

Testatfragen zu Vorlesung II

11. Skizziere die Dichte einer Gleichverteilung mit den Parametern $a = 0$ und $b = 10$.
12. Skizziere die Dichten der drei Normalverteilungen, die jeweils den Mittelwert $\mu = 0$ und die Standardabweichungen $\sigma_1 = 0.5$, $\sigma_2 = 1$ und $\sigma_3 = 2$ besitzen.
13. Die Verteilung einer metrischen Variable x wird durch die Dichtefunktion f beschrieben. Wo kann man die Wahrscheinlichkeiten $\text{Prob}\{x < 0\}$, $\text{Prob}\{0 \leq x < 1\}$ und $\text{Prob}\{x \geq 1\}$ ablesen (Skizze)?
14. Die Verteilung einer metrischen Variable x wird durch die Dichtefunktion f beschrieben. Wie wird das 97.5%-Quantil $Q_{97.5}(f)$ erklärt (Skizze)?
15. Welcher Verteilung unterliegen die Stichproben-Mittelwerte $\hat{\mu}$ einer metrischen Variablen beim wiederholten Ziehen von gleichgroßen, unabhängigen Stichproben?
16. Definiere das 95%-Konfidenzintervall für den Stichproben-Mittelwert einer Stichprobe vom Umfang N .
17. Für 500 unabhängige, gleichgroße Stichproben aus $\mathcal{P}$ wurde jeweils das 90%-Konfidenzintervall für den Stichproben-Mittelwert der metrischen Variable x berechnet. Wieviele dieser Intervalle enthalten den wahren Mittelwert μ von x in $\mathcal{P}$?
18. Was versteht man unter einer zweifach verblindeten randomisierten Interventionsstudie?
19. Erkläre: primärer Endpunkt einer Studie.
20. Welche vier wichtigsten Grundsätze sind bei der Planung und Durchführung einer randomisierten Studie zu beachten?

Anmerkungen

* *Verteilungsdichte* (p. 2-03): Eine Verteilungsdichte ist eine Funktion $f: \mathbb{R} \rightarrow [0, \infty)$ mit folgenden mathematischen Eigenschaften: Sie muß Lebesgue-integrierbar sein, darf nur nichtnegative Werte $f(x) \geq 0$ annehmen und muß im Integral auf Eins normiert sein: $\int_{-\infty}^{\infty} f(x) dx = 1$. In der Grafik auf p. 2-03 ist die Dichte der Standard-Normalverteilung dargestellt.

* *Gleichverteilung* (p. 2-04): Eine Gleichverteilung auf dem Intervall $[a, b]$ wird durch die beiden Intervallgrenzen a, b als Parameter bestimmt. Ihre Dichte lautet:

$$f(x, a, b) = \begin{cases} 0 & \text{für } x < a, \\ 1/(b - a) & \text{für } a \leq x \leq b, \\ 0 & \text{für } x > b. \end{cases}$$

Mittelwert und Median stimmen überein: $\mu = m = (a + b)/2$, und es gilt $\sigma = (b - a)/(2\sqrt{3})$.

* *Exponentialverteilung* (p. 2-04): Die Exponentialverteilung ist auf der nichtnegativen Halbachse $[0, \infty)$ definiert und wird durch genau einen Parameter $\lambda > 0$ bestimmt. Ihre Dichte lautet:

$$f(x, \lambda) = \begin{cases} \lambda \cdot e^{-\lambda x} & \text{für } x \geq 0, \\ 0 & \text{für } x < 0. \end{cases}$$

Sie ist vor allem bei der Modellierung von Zeitdauern und Lebensdauern anzutreffen (dabei ist λ die Ausfallrate). Es gilt: $\mu = 1/\lambda$, $m = \ln 2/\lambda$ und $\sigma = 1/\lambda$.

* *Normalverteilung* (pp. 2-04, 2-08): Die Normalverteilung ist auf der gesamten reellen Achse definiert und wird unmittelbar durch die zwei Parameter Mittelwert μ und Standardabweichung σ bestimmt. Ihre Dichte ("Glockenkurve") ist bezüglich der y -Achse symmetrisch und lautet:

$$f(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$$

Aufgrund des sog. zentralen Grenzwertsatzes spielt die Normalverteilung in der Wahrscheinlichkeitsrechnung eine zentrale Rolle: Summiert man viele unabhängige Zufallsvariablen auf, so entsteht unabhängig davon, wie diese verteilt sind, in guter Näherung eine normalverteilte Größe.

* *t-Verteilung* (p. 2-04): Die *t*-Verteilung hängt mit der Verteilung von Differenzen von Stichproben-Mittelwerten zusammen (siehe Vorlesung III, Lehrsatz 3.6.). Sie ist auf der gesamten reellen Achse definiert und wird unmittelbar durch einen ganzzahligen Parameter, die Anzahl der Freiheitsgrade df , bestimmt. Ihre Dichte ist bezüglich der y -Achse symmetrisch. Für $df \geq 30$ ist die *t*-Verteilung fast nicht mehr von einer Standard-Normalverteilung mit $\mu = 0$ und $\sigma = 1$ zu unterscheiden.

