7 Induktive Statistik: Schlüsse ziehen

Da wir bei der Datenanalyse in der Regel Stichproben vorliegen haben, gibt es keine gesicherten Aussagen über die Parameter des theoretische Modells, das wir dem beobachteten Phänomen zugrunde legen, also über die so genannte «Grundgesamtheit». Dennoch wollen wir Aussagen über die Grundgesamtheit tätigen, oder zumindest solche, die für eine größere Anzahl von Stichproben zutreffen. Wir werden dazu im abschließenden Kapitel das Prinzip *statistischer Tests* anschauen, das sind Tests, bei denen wir *Hypothesen* über die Parameter der Grundgesamtheit oder über die Ähnlichkeit zweier Stichproben aufstellen, und die Gültigkeit dieser Hypothesen auch so gut es geht überprüfen.

7.1 Prinzip statistischer Tests

Zunächst einmal einige Beispiele dafür, was wir mit statistischen Tests überprüfen können:

- Eine Imbisskette wirbt damit, dass in ihren Semmeln mindestens 130g Leberkäse enthalten sind. Einige Kunden vermuten aber, dass die Stücke viel kleiner sind. Zehn von ihnen wägen ihren Leberkäse nach. Es ergibt sich, dass *im Durchschnitt* dieser Stichprobe eine Portion Leberkäse nur 129.4g wiegt. Ist das nur ein stichprobenbedingter Zufall? Oder steckt da Methode dahinter, und die Stücke sind zu klein?

- Jemand möchte für eine bestimmte Entscheidung wissen, ob sich die mittlere Jahrestemperatur in Wiener Neustadt von jener in Villach unterscheidet.

- ▷ Ich möchte feststellen, ob sich bei Hamstern in zwei nebeneinanderliegenden Käfigen das Risiko einer kontaktlosen Übertragung des Arenavirus durch das Tragen von FFP2-Schutzmasken verringern lässt.

Es geht im Folgenden also darum, entweder zwei Stichproben miteinander zu vergleichen, oder eine Stichprobe mit der ihr zugrunde liegenden Grundgesamtheit. Für diese Vergleiche können wir die jeweiligen Parameter – in unseren Fällen meist Mittelwerte – heranziehen und sie mit Hilfe so genannter **Hypothesentests** überprüfen (manchmal auch als *Signifikanztest*[1] bezeichnet).

Ausgangspunkt ist dabei zunächst eine bestimmte Frage, wie zum Beispiel: «*Diese Stichprobe ergibt einen Mittelwert von 129.4 g Leberkäse pro Semmel. Erwartet hätte ich 130 g. Ist das nur zufällig oder hat auch die zugehörige Grundgesamtheit einen Erwartungswert ungleich 130 g?*»

Wir treffen dann eine Annahme, die wir **Arbeitshypothese** nennen, und die die Antwort auf die oben gestellte Frage als Behauptung formuliert. Will ich zum Beispiel wissen, ob eine Grundgesamtheit einen Erwartungswert ungleich $130g$ hat, so könnte meine Arbeitshypothese lauten: $H_A : \mu \neq 130$.

Für andere Fragestellungen benötige ich entsprechend andere Hypothesen. Dabei gibt es mathematisch-formal drei Arten von Arbeitshypothesen[2]:

Fall 1a: $H_A : \mu < 130$
Fall 1b: $H_A : \mu > 130$
Fall 2 : $H_A : \mu \neq 130$

(wobei 130 nur ein Beispiel ist und stattdessen auch ein anderer Wert stehen kann.)

Ziel des Hypothesentests ist es, eine gewählte Hypothese zu akzeptieren oder zu verwerfen. Unsere Entscheidung ist abhängig davon, ob die in der Stichprobe beobachteten Werte eher unwahrscheinlich sind, sollte die Hypothese wahr sein. Wir überprüfen also ein *Modell* (die Grundgesamtheit) anhand der *Daten* aus der Stichprobe (mitunter auch aus mehreren Stichproben): Solange Modell und Daten konsistent sind, gibt es keinen guten Grund, die Hypothese nicht zu akzeptieren. Passen sie aber nicht gut zueinander[3], dann lehnen wir sie ab.

Wenn sich unsere Daten «hypothesenkonform» zeigen und wir die Hypothese akzeptieren, heißt das aber nicht, das wir irgendwas «beweisen» konnten. Tat-

[1] vom lat. *significanter* = klar, deutlich
[2] Beachte: In unseren Fällen haben Arbeitshypothesen nie ein Gleichheitszeichen. Die stehen alle in den *Nullhypothesen*, zu denen wir gleich kommen.
[3] Was «*nicht gut zueinander passen*» bedeutet, müssen wir noch definieren…

sächlich lässt sich mit Stichproben gar nichts *beweisen*. Wenn ein Experiment mit den theoretischen Voraussagen übereinstimmt, heißt das noch nicht, dass die Theorie richtig ist. Es könnte ja auch eine andere, uns unbekannte Theorie zu diesen Ergebnissen geführt haben.

Theorien lassen sich allerdings durch ein einziges negatives Experiment widerlegen[4]. Daher gehen wir folgendermaßen vor: Zu jeder Arbeitshypothese formulieren wir noch eine zweite Hypothese, die genau das Gegenteil behauptet. Wir nennen das die **Nullhypothese**. Stellt sich dann heraus, dass die Nullhypothese nicht zutrifft, können wir daraus schließen, dass die Arbeitshypothese richtig sein muss – also genau, was wir insgeheim ohnehin zeigen wollten.

Bei Hypothesentests wird also immer eine Arbeitshypothese H_A aufgestellt, und dann die zugehörige Nullhypothese H_0 getestet. Wenn im Zuge des Tests anhand einer (oder mehrerer) Stichproben H_0 verworfen wird, können wir H_A akzeptieren.

Arbeitshypothese und Nullhypothese: Nicht jedes Ding hat zwei Seiten

Bei der Wahl der Hypothesen müssen wir unterscheiden, ob uns die Abweichungen des getesteten Parameters nach oben und unten gleich wichtig sind oder nur in eine Richtung interessieren.

