10. Schließende Statistik

Inhalt

In der ***Beschreibenden Statistik*** wird eine begrenzte Menge von Daten (z. B. Messwerten, Beobachtungen) durch ihre statistischen Eigenschaften (z. B. Mittelwert, Korrelationskoeffizient) charakterisiert.

In der ***Schließenden Statistik*** wird versucht, Aussagen über eine große Grundgesamtheit aus der Untersuchung einer (vergleichsweise) kleinen Stichprobe zu gewinnen. Die zwei wichtigsten Aufgaben sind

- Schätzen unbekannter Parameter,
- Testen von Hypothesen.

Im Rahmen der Vorlesung kann nur ein sehr kleiner Überblick über die wichtigsten Begriffe und Methoden der Schließenden Statistik gegeben werden.

10.1. Schätzen unbekannter Parameter

10.1.1. Stichprobe

Im Folgenden betrachten wir nur Stichproben mit quantitativen Merkmalen. Diese lassen sich durch Zufallsvariablen beschreiben. Wir werden uns auf univariate Merkmale beschränken. Die Methoden lassen sich jedoch auf multivariate Merkmale erweitern.

Definition 10.1 (Stichprobe)
Eine ***Stichprobe*** (engl. *sample*) vom Umfang n ist eine Folge $X_1, \ldots, X_n$ von unabhängigen, identisch verteilten Zufallsvariablen. Dabei ist $X_i \in \mathbb{R}$ die Merkmalsausprägung des i-ten Elements der Stichprobe.

Die X_i heißen ***Stichprobenvariablen***.

Eine Stichprobe entspricht demnach der n-fachen Ausführung eines Zufallsexperiments unter identischen Bedingungen.

10.1.2. Schätzer

Definition 10.2 (Stichprobenfunktion/Schätzfunktion, Schätzer)
Seien $X_1, \ldots, X_n$ eine Stichprobe vom Umfang n und $\theta \in \mathbb{R}$ ein Parameter der Grundgesamtheit, über den eine Aussage gewonnen werden soll.

Dann heißt eine Funktion

$$g : \mathbb{R}^n \to \mathbb{R}, \ (X_1, \ldots, X_n) \mapsto \hat{\theta}$$

Stichprobenfunktion (auch ***Schätzfunktion***). Die Größe $\hat{\theta} = g(X_1, \ldots, X_n)$ heißt ***Schätzer für θ*** (engl. *estimator*).

Beispiel : Parameter

Grundgesamtheit (Stichprobe)	Beispiele für Parameter θ
Bevölkerung eines Staates (befragte Personen)	Anteil Wähler einer Partei, Anzahl Käufer eines Produkts
Produktion einer Maschine (getestete Artikel)	Anzahl X defekter Artikel, Wahrscheinlichkeit $P(X > x)$
Strom-Spannungs-Kurve $I(U)$ (Messwerte U_i, I_i)	Ohmscher Widerstand R

10.1.3. Erwartungstreue und Konsistenz

Da ein Schätzer auf einer zufälligen Stichprobe basiert, ist er eine Zufallsvariable. Es ist i. A. nicht zu erwarten, dass er exakt den zu schätzenden Parameter liefert. Wir können aber fordern, dass sein Erwartungswert gleich dem zu schätzenden Parameter ist und dass er mit wachsendem Stichprobenumfang genauer wird.

Beispiel : Guter Schätzer für Erwartungswert?

Sei $X_1, \ldots, X_n$ eine Stichprobe eines Merkmals X, das zufällig verteilt ist mit Erwartungswert μ und Varianz σ^2 . Ist

$$\hat{\mu} = \overline{X} := \frac{1}{n} \sum_{i=1}^{n} X_i$$

ein geeigneter Schätzer für den Erwartungswert μ ?

Nach den Rechenregeln für den Erwartungswert und die Varianz erhalten wir (vgl. Satz 6.27)

$$\mathrm{E}(\overline{X}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(X_i) = \frac{1}{n}\, n\mu = \mu\,,$$

$$\mathrm{Var}(\overline{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{1}{n^2}\, n\sigma^2 = \frac{\sigma^2}{n} \xrightarrow[n\to\infty]{} 0\,.$$

Der Erwartungswert des Schätzers entspricht also dem zu schätzenden Parameter und die Varianz des Schätzers sinkt mit wachsendem Stichprobenumfang.

Definition 10.3 (Erwartungstreue und Konsistenz)
Sei $\hat{\theta}$ ein Schätzer für einen Parameter θ.

