Kapitel 6

Schließende Statistik

Die bisher betrachteten Wahrscheinlichkeitsverteilungen folgten aus Modellen, etwa dem des Binomialexperiments oder des Ziehens aus einer Urne; wir haben daher die Verteilung als bekannt angenommen. In der Praxis ist die Zielrichtung meist umgekehrt: Man kennt die Verteilung nicht, hat aber eine Stichprobe und möchte daraus Aussagen über die unbekannte Verteilung gewinnen. Solche Probleme sollen jetzt untersucht werden.

6.1 Mittelwert, empirische Varianz und Punktschätzung

6.1.1 Erwartungswert und Varianz von Summen unabhängiger Zufallsvariablen

In diesem Kapitel wird die Unabhängigkeit sehr wichtig werden. Insbesondere wird stets vorausgesetzt, dass Elemente einer Stichprobe voneinander unabhängig sind. Es wird dann immer wieder benutzt werden, dass Erwartungswert[1] bzw. Varianz einer Summe unabhängiger Zufallsvariablen gleich der Summe der Erwartunswerte bzw. der Varianzen der einzelnen Zufallsvariablen ist. Genauer: Für unabhängige Zufallsvariablen $X_1, \ldots, X_n$ gilt

$$\mathbb{E}\Big(\sum_{i=1}^{n} X_i\Big) = \sum_{i=1}^{n} \mathbb{E}(X_i)\,, \quad \mathbb{V}\Big(\sum_{i=1}^{n} X_i\Big) = \sum_{i=1}^{n} \mathbb{V}(X_i)\,.$$

6.1.2 Eigenschaften des Mittelwerts

Gegeben sei eine Stichprobe $(x_1, x_2, \ldots, x_n)$. Die Verteilung der zugrundeliegenden Zufallsvariablen X kennen wir nicht, insbesondere sind der Erwartungswert[2] μ und die Varianz σ^2 unbekannt und sollen aus der Stichprobe geschätzt werden.

[1] Für den Erwartungswert gilt die Additivität sogar ganz allgemein. Wir werden sie aber nur im Zusammenhang mit unabhängigen Zufallsvariablen benutzen.

[2] Der Erwartungswert hieß im ersten Teil der Vorlesung durchweg $\underline{m}$, ab jetzt nennen wir ihn μ, wie es in der Statistik üblich ist.

Der uns schon bekannte Stichprobenmittelwert

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i =: \widehat{\mu}$$

ist ein Näherungswert für μ; in der Sprache der Statistik spricht man von einem *Schätzwert* oder *Schätzer* (der Hut $\widehat{\ }$ ist zu lesen als "Schätzer für").

Im Gegensatz zu μ hängt $\widehat{\mu}$ von der zufällig gewählten Stichprobe ab. Um diese Abhängigkeit herauszuarbeiten, betrachten wir zwei Möglichkeiten, die Stichprobe $(x_1, x_2, \ldots, x_n)$ zu interpretieren:

i) als n Realisierungen der Zufallsvariablen X (z.B. Augenzahl beim einfachen Würfelexperiment)

ii) als *eine* Realisierung der Zufallsvariablen $(X_1, \ldots, X_n)$, der sogenannten *mathematischen Stichprobe*. Dabei ist X_i das Ergebnis des i-ten (Teil-) Experiments (z.B. die Augenzahl beim i-ten Wurf in einem n-fachen Würfelexperiment). Die X_i sind unabhängig und *identisch verteilt*. Letzteres heißt, sie haben alle dieselbe Wahrscheinlichkeitsverteilung (nämlich die der Größe X aus i). "Unabhängig und identisch verteilt" wird als i.i.d. ("independent and identically distributed") abgekürzt.

Mit der zweiten Interpretation ist $\overline{x}$ eine Realisierung der Zufallsvariablen

$$\overline{X} := \frac{1}{n}\sum_i X_i .$$

Ihr Erwartungswert ist gleich dem Erwartungswert von X, denn

$$\begin{aligned}
\mathbb{E}(\overline{X}) &= \mathbb{E}\left(\frac{1}{n}\sum_i X_i\right) = \frac{1}{n}\mathbb{E}\left(\sum_i X_i\right) = \frac{1}{n}\sum_i \mathbb{E}(X_i) \\
&= \frac{1}{n}\sum_i \mu = \frac{1}{n} n\,\mu = \mu\,, \qquad (6.1)
\end{aligned}$$

wobei wir die Rechenregeln für den Erwartungswert, die Additivität des Erwartungswerts und die identische Verteilung der X_i verwendet haben.

Den Erwartungswert $\mathbb{E}(\overline{X})$ kann man — salopp ausgedruückt — als Mittelwert über unendlich viele Stichprobenmittelwerte verstehen. Im Mittel liefert also der Stichprobenmittelwert den "wahren" Erwartungswert μ der zugrundeliegenden Verteilung, man bezeichnet ihn deshalb als **erwartungstreuen** (oder auch **unverzerrten**) Schätzer des Erwartungswerts. Allgemein heißt eine Zufallsvariable U erwartungstreuer Schätzer für eine Größe u, wenn $\mathbb{E}(U) = u$.

Für die Varianz des Mittelwertes erhält man

$$\begin{aligned}
\mathbb{V}(\overline{X}) &= \mathbb{V}\left(\frac{1}{n}\sum_i X_i\right) = \frac{1}{n^2}\mathbb{V}\left(\sum_i X_i\right) = \frac{1}{n^2}\sum_i \mathbb{V}(X_i) \\
&= \frac{1}{n^2}\sum_i \sigma^2 = \frac{1}{n^2} n\,\sigma^2 = \frac{1}{n}\sigma^2\,, \qquad (6.2)
\end{aligned}$$

wobei wir die Rechenregeln für die Varianz, die Additivität der Varianz bei unabhängigen Zufallsvariablen, sowie wieder die identische Verteilung der X_i verwendet haben.

Die Varianz geht also mit wachsendem Stichprobenumfang gegen 0; man sagt, der Mittelwert ist ein **konsistenter** Schätzer für den Erwartungswert. Allgemein heißt ein Schätzer U konsistent, wenn $\lim_{n\to\infty} \mathbb{V}(U) = 0$. Eigenschaften (6.1) und (6.2) zusammen besagen, dass die Schätzung mit zunehmendem Stichprobenumfang immer "besser" wird in dem Sinne, dass sie den wahren Wert immer besser trifft.

6.1.3 Zentraler Grenzwertsatz und Normalapproximation

Im vorigen Abschnitt haben wir den Mittelwert als Zufallsvariable betrachtet und deren Erwartungswert und Varianz untersucht. Für große n kann man sogar eine Aussage über die *Verteilung* des Mittelwerts machen; wir kennen sie schon in der Gestalt des zentralen Grenzwertsatzes, an den wir hier nochmal erinnern:

Satz 6.1 (Zentraler Grenzwertsatz). Seien $X_1, X_2, \ldots, X_n$ unabhängige und identisch verteilte ZV mit Erwartungswert μ mit $|\mu| < \infty$ und Varianz $0 < \sigma^2 < \infty$. Dann ist ihr Mittelwert, $\overline{X} := \frac{1}{n}\sum_{i=1}^n X_i$, für großes n näherungsweise normalverteilt mit Erwartungswert μ und Varianz σ^2/n, genauer:

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{(\overline{X}-\mu)\sqrt{n}}{\sigma} \le z\right) = \Phi(z),$$

wobei $\Phi(z)$ die Verteilungsfunktion der Standardnormalverteilung bezeichnet. □

Bemerkung. Man beachte, dass die Zufallsgröße in der Aussage des zentralen Grenzwertsatzes gerade der sogenannte **standardisierte** Stichprobenmittelwert ist:

$$\frac{(\overline{X}-\mu)\sqrt{n}}{\sigma} = \frac{\overline{X}-\mathbb{E}(\overline{X})}{\sqrt{\mathbb{V}(\overline{X})}}$$

mit Erwartungswert 0 und Varianz 1. Die analoge Transformation für normalverteilte Zufallsvariablen haben Sie schon in Gleichung (1.50) kennen gelernt.

Wie in Kapitel 4.2 bereits ausgeführt, lässt sich die Aussage des ZGWS durch Umkehrung der Standardisierung auch unmittelbar für den Stichprobenmittelwert umformulieren,

$$\overline{X} \approx \mathcal{N}(\mu, \frac{\sigma^2}{n}) \qquad (n \text{ groß}), \qquad (6.3)$$

sowie für die Summe der X_i:

$$\sum_{i=1}^{n} X_i \;=\; n\overline{X} \;\approx\; \mathcal{N}(n\mu, n\sigma^2) \qquad (n \text{ groß}). \tag{6.4}$$

Insbesondere kann für hinreichend großes n der linke Ausdruck durch den auf der rechten Seite sehr gut angenähert werden. Eine Illustration findet sich in Abb. 6.1. Diese Illustration bezieht sich auf die Approximation von $\overline{X}$ gemäß Gleichung (6.3). In anderen Quellen, insbesondere im Internet, finden sich sehr viel häufiger andere Darestllungen, etwa für die Approximation der *Summe* gemäß Gleichung (6.4). Lassen Sie sich dadurch nicht verwirren!