Hypothesen der Form[5]

$$H_A : \mu \neq 100 \tag{7.1}$$

mit der Nullhypothese

$$H_0 : \mu = 100 \tag{7.2}$$

sind so genannte *zweiseitige Fragestellungen*. Die Abweichungen des Erwartungswertes μ von 100 sind nach oben oder unten gleich wichtig, d.h. alle abweichenden Parameterwerte, seien sie größer als 100 oder kleiner, bringen die Nullhypothese zu Fall.

[4] Von *Karl Popper* (1902-1994) stammt dazu folgendes berühmte Beispiel: Nimm an, du wolltest die Theorie prüfen «Alle Raben sind schwarz». Du beobachtest 100 Raben und stellst tatsächlich fest, dass jeder Rabe schwarz ist. Ist mit diesem Ergebnis die Theorie bewiesen? Popper sagt: Es könnte auch sein, dass der 101. Rabe, den man irgendwo beobachtet, weiß ist, und die Theorie «Alle Raben sind schwarz» wäre mit einem Schlag widerlegt.

[5] Auch hier ist der Wert 100 nur ein Beispiel und an seiner Stelle kann auch ein beliebig anderer konkreter Wert stehen. Und auch das μ kann durch einen anderen Parameter ersetzt werden.

Umgekehrt sind Hypothesentests der Form

$$H_A: \quad \mu < 100 \tag{7.3}$$
$$H_0: \quad \mu \geq 100 \tag{7.4}$$

bzw.

$$H_A: \quad \mu > 100 \tag{7.5}$$
$$H_0: \quad \mu \leq 100 \tag{7.6}$$

einseitige Fragestellungen, d.h. nur die Abweichung in eine Richtung ist interessant. Testen wir zum Beispiel ein bestimmtes Qualitätsmerkmal, so bedeutet die Unterschreitung eines vorgegebenen Sollwertes eine «schlechte» Qualität und das Ausscheiden des untersuchten Merkmalsträgers. Die Überschreitung hingegen hat in dem Fall keine negativen Folgen für die Qualität und ist daher OK.

Verspricht zum Beispiel der Hersteller einer Batterie eine Lebensdauer von «100 Lichtstunden» für die Verwendung in einer bestimmten Taschenlampe, so testen wir die Nullhypothese $H_0 : \mu \geq 100$ gegen die Arbeitshypothese $H_A : \mu < 100$ (einseitiger Test) und nicht $H_0 : \mu = 100$ gegen $H_A : \mu \neq 100$ (zweiseitiger Test). Aus Konsument:innensicht heißt ja «100 Lichtstunden» *mindestens* 100 Stunden, wir sind aber mit 110 oder 130 Stunden auch zufrieden.

Betrachten wir die Abfüllanlage einer Molkerei, die in jede Packung 1 Liter Milch einfüllen soll, so werden die Konsument:innen gegebenenfalls ebenfalls eine einseitige Fragestellung testen, die Molkerei hingegen wird einen zweiseitigen Test durchführen, weil aus ihrer Sicht auch eine Abweichung nach oben (zuviel Milch) negative Konsequenzen hat.

Die möglichen Formen von Nullhypothesen und Arbeitshypothesen für einseitige und zweiseitige Hypothesentests des Erwartungswertes sind in Tab. 7.1 zusammengefasst. (Sie entsprechen den Formeln (7.1) bis (7.6), wobei der konkrete Wert 100 durch die Variable m_0 ersetzt wurde).

Fehler erster und zweiter Art

Wir hoffen natürlich, dass wir uns mit unseren Stichproben ein gutes Spiegelbild der Grundgesamtheit beschafft haben. Trotzdem: Egal wie unsere Entscheidung

H_A	H_0	Art der Fragestellung
$\mu < m_0$	$\mu \geq m_0$	einseitig
$\mu > m_0$	$\mu \leq m_0$	einseitig
$\mu \neq m_0$	$\mu = m_0$	zweiseitig

Tabelle 7.1: Arbeitshypothesen und Nullhypothesen bei ein- bzw. zweiseitigen Hypothesentests des Erwartungswertes, wobei m_0 für eine beliebige konkrete Zahl steht

bezüglich der Nullhypothese ausfällt, es verbleibt immer eine gewisse Unsicherheit. Diese Unsicherheit hängt damit zusammen, dass wir unsere Schlüsse aus einer Stichprobe schließen. Hätten wir eine andere Stichprobe «erwischt», würde das Ergebnis vielleicht ein wenig anders aussehen. Letztlich hängt die Unsicherheit unserer Entscheidung also vom Zufall ab. Damit können wir ihr aber eine Wahrscheinlichkeit zuordnen: Wir nennen sie die **Irrtumswahrscheinlichkeit** α (auch: das *Signifikanzniveau*). α ist die Wahrscheinlichkeit dafür, dass bei einem Hypothesentest die Nullhypothese H_0 abgelehnt wird, obwohl sie wahr ist. Wir nennen dies auch einen *Fehler erster Art* (siehe Tab. 7.2).

Üblicherweise[6] wählen wir für $\alpha = 0.05$. Eine Irrtumswahrscheinlichkeit von $\alpha = 0.05$ bedeutet: Wenn wir den Hypothesentest häufig durchführen, so werden wir in 5 von 100 Fällen die Nullhypothese irrtümlich ablehnen.

Die Gegenwahrscheinlichkeit $(1 - \alpha)$ heißt auch **Sicherheitswahrscheinlichkeit**. Sie gibt an, mit welcher Wahrscheinlichkeit wir eine richtige Nullhypothese als solche erkennen und nicht ablehnen.

Wir können bei einem Hypothesentest auch den Fehler begehen, eine falsche Nullhypothese nicht abzulehnen. Dies nennen wir einen *Fehler zweiter Art* und ordnen ihm die Wahrscheinlichkeit β zu.

Die Gegenwahrscheinlichkeit $(1 - \beta)$ ist die *«Teststärke»* (auch: *Macht des Testes*). Sie gibt an, mit welcher Wahrscheinlichkeit eine falsche Nullhypothese tatsächlich als solche entlarvt und abgelehnt wird. Es ist also die Wahrscheinlichkeit, einen Fehler zweiter Art zu verhindern. (Auch hier gilt wieder: Das ist nicht dasselbe wie die Frage, wie wahrscheinlich es ist, dass die Nullhypothese falsch *ist*).

Tabelle 7.2 fasst dies noch einmal zusammen.