Der Schätzer heißt ***erwartungstreu*** (auch ***unverzerrt***, engl. *unbiased*), falls sein Erwartungswert gleich dem zu schätzenden Parameter ist,

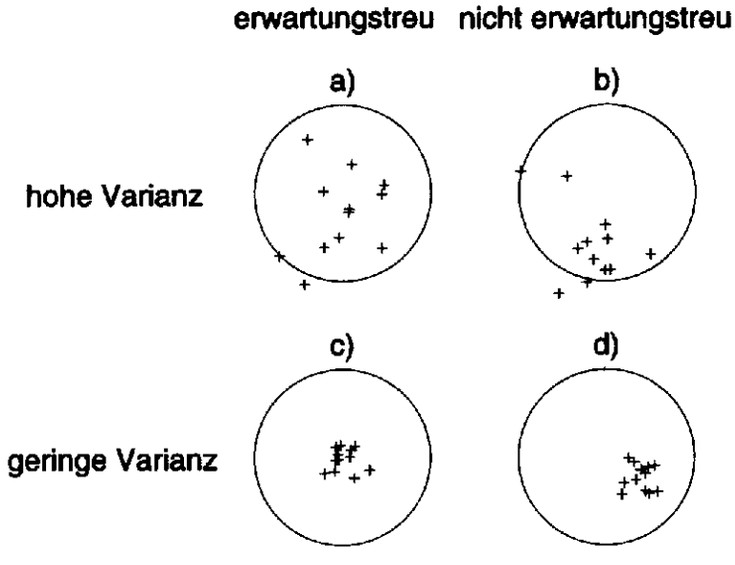
$$\mathrm{E}(\hat{\theta}) = \theta\,.$$

Der Schätzer heißt ***konsistent*** (engl. *consistent*), falls er stochastisch gegen den zu schätzenden Parameter konvergiert,

$$\forall \varepsilon > 0 : \lim_{n\to\infty} P(\left|\hat{\theta} - \theta\right| < \varepsilon) = 1\,.$$

Konsistenz bedeutet anschaulich, dass der Schätzer mit wachsendem Stichprobenumfang genauer wird. Anders gesagt lohnt sich der Aufwand, eine größere Stichprobe zu erheben.

In der folgenden Abbildung entspricht die Mitte des Kreises dem zu schätzenden Parameter θ, während die Kreuze Werte des Schätzers $\hat{\theta}$ für verschiedene Stichproben darstellen.



Bildquelle: Sachs 2018

Satz 10.4 (Schätzer für Erwartungswert, Varianz, Wahrscheinlichkeit)
Sei $X_1, \ldots, X_n$ eine Stichprobe eines Merkmals X, das zufällig verteilt ist mit Erwartungswert μ, Varianz σ^2 und der Wahrscheinlichkeit $p_A := P(X \in A)$ für $A \subseteq \mathbb{R}$.

Dann sind die folgenden Schätzer erwartungstreu und konsistent:

$$\hat{\mu} = \overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i\,,$$
$$\hat{\sigma}^2 = s_X^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\,,$$
$$\hat{p}_A = \frac{|\{i : X_i \in A\,,\ i = 1,\dots n\}|}{n}\,.$$

Für den Erwartungswert μ und die Wahrscheinlichkeit p_A folgen die Aussagen aus dem Gesetz der großen Zahlen und dem Hauptsatz der Statistik (Satz 6.28 bzw. Satz 6.29), lassen sich jedoch auch direkt nachrechnen.

Erwartungstreue der Varianz

Für die Varianz lässt sich zeigen (siehe z. B. G. Teschl und S. Teschl 2014), dass

$$\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = \frac{n-1}{n}\sigma^2$$

und

$$\mathrm{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = \frac{n}{n-1}\,\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = \sigma^2$$

gilt. Der intuitiv nahe liegende Ansatz, die Summe der n quadratischen Abweichungen vom Mittelwert durch n zu teilen, ist also nicht erwartungstreu.

Eine anschauliche Erklärung für den Divisor $n-1$ ist, dass zum Schätzen der Streuung von X (für die die Varianz ein Maß darstellt) mindestens zwei Stichprobenwerte benötigt werden. Ein Stichprobenwert dient sozusagen als Referenz, die übrigen $n-1$ Werte werden zum Schätzen der Streuung benutzt. Eine weitere Plausibilisierung für das Teilen durch $n-1$ wurde bereits im Kapitel zur Beschreibenden Statistik nach Definition 2.31 erwähnt.

Falls in der obigen Formel statt dem Schätzer $\overline{X} = \hat{\mu}$ für den Erwartungswert ein exakt bekannter Wert μ verwendet wird, ist hingegen der Ansatz mit dem Divisor n erwartungstreu. Erstens kommt dieser Fall jedoch in der Praxis kaum vor, zweitens ist der Unterschied zwischen den Ergebnissen für n und $n-1$ bei ausreichend großem Stichprobenumfang vernachlässigbar.

Satz von Gauß-Markow

Satz 10.5 (Schätzer für Parameter der Regressionsgerade)
Die bei der linearen Regression (Satz 9.13) mit Hilfe der Methode der kleinsten Quadrate bestimmten Schätzer $\hat{a}$ und $\hat{b}$ für die Parameter der Regressionsgerade sind erwartungstreu.

Dies lässt sich direkt nachrechnen (siehe z. B. Sachs 2018).

Der bereits erwähnte Satz von Gauß-Markow zeigt, dass in einem linearen Modell die Methode der kleinsten Quadrate unter häufig erfüllten Voraussetzungen optimale erwartungstreue Schätzer für die Modellparameter liefert.

Satz 10.6 (Satz von Gauß-Markow)
In einem linearen Modell ist der Kleinste-Quadrate-Schätzer der beste (minimalvariante) lineare erwartungstreue Schätzer (engl. *best linear unbiased estimator*, *BLUE*), wenn die zufälligen Fehler der Beobachtungen

- unkorreliert sind,
- Erwartungswert null und
- die gleiche Varianz haben.

