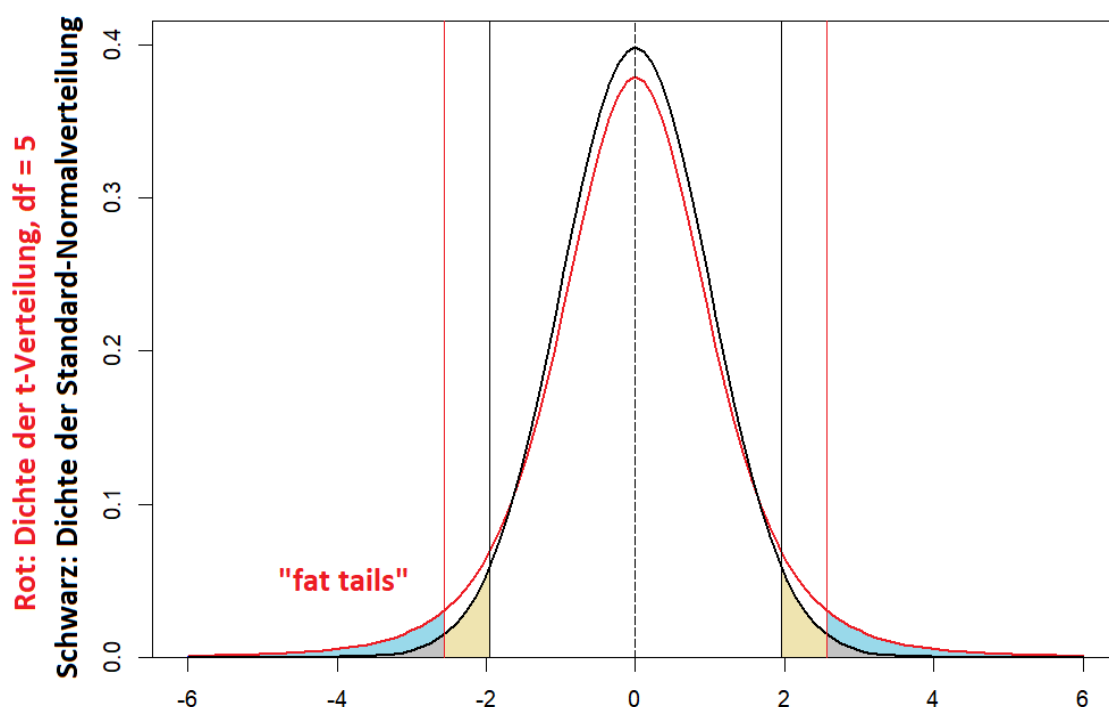


Testatfragen zu Vorlesung III

21. Beschreibe den Ablauf eines statistischen Hypothesentests.
22. Was versteht man unter dem Fehler I. Art beim Hypothesentest, und wie wird er kontrolliert?
23. Was versteht man unter der Power eines Hypothesentests?
24. Richtig oder falsch? “Der statistische Hypothesentest gegen die Nullhypothese $\mathcal{H}_0$ fiel negativ aus ($p = 0.915$). Damit ist die Nullhypothese bewiesen.” Begründe die Antwort.
25. Definiere: gepoolte empirische Standardabweichung $\hat{\sigma}_p$.
26. Wie lautet die Prüfgröße $\hat{T}$ beim klassischen t -Test, und welcher Verteilung unterliegt sie?
27. Es sei $df \geq 30$. Welche Näherungswerte darf man für die 2.5-%- und 97.5-%-Quantile $Q_{2.5}(t, df)$ und $Q_{97.5}(t, df)$ benutzen?
28. Skizziere den Akzeptanz- und Ablehnbereich für einen zweiseitigen t -Test zum Signifikanzniveau $\alpha = 0.05$.
29. Wie verhält sich der Fehler II. Art beim Hypothesentest, wenn der Stichprobenumfang anwächst?
30. Nenne ein Beispiel für zwei verbundene Stichproben.