[6] 1931 beschrieb *Ronald Fisher* (1890-1962) in seinem Buch *The Design of Experiments*, dass für viele wissenschaftliche Experimente ein α von 0.05 («1 aus 20») ein angemessener Wert für das Signifikanzniveau sei. Seitdem wurde dieser Wert von vielen Disziplinen ohne weiteres Hinterfragen übernommen. – Wir werden es ebenso tun.

Die Sicherheitswahrscheinlichkeit gibt an, mit welcher Wahrscheinlichkeit wir eine richtige Nullhypothese auch als solche erkennen – was aber nicht gleichbedeutend ist mit der Wahrscheinlichkeit, dass die Nullhypothese richtig *ist*.

Das mag beim ersten Durchlesen verwirrend klingen; vielleicht hilft folgendes Beispiel: Angenommen, du erhältst eine Mail in deinen Posteingang, bei der im Header «vera.wormser@heidelberg.com» als Absenderin angegeben ist. Tatsächlich hattest du vor 18 Jahren einmal in Heidelberg eine Freundin namens Vera Wormser. Es könnte sich aber auch um einen Fall von Mail-Spoofing handeln.

Bevor du jetzt was Unüberlegtes tust, könntest du darüber nachdenken, wie groß die Wahrscheinlichkeit ist, dass sich Vera wirklich bei dir meldet und wenn dir diese Wahrscheinlichkeit hoch genug erscheint, die Mail öffnen.

Oder du schätzt ab, wie groß die Wahrscheinlichkeit ist, dass du an verschiedenen Kriterien und Hinweisen erkennen würdest, ob die Mail tatsächlich von Vera stammt und danach entscheiden, die Mail zu öffnen (oder zu löschen…). Und diese Wahrscheinlichkeit kann ganz anders eingeschätzt werden als die oben angegebene.

Annahme oder Verwerfen der Hypothese

Nach welchen Kriterium entscheiden wir jetzt, ob wir eine Hypothese annehmen oder ablehnen?

Für die Durchführung des Hypothesentests benötigen wir eine *Testfunktion* und Kenntnisse[7] über deren Verteilung unter der Annahme, dass H_0 zutrifft. Wir

[7] *Wir* benötigen diese Kenntnisse zum Glück nicht sehr genau sondern verlassen uns auf die Vorgangsweisen, die Mathematiker:innen und Statistiker:innen in der Vergangenheit gefunden haben.

	H_0 ist richtig	H_0 ist falsch
H_0 annehmen H_A verwerfen	richtige Entscheidung $P = (1 - \alpha) =$ *Sicherheitswahrscheinlichkeit*	Fehler 2. Art H_0 annehmen, obwohl H_A gilt: $P = \beta$
H_0 verwerfen H_A annehmen	Fehler 1. Art H_A annehmen, obwohl H_0 gilt: $P = \alpha =$ *Irrtumswahrscheinlichkeit*	richtige Entscheidung $P = (1 - \beta) =$ *Teststärke*

Tabelle 7.2: Entscheidungsmöglichkeiten und zugehörige Wahrscheinlichkeiten bei einem statistischen Hypothesentest

nennen die Testfunktion allgemein $F(\mathbf{X})$.

Für eine konkrete Stichprobe können wir eine *Realisierung* von $F(\mathbf{X})$ bestimmen, d.h. eine konkrete *Prüfgröße* f ausrechnen. Mit dieser Prüfgröße sind wir nun in der Lage, die Nullhypothese zu beurteilen. Dazu müssen wir zuvor noch ein Intervall dergestalt bestimmen, dass der Wert f mit einer Wahrscheinlichkeit von $(1 - \alpha)$ in diesem Intervall enthalten ist. Die Grenzen dieses Intervalls nennen wir die *Sicherheitsgrenzen*; der Bereich, der außerhalb dieses Intervalls liegt, führt zur Ablehnung von H_0 und wir bezeichnen ihn als *kritischen Bereich*, auch: *Verwerfungsbereich*.

Liegt die Prüfgröße f innerhalb der Sicherheitsgrenzen, so wird die Nullhypothese H_0 angenommen, weil ihr die vorliegenden Stichprobendaten nicht widersprechen. Liegt die Prüfgröße im kritischen Bereich, so verwerfen wir H_0 und akzeptieren die Arbeitshypothese H_A.

Arten statistischer Hypothesen

Parameterhypothesen sind Annahmen über einen (unbekannten) Parameter des Modells wie zum Beispiel in Tabelle 7.1 für solche über den Erwartungswert. (Siehe S.148)

Unterschiedshypothesen beziehen sich auf den Unterschied zweier Stichproben und vergleichen zum Beispiel die aus ihnen erhaltenen Mittelwerte. (Siehe S.154). Wir sprechen dabei auch von Tests für *unabhängige Stichproben*.

Veränderungshypothesen sind Unterschiedshypothesen sehr ähnlich, nur vergleichen sie Daten, die aus der Messung oder Beobachtung derselben Merkmalsträger zu verschiedenen Zeitpunkten. Verglichen wird dabei zum Beispiel ein «Vorher-Zustand» mit einem «Nachher-Zustand» (z.B. $H_0 : \mu_{vorher} = \mu_{nachher}$ gegen $H_A : \mu_{vorher} \neq \mu_{nachher}$). Nachdem wir hier dieselben Merkmalsträger (zweimal) untersuchen, sprechen wir auch von Tests für *verbundene Stichproben*. (Siehe S.156)

Zusammenhangshypothesen postulieren einen Zusammenhang zwischen zwei Zufallsvariablen und testen dann zum Beispiel, ob der Korrelationskoeffizient des Modells gleich Null ist oder nicht, also[8] $H_0 : \rho = 0$ gegen $H_A : \rho \neq 0$, siehe auch S.157.

[8] Den empirischen Korrelationskoeffizienten haben wir in Formel 4.7 angegeben. Für das theoretische Pendant verwenden wir wieder – wie üblich – den entsprechenden griechischen Buchstaben, hier ein «Rho» ρ.

Parameter-, Unterschieds-, Veränderungs- und Zusammenhangshypothesen können sich auf einen einzigen Wert beziehen und werden dann auch *Punkthypothesen* oder **ungerichtete Hypothesen** genannt, oder auf einen ganzen Bereich, so genannte *Bereichshypothesen* oder **gerichtete Hypothesen**[9].