10.1.4. Intervallschätzung und Konfidenzintervall

Bisher haben wir ***Punktschätzungen*** der genauen Werte unbekannter Parameter betrachtet. Diese haben das Problem, dass nicht bekannt ist, wie nah ein bestimmter Schätzer an dem zu schätzenden Parameter liegt. Erwartungstreue und Konsistenz liefern nur statistische Aussagen.

Bei einer ***Intervallschätzung*** wird dagegen aus einer Stichprobe ein Intervall geschätzt, in dem der unbekannte Parameter mit einer vorgegebenen Wahrscheinlichkeit liegt.

Konfidenzintervall

Definition 10.7 (Konfidenzintervall, Konfidenzniveau)
Seien $X_1, \ldots, X_n$ eine Stichprobe vom Umfang n und θ ein zu schätzender Parameter der Grundgesamtheit.

Ein Intervall $[g_u(X_1, \ldots, X_n), g_o(X_1, \ldots, X_n)]$, dessen untere und obere Grenzen aus Stichprobenwerten berechnet werden, heißt ***Konfidenzintervall*** (auch ***Vertrauensintervall***, engl. *confidence interval*) zum Niveau $1-\alpha$, wenn es den Parameter θ mit der Wahrscheinlichkeit $1-\alpha$ überdeckt:

$$P(\theta \in [g_u(X_1, \ldots, X_n), g_o(X_1, \ldots, X_n)] = 1-\alpha\,.$$

$1-\alpha$ heißt ***Konfidenzniveau*** (auch ***Vertrauenswahrscheinlichkeit***, ***Sicherheit***, engl. *confidence level*).

Da der zu schätzende Parameter fest ist und das Konfidenzintervall von der Stichprobe abhängt, sagt man, dass „das Konfidenzintervall den Parameter überdeckt“, und nicht, dass „der Parameter in das Konfidenzintervall fällt“.

Erwartungswert eines normalverteilten Merkmals

Exemplarisch behandeln wir die Intervallschätzung des Erwartungswertes eines normalverteilten Merkmals mit bekannter Varianz. Im Fall unbekannter Varianz ist das Vorgehen ähnlich, nur muss statt der Normalverteilung die weiter unten in Abschnitt 10.3.1 eingeführte Student-t-Verteilung benutzt werden.

Die Schätzung normalverteilter Merkmale ist wichtig, da sie wegen des Zentralen Grenzwertsatzes bei ausreichend großem Stichprobenumfang eine Näherung für Schätzungen des Mittelwerts beliebig verteilter Merkmale darstellt; als Faustregel für die Gültigkeit dieser Näherung gilt ein Stichprobenumfang $n \gtrsim 30$.

Satz 10.8 (Erwartungswert eines normalverteilten Merkmals)
Sei $X_1, \ldots, X_n$ eine Stichprobe vom Umfang n eines $\mathrm{N}(\mu, \sigma^2)$-normalverteilten Merkmals mit bekannter Varianz σ^2 und unbekanntem Erwartungswert μ. Sei $\overline{X}$ das arithmetische Mittel der Stichprobe.

Dann ist ein Konfidenzintervall zum Niveau $1-\alpha$ gegeben durch

$$\left[\overline{X} - \frac{\tilde{z}_{1-\alpha/2}\,\sigma}{\sqrt{n}}, \overline{X} + \frac{\tilde{z}_{1-\alpha/2}\,\sigma}{\sqrt{n}}\right],$$

wobei $\tilde{z}_{1-\alpha/2}$ das $(1-\frac{\alpha}{2})$-Quantil der Standardnormalverteilung $\Phi(z)$ ist,

$$\Phi(\tilde{z}_{1-\alpha/2}) = 1 - \frac{\alpha}{2}.$$

Woher kommt das $(1-\frac{\alpha}{2})$-Quantil der Standardnormalverteilung $\Phi(z)$?

Für ein standardnormalverteiltes Merkmal Z suchen wir z, so dass

$$P(Z \in [-z, z]) = 1 - \alpha$$

Mit der Verteilungsfunktion $\Phi(z)$ erhalten wir:

$$\begin{aligned}
&\Leftrightarrow && \Phi(z) - \Phi(-z) = 1 - \alpha \\
&\Leftrightarrow && \Phi(z) - (1 - \Phi(z)) = 1 - \alpha \\
&\Leftrightarrow && 2\,\Phi(z) - 1 = 1 - \alpha \\
&\Leftrightarrow && \Phi(z) = \frac{2-\alpha}{2} = 1 - \frac{\alpha}{2}
\end{aligned}$$

Wir müssen die Tabelle der Standardnormalverteilung (Anhang A.1) also „rückwärts“, d. h. von innen nach außen lesen. Übliche Werte für $1-\alpha$ und das $(1-\frac{\alpha}{2})$-Quantil $\tilde{z}_{1-\alpha/2}$ sind (siehe Anhang A.2)

$1-\alpha$	0.90	0.95	0.99
$\tilde{z}_{1-\alpha/2}$	1.645	1.960	2.576

Beispiel 10.9 (Abfüllanlage)
Eine Abfüllanlage für 1-Liter-Flaschen habe eine bekannte Standardabweichung von 3 ml . Eine Stichprobe vom Umfang 50 liefere als Mittelwert des Flascheninhalts 999 ml .

