D_5$ wo $F(\tilde{x}_{0,5}) = \frac{1}{2}$

$\tilde{x}_{0,5} = x_{(\frac{n+1}{2})}$ für n gerade

$\tilde{x}_{0,5} = \frac{1}{2}\left[x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right]$ für n ungerade

$\tilde{x}_{0,5} = e_{m-1} + \frac{d_m}{f_m}\left[\frac{n}{2} - F_{m-1}\right]$ bei gruppierten Daten

e_{m-1} untere Klassengrenze der Medianklasse m

d_m Klassenbreite

F_{m-1} absolute kumulierten Häufigkeiten mit $F_{m-1} \leq \frac{n}{2} \leq F_m$

2.5 Modus

Ausreißerunempfindlich.

$\mathring{x} = \max_{i=1}^{k} f_i$ mit k Anzahl der Klassen

2.6 Midrange

Ist am ausreißerunempfindlichsten.

$MR = \frac{M_i + M_a}{2}$

3 Streuungsmaßzahlen

3.1 Range

Ist am ausreißerempfindlichsten.

$Range = M_a - M_i$

3.2 Inter-quartil range IQR

Ist robuster als der Range.

$IQR = Q_3 - Q_1$

3.3 Varianz

Ist ausreißerempfindlich, da auf ein ausreißerempfindliches Lagemaß ($\bar{x}$) Bezug genommen wird.

$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$

3.4 Mean absolute deviation MAD

Weniger ausreißerempfindlich.

$MAD = \frac{1}{N}\sum|x_i - \bar{x}|$

4 Symmetriemaßzahlen

4.1 Momentenkoeffizient der Schiefe

Liefert eine quantitative Aussage.

$\gamma_1 = \frac{\mu_3}{s^3}$ ist *dimensionslos* und ausreißerempfindlich

Werte von γ_1:

- $\gamma_1 < 0$ linksschief
- $\gamma_1 = 0$ symmetrisch
- $\gamma_1 > 0$ rechtsschief

4.2 Wölbung (Kurtosis)

$\gamma_2 = \frac{\mu_4}{s^4} - 3$ mit $\mu_4 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^4$

$\gamma_2 \in [-3; \infty[$

Je positiver der Wert, desto gewölbter, je negativer, desto flacher.

4.3 Quantilskoeffizient der Schiefe

Ist wenig ausreißerempfindlich.

$QS_{0,25} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1}$

4.4 Fechnersche Lageregel

Liefert eine *qualitative Aussage*.

$\bar{x} \approx \tilde{x} \approx \mathring{x}$ symmetrisch

$\bar{x} < \tilde{x} < \mathring{x}$ linksschief

$\bar{x} > \tilde{x} > \mathring{x}$ rechtsschief

5 Kontingenzmaße

Mithilfe von Kontingenzmaßen kann man den *Grad des Zusammenhangs* zweier Merkmale überprüfen und, je nach Art des Merkmals, auch eine Aussage über die *Art des Zusammenhangs* („seine Richtung“) oder die *Dynamik der Änderung* (linear, quadratisch, kubisch, ...) der abhängigen Variable bei Änderung der bestimmenden Variable treffen.

Bei *qualitativen* Daten kann man den χ^2 Wert berechnen, welcher nur eine Aussage über die Stärke des Zusammenhangs liefert, ohne Angabe über eine Art oder Dynamik der Änderung.

Bei *ordinalen* Daten bieten sich die Rang-Korrelationskoeffizienten von Kendall oder Spearman zur Berechnung von Grad und Art des Zusammenhangs an.

Für *quantitative* Daten kann man mit dem Produkt-Moment-Korrelationskoeffizient nach Bravais-Pearson zusätzlich auch eine lineare Änderungsdynamik erkennen.

5.1 χ^2 Wert für Unabhängigkeit

χ^2 ist ein primitives Maß der Kontingenz für *qualitative Merkmale*.

Bei quantitativen Merkmalen (also nicht dichotome) muss man den Teilungspunkt für die Felder selbst wählen. Hierbei eignet sich der Median besser als das arithmetische Mittel.