Die meisten Größen, die man in der Natur beobachtet, werden von einer großen Zahl zufälliger Faktoren beeinflusst. Der zentrale Grenzwertsatz sorgt dann dafür, dass deren Summe (bzw. deren Mittelwert) in guter Näherung normalverteilt ist, auch wenn das für die einzelnen Faktoren keineswegs zutrifft. (Dies gilt sogar dann, wenn die X_i nicht identisch verteilt sind, vergleiche die Bemerkungen aus Kapitel 4.) Der zentrale Grenzwertsatz ist somit die Ursache für die fundamentale Bedeutung und weite Verbreitung der Normalverteilung.

Außerdem erlaubt er die Approximation anderer Verteilungen durch die Normalverteilung, und zwar immer dann, wenn die Ereignisse aus hinreichend vielen unabhängigen Einzelereignissen zusammengesetzt sind. Darüber gibt die folgende Liste Auskunft (s. auch Kap. 4.2, wo die Approximation genauer begründet wurde).

Beispiele und "Faustregeln" für die Normalapproximation:

i) Ist X binomialverteilt, gilt also $X \sim B(n,p)$, mit $\sigma^2 = np(1-p) \geq 9$, so ist X näherungsweise normalverteilt nach $\mathcal{N}(np, np(1-p))$.

ii) Ist X Poisson-verteilt mit Parameter $\lambda = \mu = \sigma^2 \geq 10$, so ist X näherungsweise normalverteilt nach $\mathcal{N}(\lambda, \lambda)$.

iii) Ist X hypergeometrisch verteilt mit Parametern W, S, n, und gilt $n/(W+S) \leq 0.05$ sowie $np(1-p) \geq 9$ mit $p = W/(W+S)$, so ist X näherungsweise normalverteilt nach $\mathcal{N}(\mu, \sigma^2)$ mit $\mu = np$ und $\sigma^2 = np(1-p)$. (Man kann in diesem Fall die hypergeometrische Verteilung gut durch die Binomialverteilung annähern und dann die Normalapproximation aus i) benutzen.)

Bemerkung. Sind X_i für $i = 1, \ldots, n$ unabhängige, normalverteilte ZV mit $X_i \sim \mathcal{N}\big(\mu_i, \sigma_i^2\big)$, so ist ihre Summe *sogar exakt* normalverteilt, genauer:

$$\sum_i X_i \;\sim\; \mathcal{N}\Big(\sum_i \mu_i, \sum_i \sigma_i^2\Big)\,.$$

Dies ist, wie schon weiter oben erwähnt, der sogenannte *Additionssatz der Normalverteilung*.

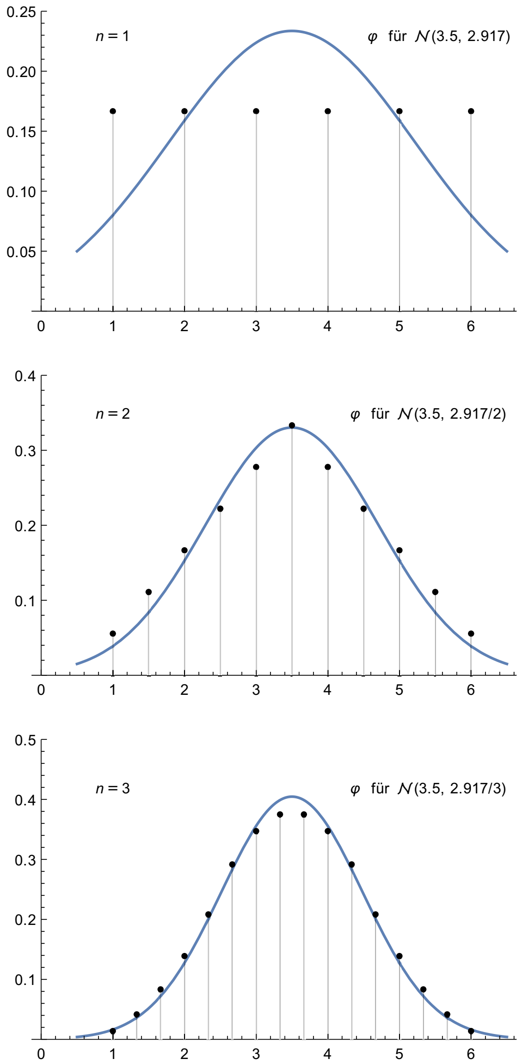

Abbildung 6.1: Exakte Verteilung der mittleren Augenzahl bei $n = 1, 2, 3$ Ausspielungen eines idealen Würfels, und deren Normalapproximation nach der lokalen Version des ZGWS, vgl. (6.3). Durchgezogene Kurve: Dichte $\varphi(x)$ der zugehörigen Normalverteilung. Beim einmaligen Würfeln hat die Augenzahl X den Erwartungswert $\mu = 3.5$ und die Varianz $\sigma^2 = 35/12 = 2.917$, der Stichprobenmittelwert $\overline{X}$ ist also gemäß (6.3) approximativ nach $\mathcal{N}(3.5, 2.917/n)$ verteilt.

6.1.4 Eigenschaften der empirischen Varianz

Die uns ebenfalls schon bekannte *empirische Varianz*

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \widehat{\sigma^2} \tag{6.5}$$

kann man – analog zur Vorgehensweise beim Mittelwert – als Realisierung der Zufallsgröße $S^2 := (1/(n-1))\sum_i (X_i - \overline{X})^2$ auffassen; für sie gilt $\mathbb{E}(S^2) = \sigma^2$, und $\lim_{n\to\infty} \mathbb{V}(S^2) = 0$. Die empirische Varianz ist somit eine erwartungstreue und konsistente Schätzung der Varianz σ^2 der der Stichprobe zugrundeliegenden Verteilung – also wieder eine Schätzung mit "guten" Eigenschaften. Ganz so selbstverständlich, wie es scheint, sind diese Eigenschaften aber nicht.

Beispiel. Die Zufallsvariable $\widetilde{S^2} := (1/n)\sum_i (X_i - \overline{X})^2$ hat Erwartungswert

$$\mathbb{E}(\widetilde{S^2}) = \mathbb{V}(X) - \mathbb{V}(\overline{X}) = (1 - (1/n))\,\sigma^2\,, \tag{6.6}$$

schätzt die Varianz also zu klein! Die Ursache ist, dass $\widetilde{S^2}$ die Schwankung um den *Stichprobenmittelwert* misst, während eigentlich die Schwankung um den wahren (aber unbekannten) Erwartungswert μ geschätzt werden soll. Da der Mittelwert aus der Stichprobe selber ermittelt wurde, fällt die Schwankung um ihn etwas geringer aus als die um μ. $\widetilde{S^2}$ liefert zwar auch einen Schätzer für die Varianz, aber dieser ist nicht erwartungstreu. Wir haben also den Grund für den Faktor $1/(n-1)$ in der Definition der empirischen Varianz gefunden.

Erwartungswert und Varianz sind die beiden wichtigsten Momente einer Verteilung; Mittelwert und empirische Varianz werden entsprechend als **Momentenschätzer** bezeichnet.

6.1.5 Schätzung weiterer Parameter der Verteilung

Momentenschätzer kann man auch benutzen, um weitere unbekannte Parameter von Verteilungen zu schätzen, etwa das p eines Bernoulliexperiments, oder den Intensitätsparameter der Poisson-Verteilung. Zu diesem Zweck werden die Beziehungen zwischen dem zu schätzenden Parameter und den Momenten der Verteilung benutzt, um den Parameter als Funktion der Momente auszudrücken. Die Momente werden dann durch die *Momentenschätzungen* ersetzt; so ergibt sich ein Schätzwert für den Parameter. Dieses Vorgehen nennt man die **Momentenmethode**.

Beispiel. i) Ist X binomialverteilt mit bekanntem n und unbekanntem p (Wahrscheinlichkeit für das Eintreffen des Ereignisses A), so ist $\mu = np$, also $p = \mu/n$. Ersetzt man nun μ durch den zugehörigen Schätzwert $\widehat{\mu}$, so erhält man einen Schätzwert $\widehat{p}$ für p:

$$\widehat{p} = \frac{\widehat{\mu}}{n} = \frac{\overline{x}}{n} = \frac{n_A}{n}, \tag{6.7}$$

wobei n die Gesamtzahl der Experimente und n_A die Zahl der Experimente mit Ausgang A ist. (Wir haben hier das n-fache Bernoulliexperiment als einfaches Binomialexperiment mit Stichprobenumfang 1 aufgefasst, also $\widehat{\mu} = \overline{x} = n_A$.) Gleichung (6.7) ähnelt verdächtig der Häufigkeitsdefinition der Wahrscheinlichkeit, siehe das Beispiel des Münzwurfs im ersten Kapitel! $\widehat{p}$ ist erwartungstreuer und konsistenter Schätzer für p.
ii) Sei X eine Poisson-verteilte ZV und $(x_1, \ldots, x_n)$ eine Stichprobe, wobei wir den Parameter λ nicht kennen. Dann ist $\mu = \lambda$ und somit

$$\widehat{\lambda} = \widehat{\mu} = \overline{x} \tag{6.8}$$

erwartungstreuer, konsistenter Schätzer des Parameters λ. Achtung: Es ist nicht zu empfehlen, anstelle des Mittelwerts die empirische Varianz zur Schätzung von λ zu verwenden, denn der resultierende Schätzer hat eine höhere Varianz als (6.8).

