Dr. Katharina Best
Sommersemester 2011
21. April 2011

ulm university universität **uulm**

Reader – Teil 2: Statistik normalverteilter Daten

WiMa-Praktikum

Die Normalverteilung spielt auf Grund des zentralen Grenzwertsatzes eine zentrale Rolle in der Statistik. Szenarien, die die Normalverteilung voraussetzen sind weitreichend untersucht.

Bevor wir beginnen, sollten wir uns die Symmetrie der Normalverteilung ins Gedächtnis zurückrufen; die Dichte φ ist achsensymmetrisch zur y-Achse, die Verteilungsfunktion Φ punktsymmetrisch zum Punkt $(0, 1/2)$. Bezeichnen wir nun mit z_α das α-Quantil der Normalverteilung. Dann gilt

$$z_\alpha = -z_{1-\alpha} \tag{1}$$

für alle $\alpha \in (0,1)$ als Folge erwähnter Symmetrien.

1 Bezug verschiedener Verteilungen auf die Normalverteilung

Bei der Untersuchung normalverteilter Daten spielen aus der Normalverteilung abgeleitete Verteilungen eine große Rolle.

1.1 χ^2-Verteilung

Wir gehen von i. i. d. $\mathcal{N}(0,1)$-verteilten Zufallsvariablen $X_1, \ldots, X_n$ aus. Die Verteilung von $Z_n = \sum_{i=1}^n X_i^2$ wird als *Chi-Quadrat-Verteilung mit n Freiheitsgraden* bezeichnet. Die Dichte ist

$$h_n(x) = \frac{1}{2^{n/2}\Gamma(n/2)}\, x^{n/2-1} e^{-x/2} \tag{2}$$

für $x > 0$ und sonst $h_n(x) = 0$. Werden zwei unabhängige χ^2-verteilte Zufallsvariablen Z_n und Z_m addiert, so ist die Summe χ^2_{n+m}-verteilt.

1.2 *F*-Verteilung

Seien X und Y zwei unabhängige Zufallsvariablen, die χ^2_m-verteilt resp. χ^2_n-verteilt sind. Dann wird die Verteilung der Zufallsvariablen

$$Z := \frac{\frac{1}{m}X}{\frac{1}{n}Y} \tag{3}$$

als *F-Verteilung mit den Freiheitsgraden m und n* bezeichnet, ihre Dichte ist

$$f_{m,n}(x) = \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}\, m^{m/2} n^{n/2} \frac{x^{\frac{m}{2}-1}}{(n+mx)^{\frac{m+n}{2}}} \tag{4}$$

für $x > 0$ und $f_{m,n} = 0$ sonst.

1.3 *t*-Verteilung

Seien X eine $\mathcal{N}(0,1)$-verteilte und Z_n eine von X unabhängige χ^2_n-verteilte Zufallsvariable. Wir betrachten die Zufallsvariable

$$T := \frac{X}{\sqrt{\frac{1}{n}Z_n}} \tag{5}$$

und sagen, dass sie *t-verteilt mit n Freiheitsgraden* ist. Die Dichte ist gegeben durch

$$g_n(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n}\,\Gamma(\frac{n}{2})} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} \tag{6}$$

für $x \in \mathbb{R}$.

1.4 Cauchy-Verteilung

Eine t-Verteilung mit nur einem Freiheitsgrad, d. h. die Verteilung der Zufallsvariable

$$T = \frac{X}{|Y|}, \tag{7}$$

wobei X, Y unabhängig und $\mathcal{N}(0,1)$ sind, wird als *Cauchy-Verteilung* bezeichnet. Ihre Dichte folgt direkt aus (6) und ist

$$g_1(x) = \frac{1}{\pi(1+x^2)} \tag{8}$$

für $x \in \mathbb{R}$.

