Vorlesung
Einführung in die mathematische Statistik

Prof. A. Antille

Sommersemester 2004

Literatur

P.J. Bickel – K.A. Doksum, Mathematical Statistics: Basic Ideas and Selected Topics (Holden-Day, 1977).

L. Breiman, Statistics: With a View Toward Applications (Houghton Mifflin, 1973).

B.L. van der Waerden, Mathematische Statistik (Grundlehren der math. Wissenschaften, Bd 87, Springer, 1971).

H. Witting, Mathematische Statistik (Teubner, 1966).

Inhalt

I. Vorbemerkungen, Statistische Modelle, Beispiele

Im Sprachgebrauch bedeutet "Statistik" die Sammlungen von Daten, welche für den Staat wichtig sind: Steuerstatistik, Sterbestatistik, Arbeitslosenstatistik, Studentenstatistik etc.

Die mathematische Statistik hat mit Fragen dieser Art wenig oder nichts mehr zu tun.

Wesentliches Merkmal: Der Schritt vom Sammeln von Daten zum Schliessen aus Daten, bzw. zum Führen von Entscheidungen nach Sichtungen von Daten.

Aufgabe der mathematischen Statistik ist es, mathematische Modelle zu entwickeln, die es erlauben, aus zufälligen Beobachtungen Entscheidungen abzuleiten. *Die wahre Verteilung der beobachteten Zufallsgrösse ist unbekannt.*

Sei $\mathcal{X}$ eine Teilmenge von $\mathbb{R}^n$, $\mathcal{A}$ eine σ-Algebra von Teilmengen von X, Θ eine Teilmenge von $\mathbb{R}^k$.

Definition Ein statistisches Modell ist ein Tripel $\big(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta}\big)$, wobei $(P_\theta)_{\theta\in\Theta}$ eine Familie von Wahrscheinlichkeiten ist.

Interpretation: Beobachtet wird eine Zufallsgrösse X mit Werten in $\mathcal{X}$. Die Verteilung von X ist unbekannt. Sie gehört aber der Familie $(P_\theta)_{\theta\in\Theta}$ an. Aufgabe der Statistik ist es, auf Grund einer Beobachtung von X, Entscheidungen über den wahren Wert von θ, d.h. über die zugrundeliegende Verteilung von X, abzuleiten.

Θ heisst Parameterraum und $\mathcal{X}$ Beobachtungsraum oder Stichprobenraum.

Beispiel 1: Um die Qualität eines Heilverfahrens zu überprüfen, werde es auf n Personen angewandt. Dabei handle es sich jeweils um unabhängige Wiederholungen ein- und desselben Experiments, wobei nur das Eintreten oder Nichteintreten von Heilerfolg (mit einer Wahrscheinlichkeit θ, $0 \le \theta \le 1$) interessiert. Hier werden Zufallsgrössen $X_1, X_2, \ldots, X_n$ verwendet, die nur zwei Werte annehmen können, nämlich 1 (für Erfolg) und 0 (für Nichterfolg) mit den Wahrscheinlichkeiten θ bzw. $1-\theta$. Demgemäss liegt eine Zufallsgrösse $X := (X_1, \ldots, X_n)$ vor, wobei $X_1, X_2, \ldots, X_n$ unabhängig sind. Das zugrundeliegende statistische Modell ist dann $\big(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in[0,1]}\big)$, wobei

$$\mathcal{X} = \big\{x := (x_1, x_2, \ldots, x_n) : x_i \in \{0,1\}\big\}, \quad \mathcal{A} = \mathcal{P}(\mathcal{X}) \quad \text{und}$$
$$P_\theta(\{x\}) = \theta^{\sum_{i=1}^n x_i} \, (1-\theta)^{n-\sum_{i=1}^n x_i}, \, \forall\, x \in \mathcal{X}.$$

Typische Fragen: 1. Wie gross ist der wahre Wert von θ (Schätzproblem!) ?
2. Ist der wahre Wert grösser als (z.B.) 0,65 (Testproblem!) ?