$\chi^2_{unabh} = \sum \frac{\left(f^o_{ij} - f^e_{ij}\right)^2}{f^e_{ij}}$ mit $f^e_{ij} = \frac{f_{i.} \cdot f_{.j}}{f_{..}}$

$\chi^2_{unabh} \in [0; \infty[$, da es keine „Richtung“ gibt.

Zwischen 0 und 3 geht man von statistischem Rauschen aus. Werte von 3 – 12 sind erklärbar, ab 12 wird die Abweichung erklärungsbedürftig.

Normierung Kontingenzkoeffizient $C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$

Cramer's $V = \sqrt{\frac{\chi^2}{N \cdot \min(r-1)(c-1)}}$ bei $df \neq 1$

Phi-Koeffizient $\Phi = \sqrt{\frac{\chi^2}{N}}$ am Besten nur bei 4-Feldertafeln, also $df = 1$

Alle Werte $\in [0; 1[$, wobei 0 *perfekt unabhängig* und 1 *perfekt abhängig* meint.

5.2 Rang-Korrelationskoeffizient

Der Rang-Korrelationskoeffizienten ermöglicht das Erkennen von gleich- oder gegensinnigen *monotonen Zusammenhängen*.

Die Koeffizienten sind $\in [-1; 1]$ wobei -1 *pefekt gegensinniger* und 1 *perfekt gleichsinniger monotoner Zusammenhang* meint. 0 suggeriert *verrauscht/unabhängig*. 0,8 – 1 stark, 0,5 – 0,8 mittlerer, 0,2 – 0,5 schwacher Zusammenhang.

5.2.1 Kendall

Die Daten müssen nach einem Merkmal aufsteigend sortiert sein. Das zweite Merkmal wird in Abhängigkeit des ersten automatisch mitsortiert.

$r_k = \frac{C - D}{C + D}$

5.2.2 Spearman

Funktioniert nur paarweise, wobei Δ_i die Differenz der Ausprägungen ist.

$r_s = 1 - \frac{6 \sum \Delta_i^2}{n \cdot (n^2 - 1)}$

5.3 Proportional Reduction of Error (PRE) Maße

PRE Maße geben an, in welchem Maß sich der Vorhersagefehler einer (abhängigen) Variable ändert, wenn zur Vorhersage eine zweite Variable hinzugezogen wird. Je größer die Reduzierung des Vorhersagefehlers, umso größer ist der Zusammenhang zwischen den beiden Variablen.

5.3.1 λ

Gibt mir die Verminderung des Fehlers bei der Vorraussage eines abhängigen qualitativen Merkmals unter Wissen über die Vorbedingung im Vergleich zum Fehler ohne Wissen.

E_1 ist der Fehler ohne Wissen, das kleinere der Zeilen-/Spaltensumme.

E_2 ist der Fehler mit Wissen, die Summe der kleinsten Ausprägungen jeder Zeile/Spalte.

$\lambda_{x\to y} = \frac{E_1 - E_2}{E_1}$

$\lambda \in [0;1]$, wobei der Wert die Prozent der Fehlerverminderung angibt.

5.3.2 Produkt-Moment-Korrelationskoeffizient r_{xy} nach Bravais-Pearson

Dieser misst lineare Zusammenhänge, kann also quadratische beispielsweise nicht erkennen.

$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}$

$s_{xy} = \frac{1}{N} \sum [(x_i - \bar{x})(y_i - \bar{y})]$ bzw. $s_{xy} = \frac{1}{N} \sum x_i y_i - (\bar{x} \cdot \bar{y})$

$r_{xy} \in [-1;1]$

1 meint einen *perfekt gleichsinnigen* und -1 einen *perfekt gegensinnigen lineare Zusammenhang.*

Bestimmtheitsmaß $B = r_{xy}^2$

B gibt die Bestimmtheit von r_{xy} an und die Verminderung des Fehlers bei Erfahren eines konkreten Wertes im Vergleich zur Schätzung über das arithmetische Mittel.

Regressionsgerade $\hat{y} = kx + d$ mit $k = \frac{s_{xy}}{s_x^2}$ und $d = \bar{y} - k \cdot \bar{x}$

Durch den Schwerpunkt der Punktewolke S($\bar{x} - \bar{y}$) geht die Gerade immer.