Machen Sie sich zu allen Verteilungen Gedanken zu:
(a) Form der Dichte (für kleine/große n)
(b) Symmetrie
(c) Momente
(d) Verhalten der Quantilfunktion
(e) ...

1.5 Stichprobenaussagen

Gehen wir nun von einer i. i. d. Zufallsstichprobe $X_1, \ldots, X_n$ aus, wobei $X_1 \sim \mathcal{N}(\mu, \sigma^2)$ ist, so gilt

- $\overline{X}_n$ und S_n^2 sind unabhängig,
- $(n-1)S_n^2/\sigma^2$ ist χ^2_{n-1}-verteilt,
- $\overline{X}_n$ ist $\mathcal{N}(\mu, \sigma^2/n)$-verteilt,
- $\sqrt{n}\frac{\overline{X}_n - \mu}{S_n}$ ist t_{n-1}-verteilt.

Diese Eigenschaften kommen im folgenden Abschnitt immer wieder zu Tragen.

2 Statistische Tests und Konfidenzintervalle unter Normalverteilungsannahme

Grundsätze zum Aufbau von Konfidenzintervallen und Tests wird im folgenden fortgesetzt.

2.1 Ein-Stichproben-Szenarien

Wir gehen in diesem Abschnitt generell von einer i. i. d. Stichprobe $X_1, \ldots, X_n$ aus, wobei $X_1 \sim \mathcal{N}(\mu, \sigma^2)$ gilt. Weitere Bedingungen an μ und σ^2 finden sich in den einzelnen Situationsbeschreibungen.

2.1.1 Untersuchung des Erwartungswertes bei bekannter Varianz

Wir gehen von $\mu \in \mathbb{R}$ als unbekannt aus, weiterhin ist $\sigma^2 > 0$ bekannt. Mit diesen vorgaben können wir de Verteilungsaussage

$$T(X_1, \ldots, X_n) = \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \sim \mathcal{N}(0,1) \tag{9}$$

treffen, die sowohl zum Aufbau eines Konfidenintervalls als auch eines kritischen Bereichs des entsprechenden stochastischen Tests verwendet werden kann.

2.1.2 Untersuchung des Erwartungswertes bei unbekannter Varianz

In diesem Fall nehmen wir sowohl $\mu \in \mathbb{R}$ als auch $\sigma^2 > 0$ als unbekannt an, interessieren uns für μ. Damit können wir

$$T(X_1, \ldots, X_n) = \sqrt{n}\frac{\overline{X}_n - \mu}{S_n} \sim t_{n-1} \tag{10}$$

formulieren.

2.1.3 Untersuchung der Varianz bei bekanntem Erwartungswert

Diese Situation setzt voraus, dass $\mu \in \mathbb{R}$ bekannt ist und wir uns für das unbekannte $\sigma^2 > 0$ interessieren. Damit bildet

$$T(X_1, \ldots, X_n) = \frac{n}{\sigma_0^2} \tilde{S}_n^2 \sim \chi_n^2 \tag{11}$$

die Grundlage unserer Betrachtung.

2.1.4 Untersuchung der Varianz bei unbekanntem Erwartungswert

Änhlich des vorletzten Szenarios sind $\mu \in \mathbb{R}$ und $\sigma^2 > 0$ unbekannt. Unser Interesse gilt diesmal μ und es gilt

$$T(X_1, \ldots, X_n) = \frac{n-1}{\sigma_0^2} S_n^2 \sim \chi_{n-1}^2. \tag{12}$$

2.2 Zwei-Stichproben-Szenarien

Sind zwei Stichproben vorhanden, so ist zu unterscheiden, ob diese voneinander unabhängig sind oder nicht. Im ersten Fall sprechen wir von unverbundenen, im letzteren von verbundenen Stichproben.