Mögliche Entscheidungen für Frage 1: Alle Werte im Intervall $[0, 1]$,

Mögliche Entscheidungen für Frage 2: Ja oder nein.

Beispiel 2: n Messungen einer Länge θ ergeben $x_1, x_2, \ldots, x_n$. Die Vorstellung ist die, dass diese Werte so zustandekommen, dass zur wahren Länge θ ein jeweils unabhängiger zufälliger Messfehler hinzukommt. Der Vektor $x := (x_1, x_2, \ldots, x_n)$ kann als als *eine Beobachtung eines Zufallsvektors* $X := (X_1, \ldots, X_n)$ interpretiert werden. Ferner gilt $X_i = \theta + Z_i$,

$i = 1, 2, \ldots, n$, wobei die Zufallsgrössen (zufällige Messfehler) $Z_1, Z_2, \ldots, Z_n$ unabhängig sind.

Typische Frage: Wie gross ist die Länge? (Schätzproblem!)

Falls $E(Z_i) = 0$, ist es üblich $\bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i$, den Mittelwert der Beobachtungen $X_1, \ldots, X_n$, als Schätzer zu nehmen. Für grosse Werte von n ist dieses Schätzverfahren (Entscheidungsverfahren), wegen der Gesetze der grossen Zahlen, sicher sinnvoll. Ob man es besser machen kann, ist eine andere Frage.

Würde man die Zufallsgrössen $Z_1, Z_2, \ldots, Z_n$ normalverteilt $N(0, \sigma^2)$ (σ^2 bekannt) voraussetzen, wäre dann das zugrundeliegende statistische Modell: $\big(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta \in \mathbb{R}}\big)$, wobei $\mathcal{X} = \mathbb{R}^n$, $\mathcal{A} = \beta_{\mathbb{R}^n}$ = Borel'sche σ-Algebra und $P_\theta(A) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \int\limits_A e^{-\sum\limits_{i=1}^{n} \frac{(x_i - \theta)^2}{2\sigma^2}} \, dx_1 dx_2 \ldots dx_n$, $\forall\, A \in \mathcal{A}$.

Beispiel 3: (Schätzproblem)
$X_1, X_2, \ldots, X_n$ seien wie im Beispiel 1. Würde man *nur* $X := \sum\limits_{i=1}^{n} X_i$ beobachten, dann wäre das zugrundeliegende Modell:

$$(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta \in [0,1]}), \quad \text{wobei} \quad \mathcal{X} = \{0, 1, \ldots, n\},$$

$$\mathcal{A} = \mathcal{P}(\mathcal{X}), \quad P_\theta(\{x\}) = \tbinom{n}{x} \theta^x (1-\theta)^{n-x}, \; \forall\, x \in \mathcal{X}$$

(X ist $B(n, \theta)$ verteilt!).

Beispiel 4: (Schätzproblem)
Eine "unendlich grosse" Urne enthält θ (unbekannt) Kugeln. Die Kugeln seien von 1 bis θ durchnumeriert. n Kugeln werden der Reihe nach zufällig (mit Zurücklegen) ausgewählt. $X := (X_1, \ldots, X_n)$ sei der Vektor der beobachteten Nummer. Das entsprechende Modell ist dann $\big(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta \in \{1,2,\ldots,\}}\big)$, wobei

$$\mathcal{X} = \{x := (x_1, \ldots, x_n) : x_i \in \{1, 2, 3, \ldots\}\}, \; \mathcal{A} = \mathcal{P}(\mathcal{X})$$
$$\text{und} \quad P_\theta(\{x\}) = \tfrac{1}{\theta^n}, \; \forall\, x \in \mathcal{X} \text{ mit } x_i \in \{1, 2, \ldots, \theta\}.$$

Gesucht ist eine Schätzung für die Anzahl der Kugeln.

n sei gross. Was meinen Sie über die zwei folgenden Vorschläge?:

1. $T(X_{1,2}, \ldots, X_n) := \max\{X_1, X_2, \ldots, X_n\}$,
2. $S(X_1, X_2, \ldots, X_n) := 2\bar{X}_n - 1$, wobei $\bar{X}_n := \dfrac{1}{n} \sum\limits_{i=1}^{n} X_i$.