Es gibt noch weitere Methoden zur Parameterschätzung, wie zum Beispiel die sogenannte *Maximum-Likelihood-Methode* oder auch die *Methode der kleinsten Quadrate*. Dies werden wir hier nicht vertiefen, siehe die entsprechenden Lehrbücher. All diese Methoden liefern für den unbekannten Parameter jeweils einen einzigen Wert; eine sogenannte **Punktschätzung**. Dieser Wert entspricht i.a. nicht dem wahren Wert. Wir werden im nächsten Abschnitt eine Methode kennenlernen, anstelle eines einzigen Schätzwertes obere und untere Grenzen anzugeben, innerhalb derer wir den Wert des Parameters vermuten dürfen.

6.2 Schätzung von Konfidenzintervallen

Im letzten Abschnitt haben wir für unbekannte Parameter der Verteilung einzelne Schätzwerte ermittelt. Basierend auf solchen *Punktschätzungen* sollen nun Intervalle ermittelt werden, innerhalb derer wir den Parameter mit einer gewissen Sicherheit vermuten dürfen – sogenannte **Konfidenzintervalle**.

Wir betrachten zunächst die Wahrscheinlichkeit, eine standardnormalverteilte Zufallsvariable Z in einem *vorgegebenen symmetrischen Intervall* $[-c, c]$ anzutreffen: Das ist

$$\mathbb{P}(Z \in [-c, c]) = \mathbb{P}(Z \leq c) - \mathbb{P}(Z \leq -c) = \Phi(c) - \Phi(-c) = 2\Phi(c) - 1,$$

da wegen der Symmetrie der Standardnormalverteilung $\Phi(-c) = 1 - \Phi(-c)$. Es soll nun umgekehrt ein symmetrisches Intervall so bestimmt werden, dass die Werte von Z mit *vorgegebener Wahrscheinlichkeit* $1 - \alpha$ hineinfallen. Dazu muss die Gleichung

$$2\Phi(c) - 1 = 1 - \alpha \iff \Phi(c) = 1 - \frac{\alpha}{2}$$

nach c aufgelöst werden. Da wir hierfür keine expliziten Formeln zur Verfügung haben, entnehmen wir den Wert der Normalverteilungstabelle (rückwärts benutzt).

Allgemein bezeichnet man denjenigen Wert z_q, für den

$$\Phi(z_q) = q \tag{6.9}$$

gilt, als das q-**Quantil** der Standardnormalverteilung, siehe Abbildung 6.2. Anders ausgedrückt ist $c = z_q = \Phi^{-1}(q)$, wobei Φ^{-1} die Umkehrfunktion von Φ bezeichnet. Die Umkehrfunktion existiert, da Φ streng monoton steigend ist.

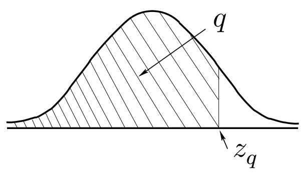

Abbildung 6.2: Der Wert z_q ist das q-Quantil der Verteilung. Die Masse der Verteilung bis z_q ist gerade q.

Mit Quantilen kann man jetzt das gesuchte Intervall explizit angeben. Wir haben

$$\mathbb{P}(Z \in [-z_{1-\frac{\alpha}{2}}, z_{1-\frac{\alpha}{2}}]) = 1 - \alpha,$$

vergleiche Abbildung 6.3.

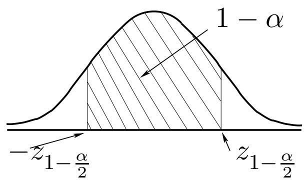

Abbildung 6.3: Die Werte der ZV Z sollen mit Wahrscheinlichkeit $1 - \alpha$ in ein symmetrisches Intervall um 0 fallen. Es sind die entsprechenden Quantile eingetragen.

Beispiel. Sei $\alpha = 0.05$ vorgegeben. Die Gleichung $\Phi(c) = 0.975$ ergibt $c = 1.96 = z_{0.975}$; diesen Wert findet man, indem man die Normalverteilungstabelle 'rückwärts' benutzt. Das gesuchte Intervall ist also $[-1.96, 1.96]$. In diesem Intervall liegen also 95% der Werte von Z.

Sei nun X eine nach $\mathcal{N}(\mu, \sigma^2)$ verteilte ZV. Als "Umkehrung der Standardisierung" schreiben wir X als $X = \mu + \sigma Z$, also gilt $Z \sim \mathcal{N}(0, 1)$. Damit erhalten wir

$$\mathbb{P}(X \in [\mu \ \pm z_{1-\frac{\alpha}{2}}\sigma]) \ = 1 - \alpha\,. \tag{6.10}$$

Wir betrachten nun – wie im letzten Abschnitt – unabhängige und identisch verteilte Zufallsvariablen $X_1, \ldots, X_n$ mit $\mathbb{E}(X_i) = \mu$ und $\mathbb{V}(X_i) = \sigma^2 > 0$ und fragen nach dem symmetrischen Intervall, das den Mittelwert $\overline{X} = \frac{1}{n}\sum_i X_i$ mit

Wahrscheinlichkeit $1 - \alpha$ einschließt. Wir nehmen an, dass n so groß ist, dass die Normalapproximation zulässig ist. Dann ist $\overline{X}$ näherungsweise nach $\mathcal{N}(\mu, \sigma^2/n)$ verteilt, und aus Gl. (6.10) folgt

$$\mathbb{P}\left(\overline{X} \in \left[\mu \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right]\right) \simeq 1 - \alpha\,. \tag{6.11}$$

Dabei bedeutet $\simeq$ ungefähre Gleichheit mit einem kleinen, nicht weiter quantifizierten Fehler. Sind die X_i *normalverteilt*, so gilt (6.11) sogar *exakt*, auch für kleine n; das folgt aus dem Additionssatz der Normalverteilung. In jedem Fall wird das Intervall, das den Mittelwert mit vorgegebener Wahrscheinlichkeit $1 - \alpha$ einschließt, immer kleiner, je größer n ist. Wir können auch sagen, wie groß n sein muss, damit das Intervall eine vorgegebene Breite nicht überschreitet. Denn aus (6.11) können wir folgern: Ist $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$, so ist die Lösung von

$$z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = \delta, \quad \text{also} \quad n = \left(\frac{z_{1-\frac{\alpha}{2}} \cdot \sigma}{\delta}\right)^2, \tag{6.12}$$

die Zahl der Versuche, die mindestens ausgeführt werden muss, damit die Abweichung des Mittelwerts vom Erwartungswert mit Wahrscheinlichkeit $1 - \alpha$ nicht größer ist als δ. Überlegungen dieser Art spielen in der *Versuchsplanung* eine außerordentlich wichtige Rolle, nicht zuletzt bei der Planung der Stichprobengröße in medizinischer Studien.

Beispiel. Wie oft muss man (mit einem idealen Würfel) mindestens würfeln, damit der Mittelwert der Augenzahl mit 95% Wahrscheinlichkeit zwischen 3.4 und 3.6 liegt? Wir kennen Erwartungswert und Varianz der Augenzahl beim einmaligen Würfeln als $\mu = 3.5$, $\sigma^2 = 2.917$. Es ist $\delta = 0.1$ und $z_{0.975} = 1.96$. Einsetzen in (6.12) ergibt $n = \frac{1.96^2 \cdot 2.917}{0.1^2} \simeq 1121$.