Und dann gibt es noch **Verteilungshypothesen**, das sind Annahmen über die Form der Verteilung des Modells (= der Grundgesamtheit), die wir auf Grund der Verteilungsform der Stichprobe aufstellen. Zum Beispiel könnten wir die Hypothese aufstellen, dass bestimmte Merkmale *normalverteilt* sind. Wir nennen Tests, die Verteilungshypothesen prüfen, **Anpassungstests**. Sie sind aber nicht Gegenstand dieser einführenden Lehrveranstaltung (oder dieser Unterlagen).

Bedenke auch, wenn du zum Beispiel in deiner Abschlussarbeit statistische Hypothesen aufstellst und überprüfst, dass du zuvor je nach zu verwendendem Test vielleicht prüfen musst, ob gleiche Varianzen der Stichproben vorliegen (so genannte *Varianzhomogenität*; getestet werden kann das mit einem *Levene-Test*) oder dass die verwendeten Daten normalverteilt sind – was man übrigens manchmal auch ausreichend durch einen graphischen Vergleich der Dichtefunktion mit einer Glockenkurve (siehe Abb.5.9 auf Seite 120) herausfinden kann. Mathematisch-statistisch genauere Tests auf Normalverteilungen sind z.B. der *Kolmogorov-Smirnov Test* (auch: *KS-Test*), ein *Shapiro-Wilk Test* oder ein *Anderson-Darling Test*). Für konkrete Beispiele dafür sei auf weiterführende Literatur verwiesen.

7.2 Testen von Parameterhypothesen

t-Test eines Mittelwerts aus normalverteilten Daten einer kleinen Stichprobe

Wenn wir zwar eine kleine Stichprobe haben, aber von einer Normalverteilung der Daten ausgehen, dann können wir den Mittelwert dieser Stichprobe mit einem so genannten **t-Test** testen. Das *t* bezieht sich auf den Namen der Wahrscheinlichkeitsverteilung, die wir bei diesem Test verwenden: Die *Student-* oder *t*-Verteilung. Wie sie aussieht, können wir in Abb.5.7 (Seite 118) rechts oben feststellen. Und wie man ihre Funktionswerte berechnet, wissen wir auch schon: Sie ist bereits in Formel 6.1 und 6.2 vorgekommen und auf Seite 131 ihre Berechnung in *Excel* oder *R* angegeben.

Beim t-Test (genauer: beim **einfachen t-Test**) eines Mittelwerts wollen wir über-

[9] Der Unterschied ist ziemlich einfach: Wenn im Hypothesenpaar H_0, H_A nur die Vergleichssymbole $=$ und $\neq$ vorkommen, handelt es sich um *ungerichtete* Hypothesen; wenn die Ungleichheitszeichen $\leq$ und $>$ bzw. $\geq$ und $<$ vorkommen, um *gerichtete*.

prüfen, ob der unbekannte Erwartungswert μ einer normalverteilten Zufallsvariablen X einen bestimmten Wert m_0 besitzt bzw. über- oder unterschreitet. m_0 kann zum Beispiel ein Sollwert bei der Herstellung eines Produkts sein. Dabei kennen wir für μ und σ nur Schätzwerte, nämlich das arithmetische Mittel $\bar{x}$ und die empirische Standardabweichung s.

Beispiel 33 *Als einfaches Beispiel können wir die Herstellung von Brotlaiben betrachten. Deren (in kg gemessene) Masse X sei normalverteilt. Das angegebene Verkaufsgewicht des Brotes sei $\mu = 2$ kg. Eine Konsumentenschutzorganisation zieht nun eine Stichprobe von $n = 20$ Brotlaiben und stellt einen Stichprobenmittelwert von $\bar{x} = 1.97$ kg und eine empirische Standardabweichung von $s = 0.1$ kg fest. Es soll überprüft werden, ob diese Stichprobe gegen die Hypothese spricht, dass die Brote der Grundgesamtheit mindestens 2 kg wiegen – dass wir also vom Verkäufer nicht bedackelt werden.*

Zunächst sind Arbeits- und Nullhypothese festzulegen (vgl. auch Tab.7.1):

- Für die *einseitige* Fragestellung lauten sie (je nachdem, welche Richtung für uns interessant ist):
  Fall 1a: $H_A : \mu < m_0 \rightarrow H_0 : \mu \geq m_0$ oder
  Fall 1b: $H_A : \mu > m_0 \rightarrow H_0 : \mu \leq m_0$

- Für eine *zweiseitige* Fragestellung:
  Fall 2: $H_A : \mu \neq m_0 \rightarrow H_0 : \mu = m_0$

Beispiel 33 *(Fortsetzung)*

Im konkreten Beispiel geht es um eine einseitige Fragestellung, weil wir ja nur unzufrieden sind, wenn das Brot weniger als 2 kg wiegt. Wir wollen also herausfinden: Deutet der Mittelwert $\bar{x} = 1.97$, den wir aus der Stichprobe erhalten haben, auf eine Grundgesamtheit hin, deren Erwartungswert μ signifikant kleiner als $m_0 = 2$ ist?

Wir wählen als Arbeitshypothese bzw. als Nullhypothese (entsprechend Fall 1a):

$$H_A : \mu < 2$$

$$H_0 : \mu \geq 2$$

Anschließend ist eine Irrtumswahrscheinlichkeit festzulegen. Wir werden den üblichen Wert von $\alpha = 0.05$ wählen.

Als Testfunktion ziehen wir folgende Funktion heran:

$$F(\mathbf{x}) = \frac{\bar{x} - m_0}{s}\sqrt{n} \tag{7.7}$$

Beispiel 33 *(Fortsetzung)*

Aus der Realisierung der Stichprobe unseres Beispiels können wir die konkrete Prüfgröße angeben:

$$f = \frac{\bar{x} - m_0}{s}\sqrt{n} = \frac{1.97 - 2}{0.1}\sqrt{20} = -1.34$$

Nun bestimmen wir den kritischen Bereich. Unter H_0 besitzt die Funktion eine *t-Verteilung* mit $(n-1)$ Freiheitsgraden. Die entsprechenden Werte für den kritischen Bereich erhalten wir aus einer Tabelle oder einem Programm, und dann können wir eine Entscheidung treffen: Die Nullhypothese wird abgelehnt, falls die Testgröße im kritischen Bereich liegt, andernfalls wird H_0 akzeptiert.