Begründung für den zweiten Vorschlag:
Die Zufallsgrössen $X_1, \ldots, X_n$ sind i.i.d. Also ist $\bar{X}_n \approx E(X_1) = \dfrac{\theta + 1}{2}$ wegen der Gesetze der grossen Zahlen und somit $\theta \approx 2\bar{X}_n - 1$.

Beispiel 5: (Testproblem)
Ein Angler fängt in seinem gewohnten Teich an einem Nachmittag durchschnittlich 6 Fische. Ein Freund überredet ihn, in einem anderen Teich zu angeln. Dort fängt er aber in der gleichen Zeit nur 4 Fische. Lohnt es sich für ihn, wenigstens noch einmal einen Versuch mit dem zweiten Teich zu machen?

Für diese Situation können wir folgendes mathematische Modell betrachten: Sei X die, beim zweiten Versuch, Anzahl der gefangenen Fische.

Teich 1: X ist Poisson-verteilt mit Parameter $\lambda_1 = 6$,

Teich 2: X ist Poisson-verteilt, aber mit *unbekanntem* λ_2.

Das statistische Problem liegt gerade darin, dass λ_2 unbekannt ist. Wenn λ_2 bekannt wäre, wüsste der Angler, wohin er nächsten Sonntag geht. Bekannt ist aber nur der Beobachtungswert $X = 4$, das von verschiedenen λ herrühren kann.

Wir werden später sehen (III), dass sich dieses Problem als Testproblem deuten lässt. Ein Testproblem ist folgendermassen beschaffen: Es soll *eine Entscheidung zwischen zwei Möglichkeiten* getroffen werden.

Beispiel 6: (Vertrauensintervalle)
X sei wie im Beispiel 2. Als Schätzer für die Länge haben wir $\bar{X}_n$ erwähnt. $\bar{X}_n$ ist eine Zufallsgrösse. In der Praxis ist aber die Wahrscheinlichkeit Null, dass $\bar{X}_n$ den wahren Wert liefert. Für grosse Werte von n weiss man nur, dass der wahre Wert in der Nähe von $\bar{X}_n$ liegt. Um ein Gefühl für die Güte von $\bar{X}_n$ zu haben, könnte man so verfahren: Man gibt sich eine Zahl β in der Nähe von 1 vor, z.B. $\beta = 0,99$. Dann sucht man ein um $\bar{X}_n$ symmetrisches Intervall $I(X)$, das den wahren Wert mit einer Wahrscheinlichkeit $= 0,99$ enthält (falls ein Intervall überhaupt existiert!). Ein solches Intervall heisst Vertrauensintervall vom Niveau $0,99$. Je "kleiner" $I(X)$, desto besser ist der Schätzer $\bar{X}_n$. Vertrauensintervalle werden wir im Kapitel II, § 3 besprechen.

II Schätztheorie

§ 1 Einige Schätzmethoden

$X_1, X_2, \ldots, X_n$ seien i.i.d. reelle *diskrete* Zufallsgrössen (d.h. mit Werten in einer abzählbaren Teilmenge $E = \{e_1, e_2, \ldots\}$) oder Zufallsgrössen mit einer *Dichte*. Beobachtet wird der Zufallsvektor $X := (X_1, \ldots, X_n)$. Die Verteilung von X gehöre einer Familie $(P_\theta)_{\theta \in \Theta \subseteq \mathbb{R}^k}$ von Wahrscheinlichkeiten an. $m_k(\theta)$ sei der k–te Moment von X_1 unter P_θ, d.h.

i) $m_k(\theta) := \sum_{i=1}^{\infty} e_i^k P_\theta(X_1 = e_i)$ falls $\sum_{i=1}^{\infty} |e_i|^k P_\theta(X_1 = e_i) < \infty)$ im diskreten Fall und

ii) $m_k(\theta) := \int x^k g_\theta(x) dx$ (falls $\int |x|^k g_\theta(x) dx < \infty$) im Falle, wo X_1 die Dichte g_θ besitzt.