Wir halten nun fest, dass die Aussage "$\overline{X} \in [\mu \pm \delta]$" gleichbedeutend ist mit "$\mu \in [\overline{X} \pm \delta]$" — denn beides besagt, dass $\overline{X}$ und μ nicht weiter als δ voneinander entfernt sind. Wir können (6.11) daher auch lesen als

$$\mathbb{P}\left(\mu \in \left[\overline{X} \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right]\right) \simeq 1 - \alpha\,. \tag{6.13}$$

Das zufällige Intervall $[\overline{X} \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$ heißt **Konfidenzintervall** (oder **Vertrauensintervall**) für den Erwartungswert μ zum Konfidenzniveau $1 - \alpha$. Ganz allgemein definiert man:

Definition 6.2 (Konfidenzintervall)**.** Seien C_1 und C_2 (mit $C_1 < C_2$) Zufallsvariablen auf einem Wahrscheinlichkeitsraum (diese können etwa über eine mathematische Stichprobe $(X_1, \ldots, X_n)$ definiert sein). Es sei u ein reeller Parameter, und $0 < \alpha < 1$ eine reelle Zahl. Das zufällige Intervall $[C_1, C_2]$ heißt **Konfidenzintervall für den unbekannten Parameter u zum Konfidenzniveau** $1 - \alpha$, wenn die Beziehung

$$\mathbb{P}(C_1 \leq u \leq C_2) = 1 - \alpha$$

erfüllt ist. In diesem Fall nennt man für Realisierungen c_1 und c_2 von C_1 und C_2 das Intervall $[c_1, c_2]$ das **beobachtete** Konfidenzintervall für u zum Konfidenzniveau $1 - \alpha$.

Bemerkung. Ein Konfidenzintervall $[C_1, C_2]$ ist ein *zufälliges* Intervall. Werden C_1 und C_2 durch die entsprechenden Realisierungen c_1 und c_2 ersetzt, die etwa über eine beobachtete Stichprobe $(x_1, \ldots, x_n)$ bestimmt sind, so erhält man das *beobachtete* Konfidenzintervall $[c_1, c_2]$. Wir werden im folgenden solche Intervalle $[c_1, c_2]$ einfach Konfidenzintervalle nennen, weil klar ist, dass es sich um beobachtete Konfidenzintervalle handelt. Die Aussage "$u \in [c_1, c_2]$" ist dann in $(1 - \alpha) \cdot 100\%$ aller Fälle richtig. Vorsicht: Dies heißt nicht, dass mit Wahrscheinlichkeit $(1-\alpha)$ der wahre Wert im Intervall $[c_1, c_2]$ liegt; solch eine Aussage macht gar keinen Sinn, weil ja u gegeben ist und c_1, c_2 zufällige Realisierungen sind.

Im Allgemeinen müssen Konfindenzintervalle nicht symmetrisch sein, wir werden es in diesem Kurs aber nur mit symmetrischen Intervallen zu tun haben. Insbesondere ist die Symmetrie immer dann eine natürliche Eigenschaft, wenn es um den Erwartungswert einer Verteilung geht, die symmetrisch um eben diesen Erwartungswert ist.

Die folgende Liste gibt über einige wichtige Anwendungen Auskunft.

1. **Konfidenzintervall für den Erwartungswert der Normalverteilung**: Gegeben sei eine Stichprobe vom Umfang n mit Mittelwert $\overline{x}$. Es wird vorausgesetzt, dass die zugrundeliegende Zufallsgröße normalverteilt ist, oder dass n so groß ist, dass die Normalapproximation anwendbar ist.

   (a) Wenn die **Varianz σ^2 von X bekannt** ist, gilt (6.11), und das (symmetrische) $1 - \alpha$-Konfidenzintervall für den Erwartungswert μ von X lautet:
   $$\left[\overline{x} \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right].$$

   (b) Wenn σ^2 unbekannt ist, muss mit der **empirischen Varianz** s^2 gearbeitet werden. In Analogie zu Z_n aus (4.4) betrachten wir zu diesem Zweck die ZV
   $$T := \frac{\overline{X} - \mu}{S/\sqrt{n}},$$
   wobei S^2 die empirische Varianz *als Zufallsvariable* bezeichnet, vgl. Abschnitt 6.1.4. Man kann zeigen: Die Zufallsvariable T ist **Student-t-verteilt** mit $k = n - 1$ Freiheitsgraden (kurz: t_{n-1}-verteilt),
   $$\mathbb{P}(T \le t) = c \int_{-\infty}^{t} \left(1 + \frac{\tau^2}{n-1}\right)^{-n/2} \mathrm{d}\tau,$$
   wobei c eine explizit bekannte, von n abhängige Normierungskonstante ist. (Wie bei der Normalverteilung kennt man keinen geschlossenen

Ausdruck für die Verteilungsfunktion und muss die zugehörigen Werte in Tabellen nachschlagen.) Wie die Standardnormalverteilung hat Die Dichte der t_{n-1}-Verteilung hat — wie die der Standardnormalverteilung — die Gestalt einer Glockenkurve und ist symmetrisch um 0. Das symmetrische $1-\alpha$-Konfidenzintervall für den Erwartungswert μ ist dann

$$\left[\overline{x} \pm t_{n-1,1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right], \tag{6.14}$$

wobei $t_{n-1,q}$ das q-Quantil der t_{n-1}-Verteilung ist, also derjenige Wert, der

$$\mathbb{P}(T \leq t_{n-1,q}) = q$$

erfüllt. Quantilwerte können der (Quantil-)Tabelle der Student-t-Verteilung entnommen werden. Die t-Verteilung hat eine größere Varianz als die Standardnormalverteilung (die Glockenkurve ist etwas 'breiter'), daher ist $t_{n-1,q}$ für endliche n immer größer als z_q (so ist z.B. $t_{2,0.975} = 4.3 > 1.96 = z_{0.975}$); entsprechend größer fällt das Intervall aus. Es gilt jedoch $\lim_{n\to\infty} t_{n-1} = \mathcal{N}(0,1)$ und somit auch $\lim_{n\to\infty} t_{n-1,q} = z_q$. Tatsächlich ist für große Werte $(n \geq 30)$ die Approximation durch die Normalverteilung eine gute Näherung. Dies kann man auch schon mit der expliziten Darstellung der Verteilungsfunktion erkennen (vgl. die Herleitung der Poisson-Verteilung und den Beweis des ZGWS).

2. **Konfidenzintervall für den Parameter p eines Bernoulli-Experiments.** Bei n Wiederholungen eines Bernoulli-Experiments wurde n_A mal das Ereignis A beobachtet, woraus sich $\overline{x} = \widehat{p} = n_A/n$ als Schätzer für p, die Wahrscheinlichkeit von A, ergibt. Für großes n erhält man als Schätzer für die Varianz des (einfachen) Bernoulli-Experiments

$$\begin{aligned} s^2 &= \frac{1}{n-1}\left[\Big(\sum_{i=1}^{n} x_i^2\Big) - n\overline{x}^2\right] = \frac{1}{n-1}(n\overline{x} - n\overline{x}^2) = \frac{n}{n-1}\overline{x}(1-\overline{x}) \\ &= \frac{n}{n-1}\widehat{p}(1-\widehat{p}) \simeq \widehat{p}(1-\widehat{p}), \end{aligned}$$

wobei wir im zweiten Schritt benutzt haben, dass $x_i^2 = x_i$ da $x_i \in \{0,1\}$. Für $n\widehat{p}(1-\widehat{p}) \geq 9$ ist die Normalapproximation zulässig, so dass das Konfidenzintervall für den Erwartungswert $\mu = p$ des Bernoulliexperiments durch (6.14) gegeben ist. Einsetzen von $\overline{x}$ und s^2 in (6.14) liefert somit

$$\left[\widehat{p} \pm z_{1-\frac{\alpha}{2}} \frac{\sqrt{\widehat{p}\,(1-\widehat{p})}}{\sqrt{n}}\right] \tag{6.15}$$

als ist das symmetrische $1-\alpha$-Konfidenzintervall für p. (Falls (6.15) Intervallgrenzen < 0 bzw. > 1 liefert, müssen diese natürlich durch 0 bzw. 1

ersetzt werden.) Beachten Sie, dass in diesem Beispiel die Varianz unbekannt ist und deshalb eigentlich mit der t-Verteilung gerechnet werden muss; weil n groß ist, verwenden wir die Approximation von $t_{n-1,1-\frac{\alpha}{2}}$ durch $z_{1-\frac{\alpha}{2}}$.

Beispiel. Die Geburtsstatistik der Schweiz weist zwischen 1950 und 1970 $n = 1944700$ Geburten aus, davon $n_J = 997600$ Jungengeburten. Es sollen das 99% und das 99.9% Konfidenzintervall für die Wahrscheinlichkeit p einer Jungengeburt ermittelt werden. Man erhält $\widehat{p} = n_J/n = 0.512984$ und $s^2 = 0.249831$. Weiterhin ist, da n riesig ist, die Normalapproximation anwendbar, die relevanten Quantilwerte sind $z_{0.995} = 2.575$ und $z_{0.9995} = 3.300$. Mit diesen Zahlen ergeben sich die 99% und 99.9% Konfidenzintervalle als $[0.51206, 0.51391]$ bzw. $[0.51190, 0.51416]$. Keines schließt den Wert 0.5 ein. Wir können daher mit Recht behaupten, dass Jungen- und Mädchengeburten nicht gleich häufig sind: Wir wissen dies zwar nicht mit Sicherheit, haben aber ein Verfahren benutzt, das mit 99.9%iger Sicherheit wahre Behauptungen liefert.