H_A	H_0	Prüfgröße	Entscheidung
$\mu < m_0$	$\mu \geq m_0$	$f < -t_{(1-\alpha,n-1)}$	H_0 ablehnen, H_A akzeptieren
		$f \geq -t_{(1-\alpha,n-1)}$	H_0 akzeptieren, H_A ablehnen
$\mu > m_0$	$\mu \leq m_0$	$f > t_{(1-\alpha,n-1)}$	H_0 ablehnen, H_A akzeptieren
		$f \leq t_{(1-\alpha,n-1)}$	H_0 akzeptieren, H_A ablehnen
$\mu \neq m_0$	$\mu = m_0$	$\lvert f\rvert > t_{(1-\alpha/2,n-1)}$	H_0 ablehnen, H_A akzeptieren
		$\lvert f\rvert \leq t_{(1-\alpha/2,n-1)}$	H_0 akzeptieren, H_A ablehnen

Tabelle 7.3: Mögliche Ergebnisse eines t-Tests. Den Wert für f berechnen wir aus Formel 7.7, den Wert für t entnehmen wir aus einer Tabelle oder eine Programm. Beachte: Bei zweiseitiger Fragestellung wird t für $(1-\alpha/2)$ berechnet, bei einseitiger Fragestellung hingegen für $(1-\alpha)$. Die Anzahl der Freiheitsgrade ist immer $n-1$.

In *MS Excel* heißt der Befehl zur Berechnung von t für eine einseitige Fragestellung `=T.INV(1-alpha;n-1)`, wobei für `alpha` und `n-1` die jeweiligen Werte für α und n einzusetzen sind. In *R* lautet der Befehl `qt((1-alpha),n-1)`.

Für eine zweiseitige Fragestellung müssen wir eingeben: `=T.INV(1-alpha/2;n-1)` bzw. `qt((1-alpha/2),n-1)`.

Achtung auf einseitige und zweiseitige Fragestellungen

Beispiel 33 *(Fortsetzung)*

In unserem Beispiel benötigen wir also noch den t-Wert an der Stelle $(0.95, 19)$*. Das Quantil der t-Verteilung ist (laut Excel oder R) an dieser Stelle gleich* 1.729*. Nach Tab.7.3 benötigen wir den negativen Wert davon, also* $-t = -1.729$*.*

Die Prüfgröße f *haben wir schon ausgerechnet; jetzt können wir vergleichen:*

Da $-1.3 > -1.7$*, also* $f > -t$*, werden wir* ***die Nullhypothese annehmen*** *und können* ***die Arbeitshypothese nicht mehr aufrechterhalten****.*

Das bedeutet: Die in der Stichprobe beobachtete mittlere Masse von 1.97 *ist zwar kleiner als der Sollwert* 2 *kg, diese Abweichung ist allerdings statistisch nicht signifikant sondern vermutlich zufällig bedingt. Die Wahrscheinlichkeit, aus einer Grundgesamtheit mit* $\mu = 2$ *eine Stichprobe mit einem Mittelwert von höchstens* 1.97 *zu erhalten, ist größer als* 5%*. Es gibt daher – aus Sicht der Statistik – keinen Grund, das angegebene Verkaufsgewicht von* 2 *kg zu beanstanden.*

An dieser Stelle noch ein Hinweis zum Aufstellen der Hypothesen: Gehen wir, so wie im eben gerechneten Beispiel, von einer gerichteten Hypothese aus, und schon der empirische Wert aus der Stichprobe geht in die andere Richtung, muss man schon sehr gut argumentieren (können), warum man dennoch eine gegenteilige Arbeitshypothese aufstellt. Hätten wir also aus der «Brot-Stichprobe» einen Mittelwert von $\bar{x} = 2.1\ kg$ erhalten, spricht nicht viel dafür, eine Arbeitshypothese der Form $H_A : \mu < 2$ aufzustellen. Bei einer zweiseitigen Fragestellung geht das aber nicht: Hier können wir nicht einfach Null- und Arbeitshypothese umdrehen. Wie schon in der Fußnote auf Seite 142 beschrieben: Das Gleichheitszeichen steht immer bei der Nullhypothese, nie bei der Arbeitshypothese.

Gaußtest eines Mittelwerts aus Daten einer großen Stichprobe

Wir sind beim t-Test davon ausgegangen, dass die Daten normalverteilt sind. Wenn wir darüber nicht sicher sind, aber der Stichprobenumfang n größer als 30 ist, dann können wir einen ähnlichen Test anwenden, den so genannten *Gaußtest* bzw. streng genommen den **approximativen Gaußtest**[10].

Wenn wir den Mittelwert testen, ist die Testfunktion beim Gaußtest dieselbe wie beim t-Test des Mittelwerts (Formel 7.7). Der Vollständigkeit halber schreiben

[10] benannt nach Johann Friedrich Carl Gauß, den wir schon in Fußnote 12 auf Seite 119 kennengelernt haben.

H_A	H_0	Prüfgröße	Entscheidung
$\mu < m_0$	$\mu \geq m_0$	$f < -z_{(1-\alpha)}$	H_0 ablehnen, H_A akzeptieren
		$f \geq -z_{(1-\alpha)}$	H_0 akzeptieren, H_A ablehnen
$\mu > m_0$	$\mu \leq m_0$	$f > z_{(1-\alpha)}$	H_0 ablehnen, H_A akzeptieren
		$f \leq z_{(1-\alpha)}$	H_0 akzeptieren, H_A ablehnen
$\mu \neq m_0$	$\mu = m_0$	$\lvert f \rvert > z_{(1-\alpha/2)}$	H_0 ablehnen, H_A akzeptieren
		$\lvert f \rvert \leq z_{(1-\alpha/2)}$	H_0 akzeptieren, H_A ablehnen

Tabelle 7.4: Mögliche Ergebnisse eines approximativen Gaußtests. Den Wert für f berechnen wir aus Formel 7.8, den Wert für z entnehmen wir aus einer Tabelle oder eine Programm.

wir sie hier nochmal auf:

$$F(\mathbf{x}) = \frac{\bar{x} - m_0}{s}\sqrt{n} \tag{7.8}$$

Den kritischen Bereich erhalten wir jetzt aber nicht aus der t-Verteilung, sondern aus einer *Standardnormalverteilung* und treffen die Entscheidung über die Annahme der Nullhypothese nach Tab.7.4. Die dafür benötigten Werte z können wir wieder aus einem Programm entnehmen.

