Tutorium

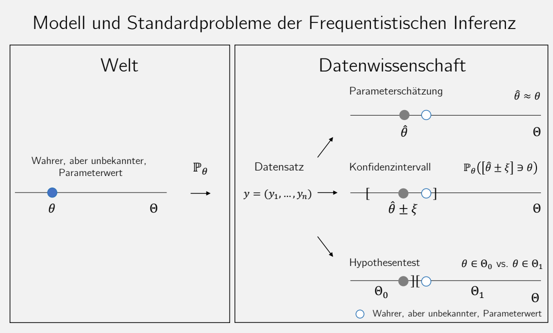
Allgemeines Lineares Modell

BSc Psychologie SoSe 2023

9. Termin: (7) T-Statistiken

Sean Mulready

Wiederholung - Frequentistisches Weltbild



- Wir nehmen an, dass wahre, aber unbekannte Parameter existieren.
- Wir nehmen weiterhin an, dass probabilistische Prozesse existieren, die, gegeben dieser wahren, aber unbekannten Parameter, Datensätze generieren können.
- Für diese probabilistischen Prozesse gehen wir davon aus, dass ihnen bestimmte Verteilungen bzw. Wahrscheinlichkeitsdichten zugrundeliegen (z.B. Normalverteilung der Zufallsfehler)
- Wir verwenden erhobene Daten dafür, Parameterwerte zu schätzen (Wir berechnen eine "Einschätzung", was der wahre Wert sein könnte, den wir nicht beobachten können).
- Dabei bilden die angenommenen Verteilungen bzw. Warhscheinlichkeitsdichten der probabilistischen Prozesse (die, wie wir annehmen, die Daten generiert haben), die Grundlage für Parameterschätzung und Modellevaluation.

Wiederholung - Standardannahmen frequentistischer Inferenz

- Gegeben sei ein statistisches Modell $\mathbb{P}_\theta$ (wir betrachten das ALM), in dem probabilistische Prozesse definiert sind, die Datensätze generieren können.
- Es wird angenommen, dass ein vorliegender Datensatz eine der möglichen Realisierungen der Daten des Modells ist. Eine mögliche Realisierung wäre $y^{(1)}$, $y^{(2)}$ wäre eine andere, $y^{(3)}$ wäre nochmals eine andere.
- Aus frequentistischer Sicht kann man unendlich oft Datensätze basierend auf einem Modell generieren und zu jedem Datensatz Schätzer oder Statistiken auswerten, z.B. den Betaparameterschätzer (z.B. $\hat{\beta}^{(1)}$ für Datensatz $y^{(1)}$)

Datensatz (1) : $y^{(1)} = \left(y_1^{(1)}, y_2^{(1)}, \dots, y_n^{(1)}\right)^T$ mit $\hat{\beta}^{(1)} = (X^TX)^{-1}X^Ty^{(1)}$

Datensatz (2) : $y^{(2)} = \left(y_1^{(2)}, y_2^{(2)}, \dots, y_n^{(2)}\right)^T$ mit $\hat{\beta}^{(2)} = (X^TX)^{-1}X^Ty^{(2)}$

Datensatz (3) : $y^{(3)} = \left(y_1^{(3)}, y_2^{(3)}, \dots, y_n^{(3)}\right)^T$ mit $\hat{\beta}^{(3)} = (X^TX)^{-1}X^Ty^{(3)}$

Datensatz (4) : $y^{(4)} = \left(y_1^{(4)}, y_2^{(4)}, \dots, y_n^{(4)}\right)^T$ mit $\hat{\beta}^{(4)} = (X^TX)^{-1}X^Ty^{(4)}$