Füllt die Anlage systematisch zu wenig ab, oder lässt sich der Wert durch Zufall erklären?

Wir beschreiben das Vorgehen Schritt für Schritt.

1. Wir wählen das Konfidenzniveau $1-\alpha = 0.95$, entsprechend $\alpha = 0.05$.
2. Aus der Tabelle in Anhang A.2 lesen wir das 0.975-Quantil $\tilde{z}_{1-\alpha/2} = 1.960$ ab.
3. Für eine Stichprobe vom Umfang $n = 50$ haben wir den Mittelwert $\bar{x} = 999$ berechnet.
4. Als Konfidenzintervall erhalten wir mit der bekannten Standardabweichung $\sigma = 3$

$$\left[999 - \frac{1.960 \cdot 3}{\sqrt{50}}, 999 + \frac{1.960 \cdot 3}{\sqrt{50}}\right] \approx [998.17, 999.83] .$$

Das Konfidenzintervall überdeckt den tatsächlichen Erwartungswert μ mit einer Wahrscheinlichkeit von 0.95 . Da der Sollwert $\mu_{\text{soll}} = 1000$ nicht in dem Konfidenzintervall enthalten ist, können wir mit 95 %-iger Sicherheit schließen, dass die Anlage zu wenig abfüllt.

Folgerung 10.10 (Erwartungswert eines normalverteilten Merkmals)
(i) Die Breite b des Konfidenzintervalls ergibt sich als

$$b = 2\,\frac{\tilde{z}_{1-\alpha/2}\,\sigma}{\sqrt{n}} .$$

Soll die Breite des Konfidenzintervalls $b \leq B$ sein, folgt für den Stichprobenumfang

$$n \geq \left(\frac{2\,\tilde{z}_{1-\alpha/2}\,\sigma}{B}\right)^2 .$$

(ii) Je kleiner α ist, um so größer sind $\tilde{z}_{1-\alpha/2}$ und damit die Breite b des Konfidenzintervalls.

Der erste Teil gibt die Mindestgröße einer Stichprobe für ein gewünschtes Konfidenzintervall an. Der zweite Teil sagt aus, dass das Konfidenzniveau nicht beliebig gewählt werden kann; im Extremfall $1-\alpha = 1$ ergibt sich das Konfidenzintervall $(-\infty, \infty)$. Dies ist anschaulich nachvollziehbar: Mit hundertprozentiger Sicherheit kann nur gesagt werden, dass die zu schätzende Größe (hier der Erwartungswert) irgendeinen reellen Wert besitzt.

10.2. Testen von Hypothesen

Mit Hilfe eines statistischen Tests wird untersucht, ob eine ***Hypothese*** (Vermutung) gültig ist. Dazu wird anhand einer Stichprobe geprüft, ob die gegenteilige ***Nullhypothese*** (die meistens dem „Normalzustand“ entspricht) gültig ist oder zugunsten der ursprünglichen Hypothese (der ***Alternative***) verworfen werden muss. Anwendungen finden sich z. B. in folgenden Zusammenhängen.

- Medizinische Tests: Hat eine Person Tuberkulose?
- Qualitätssicherung: Füllt eine Abfüllanlage die korrekte Menge Saft/Bier/. . . ab?
- Nutzerstudien:
  - Verringert eine neue Benutzeroberfläche die Häufigkeit von Fehlbedienungen?
  - Erhöht sie die Kundenzufriedenheit? Erhöht sie den Umsatz?

10.2.1. Nullhypothese, Teststatistik und Signifikanzniveau

Definition 10.11 (Nullhypothese, Alternative, Teststatistik, Signifikanzniveau)
Ein ***statistischer Test*** besteht aus

(i) einer ***Nullhypothese*** (engl. *null hypothesis*) H_0 und einer ***Alternative*** H_1, die sich gegenseitig ausschließen;

(ii) einer Stichprobe $X_1, \ldots, X_n$ vom Umfang n;

(iii) einer ***Teststatistik*** (auch ***Prüfgröße***) $T(X_1, \ldots, X_n)$, die sensibel für das Testproblem ist und deren Verteilung unter der Nullhypothese bekannt ist;

(iv) einem ***Signifikanzniveau*** (engl. *significance level*) α.

Der ***kritische Bereich*** (auch ***Verwerfungsbereich***, engl. *region of rejection*) des Tests besteht aus den Werten der Teststatistik T, die für die Alternative H_1 sprechen und die unter der Hypothese H_0 mit einer Wahrscheinlichkeit $\leq \alpha$ auftreten.

Das folgende Beispiel behandelt die gleiche Problemstellung wie das Beispiel 10.9 der Flaschenabfüllanlage. Es wird hier jedoch als statistischer Test formuliert.

Beispiel 10.12 (Schokoladenherstellung)
Eine Maschine zur Herstellung von 100-Gramm-Tafeln Schokolade habe eine bekannte Standardabweichung von 2 g. Eine Stichprobe vom Umfang 10 ergebe ein arithmetisches Mittel von 98.9 g.