Das Grundprozedere bei unverbundenen Stichproben ist, dass wir zwei i. i. d. Stichproben $X_1, \ldots, X_n$ und $\widetilde{X}_1, \ldots, \widetilde{X}_m$ haben, wobei $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ und $\widetilde{X}_1 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ gilt. Desweiteren sind X_i und X_j für $1 \leq i \leq n$ und $1 \leq j \leq n$ unabhängig.

Bei verbundenen Stichproben haben wir im Gegensatz dazu $X_1, \ldots, X_n$ und $\widetilde{X}_1, \ldots, \widetilde{X}_n$ Stichproben gleicher Länge, sodass $\begin{pmatrix} X_1 \\ \widetilde{X}_1 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ \widetilde{X}_n \end{pmatrix}$ als i. i. d. Stichprobe angesehen wird. Über die Verteilung der eindimensionalen wird keine nähere Aussage getroffen.

2.2.1 Untersuchung der Differenz zweier Erwartungswerte bei bekannten Varianzen und unverbundenen Stichproben

Wir gehen davon aus, dass $\sigma_1^2, \sigma_2^2 > 0$ bekannt und $\mu_1, \mu_2 \in \mathbb{R}$ unbekannt sind. Damit können wir

$$T(X_1, \ldots, X_n, \widetilde{X}_1, \ldots, \widetilde{X}_m) = \overline{X}_n - \overline{\widetilde{X}}_m \sim \mathcal{N}\left(\mu_1 - \mu_2, \sigma_1^2/n + \sigma_2^2/m\right) \tag{13}$$

angeben.

2.2.2 Untersuchung der Differenz zweier Erwartungswerte bei unbekannten aber identischen Varianzen und unverbundenen Stichproben

In dieser Situation setzen wir voraus, dass $\mu_1, \mu_2 \in \mathbb{R}$ unbekannt sind. Weiterhin wissen wir, dass $\sigma^2 := \sigma_1^2 = \sigma_2^2 > 0$ gilt, jedoch nicht näher bekannt ist. Mit Hilfe der *gepoolten Stichprobenvarianz*

$$S_p^2 := \frac{(n-1)S_n^2 + (m-1)S_m^2}{n+m-1}, \tag{14}$$

die die Stichprobenvarianzen S_n^2 von $X_1, \ldots, X_n$ und S_m^2 von $\widetilde{X}_1, \ldots, \widetilde{X}_n$ verwendet, ist

$$T(X_1, \ldots, X_n, \widetilde{X}_1, \ldots, \widetilde{X}_m) = \frac{\overline{X}_n - \overline{\widetilde{X}}_m - (\mu_1 - \mu_2)}{S_p\sqrt{1/n + 1/m}} \sim t_{n+m-2}. \tag{15}$$

2.2.3 Behrens-Fisher-Problem *oder* Untersuchung der Differenz zweier Erwartungswerte bei unbekannten, nicht identischen Varianzen und unverbundenen Stichproben

Haben wir es mit normalverteilten Stichproben zu tun, deren Varianzen nicht identisch sind, und mindenstens eine unbekannt ist, so ist die Formulierung einer Teststatistik nur mir approximativer Verteilungsaussage möglich. Betrachten wir die Teststatistik

$$T(X_1, \ldots, X_n, \widetilde{X}_1, \ldots, \widetilde{X}_m) = \frac{\overline{X}_n - \overline{\widetilde{X}}_m - (\mu_1 - \mu_2)}{\sqrt{S_n^2/n + S_m^2/m}} \tag{16}$$

mit Stichprobenvarianzen S_n^2 und S_m^2 wie im vorhergehenden Fall, so ist T approximativ t-verteilt. Die Anzahl der Freiheitsgrade ist

$$\left\lceil \frac{\left(\frac{S_n^2}{n} + \frac{S_m^2}{m}\right)^2}{\frac{\left(\frac{S_n^2}{n}\right)^2}{n-1} + \frac{\left(\frac{S_m^2}{m}\right)^2}{m-1}} \right\rceil, \tag{17}$$

also zufällig. Der resultierende Test wird *Welch-Test* genannt.