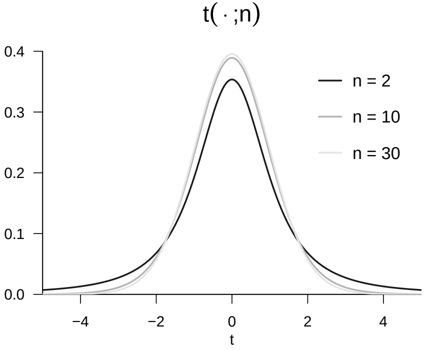
Datensatz (5) : $y^{(5)} = \dots$

- Um die Qualität statistischer Methoden zu beurteilen betrachtet die frequentistische Statistik die Wahrscheinlichkeitsverteilungen von Schätzern und Statistiken unter Annahme der Datenverteilung
- Was ist zum Beispiel die Verteilung (möglicher) Betaparameterschätzer ($\hat{\beta}^{(1)}$, $\hat{\beta}^{(2)}$, $\hat{\beta}^{(3)}$, $\hat{\beta}^{(4)}$), ... also die Verteilung der Zufallsvariable $\hat{\beta} := (X^TX)^{-1}X^T\upsilon$?
- eine statistische Methode ist im Sinne der frequentistischen Standardannahmen "gut" , wenn sie bei häufiger Anwendung "im Mittel gut" ist.
- Im Einzefall (nur ein Datensatz) kann sie auch "schlecht" sein

Selbstkontrollfragen

1. Skizzieren Sie die WDFen von T-Zufallsvariablen mit 2,10 und 30 Freiheitsgraden.
2. Skizzieren Sie die WDFen von nichtzentralen T-Zufallsvariablen mit Nichtzentralitätsparameter 0,5 und 15.
3. Geben Sie die Definition der T-Statistik wieder.
4. Erläutern Sie für die T-Statistik die Bedeutung der Wahl von $c \in \mathbb{R}^p$.
5. Erläutern Sie für die T-Statistik die Bedeutung der Wahl von $\beta_0 \in \mathbb{R}^p$.
6. Wann und warum kann die T-Statistik als Signal-zu-Rauschen-Verhältnis interpretiert werden?
7. Geben Sie das Theorem zur T-Statistik wieder.
8. Geben Sie die Form der T-Statistik im Szenario von n u.i.n.v. Zufallsvariablen wieder.
9. Erläutern Sie den Zusammenhang der T-Statistik und Cohen's d.
10. Geben Sie das Theorem zu Konfidenzintervallen für Betaparameterkomponenten wieder.

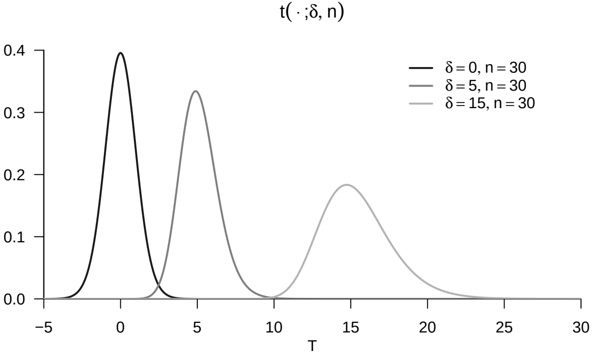
1.Skizzieren Sie die WDFen von T-Zufallsvariablen mit 2,10 und 30 Freiheitsgraden.



Anmerkungen:

- Die Verteilung ist um 0 symmetrisch
- Steigendes n verschiebt Wahrscheinlichkeitsmasse aus den Ausläufen zum Zentrum

2.Skizzieren Sie die WDFen von nichtzentralen T-Zufallsvariablen mit Nichtzentralitätsparameter 0,5 und 15.



3.Geben Sie die Definition der T-Statistik wieder.

Definition (T-Statistik)

Es sei

$$\upsilon = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0_n, \sigma^2 I_n) \tag{1}$$

das ALM. Weiterhin seien

$$\hat{\beta} := (X^T X)^{-1} X^T \upsilon \text{ und } \hat{\sigma}^2 := \frac{(\upsilon - X\hat{\beta})^T(\upsilon - X\hat{\beta})}{n-p} \tag{2}$$

die Betaparameter- und Varianzparameterschätzer, respektive. Dann ist für einen *Kontrastgewichtsvektor* $c \in \mathbb{R}^p$ und einen Parameter $\beta_0 \in \mathbb{R}^p$ die *T-Statistik* definiert als

$$T := \frac{c^T\hat{\beta} - c^T\beta_0}{\sqrt{\hat{\sigma}^2 c^T (X^T X)^{-1} c}}. \tag{3}$$

4.Erläutern Sie für die T-Statistik die Bedeutung der Wahl von $c \in \mathbb{R}^p$.