Wir haben in diesem Abschnitt Konfidenzintervalle für Erwartungswerte und verwandte Größen bestimmt, unter der Normalverteilungsannahme. Gelegentlich benötigt man auch Konfidenzintervalle für die Varianz oder für andere Parameter einer (unbekannten) Verteilung. Diese Probleme kann man mit den hier besprochenen Methoden behandeln, siehe die angeführte Literatur.

6.3 Das Testen von Hypothesen

Experimentelle Daten dienen letzten Endes immer dem Zweck, bestimmte Vermutungen zu bestätigen oder zu widerlegen. Bei zufälligen Größen kann man solche Entscheidungen nie mit letzter Sicherheit treffen; man benötigt hier das Konzept des **statistischen Tests**.

Eine **Hypothese** ist in diesem Zusammenhang ganz allgemein eine Annahme über eine Zufallsvariable. Ein **Test** einer Hypothese ist ein Prüfverfahren, das man anwendet, um festzustellen, ob die Hypothese abgelehnt werden soll oder nicht. Dabei stehen jeweils zwei Hypothesen zur Debatte: Die zu prüfende oder **Nullhypothese** H_0, und die **Alternativhypothese** H_A, für die man sich bei Ablehnung von H_0 entscheidet. Das Konzept des statistischen Tests soll nun am Beispiel des sogenannten **Gauß-Tests** illustriert werden, s. Abb. 6.4.

Sei $X \sim \mathcal{N}(\mu, \sigma^2)$, wobei σ^2 *bekannt*, μ aber *unbekannt* ist. Mit Hilfe einer Stichprobe vom Umfang n möchte man $H_0 : \mu = \mu_0$ gegen $H_A : \mu \neq \mu_0$ überprüfen.

Als **Prüfgröße** dient der Stichprobenmittelwert $\overline{X}$. Wir geben (vor der Durchführung des Tests!) eine kleine positive Zahl α vor, das **Signifikanzniveau** des Tests oder die **Irrtumswahrscheinlichkeit**. Wir überlegen dann, welcher Verteilung der Stichprobenmittelwert folgen würde, wenn die H_0 richtig wäre: Unter H_0 wäre $\overline{X} \sim \mathcal{N}(\mu_0, \sigma^2/n)$. Wir lehnen die H_0 dann ab, wenn der beobachtete Stichprobenmittelwert so weit von μ_0 abweicht, dass es "zu unwahrscheinlich" ist, dass es

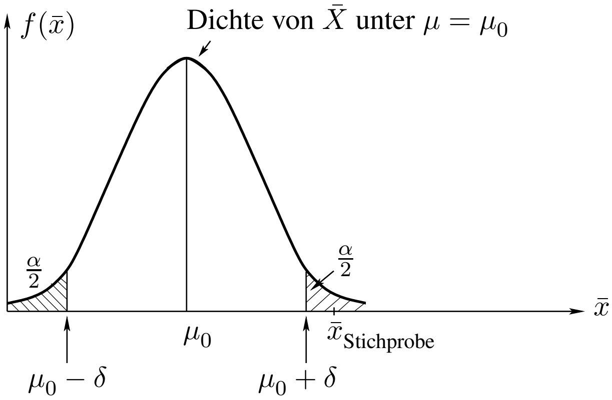

Abbildung 6.4: Prinzip des einfachen, zweiseitigen Gauß-Tests. μ_0 ist der Erwartungswert unter H_0. Die Nullhypothese wird abgelehnt, wenn der Stichprobenmittelwert, hier mit $\bar{x}_{\text{Stichprobe}}$ bezeichnet, in den Ablehungsbereich fällt.

sich um eine zufällige Schwankung handelt. Genauer: H_0 wird abgelehnt, wenn $\overline{x}$ in den **Ablehnungsbereich** $\{\overline{x} : |\overline{x} - \mu_0| > \delta\}$ fällt. Dabei bestimmen wir δ so, dass

$$\mathbb{P}(|\overline{X} - \mu_0| > \delta) = \alpha, \quad \text{wenn} \quad \mu = \mu_0 \,. \tag{6.16}$$

Der Ablehnungsbereich besteht also gerade aus den — unter H_0 — $\alpha \cdot 100\%$ extremsten Werten. Nach Gl. (6.12) muss δ dann gerade als

$$\delta = z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \tag{6.17}$$

gewählt werden. Wir verwerfen H_0 also, falls

$$\frac{|\overline{x} - \mu_0|}{\sigma} \sqrt{n} > z_{1-\frac{\alpha}{2}} \,.$$

Andernfalls behalten wir die Nullhypothese bei. Wir können dann allerdings nicht schließen, dass H_0 richtig ist. Wir können lediglich folgern, dass wir aufgrund der Stichprobe keine signifikante Abweichung zum Signifikanzniveau $1-\alpha$ feststellen können.

Noch ein Wort zur Wahl der Nullhypothese. Die Faustregel ist, dass die Nullhypothese "konservativ" ist: Sie verneint in der Regel Unterschiede, Veränderungen usw., deren Feststellung das bisher Bekannte in Frage stellen könnte. Will man umgekehrt einen vermuteten Zusammenhang erhärten, so wählt man als H_0 die Verneinung des Zusammenhangs; wird diese dann aufgrund der Beobachtungen abgelehnt, so darf der Zusammenhang als statistisch signifikant angesehen werden.

Beispiel. Für das *Geburtenbeispiel* wird H_0: $p = 1/2$ gegen H_A: $p \neq 1/2$ zum Signifikanzniveau $\alpha = 0.001$ getestet. Unter der Nullhypothese ist — für jede einzelne Geburt — $\mu_0 = p = 1/2$ und $\sigma^2 = p(1-p) = 1/4$. Da $\frac{|\overline{x}-\mu_0|}{\sigma}\sqrt{n} = 36.20 > z_{0.9995} = 3.3$, wird H_0 abgelehnt.

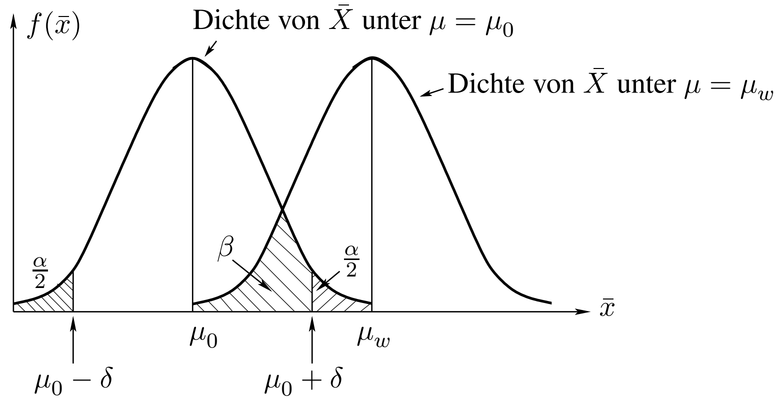

Abbildung 6.5: Fehler 1. Art und Fehler 2. Art im einfachen, zweiseitigen Gauß-Test. μ_0 ist der Erwartungswert unter H_0, μ_w ist der wahre (aber unbekannte) Erwartungswert.

Wir sind hier — wie beim Konfidenzintervall — wieder mit der Tatsache konfrontiert, dass man statistische Aussagen nie mit letzter Sicherheit treffen kann. Denn selbst der Wert $\overline{x} = 0.512984$ im Geburtenbeispiel, der — relativ zur Standardabweichung von $\overline{X}$ unter H_0 — eine sehr große Abweichung vom Erwartungswert darstellt, ist unter H_0 *möglich*; allerdings ist dieser oder ein noch extremerer Wert äußerst unwahrscheinlich. Wir müssen uns deshalb genauer mit den Fehlern auseinander setzen, die beim Testen auftreten können. Es sind dies:

Fehler 1. Art: Die Nullhypothese wird abgelehnt, obwohl sie richtig ist; der Test ist gerade so konstruiert, dass dies mit der vorgegebenen Irrtumswahrscheinlichkeit α geschieht (bei einseitigen Fragestellungen (s.u.) mit Wahrscheinlichkeit $\leq \alpha$).
Fehler 2. Art: Die Nullhypothese wird beibehalten, obwohl sie falsch ist. Die Wahrscheinlichkeit β, mit der dies geschieht, ist in Abb. 6.5 illustriert. Sie hängt von der wahren (aber unbekannten) Verteilung ab und ist deshalb sehr viel schwerer zu quantifizieren als der Fehler 1. Art. Allgemein lässt sich aber sagen, dass die Wahl eines kleineren α zu einem größeren β führt und umgekehrt.

Dies ist in der folgenden Tabelle noch einmal zusammengefasst, vergleiche auch Abbildung 6.5.