In *MS Excel* heißt der Befehl zur Berechnung von z für eine einseitige Fragestellung `=NORM.S.INV(1-alpha)`. In *R* lautet der Befehl `qnorm(1-alpha)`.

Für eine zweiseitige Fragestellung müssen wir eingeben: `=NORM.S.INV(1-alpha/2)` bzw. `qnorm(1-alpha/2)`.

Aufgabe 28 Deine Ärztin empfiehlt dir, aus gesundheitlichen Gründen deinen Kaffeekonsum auf fünf Tassen pro Tag einzuschränken. Du bist dir eigentlich sicher, dass du das im Schnitt ohnehin nicht überschreitest ($H_0 : \mu \leq 5$), dein besorgter Ehepartner schenkt dem aber keinen so rechten Glauben und behauptet, dass es mehr als fünf Tassen pro Tag sind ($H_A : \mu > 5$). Ihr vereinbart, eine Zeitlang darüber Buch zu führen. Nach 40 Tagen hast du 210 Tassen Kaffee getrunken. Im Schnitt ergab das Experiment also 5,25 Tassen pro Tag, und das bei einer Standardabweichung von 0,25. Wessen Hypothese kann bei diesem Ergebnis aufrecht erhalten werden?

Test eines Anteilswertes

Sowohl der einfache t-Test als auch der approximative Gaußtest kann angewendet werden, wenn wir einen *Anteilswert* p testen wollen. Anteilswerte haben wir bereits auf Seite 136 untersucht – dort ging es um Konfidenzintervalle für Anteilswerte[11]. Konkret verwenden wir einen t-Test, wenn $n < 30$ und einen Gaußtest wenn $n \geq 30$.

Am Beginn beider Tests steht wie üblich die Formulierung der Hypothesen. Das Schema möglicher Hypothesenpaare kennen wir ja mittlerweile bereits; wir können einen der folgenden drei Fälle behandeln:

Fall 1a:	$H_A : p < p_0$	$\rightarrow$	$H_0 : p \geq p_0$
Fall 1b:	$H_A : p > p_0$	$\rightarrow$	$H_0 : p \leq p_0$
Fall 2 :	$H_A : p \neq p_0$	$\rightarrow$	$H_0 : p = p_0$

Die Testfunktion für den Hypothesentest des Anteilswertes unterscheidet sich ein wenig von jener für den Mittelwert. Wenn $\hat{p}$ der Anteilswert ist, den wir aus der Stichprobe erhalten haben, dann lautet sie:

$$F(\mathbf{x}) = \frac{\hat{p} - p_0}{\sqrt{p_0 \, (1 - p_0)}} \cdot \sqrt{n} \tag{7.9}$$

Nachdem wir den konkreten Funktionswert f ausgerechnet haben, ziehen wir für den Vergleich wieder entweder die Quantile der t-Verteilung (wenn $n - 1 \leq 30$) oder jene der Normalverteilung heran (wenn $n > 30$), siehe Tab.7.5.

Aufgabe 29 Eine Software-Entwicklerin überlegt, ihre Software als *Donationware* zur Verfügung stellen, d.h. sie kann grundsätzlich kostenlos verwendet werden, sie bittet aber um (freie) Spenden, damit auf Sicht wenigstens die bei ihr entstehenden Drittkosten abgedeckt sind. Sie schätzt: Nur wenn mindestens 25% der User bereit sind, 10 € zu spenden, werde ich kein Geld verlieren.
Sie startet einmal probehalber und stellt nach den ersten 500 Downloads fest: 145 User haben tatsächlich 10 € gespendet. Das sind sogar 29%. Kann sie (aus statistischer Sicht) optimistisch sein, dass der Anteil der spendenfreudigen User tatsächlich größer als 25% ist?

[11] Und wie auf Seite 138 angegeben gilt auch hier: In aller Strenge gelten die Formeln dieses Abschnitts nur, wenn $n \cdot \hat{p} > 5$ und $n \cdot (1 - \hat{p}) > 5$.

H_A	H_0	Prüfgröße wenn $n < 30$	Prüfgröße wenn $n \geq 30$	Entscheidung
$p < p_0$	$p \geq p_0$	$f < -t_{(1-\alpha,n-1)}$	$f < -z_{(1-\alpha)}$	H_0 ablehnen
		$f \geq -t_{(1-\alpha,n-1)}$	$f \geq -z_{(1-\alpha)}$	H_0 akzeptieren
$p > p_0$	$p \leq p_0$	$f > t_{(1-\alpha,n-1)}$	$f > z_{(1-\alpha)}$	H_0 ablehnen
		$f \leq t_{(1-\alpha,n-1)}$	$f \leq z_{(1-\alpha)}$	H_0 akzeptieren
$p \neq p_0$	$p = p_0$	$\|f\| > z_{(1-\alpha/2)}$	$\|f\| > z_{(1-\alpha/2)}$	H_0 ablehnen
		$\|f\| \leq t_{(1-\alpha/2,n-1)}$	$\|f\| \leq z_{(1-\alpha/2)}$	H_0 akzeptieren

Tabelle 7.5: Entscheidungsalternativen beim Test eines Anteilswertes. Dabei gilt: Wenn H_0 abgelehnt wird, kann H_A angenommen werden.
($n < 30$: t-Test. $n \geq 30$: Gaußtest)

7.3 Testen von Unterschiedshypothesen

In diesem Abschnitt geht es um das Testen von Hypothesen über zwei unabhängige Stichproben.

Vergleich zweier Mittelwerte mittels Gaußtest bei großen Stichproben

Wenn wir zwei Stichproben mit Stichprobenumfang n_1 bzw. n_2 (die jeweils ≥ 30 sind) und den Mittelwerten $\bar{x}_1$ und $\bar{x}_2$ erhoben haben, können wir die Differenz $\bar{x}_1 - \bar{x}_2$ bilden und für den Fall, dass diese Differenz ungleich Null ist, untersuchen, ob es auch eine von Null verschiedene Differenz der beiden Erwartungswerte der jeweils zugrundeliegenden Modelle $\mu_1 - \mu_2$ gibt, de facto also zwei unterschiedliche Modelle vorliegen, aus denen die jeweiligen Stichproben stammen.