- der *Kontrastgewichtsvektor* $c \in \mathbb{R}^p$ projiziert $\hat{\beta}$ auf einen Skalar $c^T \hat{\beta} \in \mathbb{R}$
- Die Wahl p-dimensionaler Einheitsvektoren für c erlaubt die Auswahl einzelner Komponenten von $\hat{\beta}$ bzw β_0
  (Anmerkung: Einheitsvektoren e_i sind Vektoren, die an der Stelle i eine 1 und sonst nur 0en haben)
- Eine generelle Wahl von c erlaubt die Evaluation beliebiger Linearkombinationen von $\hat{\beta}$ bzw β_0

Beispiel: Wir nehmen an, wir haben einen Betaparametervektor $\beta \in \mathbb{R}^p$ mit $p = 3$, also $\begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$.

Dann kann ich mit den p-dimensionalen Einheitsvektoren

$$e_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, e_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, e_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

β_1 oder β_2 oder β_3, respektive, einzeln auswählen.

Weitere Beispiele:

$c = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$ gewichtet alle Betakomponenten gleich, $c = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}$ gewichtet β_1 doppelt, β_2 einfach und β_3 gar nicht.

5.Erläutern Sie für die T-Statistik die Bedeutung der Wahl von $\beta_0 \in \mathbb{R}^p$.

- Wählt man $\beta_0 := 0_p$, so erhält man eine Deskriptivstatistik, die es erlaubt, geschätzte Regressoreffekte (Komponenten/Linearkombinationen (abhängig von c) von $\hat{\beta}$) im Sinne eines Signal-zu-Rauschen Verhältnisses in Bezug zu der durch $\hat{\sigma}^2$ quantifizierten Residualdatenvariabilität zu setzen.
- Wählt man für $\beta_0 = \beta$, also den w.a.u. Betaparameterwerte, so können mit der T-Statistik Konfidenzintervalle für einzelene Komponenten des Betaparametervektors bestimmt werden.
- Sagt man, $\beta_0 \in \Theta_0$ im Kontext eines Testszenarios als das Element der Nullhypothese Θ_0, so eröffnet die T-Statistik die hypothesentestbasierte Inferenz über Betaparameterkomponenten und ihrer Linearkombinationen des ALMs (abhängig von c)

6.Wann und warum kann die T-Statistik als Signal-zu-Rauschen-Verhältnis interpretiert werden?

- Wenn $\beta_0 := 0_p$, dann ergibt sich die T-Statistik zu

$$T = \frac{c^T \hat{\beta}}{\sqrt{\hat{\sigma}^2 c^T (X^T X)^{-1} c}} = \frac{\text{Geschätzte Effektstärke}}{\text{Geschätzte stichprobenumfangskalierte Datenvariabilität}} \tag{4}$$

- Im Zähler steht der Betaparameterschätzer, im Nenner der Varianzparameterschätzer
- Im Nennen der T-Statistik wird vor allem sichergestellt, dass die adäquate (Co-)Standardabweichung (durch multiplikation mit c) als Bezugsgröße dient, da es sich bei $\hat{\sigma}^2 (X^T X)^{-1}$ um die Kovarianz des Betaparameterschätzers handelt
- Der Effekt ist also das (ëigentliche") Signal, das laut dem formulierten Modell "übertragen"werden sollte und die Fehlervarianz ist das Restrauschen"
- Damit bestimmt die T-Statistik für $\beta_0 = 0_p$ das Verhältnis von Signal zu Rauschen und somit auch Unsicherheit.
- eine weitere Sichtweise für diesen Fall ist die T-Statistik als in Varianzeinheiten ausgedrücke geschätzte Effektstärke

7.Geben Sie das Theorem zur T-Statistik wieder.