	Das Testergebnis lautet	
	H_0 ablehnen	H_0 beibehalten
H_0 ist richtig	Fehler 1. Art α	kein Fehler $1 - \alpha$
H_0 ist falsch	kein Fehler $1 - \beta$	Fehler 2. Art β

Einseitige und zweiseitge Tests. Der bisher betrachtete Gauß-Test war ein sogenannter *zweiseitiger* Test. Hier haben wir nur danach gefragt, ob zwei Werte gleich oder verschieden sind, wobei kein Anfangsverdacht geäußert wurde, welcher der größere ist, wir hatten also eine *zweiseitige Fragestellung*. Entsprechend lautete das Hypothesenpaar $H_0 : \mu = \mu_0$ versus $H_A : \mu \neq \mu_0$, und große Abweichungen der Prüfgröße nach oben *oder* nach unten (bezüglich des Idealwerts μ_0) führten gleichermaßen zur Ablehnung der Nullhypothese (wie im GEBURTENBEISPIEL). Ist man dagegen nur an Abweichungen in einer Richtung interessiert, so ist die Fragestellung *einseitig*. Typische Beispiele sind Fragen wie: Werden Stickoxidgrenzwerte überschritten? Ist der Cholesterinspiegel zu hoch? Sind nur Abweichungen nach oben (unten) relevant, so lautet das Hypothesenpaar $H_0 : \mu \leq \mu_0$ versus $H_A : \mu > \mu_0$ ($H_0 : \mu \geq \mu_0$ versus $H_A : \mu < \mu_0$). Der Ablehnungsbereich besteht dann aus den (unter H_0) $\alpha \cdot 100\%$ *größten (kleinsten)* Werten; vgl. Abb. 6.6. Genauer wählt man für das Hypothesenpaar $H_0 : \mu \leq \mu_0$ versus $H_A : \mu > \mu_0$ unser δ so, dass, in Analogie zu (6.16):

$$\mathbb{P}(\overline{X} - \mu_0 > \delta) = \alpha, \quad \text{wenn} \quad \mu = \mu_0 \tag{6.18}$$

und lehnt H_0 ab, wenn $\overline{X} - \mu_0 > \delta$; für das Hypothesenpaar $H_0 : \mu \geq \mu_0$ versus $H_A : \mu < \mu_0$ wählt man δ so, dass

$$\mathbb{P}(\overline{X} - \mu_0 < -\delta) = \alpha, \quad \text{wenn} \quad \mu = \mu_0 \tag{6.19}$$

und lehnt H_0 ab, wenn $\overline{X} - \mu_0 < -\delta$. Dies garantiert, dass der Fehler erster Art höchstens gleich dem vorgegebenen α ist:

$$\mathbb{P}(H_0 \text{ wird abgelehnt wenn wahr}) \leq \alpha.$$

Die Ungleichung rührt daher, dass wir die H_0 hier als Ungleichung formuliert haben. Gleichheit gilt gerade dann, wenn $\mu = \mu_0$; andernfalls ist der Fehler erster Art kleiner als α. Ganz allgemein arbeitet man bei einseitigen Fragestellungen mit der Verteilung der Prüfgröße im 'ungünstigsten' Fall, d.h. in dem Fall, der den größten Fehler 1. Art liefert; dieser entspricht dem Gleichheitszeichen.

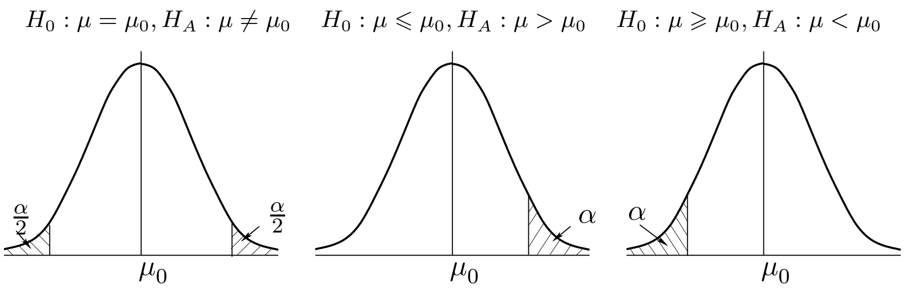

Abbildung 6.6: Zweiseitiger Test (links) und einseitige Tests (Mitte und rechts).

Allgemeines Prinzip: Das bisher beschriebene Beispiel illustriert ein ganz allgemeines Prinzip, nach dem für sehr allgemeine Situationen (nicht nur die hier besprochenen) Tests hergeleitet werden können. Man geht in vier Schritten vor:

1. Hypothesenpaar formulieren (betrifft oft Erwartungswerte, oder andere Parameter von Verteilungen) und α festlegen

2. Prüfgröße festlegen (eine geeignete, aus der Stichprobe ermittelte Größe; oft der Stichprobenmittelwert, aber keineswegs immer)

3. Verteilung der Prüfgröße unter H_0 bestimmen (oft eine Normalverteilung, aber keineswegs immer)

4. H_0 ablehnen, wenn der gemessene Wert der Prüfgröße in den Ablehnungsbereich fällt, das sind die (unter H_0 höchstens) $\alpha\cdot$ 100 % extremsten Werte der Prüfgröße. Manchmal ist die explizite Konstruktion des Ablehnungsbereichs gar nicht notwendig, wie Sie im folgenden Beispiel sehen werden.

Beispiel. Louis Pasteur impfte im Jahr 1880 sechs Hühner mit einem Serum gegen Cholera. Als er diese und sechs ungeimpfte Hühner mit Cholera-Erregern infizierte, waren am nächsten Tag sechs Hühner tot. Diese 6 toten Hühner waren gerade die 6 ungeimpften.

Wir wollen nun statistisch nachweisen, dass das Serum eine Wirkung hatte und gehen dabei gemäß obiger 'Gebrauchsanleitung' vor:

1. Die Fragestellung ist einseitig. Wir testen die Nullhypothese 'Tod durch Cholera ist bei ungeimpften Hühnern nicht häufiger als bei geimpften' gegen die Alternativhypothese 'Tod durch Cholera ist bei ungeimpften Hühnern häufiger' ($\alpha = 0.05$).

2. Als Prüfgröße nehmen wir X, die Anzahl der ungeimpften Hühner unter den 6 toten.

3. Falls das Serum unwirksam gewesen wäre, wären die 6 toten Hühner eine zufällige Auswahl aus den 6 geimpften und den 6 ungeimpften gewesen, also $X \sim \text{hyper}(6, 6, 6)$. Dies ist die Verteilung der Prüfgröße unter H_0 ('mit dem Gleichheitszeichen').

4. Wir konstruieren zunächst den Ablehungsbereich. Wegen der einseitigen Fragestellung interessieren uns die unter H_0 höchstens 5 % *größten* Werte. Wir erhalten

$$\begin{aligned}
\mathbb{P}(X = 6) &= \frac{\binom{6}{6}\binom{6}{0}}{\binom{12}{6}} = \frac{1 \cdot 1}{924} = 0.0011 \\
\mathbb{P}(X = 5) &= \frac{\binom{6}{5}\binom{6}{1}}{\binom{12}{6}} = \frac{6 \cdot 6}{924} = 0.038 \\
\mathbb{P}(X = 4) &= \frac{\binom{6}{4}\binom{6}{2}}{\binom{12}{6}} = \frac{15 \cdot 15}{924} = 0.244.
\end{aligned}$$

Da $\mathbb{P}(X = 5) + \mathbb{P}(X = 6) = 0.039 < 0.05 = \alpha < 0.283 = \mathbb{P}(X = 4) + \mathbb{P}(X = 5) + \mathbb{P}(X = 6)$, besteht der Ablehungsbereich aus den Werten 5 und 6. Man wird also dem Serum eine Wirkung zuschreiben, wenn mindestens 5 der toten Hühner ungeimpft waren. Im vorliegenden Fall wird die H_0 also abgelehnt, die Wirkung des Serums ist statistisch signifikant.

In diesem Beispiel kann man sich die Konstruktion des Ablehungsbereichs allerdings auch sparen. Denn es reicht, folgendermaßen zu argumentieren: Wir haben $X = 6$ beobachtet. Die Wahrscheinlichkeit unter H_0 für diesen oder einen noch größeren Wert ist $\mathbb{P}(X = 6) = 0.0011$ (in diesem Fall ist ja gar kein größerer Wert möglich, aber Vorsicht: Im Allgemeinen sind hier *alle* Werte zu beachten, die mindestens so extrem sind wie der beobachtete; hätte man also $X = 5$ beobachtet, wäre die relevante Wahrscheinlichkeit $\mathbb{P}(X = 5) + \mathbb{P}(X = 6) = 0.039$). Da $0.0011 < \alpha$, gehört der beobachtete Wert offenbar zum Ablehnungsbereich.