* *Chi-Quadrat-Verteilung* (p. 2-04): Grob gesprochen, ergibt sich eine χ^2 -Verteilung mit $df = N$ Freiheitsgraden, wenn man die Quadrate von N metrischen Variablen aufsummiert, die alle einer Standard-Normalverteilung mit $\mu = 0$ und $\sigma = 1$ unterliegen. Sie ist nur auf der nichtnegativen Halbachse $[0, \infty)$ definiert. Für $df \geq 100$ Freiheitsgrade ist die χ^2 -Verteilung fast nicht mehr von einer Normalverteilung mit $\mu = N$ und $\sigma = \sqrt{2N}$ zu unterscheiden.

* *Fisher-Verteilung* (p. 2-04): Eine Fisher-Verteilung mit (df_1, df_2) Freiheitsgraden entsteht als Verteilung des Quotienten von zwei Chi-Quadrat-verteilten metrischen Variablen mit df_1 bzw. df_2 Freiheitsgraden. Sie ist nur auf der nichtnegativen Halbachse $[0, \infty)$ definiert.

* *Beliebige p-Quantile* (p. 2-07): Eine andere Bezeichnung für Quantil ist “Perzentil”.

* *Lehrsatz 2.9.* (p. 2-12) ist ebenfalls eine Konsequenz des zentralen Grenzwertsatzes. Das Theorem kann wie folgt umformuliert werden:

Lehrsatz 2.9.a: Eine metrische Variable x in der Gesamtpopulation $\mathcal{P}$ möge den Mittelwert μ und die Standardabweichung σ besitzen. Mit den Stichproben-Mittelwerten $\hat{\mu}$ von x (für unabhängige Stichproben vom gleichen Umfang $N \geq 30$) konstruieren wir die metrische Variable

$$v = \frac{\hat{\mu} - \mu}{\sigma/\sqrt{N}}.$$

Unabhängig davon, auf welche Weise x in $\mathcal{P}$ verteilt ist, unterliegt die Variable v in sehr guter Näherung einer Standard-Normalverteilung mit Mittelwert 0 und Standardabweichung 1.

* *Bezeichnungen* auf pp. 2-13 ff.: Die Bezeichnungen der Quantile der Normalverteilung auf den folgenden Seiten sind verkürzt: Statt $Q_p(\mathcal{N})$ müßten wir eigentlich $Q_p(\mathcal{N}, \mu, \sigma/\sqrt{N})$ schreiben, um die Abhängigkeit der Normalverteilung $\mathcal{N}(\mu, \sigma/\sqrt{N})$ von den Parametern μ und $\sigma/\sqrt{N}$ zum Ausdruck zu bringen.

* *Boston Housing Dataset* (p. 2-15): Der Datensatz wurde im Zusammenhang mit [HARRISON JR./RUBINFELD (1978)] publiziert und enthält Variablen, die seinerzeit die Preisbildung für Wohnhäuser im Großraum Boston beeinflusst haben (abrufbar z.B. unter faculty.tuck.dartmouth.edu/images/uploads/faculty/business-analytics/Boston_Housing.xlsx (09.10.2026)). Zur damaligen Zeit galt ein Datensatz mit $N = 506$ Messungen für 14 Variablen bereits als “big data”; in der Folge wurden die Daten immer wieder als Beispiel herangezogen. Die durchschnittliche Klassenstärke, definiert als “pupil-teacher ratio by town school district”, kann als ein Maß für die Lebensqualität im betreffenden Stadtbezirk aufgefaßt werden. Die Konfidenzintervalle wurden mit σ bestimmt und haben daher alle die gleiche Breite.

* *Zusatz zu Definition 2.10.* (p. 2-16): Wenn man bei der Bildung des Konfidenzintervalls anstelle von σ die empirische Standardabweichung $\hat{\sigma}$ benutzt, kann man sich nicht mehr auf Lehrsatz 2.9. berufen. Stattdessen gilt:

Lehrsatz 2.9.b: Eine metrische Variable x in der Gesamtpopulation $\mathcal{P}$ möge den Mittelwert μ und die Standardabweichung σ besitzen. Mit den Stichproben-Mittelwerten $\hat{\mu}$ und den empirischen Standardabweichungen $\hat{\sigma}$ von x (für unabhängige Stichproben vom gleichen Umfang $N \geq 30$) konstruieren wir die metrische Variable

$$v = \frac{\hat{\mu} - \mu}{\hat{\sigma}/\sqrt{N}}.$$

Unabhängig davon, auf welche Weise x in $\mathcal{P}$ verteilt ist, unterliegt auch die Variable v in sehr guter Näherung einer Standard-Normalverteilung mit Mittelwert 0 und Standardabweichung 1.

Die Breite der Konfidenzintervalle zu verschiedenen Stichproben gleichen Umfangs ist dann allerdings nicht mehr konstant, sondern zusammen mit $\hat{\sigma}$ variabel.