Stellt die Maschine zu kleine Tafeln her oder ist der Wert durch den Zufall zu erklären?

Wir erläutern das Vorgehen wieder Schritt für Schritt.

1. Als Nullhypothese nehmen wir an, dass der Erwartungswert wie vorgesehen $\mu_0 = 100$ ist, als Alternative, dass er davon abweicht (***zweiseitiger Gauß-Test***),

$$H_0 : \ \mu = \mu_0 = 100\,, \quad H_1 : \ \mu \neq \mu_0\,.$$

2. Als Teststatistik (Prüfgröße) verwenden wir die standardisierte Abweichung des arithmetischen Mittels $\overline{X}$ der Stichprobe vom Sollwert μ_0,

$$T(X_1,\ldots,X_n) = Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}\,.$$

3. Als Signifikanzniveau wählen wir $\alpha = 0.05$.

4. Als Stichprobenumfang wählen wir $n = 10$.

5. Nach Festlegen der obigen Parameter erheben wir die Stichprobe $X_1,\ldots,X_n$. Als konkreten Wert der Prüfgröße Z erhalten wir in unserem Beispiel

$$z = \frac{98.8 - 100.0}{2.0/\sqrt{10}} \approx -1.897\,.$$

6. Der kritische Bereich besteht aus den Werten der Prüfgröße, die großen Abweichungen vom unter der Nullhypothese angenommenen Erwartungswert $\mu_0 = 100$ entsprechen und mit der Wahrscheinlichkeit

$$P(|Z| > c) = \alpha$$

auftreten. Analog zur Berechnung des Konfidenzintervalls entspricht c dem Quantil $\tilde{z}_{1-\alpha/2}$ der Standardnormalverteilung (siehe Anhang A.2):

$$c = \tilde{z}_{1-\alpha/2} = \tilde{z}_{0.975} \approx 1.960\,.$$

7. Entscheidungsregel: Die Nullhypothese ist zu verwerfen, wenn für den beobachteten Wert z der Prüfgröße gilt

$$|z| > \tilde{z}_{1-\alpha/2}\,.$$

Für unser Beispiel erhalten wir

$$|z| \approx 1.897 \leq 1.960 = \tilde{z}_{1-\alpha/2}\,.$$

Da der Wert der Prüfgröße außerhalb des kritischen Bereichs liegt, behalten wir die Nullhypothese bei, dass die Maschine in Ordnung ist. Läge der Wert innerhalb des kritischen Bereichs, müssten wir die Nullhypothese zugunsten der Alternative verwerfen.

Entscheidend ist, anders als im vorliegenden Beispiel zuerst den Test zu planen, insbesondere das Signifikanzniveau festzulegen, und erst danach die Stichprobe zu erheben. Bei umgekehrtem Vorgehen ist die Versuchung groß, das Signifikanzniveau anhand der erhobenen Daten an das gewünschte Ergebnis anzupassen.

Definition 10.13 (parametrischer Test, einfach/zusammengesetzt, zwei-/einseitig)
Ein statistischer Test, der Aussagen über einen Parameter θ eines Merkmals der Grundgesamtheit macht, heißt ***parametrischer Test***.

Eine Hypothese eines parametrischen Tests heißt

- ***einfach***, wenn Sie nur aus einem Parameterwert besteht, z. B. $\theta = \theta_0$;
- ***zusammengesetzt***, wenn Sie aus mehreren Parameterwerten besteht, z. B. $\theta \neq \theta_0$ oder $\theta \geq \theta_0$.

Ein Test heißt

- ***zweiseitig***, wenn er Abweichungen vom Sollwert in beide Richtungen prüft, z. B. $\theta = \theta_0$ vs. $\theta \neq \theta_0$;
- ***einseitig***, wenn er Abweichungen vom Sollwert in eine Richtung prüft, z. B. $\theta \geq \theta_0$ vs. $\theta < \theta_0$.

Nicht-parametrische Tests werden z. B. verwendet, um zu entscheiden, ob sich ein Problem durch ein bestimmtes Modell beschreiben lässt.

Für einseitige Tests ändert sich die Berechnung des kritischen Bereichs. Ob ein einseitiger oder zweiseitiger Test zu verwenden ist, hängt von der Fragestellung ab: Während einer Schokoladenfirma daran gelegen ist, dass möglichst genau 100 g abgefüllt werden (zweiseitiger Test), möchte eine Verbraucherschutzorganisation möglicherweise nur sicherstellen, dass nicht weniger als 100 g enthalten sind (einseitiger Test).

Entscheidend für den Nutzen eines Tests ist die Wahl einer passenden Teststatistik. Gebräuchliche Tests (mit zugehörigen Teststatistiken) für verschiedene Problemstellungen sind der Student-t-Test, der F-Test und der χ^2-(Chi-Quadrat-)Test. Auf Einzelheiten kann im Rahmen der Vorlesung nicht eingegangen werden.

10.2.2. p-Wert

Während der kritische Bereich zu einem gegebenen Signifikanzniveau im Voraus berechnet werden und für mehrere Stichproben verwendet werden kann, lässt sich andersherum für eine gegebene Beobachtung der Prüfgröße fragen, wie wahrscheinlich diese unter der angenommenen Verteilung ist.