Die Arbeitshypothesen können dann lauten:

Fall 1a: $H_A : \mu_1 - \mu_2 < 0$
Fall 1b: $H_A : \mu_1 - \mu_2 > 0$
Fall 2 : $H_A : \mu_1 - \mu_2 \neq 0$

wobei wir das meist umformen zu:

Fall 1a: $H_A : \mu_1 < \mu_2$
Fall 1b: $H_A : \mu_1 > \mu_2$
Fall 2 : $H_A : \mu_1 \neq \mu_2$

Zum Beispiel könnten wir das Einkommen der Angestellten in unserem Unter-

H_A	H_0	Prüfgröße	Entscheidung
$\mu_1 < \mu_2$	$\mu_1 \geq \mu_2$	$f < -z_{(1-\alpha)}$	H_0 ablehnen, H_A akzeptieren
		$f \geq -z_{(1-\alpha)}$	H_0 akzeptieren, H_A ablehnen
$\mu_1 > \mu_2$	$\mu_1 \leq \mu_2$	$f > z_{(1-\alpha)}$	H_0 ablehnen, H_A akzeptieren
		$f \leq z_{(1-\alpha)}$	H_0 akzeptieren, H_A ablehnen
$\mu_1 \neq \mu_2$	$\mu_1 = \mu_2$	$\lvert f \rvert > z_{(1-\alpha/2)}$	H_0 ablehnen, H_A akzeptieren
		$\lvert f \rvert \leq z_{(1-\alpha/2)}$	H_0 akzeptieren, H_A ablehnen

Tabelle 7.6: Entscheidungsalternativen beim Gaußtest zweier Mittelwerte aus Stichproben mit $n \geq 30$

nehmen untersuchen und dabei unter den Frauen und den Männer jeweils eine Stichprobe ziehen. Die Mittelwerte werden da vermutlich nicht genau gleich sein, aber ist dieser Unterschied nur zufällig und wir haben im Grunde eine Equal Pay-Situation im Unternehmen, oder ist der Unterschied so signifikant, dass wir von einem Gender-Pay-Gap sprechen müssen?

Die formale Vorgangsweise beim Testen zweier Mittelwerte läuft «wie immer» ab – mittlerweile kennen wir das ja schon:

Wir entscheiden, welche Arbeits- und Nullhypothese wir aufstellen:

$$\begin{array}{lllll} \text{Fall 1a:} & H_A : \mu_1 < \mu_2 & \rightarrow & H_0 : \mu_1 \geq \mu_2 \\ \text{Fall 1b:} & H_A : \mu_1 > \mu_2 & \rightarrow & H_0 : \mu_1 \leq \mu_2 \\ \text{Fall 2 :} & H_A : \mu_1 \neq \mu_2 & \rightarrow & H_0 : \mu_1 = \mu_2 \end{array}$$

Ist zum Beispiel $\bar{x}_1$ um einiges größer als $\bar{x}_2$, dann gilt vermutlich auch $\mu_1 > \mu_2$; ist die Differenz $\bar{x}_1 - \bar{x}_2 \approx 0$, dann testen wir eher Fall 2.

Die Testfunktion lautet:

$$F(\mathbf{x}) = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2 n_2 + s_2^2 n_1}} \sqrt{n_1 n_2} \tag{7.10}$$

Und das Ergebnis können wir nach dem Schema in Tabelle 7.6 finden.

Hinweis: Auch für den Zweistichprobentest gibt es wieder eine t-Test Variante, wenn wir nur kleine Stichproben vorliegen haben. Die Formel für die Testfunktion ist dann um einiges komplexer als wir uns in dieser Einführungslehrveranstaltung «zumuten» wollen…

Aufgabe 30 Nach dem ersten Studienjahr wurde für alle Studierenden eines Jahrgangs ein nach ECTS gewichteter Notenschnitt (GPA) berechnet und daraus dann ein Durchschnittswert für jeweils alle männlichen ($n_m = 84$) und alle weiblichen ($n_w = 36$) Studierenden angegeben. Für die männlichen Studierenden beträgt er $\bar{x}_m = 1.6$, für die weiblichen $\bar{x}_w = 1.4$. Ermittelt wurden auch die zugehörigen empirischen Standardabweichungen: $s_m = 0.6$, $s_w = 0.4$.

Der empirisch ermittelte GPA der Frauen unterscheidet sich offenbar von dem der Männer. Ist dieser Unterschied signifikant bzw. inwiefern lässt sich aus den obigen Daten die Hypothese $H_A : \mu_m - \mu_w \neq 0$ behaupten?

Vergleich zweier Anteilswerte

Fall 1a:	$H_A : p_1 < p_2$	$\rightarrow$	$H_0 : p_1 \geq p_2$
Fall 1b:	$H_A : p_1 > p_2$	$\rightarrow$	$H_0 : p_1 \leq p_2$
Fall 2 :	$H_A : p_1 \neq p_2$	$\rightarrow$	$H_0 : p_1 = p_2$

Testfunktion:

$$F(\mathbf{x}) = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \tag{7.11}$$

Entscheidung:

H_A	H_0	Prüfgröße	Entscheidung
$p_1 < p_2$	$p_1 \geq p_2$	$f < -z_{(1-\alpha)}$	H_0 ablehnen, H_A akzeptieren
		$f \geq -z_{(1-\alpha)}$	H_0 akzeptieren, H_A ablehnen
$p_1 > p_2$	$p_1 \leq p_2$	$f > z_{(1-\alpha)}$	H_0 ablehnen, H_A akzeptieren
		$f \leq z_{(1-\alpha)}$	H_0 akzeptieren, H_A ablehnen
$p_1 \neq p_2$	$p_1 = p_2$	$\lvert f \rvert > z_{(1-\alpha/2)}$	H_0 ablehnen, H_A akzeptieren
		$\lvert f \rvert \leq z_{(1-\alpha/2)}$	H_0 akzeptieren, H_A ablehnen

Tabelle 7.7: Entscheidungsalternativen beim Vergleich zweier Anteilswerte

7.4 Testen von Veränderungshypothesen

Dabei handelt es sich um das Testen von Parametern von zwei *verbundenen* Stichproben. Wir könnten zum Beispiel von einer Gruppe von Personen am Faschingdienstag das Gewicht ermitteln und als Stichprobe vorhalten. 40 Tage später ermitteln wir dann das Gewicht *derselben* Personen erneut, das ist dann die mit

der ersten *verbundene* Stichprobe. Uns interessiert jetzt zum Beispiel, ob die Differenz der beiden Stichprobenmittelwerte gleich Null ist. Wir sprechen daher für diese Art von Hypotesentest auch von einem **Differenzentest**.