Anmerkungen

- * *Hypothesentest* (pp. 3-01 ff.): Zuerst vorgeschlagen von [NEYMAN/PEARSON (1933)]: “Here, for example, would be such a ‘rule of behaviour’: to decide, whether a hypothesis, H , of a given type be rejected or not, calculate a specified character, x , of the observed facts; if $x > x_0$ reject H , if $x \leq x_0$ accept H . Such a rule tells us nothing as to whether in a particular case H is true when $x \leq x_0$ or false when $x > x_0$. But it may often be proved that if we behave according to such a rule, then in the long run we shall reject H when it is true not more, say, than once in hundred times, and in addition we may have evidence that we shall reject H sufficiently often when it is false.” (ebd., p. 291)
- * *Probleme des Hypothesentest-Konzepts* (p. 3-08): Ausführlich und nicht nur für Mediziner verständlich verständlich dargelegt in [MARK ET AL. (2016)].
- * *Beispiel 3.1.* (p. 3-10): Die Werte stammen wiederum aus dem Datensatz zu [WALTHER ET AL. (2011)].
- * *Körperoberfläche* (p. 3-10): Auf zwei verschiedene Möglichkeiten, die Körperoberfläche aus Größe und Gewicht zu berechnen (oder besser gesagt, zu schätzen), kommen wir weiter unten zurück.
- * *Cohen’s d* (p. 3-12 f.): Soweit ich sehe, wurde die Größe bereits in [COHEN (1977)] eingeführt. Wir berechnen sie wieder aus den Kenngrößen der Stichprobe und bezeichnen sie daher mit einem Dach.
- * *Lehrsatz 3.6., 1)* (p. 3-14): Ich habe diese Aussage hier mit angeführt, um die Analogie zu Lehrsatz 2.9. zu verdeutlichen. Praktisch entscheidend ist jedoch Teil 2) des Satzes.
- * *Lehrsatz 3.6., 2)* (p. 3-14): Das war Gossets Entdeckung [STUDENT (1908)]. Gosset publizierte unter dem Pseudonym Student; deshalb heißt die t -Verteilung bis heute auch Student’sche Verteilung.
- * *Symmetrie der t -Verteilung* (p. 3-16): Aus der Symmetrie der t -Verteilung folgt wie bei der Normalverteilung die symmetrische Lage der Quantile bezüglich der 50-%-Marke bei Null: $Q_{(100-p)}(t, df) = -Q_p(t, df)$ für alle $0 < p < 100$.



Schwarz: Quantile $Q_{2.5}(\mathcal{N}, 0, 1)$ und $Q_{97.5}(\mathcal{N}, 0, 1)$ der Standardnormalverteilung; rot: Quantile $Q_{2.5}(t, 5)$ und $Q_{97.5}(t, 5)$ der t -Verteilung mit $df = 5$.

* *Einseitige Fragestellung* (p. 3-18): In der Praxis sind einseitige Fragestellungen eher die Ausnahme. Achtung: Der Fall, daß μ' und μ'' gerade die umgekehrte strenge Ungleichung erfüllen, kann beim einseitigen Test nicht aufgedeckt werden, sondern geht in der Nullhypothese auf.

* *Fehlerbalken-Diagramm* (p. 3-20): Man könnte die Konfidenzintervalle für die einzelnen Gruppen und Zeitpunkte auch ohne Bezug auf den zeitlichen Verlauf nebeneinander eintragen. Zu einem derartigen Diagramm gehört immer die Information über das Niveau, zu dem die Konfidenzintervalle gebildet wurden. Sehr selten wird statt des Konfidenzintervalls das Intervall $[\hat{\mu} - \hat{\sigma}, \hat{\mu} + \hat{\sigma}]$ geplottet.

* *Simulationen zur Schätzung des Fehlers II. Art* (p. 3-22 f.): Bei diesem Vorgehen erhält man für die Fehler II. Art zwar keine exakten Werte, aber dennoch ziemlich realistische Schätzungen. Das Verhalten des Fehlers wird in den Tabellen korrekt widerspiegelt: der Fehler II. Art verringert sich mit wachsender Mittelwertdifferenz, mit abnehmender Standardabweichung und mit wachsendem Umfang der Stichproben. Die geschätzte Power des Tests beträgt jeweils $(100 - p)$ Prozent.