Dazu greifen wir noch einmal das Beispiel 10.12 mit einer etwas anderen Fragestellung auf.

Beispiel : Schokoladenherstellung

Eine Maschine zur Herstellung von 100-Gramm-Tafeln Schokolade habe eine bekannte Standardabweichung von 2 g. Eine Stichprobe vom Umfang 10 ergebe ein arithmetisches Mittel von 98.9 g .

Ab welchem Signifikanzniveau α müsste die Nullhypothese (korrektes Tafelgewicht) bei diesem Ergebnis verworfen werden?

Die Nullhypothese wird verworfen, wenn gilt

$$|z| > \tilde{z}_{1-\alpha/2}\,.$$

Wir suchen also den Wert α des Signifikanzniveaus mit

$$\tilde{z}_{1-\alpha/2} = |z| \approx 1.897\,,$$

der als ***p*-Wert** (engl. *p-value*) $p = \alpha$ bezeichnet wird.

In der Tabelle der Standardnormalverteilung (Anhang A.1) lesen wir ab

$$1 - p/2 \approx \Phi(1.90) \approx 0.9713 \quad \Rightarrow \quad p \approx 0.0574\,.$$

Falls im Voraus ein Signifikanzniveau $\alpha > p \approx 0.0574$ gewählt worden wäre, müsste die Nullhypothese verworfen werden.

Definition 10.14
Der ***p*-Wert** gibt die Wahrscheinlichkeit an, unter gültiger Nullhypothese H_0 eine zufällige Stichprobe zu erhalten, welche den beobachteten Prüfwert oder einen in Richtung der Alternative H_1 extremeren Wert aufweist.

Entscheidungsregel: H_0 wird verworfen, falls der p-Wert kleiner als das zuvor gewählte Signifikanzniveau α ist.

Der p-Wert bietet im Gegensatz zum Prüfwert eine genauere und vor allem vergleichbare Einschätzung, wie deutlich oder knapp die Nullhypothese verworfen bzw. beibehalten wird. Umso wichtiger ist es auch hier, das Signifikanzniveau im Voraus festzulegen.

Ein häufiges Missverständnis ist, den p-Wert als Wahrscheinlichkeit zu deuten, mit der die Nullhypothese angesichts der gegebenen Stichprobe gültig ist. Stattdessen gibt der p-Wert die Wahrscheinlichkeit an, unter der Nullhypothese eine zufällige Stichprobe zu erhalten, welche dieselben Werte der betrachteten statistischen Kenngrößen (im obigen Beispiel des arithmetischen Mittels) aufweist wie die gegebene Stichprobe.

10.2.3. Fehler 1. und 2. Art

Wir greifen nochmals die bereits in Abschnitt 5.4.2 eingeführten Fehler 1. und 2. Art auf.

Definition 10.15 (Fehler 1. und 2. Art)
Bei einem statistischen Test können zwei Arten von Fehlern auftreten.

Bei einem ***Fehler 1. Art*** (falsch positiv) wird die Nullhypothese H_0 irrtümlich verworfen, obwohl sie wahr ist.

Bei einem ***Fehler 2. Art*** (falsch negativ) wird die Nullhypothese H_0 irrtümlich beibehalten, obwohl sie falsch ist.

	H_0 wird beibehalten	H_0 wird verworfen
H_0 wahr	okay	Fehler 1. Art (falsch positiv)
H_1 wahr	Fehler 2. Art (falsch negativ)	okay

Die Nullhypothese beschreibt meistens dem „Normalzustand“. Bei medizinischen Tests (z. B. auf Krankheiten) entspricht dies der Aussage „Patient/in ist gesund“. Fehler 1. Art werden in dem Zusammenhang als *falsch positiv* bezeichnet (Krankheit wird erkannt, obwohl Patient/in gesund ist), Fehler 2. Art als *falsch negativ* (Krankheit wird nicht erkannt, obwohl Patient/in krank ist).

Die Wahrscheinlichkeit für einen Fehler 1. Art wird vor dem Test durch die Wahl des Signifikanzniveaus α nach oben beschränkt. Die Wahrscheinlichkeit für einen Fehler 2. Art hängt dagegen von der Wahl der Teststatistik ab und ist entsprechend schwieriger zu kontrollieren.

Des Weiteren ist es i. A. nicht möglich, Fehler 1. und 2. Art gleichzeitig zu minimieren. Als Beispiel sei ein Feuermelder genannt, der gleichzeitig jede noch so kleine Rauchentwicklung erkennen und niemals einen Fehlalarm auslösen soll.

10.3. Nutzerstudien

In diesem Abschnitt wird ein kurzer Überblick über Nutzerstudien von Software-Anwendungen gegeben. Im Detail werden Nutzerstudien in der Lehrveranstaltung *Usability* des Studiengangs Mediendesigninformatik behandelt.