Die Verteilung von X ist unbekannt und wir möchten sie schätzen. Da die Verteilung durch den Parameter θ eindeutig bestimmt ist, besteht die Aufgabe darin, dass man den wahren Wert θ_0 von θ schätzt.

1.1. Die Methode der Momente

Nehmen wir nun an, dass $q(\theta) = h\big(m_1(\theta), \ldots, m_r(\theta)\big)$, wobei h eine stetige Funktion ist.

Methode der Momente: Als Schätzer für $q(\theta_0)$ wählt man $T_n(X) := h\big(M_1(X), \ldots, M_r(X)\big)$, wo $M_k(X) := \frac{1}{n}\sum_{i=1}^{n} X_i^k$, d.h. man ersetzt in der Funktion h die Momente $m_k(\theta)$ durch die sogenannten empirischen Momente $M_k(X)$.

Dieses Schätzverfahren beruht auf den starken Gesetzen der grossen Zahlen: Falls $m_i(\theta_0)$ existiert, gilt $M_i(X) = M_i(X_1, \ldots, X_n) \xrightarrow[n\to\infty]{\text{f.s.}} m_i(\theta_0)$ und wegen der Stetigkeit von h,

$$T_n(X) \xrightarrow[n\to\infty]{\text{f.s.}} h\big(m_1(\theta_0), \ldots, m_r(\theta_0)\big) = q(\theta_0)\,.$$

Beispiel 1: $X_1, X_2, \ldots, X_n$ seien i.i.d. mit einer Normalverteilung $N(\mu, \sigma^2)$. Hier ist $\theta := (\mu, \sigma^2) \in \Theta = \mathbb{R} \times R_+$. Da $\theta = \big(m_1(\theta), m_2(\theta) - m_1^2(\theta)\big)$, bekommen wir als Schätzer für θ,

$$T_n(X_1, X_2, \ldots, X_n) = \big(M_1(X), M_2(X) - M_1^2(X)\big) = \big(\bar{X}_n, \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\big),$$

wobei $\bar{X}_n := \dfrac{1}{n}\sum_{!`=1}^{n} X_i$.

Beachte: $\dfrac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{X}_n)^2 = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$.

Beispiel 2: $X_1, \ldots, X_n$ seien wie im Kapitel I, Beispiel 1. Die Methode liefert der Schätzer $T_n(X) = \bar{X}_n$, denn $\theta = m_1(\theta)$.

Wäre die Grösse $\psi(\theta) := \theta(1-\theta)$ relevant, könnte man $S_n(X) := T_n(X)\big(1 - T_n(X)\big) = \bar{X}_n(1 - \bar{X}_n)$ als Schätzer von $\psi(\theta)$ vorschlagen.

Bemerkung: $\psi(\theta) = \mathrm{Var}_\theta(X_1) = E_\theta(X_1^2) - \big(E_\theta(X_1)\big)^2 = m_2(\theta) - m_1^2(\theta)$.

Also ist $T_n(X) = M_2(X) - M_1^2(X) = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$.

Beispiel 3: $X_1, X_2, \ldots, X_n$ seien wie im Kapitel I, Beispiel 2.

Es gilt $E_\theta(X_1) = \sum_{i=1}^{\theta} \dfrac{i}{\theta} = \dfrac{\theta(\theta+1)}{2\theta} = \dfrac{\theta+1}{2}$. Also ist $\theta = 2m_1(\theta) - 1$. Die Methode der Momente liefert dann den Schätzer

$$S(X_1, \ldots, X_n) := 2M_1(X) - 1 = 2\bar{X}_n - 1.$$

Beachte: Die Schätzung $S(X_1, \ldots, X_n)$ ist sinnlos, wenn $2\bar{X}_n - 1 < \max\{X_1, \ldots, X_n\}$.