2.2.4 Untersuchung des Quotienten zweier Varianzen bei unbekannten Erwartungswerten und unverbundenen Stichproben

Hier des der Bekanntheitsgrad der Parameter komplementär zum vorhergehenden Problem. Wir kennen $\mu_1, \mu_2 \in \mathbb{R}$, während $\sigma_1^2, \sigma_2^2 > 0$ unbekannt sind. Das führt dazu, dass

$$T(X_1, \ldots, X_n, \widetilde{X}_1, \ldots, \widetilde{X}_m) = \frac{S_m^2/\sigma_2^2}{S_n^2/\sigma_1^2} \sim F_{m-1,n-1} \tag{18}$$

ist, womit wir das erste mal die F-Verteilung ins Spiel gebracht haben.

2.2.5 Untersuchung der Differenz zweier Erwartungswerte bei unbekannten Varianzen und verbundenen Stichproben

Zur Verteilung der Stichprobe wissen wir, dass die Differenz der Zufallsvektorkoordinaten normalverteilt ist, d. h. $Y_i := X_i - \widetilde{X}_i \sim \mathcal{N}(\mu, \sigma^2)$ für alle $1 \leq i \leq n$ ist, wobei $\mu \in \mathbb{R}$ unbekannt, $\sigma^2 > 0$ jedoch

bekannt ist. Mit dem Stichprobenmittel $Y_1, \ldots, Y_n$ haben wir das entsprechende Ein-Stichproben-Szenario und

$$T(X_1, \ldots X_n, \widetilde{X}_1, \ldots, \widetilde{X}_n) = \sqrt{n}\frac{\overline{Y}_n - \mu}{\sigma} \sim \mathcal{N}(0,1) \tag{19}$$

als Verteilungsaussage.

2.2.6 Untersuchung der Differenz zweier Erwartungswerte bei unbekannten Varianzen und verbundenen Stichproben

Ganz ähnlich dem vorhergehenden Fall gilt $Y_i := X_i - \widetilde{X}_i \sim \mathcal{N}(\mu, \sigma^2)$ für alle $1 \leq i \leq n$. Diesmal sind $\mu \in \mathbb{R}$ und $\sigma^2 > 0$ beide unbekannt. Wieder können wir auf das entsprechende Ein-Stichproben-Szenario zurückgreifen und mit dem Stichprobenmittel und der Stichprobenvarianz von $Y_1, \ldots, Y_n$ haben wir

$$T(X_1, \ldots X_n, \widetilde{X}_1, \ldots, \widetilde{X}_n) = \sqrt{n}\frac{\overline{Y}_n - \mu}{S_n} \sim t_{n-1}. \tag{20}$$

Machen Sie sich zu allen Szenarien Gedanken zu:
(a) Formulierung des Tests
(b) Formulierung des Konfidenzintervalls
(c) einseitige und zweiseitige Fragestellungen
(d) symmetrische Formulierung vs. asymmetrischer
(e) ...

3 p-Wert

Statistik-Software-Pakete wie SAS oder R geben bei Tests den p-Wert an. Er berechnet sich, ausgehend von der Nullhypothese als und den vorhandenen Daten $x_1, \ldots, x_n$ durch

$$\begin{aligned} p_r &= p_r(x_1, \ldots, x_n) = \mathbb{P}_{H_0}\left(T(X_1, \ldots, X_n) \geq z\right), \quad \text{(bei rechtsseitigem Test)} \\ p_l &= p_l(x_1, \ldots, x_n) = \mathbb{P}_{H_0}\left(T(X_1, \ldots, X_n) \leq z\right), \quad \text{(bei linksseitigem Test)} \end{aligned}$$

und bei beidseitigem Test

$$p = 2 \cdot \min\{p_l, p_r\},$$

wenn $z = T(x_1, \ldots, x_n)$ ist.