* *Fallzahlplanung* (p. 3-23): Für diese Aufgabe steht spezielle Software zur Verfügung, zum Beispiel das Freeware-Programm G*Power 3.1. (<https://gpower.software.informer.com/download/>) (09.10.2026). In diesem Programm sind exakte Fallzahlberechnungen in Abhängigkeit von Test, Daten, α , β , ... , implementiert. In der Vorlesung wurde stattdessen nochmals eine Simulation programmiert.

* *Welch-Test* (p. 3-24): Die Originalpublikation [WELCH (1947)] ist ein gutes Beispiel für den mathematischen Tiefgang, der hinter den "Rezepten" in unserem Kurs steckt.

* *Robustheit des t -Tests* (p. 3-25): [SAWILOWSKY/BLAIR (1992)] betonen, daß die Voraussetzung über die Normalverteilung der Variablen x in $\mathcal{P}$ in der Praxis fast niemals eingehalten wird. Umso wichtiger ist ihre Beobachtung, daß der t -Test auch für Stichproben aus den dargestellten (und weiteren, ebenso "krummen") Verteilungen sinnvoll bleibt. Aus der Veröffentlichung geht leider nicht hervor, ob der klassische t -Test oder

der Welch-Test angewendet wurde. Unter unseren Voraussetzungen wäre in allen Fällen der Welch-Test durchzuführen.

* *Drei-Äpfel-Diät* (p. 3-26): Die amerikanische Ernährungsberaterin Tammi Flynn hat vorgeschlagen, zwecks Gewichtsreduktion vor jeder Hauptmahlzeit einen Apfel zu essen; siehe [FLYNN/DRURY (2003)].

* *HDL-Cholesterol* (p. 3-26): Eine Erhöhung des HDL-Cholesterolspiegels ist mit der Verringerung des Risikos für Arteriosklerose und kardiovaskuläre Erkrankungen assoziiert.

* *Fallkontrollstudie* (p. 3-26): Die Aussagekraft einer solchen Studie hängt entscheidend davon ab, wie sorgfältig die Zuordnung (“Matching”) zwischen erkrankten bzw. behandelten Personen und Kontrollpersonen vorgenommen wird. Im Idealfall stammen beide Gruppen aus der gleichen Bevölkerungskohorte. Fallkontrollstudien können retrospektiv durchgeführt werden und sind deshalb oft wesentlich kostengünstiger als Interventionsstudien (im Beispiel würde man die Daten einfach aus den Patientenakten erheben). Weil die Stichproben aufgrund der Matchingprozedur nicht unabhängig sind, darf der gewöhnliche t -Test nicht angewendet werden.

* *Zwei Formeln zur Berechnung der Körperoberfläche* (pp. 3-26 ff.): Die beiden Formeln wurden in [MOSTELLER (1987)] und [LIVINGSTON/LEE (2001)] publiziert. Die einprägsame Mosteller-Formel ist ausdrücklich auch für Kinder geeignet; die Livingston-Lee-Formel gilt in der angegebenen Form für Gewichte ab 10 kg.

* *Freiheitsgrade der t -Verteilung bei der Prüfgröße für Paardifferenzen* (p. 3-28): Die Stichprobe mit den Kenngrößen $(N^\Delta, \hat{\mu}^\Delta, \hat{\sigma}^\Delta)$ trägt $(N^\Delta - 1)$ Freiheitsgrade bei, die formale zweite Stichprobe mit $(N^\Delta, 0, 0)$ muß ausschließlich aus Nullen bestehen, so daß keine weiteren Freiheitsgrade hinzukommen.

* *Mangelnde Praxisrelevanz eines signifikant ausgefallenen Hypothesentests* (p. 3-29): Signifikanz darf niemals mit praktischer Relevanz verwechselt werden. Umgekehrt kann ein Mittelwertunterschied auch dann, wenn ein Hypothesentest nicht signifikant ausfällt, erhebliche praktische Bedeutung haben!