6.3.1 Der Zoo der Erwartungswerttests

Nachdem wir das Prinzip des statistischen Tests kennengelernt haben, soll jetzt eine Übersicht über einige Tests folgen, die sich auf Erwartungswerte von Verteilungen beziehen (sogenannte Erwartungswerttests). Wir besprechen hier solche Tests, in denen die betrachteten Zufallsvariablen normalverteilt sind oder die Stichprobe so groß ist, dass die Normalapproximation anwendbar ist. Je nach Situation sind verschiedene Tests zu verwenden. Um den jeweils “richtigen” Test zu finden, geht man anhand folgender Kriterien vor:

1. Hat man es mit einer oder mit zwei Zufallsvariablen zu tun? Im ersten Fall vergleicht man seine Beobachtungen mit einem vorgegebenen “Idealwert” μ_0 einer Zufallsvariablen X (wie im GEBURTENBEISPIEL). Der Idealwert stammt typischerweise aus der Theorie, oder aus langjähriger Erfahrung (‘Lehrbuchwert’). Das Hypothesenpaar lautet dann z.B. $H_0 : \mu = \mu_0$ versus $H_A : \mu \neq \mu_0$, und man macht einen **Ein-Stichproben-Test** (manchmal auch **einfacher Test** genannt). Im zweiten Fall vergleicht man zwei Zufallsgrößen X und Y miteinander (oft sind das Daten aus “Experiment” und “Kontrolle”, z.B. das Wachstum von Pflanzen mit und ohne Wachstumshormon); hier lautet das typische Hypothesenpaar $H_0 : \mu_X = \mu_Y$ versus $H_A : \mu_X \neq \mu_Y$, und man macht **Zwei-Stichproben-Tests** (auch **doppelte Tests** genannt).

2. Ist die Varianz der Zufallsgröße (unter der Nullhypothese) bekannt, oder muss sie aus der Stichprobe geschätzt werden? Ist die Varianz bekannt, so verwendet man **Gauß-Tests**, ansonsten **t-Tests**.

3. Ist die Fragestellung **einseitig** oder **zweiseitig**? Diese Unterscheidung haben wir weiter oben schon besprochen.

Bei dem eingangs besprochenen Beispiel (und insbesondere beim Geburtenbeispiel) handelt es sich also um einen einfachen, zweiseitigen Gauß-Test.

Es folgt nun eine Übersicht über wichtige Erwartungswerttests.

I. Ein-Stichproben-Tests

Die Frage, ob der unbekannte Erwartungswert μ einer normalverteilten Zufallsvariablen X vom vermuteten Wert μ_0 abweicht, soll zum Signifikanzniveau α beantwortet werden. Gegeben sei eine Stichprobe vom Umfang n mit Mittelwert $\overline{x}$ und empirischer Varianz s^2.

1. **Einfacher Gauß-Test:** Die Varianz σ^2 (unter H_0) wird als bekannt vorausgesetzt. Der Test beruht darauf, dass die ZV (die *standardisierte Prüfgröße*)

   $$Z := \frac{\overline{X} - \mu_0}{\sigma}\sqrt{n}$$

   standardnormalverteilt ist, falls $\mu = \mu_0$ gilt.

   (a) Zweiseitiger Test: $H_0 : \mu = \mu_0$ gegen $H_A : \mu \neq \mu_0$. Da unter H_0 gilt

   $$\mathbb{P}(|Z| > z_{1-\frac{\alpha}{2}}) = \alpha$$

   (vergleiche (6.16) und (6.17)), verwirft man H_0, falls

   $$\frac{|\overline{x} - \mu_0|}{\sigma}\sqrt{n} > z_{1-\frac{\alpha}{2}}\,,$$

   wobei $z_{1-\frac{\alpha}{2}}$ wie immer das $1 - \frac{\alpha}{2}$-Quantil der Standardnormalverteilung bezeichnet. (Dieser Test wurde schon weiter oben ausführlich besprochen.)

   (b) Einseitiger Test: $H_0 : \mu \leq \mu_0$ gegen $H_A : \mu > \mu_0$. Da unter H_0 (ausgewertet am Gleichheitszeichen) gilt

   $$\mathbb{P}(Z > z_{1-\alpha}) = \alpha$$

   (vergleiche (6.18) sowie die Definition des Quantils (6.9)), verwirft man H_0, falls

   $$\frac{\overline{x} - \mu_0}{\sigma}\sqrt{n} > z_{1-\alpha}\,.$$

   (c) Einseitiger Test: $H_0 : \mu \geq \mu_0$ gegen $H_A : \mu < \mu_0$. Da nun unter H_0 (ausgewertet am Gleichheitszeichen) gilt

   $$\mathbb{P}(Z < -z_{1-\alpha}) = \alpha$$

   (vergleiche (6.19)), verwirft man H_0, falls

   $$\frac{\overline{x} - \mu_0}{\sigma}\sqrt{n} < -z_{1-\alpha}\,.$$

2. **Einfacher t-Test:** Die Varianz σ^2 ist *unbekannt*; daher muss die *empirische Varianz* s^2 der Stichprobe herangezogen werden. Der Test beruht darauf, dass die ZV (die standardisierte Prüfgröße)

$$T := \frac{\overline{X} - \mu_0}{S}\sqrt{n}$$

unter $\mu = \mu_0$ Student-t-verteilt ist mit $k = n - 1$ Freiheitsgraden (dies ist ganz analog zur Überlegung zu Konfidenzintervallen bei unbekannter Varianz).

   (a) Zweiseitiger Test: $H_0 : \mu = \mu_0$ gegen $H_A : \mu \neq \mu_0$. Da unter H_0 ('ausgewertet mit dem Gleichheitszeichen')

$$\mathbb{P}(|T| > t_{n-1,1-\frac{\alpha}{2}}) = \alpha,$$

   ist H_0 abzulehnen, falls

$$\frac{|\overline{x} - \mu_0|}{s}\sqrt{n} > t_{n-1,1-\frac{\alpha}{2}}\,,$$

   wobei $t_{n-1,1-\frac{\alpha}{2}}$ das $1 - \frac{\alpha}{2}$-Quantil der Student-t-Verteilung mit $k = n - 1$ Freiheitsgraden bezeichnet.

   (b) Einseitige Tests: $H_0 : \mu \lesseqqgtr \mu_0$ gegen $H_A : \mu \gtrless \mu_0$. Da unter H_0 (jeweils für $\mu = \mu_0$ ausgewertet) gilt, dass

$$\mathbb{P}(T \gtrless \pm t_{n-1,1-\alpha}) = \alpha,$$

   ist H_0 abzulehnen, falls

$$\frac{\overline{x} - \mu_0}{s}\sqrt{n} \gtrless \pm t_{n-1,1-\alpha}\,.$$

   Beispiel. Ist der Cholesterinwert von Vegetariern signifikant ($\alpha = 0.01$) geringer als der Normalwert (180 mg/100ml)? Man prüft $H_0 : \ \mu \geq 180$ gegen $H_A : \ \mu < 180$. Bei einer Gruppe von 9 Vegetariern werden folgende Werte gemessen: 154, 119, 177, 150, 138, 185, 167, 158, 174. Man erhält $\overline{x} = 158$, $s = 20.7$, und wegen $\frac{158-180}{20.7}\sqrt{9} = -3.19 < -2.90 = -t_{8,0.99}$ wird H_0 abgelehnt. Der Cholesterinwert von Vegetariern ist also signifikant niedriger als der Normalwert.

II. Zwei-Stichproben-Tests

1. **Gepaarte Stichproben: t-Differenzentest.** An n unabhängigen Objekten werden jeweils die Merkmale X und Y gemessen, die als normalverteilt vorausgesetzt werden. Wir erhalten also n Paare $(X_1, Y_1), \ldots, (X_n, Y_n)$ mit Realisierungen $(x_1, y_1), \ldots, (x_n, y_n)$. Hier ein paar Beispiele: An n Personen wird jeweils die Länge X des linken und die Länge Y des rechten Beins

gemessen; bei n verschiedengeschlechtlichen Zwillingsgeburten wird das Geburtsgewicht X des Mädchens und das Geburtsgewicht Y des Jungen registriert; oder in $n = 12$ Kalendermonaten bezeichnen X und Y die Arbeitslosigkeit in zwei verschiedenen Jahren. X_i und Y_i sind dabei *nicht* unabhängig (wohl aber X_i von X_j und Y_j für $j \neq i$) — man spricht auch von *verbundenen* Stichproben). Die Erwartungswerte $\mu_X := \mathbb{E}(X)$ und $\mu_Y := \mathbb{E}(Y)$ seien unbekannt, ebenso die Varianzen. Man testet auf Gleichheit der Erwartungswerte, also $H_0 : \mu_X = \mu_Y$ gegen $H_A : \mu_X \neq \mu_Y$. Wir betrachten nun die neue ZV $D := X - Y$ und nehmen an, dass diese normalverteilt ist. Dann ist D unter H_0 nach $\mathcal{N}(0, \sigma_D^2)$ verteilt, wobei wir die Varianz σ_D^2 nicht kennen. Der Test lässt sich also auf einen einfachen t-Test zurückführen, und zwar $H_0 : \mu_D = 0$ versus $H_A : \mu_D \neq 0$. Die Prüfgröße ist dann der Stichprobenmittelwert von D, also $\overline{D} := \frac{1}{n}(D_1 + \ldots + D_n) = \overline{X} - \overline{Y}$. Unter H_0 ist der standardisierte Stichprobenmittelwert

$$\frac{\overline{D}}{S_D}\sqrt{n}$$

nach t_{n-1} verteilt, wobei S_D^2 die empirische Varianz von D *als Zufallsvariable* bezeichnet.