6 Testtheorie

Vorgehensweise:

1. *Nullhypothese H_0 aufstellen.* Einige Richtlinien sind hier:
   - Als Nullhypothese nimmt man jene, deren irrtümliche Verwerfung die *gravierenderen Konsequenzen* besitzt, da der Fehler 1. Art durch das Signifikanzniveau des Testes kontrolliert werden kann.
   - Als Nullhypothese nimmt man oft jene Hypothese, deren *Glaubwürdigkeit* durch die Stichprobe *erschüttert* werden soll.
   - In manchen Fällen ist es zweckmäßig, die *mathematisch einfachere* Hypothese als Nullhypothese anzusetzen.
2. *Fehlerniveau festlegen* (z. B. $\alpha = 0,05$). Dies legt gleichzeitig die Größe des Fehlers 1. Art α (fälschliche Ablehnung einer richtigen H_0) fest. Je kleiner α, desto größer wird der Fehler 2. Art β, nämlich das irrtümliche Annehmen einer falschen H_0.
3. *Testgröße berechnen*: z. B. den χ^2-Wert
4. *Vergleichsgröße* bei Fehlerniveau (z. B. in einer Tabelle) mit der Testgröße *vergleichen.* Ist die Testgröße größer als die Vergleichsgröße, so wird die H_0 abgelehnt, ist sie kleiner, kann man H_0 nicht nicht ablehnen (aber auch nicht annehmen). Bei gleichen Werten, ist es am Besten man vergrößert den Stichprobenumfang.

Hinsichtlich dem in der Tabelle mit den Vergleichsgrößen zu wählenden Wert in Abhängigkeit vom Fehlerniveau muss man zwischen folgenden Arten von Tests unterscheiden:

- *einseitig*: beim Testen auf Gleichheit bzw. Ungleichheit zweier Kenngrößen
- *zweiseitig*: beim Testen auf arithmetische Ordnung ($<, \leq, >, \geq$) zweier Kenngrößen

6.1 Tests mit χ^2

6.1.1 χ^2-Unabhängigkeitstest

$\chi^2_{unabh} = \sum \frac{\left(f^o_{ij} - f^e_{ij}\right)^2}{f^e_{ij}}$ mit $f^e_{ij} = \frac{f_{i.} \cdot f_{.j}}{f_{..}}$

$df = (r-1)(c-1)$

6.1.2 χ^2-Symmetrietest

$\chi^2_{symm} = \frac{\left(f^o_{ij} - f^e_{ji}\right)^2}{f^o_{ij} + f^e_{ji}}$

$df = \frac{n(n-1)}{2}$

6.1.3 χ^2-Anpassungstest

$\chi^2_{anpass} = \sum \frac{\left(f^o_i - f^e_i\right)^2}{f^e_i}$

Ein Wert von 0 zeigt, dass die Verteilung perfekt erfüllt wird.

$df = k - 1$ mit k als Klassenzahl

Es sollen mindestens 5 Klassen vorhanden sein. Die Zahl der Freiheitsgrade verringert sich noch um die Zahl der aus der Stichprobe zu schätzenden Parameter der Verteilung.

6.2 Kolmogorov-Smirnov-Test für Anpassung

Der Колмогоров-Смирнов-Test testet auf eine Anpassung an eine *beliebige Verteilung*. Er wird verwendet, wenn man zu wenig Zellen für einen χ^2-Anpassungstest erhalten würde, oder die Besetzungen der Zellen teilweise < 5 sind. Außerdem ister *trennschärfer* als der χ^2-Anpassungstest. Da man absolute Werte vergleicht, ist er *sehr ausreißerempfindlich*.

$KS = \max_{i=1}^{k} \left|\hat{F}(x_i) - F(x_i)\right|$ bei einer diskreten Verteilung

$KS = \max_{i=1}^{k} \left\{\left|\hat{F}(x_i) - F(x_i)\right| ; \left|\hat{F}(x_i) - F(x_{i-1})\right|\right\}$ bei einer stetigen Verteilung

k ist die Anzahl der Klassen

Es gibt zwei verschiedene Tabellen für die Vergleichswerte. Mit den hier angegebenen Formeln verwendet man jene mit den absteigenden Werten und den Stichprobenumfang n als Freiheitsgrade.