* *Äquivalenztest* (p. 3-31): Ein beliebter Fehlschluß beim Testen besagt, daß die Nullhypothese wahr ist, sobald der Test nicht signifikant ausfällt. Dabei wird übersehen, daß — unabhängig von α — die Beibehaltung von $\mathcal{H}_0$ mit Wahrscheinlichkeit β ein falsch-negatives Testergebnis darstellt. Deshalb kann die annähernde Übereinstimmung $\mu' \approx \mu''$ der wahren Mittelwerte nicht dadurch belegt werden, daß ein zweiseitiger Hypothesentest gegen die Nullhypothese $\mu' = \mu''$ insignifikant ausfällt. Vielmehr muß die Nullhypothese “ $|\mu' - \mu''| \geq v$ ” lauten, wobei der maximal akzeptierte Abstand v anhand von praktischen Beobachtungen und Erfordernissen festzulegen ist. (In Beispiel 3.12. würde die Fehlerrechnung den Abstand $v \approx 0.02 \text{ m}^2$ nahelegen.) Da es sich jetzt um eine zusammengesetzte Nullhypothese handelt, müssen nacheinander zwei einseitige Tests ausgeführt werden (“TOST” bedeutet “two one-sided tests”), und zwar beide zum gleichen Signifikanzniveau α (keine Korrektur nötig). $\mathcal{H}_0$ wird verworfen, wenn beide Tests signifikant ausfallen. Siehe z.B. [WALKER/NOWACKI (2011)].

Quellenverzeichnis

- [COHEN (1977)] Cohen, J.: *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates; Hillsdale, N.J. 1977
- [FLYNN/DRURY (2003)] Flynn, T.; Drury, J.: *The “3 Apple-a-Day” Plan: Your Foundation for Permanent Weight Loss*. Get Fit Publishing; 2003
- [LIVINGSTON/LEE (2001)] Livingston, E. H.; Lee, S.: *Body surface area prediction in normal-weight and obese patients*. Am. J. Physiol. Endocrinol. Metab. **281** (2001) : E586 – E591
- [MARK ET AL. (2016)] Mark, D.B.; Lee, K. L.; Harrell jr., F. E.: *Understanding the role of p values and hypothesis tests in clinical research*. JAMA Cardiol. **1** (2016) : 1048 – 1054

- [MOSTELLER (1987)] Mosteller, R. D.: *Simplified calculation of body-surface area*. New England J. Med. **317** (1987) : 1098
- [NEYMAN/PEARSON (1933)] Neyman, J.; Pearson, E. S.: *IX. On the problem of the most efficient tests of statistical hypotheses*. Philosophical Transactions of the Royal Society of London, Series A: Mathematical, Physical and Engineering Sciences **231** (1933) : 289 – 337
- [SAWILOWSKY/BLAIR (1992)] Sawilowsky, S. S.; Blair, R. C.: *A more realistic look at the robustness and Type II error properties of the t test to departures from population normality*. Psych. Bulletin **111** (1992) : 352 – 360
- [STUDENT (1908)] Student [Gosset, W. S.] : *The probable error of a mean*. Biometrika **6** (1908) : 1 – 25
- [WALKER/NOWACKI (2011)] Walker, E.; Nowacki, A. S.: *Understanding equivalence and noninferiority testing*. J. Gen. Intern. Med. **26** (2011) : 192 – 196
- [WALTHER ET AL. (2011)] Walther, C.; Mende, M.; Gaede, L.; et al.: *Einfluss eines täglichen Schulsportunterrichts auf das kardiovaskuläre Risiko — 2-Jahres Ergebnisse einer Cluster-randomisierten Studie*. Deutsche Med. Wochenschr. **136** (2011) : 2348 – 2354
- [WELCH (1947)] Welch, B. L.: *The generalization of ‘Student’s’ problem when several different population variances are involved*. Biometrika **34** (1947) : 28 – 35

Bildnachweis

* p. 3-25: Oben links: [SAWILOWSKY/BLAIR (1992)], p. 354, Fig. 1; oben rechts: ebd., p. 355, Fig. 6; unten links: ebd., p. 355, Fig. 7; unten rechts: ebd., p. 354, Fig. 4 (alle geringfügig verändert)