Die Vorgangsweise ist relativ einfach: Aus den beiden Stichproben mit den jeweils untersuchten Zufallsvariablen X bzw. Y bilden wir die paarweisen Differenzen

$$d_i = x_i - y_i \tag{7.12}$$

und danach aus allen d_i den arithmetischen Mittelwert $\bar{d}$ und die Standardabweichung s_d.

Unsere neue Zufallsvariable $D = X - Y$ hat – wie jede andere Zufallsvariable – einen Erwartungswert μ und wir können verschiedene Hypothesen über μ aufstellen:

Fall 1a: $H_A : \mu < 0 \rightarrow H_0 : \mu \geq 0$
Fall 1b: $H_A : \mu > 0 \rightarrow H_0 : \mu \leq 0$
Fall 2: $H_A : \mu \neq 0 \rightarrow H_0 : \mu = 0$

Als Testfunktion ziehen wir folgende Funktion heran:

$$F(\mathbf{x}) = \frac{\bar{d}}{s_d}\sqrt{n} \tag{7.13}$$

Je nachdem, ob wie eine große oder kleine Stichprobe vorliegen haben, ist die weitere Vorgangsweise dieselbe wie beim *t-Test* eines Mittelwerts aus normalverteilten Daten einer kleinen Stichprobe (S. 148) oder eines approximativen Gaußtests (S. 151).

7.5 Testen von Zusammenhangshypothesen

Test einer Korrelationshypothese

Fall 1a:	$H_A : \rho < 0$	$\rightarrow$	$H_0 : \rho \geq 0$
Fall 1b:	$H_A : \rho > 0$	$\rightarrow$	$H_0 : \rho \leq 0$
Fall 2 :	$H_A : \rho \neq 0$	$\rightarrow$	$H_0 : \rho = 0$

Testfunktion:

$$F(\mathbf{x}) = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}} \tag{7.14}$$

Entscheidung:

H_A	H_0	Prüfgröße	Entscheidung
$\rho < 0$	$\rho \geq 0$	$f < -t_{(1-\alpha,n-2)}$	H_0 ablehnen, H_A akzeptieren
		$f \geq -t_{(1-\alpha,n-2)}$	H_0 akzeptieren, H_A ablehnen
$\rho > 0$	$\rho \leq 0$	$f > t_{(1-\alpha,n-2)}$	H_0 ablehnen, H_A akzeptieren
		$f \leq t_{(1-\alpha,n-2)}$	H_0 akzeptieren, H_A ablehnen
$\rho \neq 0$	$\rho = 0$	$\lvert f \rvert > t_{(1-\alpha/2,n-2)}$	H_0 ablehnen, H_A akzeptieren
		$\lvert f \rvert \leq t_{(1-\alpha/2,n-2)}$	H_0 akzeptieren, H_A ablehnen

Tabelle 7.8: Entscheidungsalternativen beim Testen einer Korrelation

7.6 Abschließende Hinweise

Zunächst sei noch einmal auf die richtige Reihenfolge beim Hypothesentest verwiesen:

1. Man stellt eine Arbeitshypothese und eine Nullhypothese auf
2. Man gibt die Irrtumswahrscheinlichkeit vor und bestimmt damit einen Ablehnungsbereich
3. *Danach* wird die Stichprobe gezogen
4. Dann wird der Hypothesentest durchgeführt und entweder die Nullhypothese oder die Arbeitshypothese angenommen

Völlig unzulässig ist es, zuerst die Stichprobe zu ziehen, in den Stichprobendaten dann verschiedene Hypothesen auszuprobieren – womöglich unter mehrfacher, abwechslungsreicher Wahl von α , und dann diejenige auszuwählen, die am Besten zu meinen Daten «passt». Statistische Tests dürfen nie so ablaufen, dass die eigentliche Fragestellung erst nach der Beobachtung der Stichprobe aufgestellt wird!

Wir haben in diesem Kapitel «Einstichprobentests» (= eine Stichprobe wird gegen eine Grundgesamtheit getestet) und «Zweistichprobentests» (= jeweils zwei Stichproben werden miteinander verglichen) angesehen. Wollen wir eine Hypothese über mehr als zwei Stichproben testen, so verwenden wir eine **einfache Varianzanalyse**. Dazu sei auf weiterführende Literatur verwiesen.

Zum Begriff «Signifikanz»:

Die Trennung zwischen *signifikant* und *nicht signifikant* (also: *nicht-zufällig* versus *zufällig*) an einer bestimmten, scharfen Grenze festzumachen ist zugegebenermaßen eine etwas vereinfachte Sicht auf die Welt. Während es in technischen Anwendungen wie der Qualitätskontrolle in Produktionsprozessen vielleicht noch nachvollziehbar ist, dass man ab einem bestimmten zahlenmäßigen Wert zum Beispiel eine Toleranzgrenze überschritten hat, ist das bei der Untersuchung menschlichen Verhaltens vielleicht nicht immer ganz argumentierbar. Menschliches Verhalten ist ziemlich komplex und Menschen sind manchmal ziemlich kompliziert, und eine einfache Schwarz-Weiß-Einteilung wird dem vielleicht nicht immer gerecht. Aktuell gibt es aber kein «kontinuierliches» Signifikanzmaß, das zum Beispiel in abgestufter Form die «Plausibilität für die Zufälligkeit» angibt oder ein Intervall, in dem die Daten sowohl mit der Idee kompatibel sind, dass sie zufällig so zustandegekommen sind als auch mit der Idee, dass hier eine Signifikanz vorliegt.