Um uns Schätzwerte für $\overline{D}$ und σ_D^2 zu verschaffen, brauchen wir eine Stichprobe für D. Die ergibt sich aus der ursprünglichen Stichprobe über die Differenzen

$$d_i = x_i - y_i \, , \quad i = 1, \ldots, n \, .$$

Der zugehörige Stichprobenmittelwert ist

$$\bar{d} = \frac{1}{n} \sum_{i=1}^{n} d_i = \overline{x} - \bar{y},$$

und die empirische Varianz lautet

$$s_d^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(d_i - \bar{d}\right)^2 \, .$$

H_0 wird abgelehnt, falls

$$\frac{|\bar{d}|}{s_d}\sqrt{n} > t_{n-1,1-\frac{\alpha}{2}} \, .$$

Ganz analog werden entsprechende einseitige Tests durchgeführt.

Beispiel. Der Ertrag X (in Zentnern) von 8 Kirschbäumen im Jahr 1990 wird mit dem Ertrag Y *derselben* 8 Bäume im Jahr 1991 verglichen. Es soll die Nullhypothese “Der Ertrag war in beiden Jahren gleich” gegen die Alternativhypothese “Der Ertrag ist verschieden” zum Signifikanzniveau 0.05 getestet werden.

Baum-Nr. i	1	2	3	4	5	6	7	8	
x_i	3.6	3.1	3.4	3.2	3.5	3.1	3.2	3.5	$\overline{x} = 3.33$
y_i	3.5	3.5	3.4	3.6	4.0	3.5	3.3	3.1	$\bar{y} = 3.48$
d_i	0.1	-0.4	0.	-0.4	-0.5	-0.4	-0.1	0.4	$\bar{d} = -0.15$

Es ergibt sich $s_d^2 = 0.104$ und somit $\frac{|\bar{d}|}{s_d}\sqrt{n} = 1.32 < 2.365 = t_{7,0.975}$, also kein signifikanter Unterschied.

2. **Zwei unabhängige Stichproben: Doppelter t-Test.** X und Y seien nun *unabhängige*, normalverteilte ZV mit unbekannten Erwartungswerten $\mu_X := \mathbb{E}(X)$ und $\mu_Y := \mathbb{E}(Y)$. Die Varianzen sind ebenfalls unbekannt, werden aber als *gleich* vorausgesetzt: $\sigma^2 = \mathbb{V}(X) = \mathbb{V}(Y)$. (Im Fall ungleicher Varianzen verwendet man stattdessen den sogenannten Berens–Fisher–Test.) Gegeben sind Stichproben von X und Y vom Umfang n_x bzw. n_y. Ihre Mittelwerte sind $\overline{x}$ bzw. $\bar{y}$, und ihre empirischen Varianzen s_x^2 bzw. s_y^2.

   Wir testen $H_0 : \mu_X = \mu_Y$ gegen $H_A : \mu_X \neq \mu_Y$. Prüfgröße ist wieder die *Differenz der Stichprobenmittel*,

$$\overline{D} := \overline{X} - \overline{Y}\,.$$

   Wegen der Unabhängigkeit von X und Y folgt aus dem Additionssatz der Normalverteilung, dass $\overline{D}$ unter H_0 eine normalverteilte ZV ist mit $\mathbb{E}(\overline{D}) = 0$ und

$$\sigma_{\overline{D}}^2 := \mathbb{V}(\overline{D}) = \mathbb{V}(\overline{X}) + \mathbb{V}(\overline{Y}) = \frac{1}{n_x}\mathbb{V}(X) + \frac{1}{n_y}\mathbb{V}(Y) = \frac{n_x + n_y}{n_x n_y}\sigma^2. \tag{6.20}$$

   Schätzwert für $\overline{D}$ ist natürlich wieder $\bar{d} := \overline{x} - \bar{y}$. Um diesen standardisieren zu können, müssen wir uns nun einen Schätzwert für $\sigma_{\overline{D}}^2$ verschaffen. Aus (6.20) folgt, dass es ausreicht, einen Schätzwert für das unbekannte σ^2 zu gewinnen. Diesen erhalten wir aus den beiden Stichproben als *gewichtetes Mittel* von s_x^2 und s_y^2 mit Gewichtsfaktoren $n_x - 1$ und $n_y - 1$:

$$\widehat{\sigma}^2 = s^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}\,.$$

   Somit ergibt sich aus (6.20) der Schätzwert für die Varianz von $\overline{D}$ als

$$\widehat{\sigma}_{\overline{D}}^2 = \frac{n_x + n_y}{n_x n_y} s^2\,.$$

   Es läßt sich nun zeigen, dass die standardisierte Prüfgröße

$$\frac{\overline{D}}{S}\sqrt{\frac{n_x n_y}{n_x + n_y}}$$

unter H_0 Student-t-verteilt mit $k = n_x + n_y - 2$ Freiheitsgraden. H_0 ist also zu verwerfen, falls

$$\frac{|\bar{d}|}{s}\sqrt{\frac{n_x n_y}{n_x + n_y}} > t_{k,1-\frac{\alpha}{2}} \,.$$

Analog konstruiert man entsprechende einseitige Tests.

Beispiel. In einer Kölner Klinik wurden im Jahr 1985 $n_x = 269$ Mädchen und $n_y = 288$ Jungen geboren, mit Durchschnittsgewicht $\overline{x} = 3050$ g bzw. $\bar{y} = 3300$ g. Die zugehörigen empirischen Standardabweichungen waren $s_x = 460$ g und $s_y = 470$ g. Es soll die Nullhypothese "Das Geburtsgewicht von Jungen und Mädchen ist *gleich*" zum Signifikanzniveau $\alpha = 0.01$ getestet werden gegen die Alternativhypothese "Das Geburtsgewicht ist verschieden". Es ist $|\bar{d}| = |\overline{x} - \bar{y}| = |3050 - 3300| = 250$, $k = n_x + n_y - 2 = 555$, $s^2 = \frac{268 \cdot 460^2 + 287 \cdot 470^2}{555} = 216409.2$, $\frac{|\overline{x}-\bar{y}|}{s}\sqrt{\frac{n_x n_y}{n_x+n_y}} = 6.34 > 2.576 = z_{0.995} \simeq t_{555,0.995}$. Die Nullhypothese wird also abgelehnt.

Bemerkung. Die Voraussetzung einer normalverteilten Grundgesamtheit kann man bei allen behandelten Tests fallenlassen, wenn die Stichprobe so groß ist, dass die zu vergleichenden Mittelwerte in guter Näherung normalverteilt sind, wie das beim GEBURTENBEISPIEL der Fall war.

6.3.2 Der Chi-Quadrat-Anpassungstest im endlichen Fall

Die bisher behandelten statistischen Tests betrafen Erwartungswerte von Verteilungen. Manche Fragen kann man anhand des Erwartungswerts allein aber nicht entscheiden. Zum Beispiel kann die erwartete Augenzahl eines gefälschten Würfels mit der eines idealen Würfels identisch sein. Um die beiden Würfel zu unterscheiden, reicht es also nicht, die Erwartungswerte zu vergleichen – man muss vielmehr die gesamte *Verteilung* untersuchen. Dies geschieht durch einen sogenannten **Anpassungstest**.

Wir betrachten nur den Fall diskreter Zufallsvariablen. Sei Y eine diskrete Zufallsvariable, die r verschiedene Werte $y_1, y_2, \ldots, y_r$ annehmen kann. (Es reicht, wenn diese nominalskaliert sind!) Die Nullhypothese behauptet, dass die Verteilung von Y durch vorgegebene Wahrscheinlichkeiten $p_1, p_2, \ldots, p_r$ beschrieben wird:

$$H_0: \text{ Es gilt } \quad \mathbb{P}(Y = y_i) = p_i \quad \text{für alle} \quad i = 1, \ldots, r.$$

Die Alternativhypothese H_A besagt, dass Y eine andere als die von der Nullhypothese behauptete Verteilung hat. Beachten Sie noch einmal, dass wir mit einem statistischen Test zunächst nur den Fehler 1. Art kontrollieren können, dass nämlich die Alternativhypothese akzeptiert wird, obwohl sie falsch ist. Also ist der χ^2-Anpassungstest eigentlich ein Test darauf, dass eine Zufallsvariable einer bestimmten Verteilung *nicht* folgt, als dass sie dieser Verteilung folgt. Wir können