Evaluierungsmethoden für Software-Anwendungen und Nutzeroberflächen:

Qualitativ

- Befragung von Nutzer*innen zur Zufriedenheit o. ä.
- Aufdecken von Design-Fehlern
- Methode: Fragebögen, z. B. ***System Usability Scale***

Quantitativ

- Messung von Bearbeitungszeiten o. ä.
- Vergleich zweier Design-Entwürfe
- Methode: statistische Tests, z. B. ***t-Test***

Aufteilung von Testnutzer*innen (Proband*innen) beim Vergleich mehrerer Entwürfe (z. B. A/B Testing):

Between-Group Design

- Jede Person testet nur einen Entwurf (gehört zu einer Gruppe).
- Vorteil: keine Beeinflussung durch anderen Entwurf (*single blind/double blind*)
- Nachteil: höhere Zahl benötigter Proband*innen

Within-Subject Design

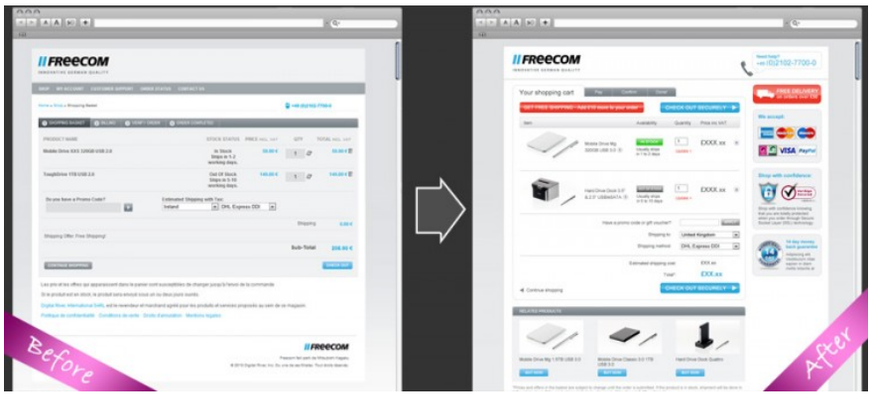
- Jede Person testet beide/alle Entwürfe
- Vorteil: geringere Zahl benötigter Proband*innen, Vergleichbarkeit
- Nachteil: mögliche Beeinflussung durch anderen Entwurf

10.3.1. Beispiel: Bewertung eines Web-Interfaces

Im Folgenden soll beispielhaft der Ablauf einer Nutzerstudie zur Bewertung einer neuen Nutzeroberfläche eines Online-Shops vorgestellt werden. Dabei lernen wir den häufig verwendeten ***t-Test*** (auch ***Student-t-Test***) kennen.

Beispiel : Bewertung eines Web-Interfaces

Ein Online-Shop möchte durch ein klarer gestaltetes Web-Interface die Bestelldauer verringern und damit die Kundenzufriedenheit erhöhen. In einer Nutzerstudie soll getestet werden, ob das Ziel mit einem gegebenen Entwurf erreicht wird.



Bildquelle: http://bexwhitedesign.com/web-design-portfolio/before-and-after-shopping-cart-optimisation-design/

Wir erläutern das Vorgehen wieder Schritt für Schritt.

1. Es sei bekannt, dass eine Bestellung von zwei Artikeln über das alte Web-Interface im Mittel $4.5\,\text{min} = 270\,\text{s}$ dauert.

   Da sich die Standardabweichung nicht zwingend von dem alten auf das neue Interface übertragen lässt, müssen wir diese als empirische Standardabweichung s_x aus der Stichprobe schätzen. In diesem Fall (unbekannte Standardabweichung) benötigen wir den ***t-Test*** (auch ***Student-t-Test***).

2. Als Nullhypothese nehmen wir an, dass der Erwartungswert der Bestelldauer für das neue Interface nicht kleiner als der für das alte Interface ist, als Alternative, dass er kleiner ist (***einseitiger t-Test***),

$$H_0 : \ \mu \geq \mu_0 = 270\,, \quad H_1 : \ \mu < \mu_0\,.$$

3. Als Teststatistik (Prüfgröße) verwenden wir die gemäß der ***t*-Verteilung** standardisierte Abweichung des arithmetischen Mittels $\bar{x}$ der Stichprobe vom Sollwert μ_0,

$$T(x_1,\ldots,x_n) = t = \frac{\bar{x}-\mu_0}{s_x/\sqrt{n}}\,.$$

4. Wie sind Signifikanzniveau α und Stichprobenumfang n (teilnehmende Personen) zu wählen?

   Empfehlung von JAKOB NIELSEN (Nielsen Norman Group, 2006):

   - Signifikanzniveau $\alpha = 0.10$
   - Stichprobenumfang $n = 20$

   Quelle: https://www.nngroup.com/articles/quantitative-studies-how-many-users/

5. Nach Festlegen der obigen Parameter erheben wir die Stichprobe $x_1,\ldots,x_{20}$ und berechnen $\bar{x} = 240$ und $s_x = 72$.