Theorem (T-Teststatistik)

Es sei

$$\upsilon = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0_n, \sigma^2 I_n) \tag{5}$$

das ALM. Weiterhin seien

$$\hat{\beta} := (X^T X)^{-1} X^T \upsilon \text{ und } \hat{\sigma}^2 := \frac{(\upsilon - X\hat{\beta})^T(\upsilon - X\hat{\beta})}{n-p} \tag{6}$$

die Betaparameter- und Varianzparameterschätzer, respektive. Schließlich sei für einen Kontrastgewichtsvektor $c \in \mathbb{R}^p$ und einen Parameter $\beta_0 \in \mathbb{R}^p$

$$T := \frac{c^T\hat{\beta} - c^T\beta_0}{\sqrt{\hat{\sigma}^2 c^T (X^T X)^{-1} c}}. \tag{7}$$

die T-Statistik. Dann gilt

$$T \sim t(\delta, n-p) \text{ mit } \delta := \frac{c^T\beta - c^T\beta_0}{\sqrt{\sigma^2 c^T (X^T X)^{-1} c}}. \tag{8}$$

8.Geben Sie die Form der T-Statistik im Szenario von n u.i.n.v. Zufallsvariablen wieder.

Wir haben das ALM Szenario unabhängiger und identisch normalverteilter Zufallsvariablen

$$\upsilon \sim N(X\beta, \sigma^2 I_n) \text{ mit } X := 1_n \in \mathbb{R}^{n\times 1}, \beta := \mu \in \mathbb{R} \text{ und } \sigma^2 > 0.$$

Weiterhin sei $c := 1$ und $\beta_0 = \mu_0$. Dann gilt für die T-Statistik

$$T = \frac{c^T\hat{\beta} - c^T\mu_0}{\sqrt{\hat{\sigma}^2 c^T (X^T X)^{-1} c}} = \frac{1^T\bar{\upsilon} - 1^T\mu_0}{\sqrt{s_\upsilon^2 1^T (1_n^T 1_n)^{-1} 1}} = \sqrt{n}\frac{\bar{\upsilon} - \mu_0}{s_\upsilon}$$

was der Einstichproben-T-Teststatistik für den Fall $\mu_0 = 0$ entspricht (vgl. Einheit(12) Hypothesentests in WTFI und Einheit (9) T-Tests in ALM.)

9.Erläutern Sie den Zusammenhang der T-Statistik und Cohen's d.

Die T-Statistik ist im ALM Szenario u.i.v. Zufallsvariablen gegeben durch

$$T = \sqrt{n}\frac{\bar{\upsilon} - \mu_0}{s_\upsilon} \tag{9}$$

Die T-Statistik nimmt hohe Werte für hohe Werte $\bar{\upsilon}$ (Effekt), kleine Werte von s_υ^2 (Datenvariabilität) und hohe Werte von n (Stichprobenumfang) an.

Cohen's d als *Effekstärkenmaß* ist gegeben als

$$d := \frac{\bar{\upsilon}}{s_\upsilon}, \tag{10}$$

so dass für $\mu_0 := 0$ gilt, dass

$$T = \sqrt{n}d \text{ bzw. } d = \frac{1}{\sqrt{n}}T. \tag{11}$$

Cohen's d ist also ein stichprobenumfang**unabhängiges** Signal-zu-Rauschen-Verhältnis.

10. Geben Sie das Theorem zu Konfidenzintervallen für Betaparameterkomponenten wieder.

Theorem (Konfidenzintervalle für Betaparameterkomponenten)

Es sei

$$\upsilon = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0_n, \sigma^2 I_n) \tag{12}$$

das ALM, $\hat{\beta}$ *und* $\hat{\sigma}^2$ *seien die Betaparameter- und Varianzparameterschätzer, respektive und für ein* $\delta \in]0,1[$ *sei*

$$t_\delta := \Psi^{-1}\left(\frac{1+\delta}{2}; n-p\right). \tag{13}$$

Schließlich seit für $j = 1, ..., p$

$$\lambda_j = \left((X^T X)^{-1}\right)_{jj} \text{ das } j\text{te Diagonalelement von } (X^T X)^{-1}. \tag{14}$$

Dann ist für $j = 1, ..., p$

$$\kappa_j := \left[\hat{\beta}_j - \hat{\sigma}\sqrt{\lambda_j}t_\delta, \hat{\beta}_j + \hat{\sigma}\sqrt{\lambda_j}t_\delta\right] \tag{15}$$

ein δ*-Konfidenzintervall für die* j*te Komponente* β_j *des Betaparameters* $\beta = (\beta_1, ..., \beta_p)^T$.

Anmerkung:

- Ψ ist die kumulative Verteilungsfunktion der T-Zufallsvariable und damit Ψ^{-1} ihre Inverse
- vgl auch (11) Konfidenzintervalle in WTFI