6.3 Varianz-Test (F-Test)

Beim Varianz-Test wird auf die *Gleichheit der Varianzen* zweier durch zwei Stichproben gegebenen Grundgesamtheiten getestet. Man benötigt dies unter anderem zur Entscheidung für eine Formel bei den t-Tests. Als Testverteilung wird die F-Verteilung (3 Parameter: 2 mal Freiheitsgrad und 1 mal Fehlerniveau) verwendet.

$F_{df_1,df_2} = \frac{s_x^2}{s_y^2}$ mit $s_x^2 \geq s_y^2$

$df_1 = n_x - 1$

$df_2 = n_y - 1$

6.4 Lagevergleich (t-Test)

t-Tests testen auf die Lage der Verteilung(en) unter Verwendung des *Mittelwerts*. Man unterscheidet:

- *Einstichproben-Test*: mit der Fragestellung $\mu \stackrel{?}{=} A$
- *Zweistichproben-Tests*: mit der Fragestellung $\mu_x \stackrel{?}{=} \mu_y$
  - *verbunden*: vereinfacht wenn z. B. gleiche Maschinen, Personen etc. die Stichprobe verursachen
  - *unverbunden*: automatisch bei ungleichem Stichprobenumfang
    - *varianz-homogen* (beides mit dem Varianz-Test überprüfbar)
    - *varianz-heterogen*

Allgemein kann man sagen, dass verbundene Stichproben aussagekräftigere Ergebnisse liefern.

Die genauen Formeln finden sich auf dem Beiblatt „Lokationsvergleich zweier Messreihen (Stichproben)“

6.5 Testen auf Korrelation

6.5.1 Produkt-Moment-Korrelationskoeffizient r_{xy} (t-verteilt)

H_0 = zwischen Stichprobe x und Stichprobe y besteht kein signifikanter linearer Zusammenhang

$t = r_{xy} \cdot \sqrt{\frac{n-2}{1-r_{xy}^2}}$ $df = n - 2$

6.5.2 Spearman (näherungsweise t-verteilt)

$t \approx r_s \cdot \sqrt{\frac{n-2}{r-r_s^2}}$ $df = n - 2$

Anmerkung: muss deutlich *nicht* Ablehnen, da dies nur eine Näherung ist

6.5.3 Kendall (näherungsweise standard-normalerteilt d. h. $\mu = 0$ und $\sigma = 1$)

$z \approx r_k \cdot \sqrt{\frac{9n \cdot (n-1)}{2 \cdot (2n+5)}}$

6.6 Streuungszerlegung (Varianzanalyse)

Bei der Streuungszerlegung wird auf die Nullhypothese, dass Stichproben keine Aussage über die Unterschiedlichkeit der Grundgesamtheit zulassen, getestet. Man vergleicht zwei oder mehr Stichproben, die auch unterschiedlich groß sein können.

k Anzahl der Stichproben

n_i Umfang der Stichprobe i

$\bar{x}_i$ Mittelwert der Stichprobe i

$\bar{\bar{x}} = \frac{1}{N}\left(n_1 \cdot \bar{x}_1 + \ldots + n_i \cdot \bar{x}_i\right)$ gewichtetes Mittel der Mittelwerte der einzelnen Stichproben

$SST = \sum_{i=1}^{k} \left(\bar{x}_i - \bar{\bar{x}}\right)^2$ Sum of square treatments

$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_i\right)^2$ Sum of square errors

$F = \frac{\frac{SST}{k-1}}{\frac{SSE}{N-k}}$ $df_1 = k - 1$ $df_2 = N - k$

Bei Ablehnung der Nullhypothese gibt es scheinbar Unterschiede. Bei einer starken Streuung der Stichproben in sich, kann man keine genaue Aussage treffen. Kleines SSE und großes SST suggeriert Unterschiede. Man kann auch erkennen, welche Stichprobe zu SST am meisten beiträgt und somit ein „Ausreißer“ ist.