   Für die Prüfgröße erhalten wir

$$t = \frac{240-270}{72/\sqrt{20}} \approx -1.863\,.$$

6. Der kritische Bereich besteht aus den Werten der Prüfgröße, die großen negativen Abweichungen vom unter der Nullhypothese angenommenen Erwartungswert $\mu_0 = 270$ entsprechen und mit der Wahrscheinlichkeit

$$P(t < c) = \alpha$$

   auftreten. Analog zur Berechnung des Konfidenzniveaus entspricht c dem negativen Quantil $-t_{n-1,1-\alpha}$ der ***t*-Verteilung mit $n-1$ Freiheitsgraden** (siehe Anhang A.3),

$$c = -t_{n-1,1-\alpha} = -t_{19,0.9} \approx -1.328\,.$$

7. Entscheidungsregel: Die Nullhypothese ist zu verwerfen, wenn gilt

$$t < -t_{n-1,1-\alpha}\,.$$

   Für unser Beispiel erhalten wir

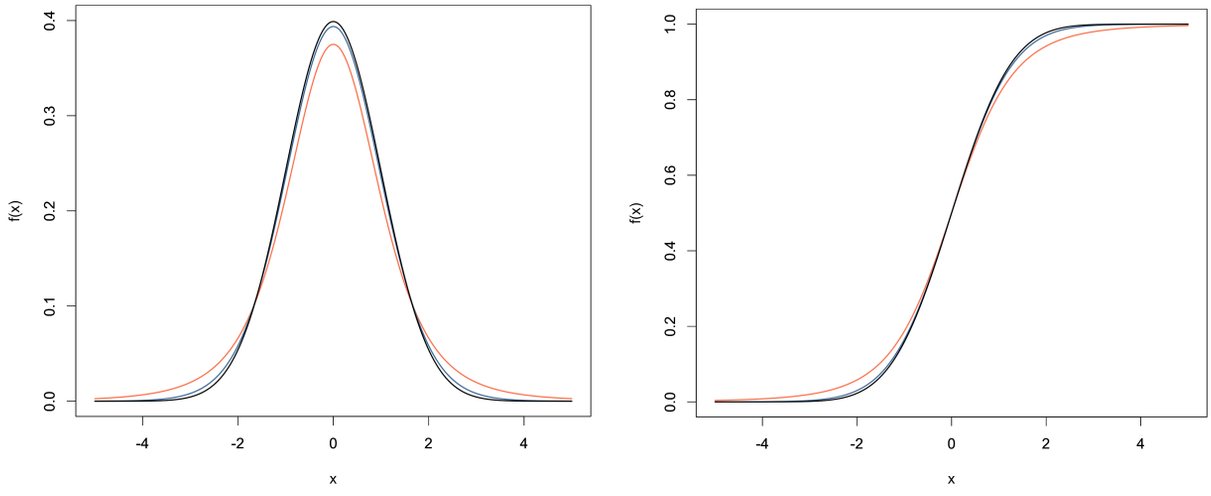
$$t \approx -1.834 < -1.328 = -t_{n-1,1-\alpha}\,.$$

Da der Wert der Prüfgröße innerhalb des kritischen Bereichs liegt, verwerfen wir die Nullhypothese und stellen fest, dass die Bestelldauer für das neue Web-Interface geringer als für das alte ist.

10.3.2. t-Verteilung

Die ***t-Verteilung*** wurde 1908 von einem Mitarbeiter der Guinnes-Brauerei in Dublin entwickelt (nach früheren Vorabeiten) und unter dem Pseudonym STUDENT veröffentlicht. Daher wird sie auch als ***Student-t-Verteilung*** bezeichnet. Sie kommt immer dann zum Einsatz, wenn die Standardabweichung nicht bekannt und der Stichprobenumfang klein ist; über die ***Freiheitsgrade*** geht der Stichprobenumfang (minus 1) direkt als Parameter in die Verteilung ein. Für wachsenden Stichprobenumfang nähert die ***t-Verteilung*** sich der Standardnormalverteilung an.

Dichte (links) und Verteilungsfunktion (rechts) der t-Verteilung mit 4 (rot) bzw. 19 (blau) Freiheitsgraden im Vergleich mit der Standardnormalverteilung (schwarz).



10.4. Schätzen und Testen in R

Die Sprache R bietet umfangreiche Funktionen zum Schätzen und Testen: Die Schließende Statistik ist das Hauptanwendungsgebiet der Software. Auf Einzelheiten kann hier nicht eingegangen werden, stattdessen wird auf die R-Literatur verwiesen. Einen guten Einstieg mit Erläuterungen zu den statistischen Grundlagen verschiedener Tests und vielen Beispielen bietet z. B. das RRZN-Handbuch Hain 2011.

A.3. p-Quantile der t-Verteilung

$t_{m,1-p}$ mit m Freiheitsgraden

$m \setminus p$	0.9	0.95	0.975	0.99	0.995	0.999
1	3.078	6.314	12.71	31.82	63.66	318.3
2	1.886	2.920	4.303	6.965	9.925	22.33
3	1.638	2.353	3.182	4.541	5.841	10.21
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.610
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.787	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
31	1.309	1.696	2.040	2.453	2.744	3.375
32	1.309	1.694	2.037	2.449	2.738	3.365
33	1.308	1.692	2.035	2.445	2.733	3.356
34	1.307	1.691	2.032	2.441	2.728	3.348
35	1.306	1.690	2.030	2.438	2.724	3.340
36	1.306	1.688	2.028	2.434	2.719	3.333
37	1.305	1.687	2.026	2.431	2.715	3.326
38	1.304	1.686	2.024	2.429	2.712	3.319
39	1.304	1.685	2.023	2.426	2.708	3.313

Symmetrie: $t_{m,1-p} = -t_{m,p}$.