* *Konfidenzintervalle zu beliebigen Prozentzahlen* (p. 2-16): Wegen der Symmetrie der Normalverteilung gilt

$$\frac{Q_{95}(\mathcal{N}, \mu, \sigma/\sqrt{N}) - Q_5(\mathcal{N}, \mu, \sigma/\sqrt{N})}{2} = Q_{95}(\mathcal{N}, 0, 1) \cdot \frac{\sigma}{\sqrt{N}}$$

mit $Q_{95}(\mathcal{N}, 0, 1) = -Q_5(\mathcal{N}, 0, 1) = 1.645$. Allgemein gilt

$$\frac{Q_{(100-p/2)}(\mathcal{N}, \mu, \sigma/\sqrt{N}) - Q_{p/2}(\mathcal{N}, \mu, \sigma/\sqrt{N})}{2} = Q_{(100-p/2)}(\mathcal{N}, 0, 1) \cdot \frac{\sigma}{\sqrt{N}}$$

mit $Q_{(100-p/2)}(\mathcal{N}, 0, 1) = -Q_{p/2}(\mathcal{N}, 0, 1)$, so daß es in jedem Fall nur auf die Quantile der Standard-Normalverteilung $\mathcal{N}(0, 1)$ ankommt. Für $N \geq 30$ darf man erneut $\hat{\sigma}$ anstelle von σ verwenden.

* $\mathcal{S}^{(1)}$, $\mathcal{S}^{(2)}$ und $\mathcal{S}^{(3)}$ (p. 2-17) wurden als unabhängige “Stichproben aus der ursprünglichen Stichprobe” gezogen.

* *Definition 2.14.* (p. 2-19): Völlig analog zu Definition 2.14. werden q -%-Konfidenzintervalle für die relative Häufigkeit $\hat{p}$ in der Stichprobe zu einem beliebigen Prozentsatz $0 < q < 100$ erklärt:

$$\left[\hat{p} - Q_{(100-q/2)}(\mathcal{N}, 0, 1) \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{N}}, \hat{p} + Q_{(100-q/2)}(\mathcal{N}, 0, 1) \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{N}} \right].$$

* *Konfidenzintervalle für andere Kenngrößen* (p. 2-20): Zum Beispiel lautet das 95%-Konfidenzintervall für die empirische Standardabweichung $\hat{\sigma}$ in einer Stichprobe vom Umfang N :

$$\left[\hat{\sigma} \cdot \sqrt{\frac{N-1}{Q_{97.5}(\chi^2, N-1)}}, \hat{\sigma} \cdot \sqrt{\frac{N-1}{Q_{2.5}(\chi^2, N-1)}} \right].$$

Die Quantile beziehen sich auf eine Chi-Quadrat-Verteilung mit $df = N - 1$ Freiheitsgraden. Achtung: dieses Konfidenzintervall ist bezüglich $\hat{\sigma}$ nicht symmetrisch!

* *Randomisierung* (pp. 2-21, 2-23, 2-30) bedeutet die zufällige Zuordnung der Teilnehmer zu den verschiedenen Studienarmen. Diese Aufgabe wird heute von speziellen Computerprogrammen übernommen.

* *Abkürzungen* (p. 2-23): Ges — Gesamtpopulation, St — Stichprobe, Rnd — Randomisierung, Z — Zielmerkmal.

* *Crossover-Design* (p. 2-23): Die korrekte Auswertung einer Crossover-Studie mit einer metrischen Zielgröße wird in [WELLEK/BLETTNER (2012)] beschrieben.

* *“Schülersportstudie”* (pp. 2-26 ff.): Neben der Publikation [WALTHER ET AL. (2011)] haben wir Zugang zu einem Teil der Studiendaten. In dem uns zur Verfügung gestellten Datensatz sind jedoch aus Gründen des Datenschutzes einige Einträge verändert worden.

* *Daten* auf p. 2-32: Die Daten in der Tabelle stammen aus dem Datensatz, Ausgangsposition und Ergebnis aus [WALTHER ET AL. (2011)], pp. 2351.

Quellenverzeichnis

- [HARRISON JR./RUBINFELD (1978)] Harrison jr., D.; Rubinfeld, D. L.: *Hedonic house prices and the demand for clean air*. J. of Environmental Economics and Management **5** (1978) : 81 – 102
- [STUESSI/RINDERER (2019)] Stuessi, C.; Rinderer, B.: *Kardiopulmonale Leistungsdiagnostik bei Kindern und Jugendlichen*. Swiss Sports & Exercise Medicine **67** (2019) : 38 – 43
- [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
- [WELLEK/BLETTNER (2012)] Wellek, S.; Blettner, M.: *Vom richtigen Umgang mit dem Crossover-Design in klinischen Studien*. Teil 18 der Serie zur Bewertung wissenschaftlicher Publikationen. Deutsches Ärzteblatt **109** (2012) : 276 – 281

Bildnachweis

- * p. 2-15: Erzeugt mit shinyapps.wiwi.hu-berlin.de/mmstat_en/confidence_mean (20.09.2021) (verändert)
- * p. 2-17 f.: Logo des Leipziger Forschungszentrums für Zivilisationskrankheiten, www.life.uni-leipzig.de/de/life_forschungszentrum.html (16.07.2021)
- * p. 2-26: [STUESSI/RINDERER (2019)], p. 41, Figure