1.2. Die Maximum–Likelihood Methode

A. Diskreter Fall:
$X_1, X_2, \ldots, X_n$ seien i.i.d. Zufallsgrössen mit Werten in $E = \{x_1, x_2, \ldots\}$ und möglichen Wahrscheinlichkeiten $P_\theta(\{x_i\})$, $\theta \in \Theta \subseteq \mathbb{R}^k$ (statistisches Modell!)

Die Maximum–Likelihood Methode:
Als Schätzer für θ wählt man den (einen) Wert $\hat{\theta}_n$ so, dass

$$L(X_1, X_2, \ldots, X_n; \hat{\theta}_n) = \max_{\theta \in \Theta} L(X_1, X_2, \ldots, X_n, \theta),$$

wobei

$$L(x_1, \ldots, x_n; \theta) := P_\theta(\{x_1\}) \ldots P_\theta(\{x_n\}), \, \forall (x_1, \ldots, x_n) \in E^n := \underbrace{E \times \ldots \times E}_{n-\text{mal}} .$$

Begründung: Wenn $X_1 = x_1, \ldots, X_n = x_n$ beobachtet wurden, ist die Wahrscheinlichkeit dafür

$$P_\theta(\{x_1\}) \cdot P_\theta(\{x_2\}) \ldots \cdot P_\theta(\{x_n\}) = L(x_1, x_2, \ldots, x_n; \theta).$$

Falls dieser Wert sehr klein ist bei einem θ, ist die Beobachtung unwahrscheinlich. Die Methode besteht darin, dass man als Schätzer denjenigen Wert $\hat{\theta}_n$ wählt, für welchen die Beobachtung am wahrscheinlichsten ist.

Beispiel 1: $X_1, X_2, \ldots, X_n$ seien wie im Kapitel I, Beispiel 1. In diesem Fall ist $E = \{0, 1\}$. Ferner gilt

$$L(x_1, \ldots, x_n; \theta) = \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n - \sum_{i=1}^n x_i}, \forall \, x = (x_1, \ldots, x_n) \in E^n.$$

Gesucht ist nun der Wert $\hat{\theta}_n$, für welchen $L(x_1, \ldots, x_n; \theta)$ maximal wird:

$$L(x_1, \ldots, x_n; \theta) \quad \text{maximal} \quad \Longleftrightarrow \quad \ln\big(L(x_1, \ldots, x_n; \theta)\big) \quad \text{maximal}.$$

Eine notwendige Bedingung dafür ist:

$$\frac{d \ln(L)}{d\theta} = (\sum_{i=1}^n x_i) \frac{d \ln(\theta)}{d\theta} + (n - \sum_{i=1}^n x_i) \frac{d \ln(1-\theta)}{d\theta} = (\sum_{i=1}^n x_i) \frac{1}{\theta} - (n - \sum_{i=1}^n x_i) \frac{1}{1-\theta} = 0.$$

Der Maximum-Likelihood Schätzer ist also $\hat{\theta}_n = \bar{X}_n$.

Beispiel 2: $X_1, X_2, \ldots, X_n$ seien wie im Kapitel I, Beispiel 4. Mit $E = \{1, 2, \ldots\}$ gilt $P_\theta(\{x\}) = \dfrac{1}{\theta^n}$ für alle $x = (x_1, \ldots, x_n) \in E^n$, falls $\max\{x_1, \ldots, x_n\} \leq \theta$ und $P_\theta(\{x\}) = 0$ sonst. Somit ist der Maximum-Likelihood Schätzer $\hat{\theta}_n = \max\{X_1, \ldots, X_n\}$.

B. Der Fall mit einer Dichte
$X_1, X_2, \ldots, X_n$ seien i.i.d. reelle Zufallsgrössen mit einer Dichte g_θ, wo $\theta \in \Theta \subseteq \mathbb{R}^k$. Man definiert L als

$$L(x_1, \ldots, x_n; \theta) = \prod_{i=1}^n g_\theta(x_i), \, \forall \, x = (x_1, \ldots, x_n) \in \mathbb{R}^n .$$

Maximum-Likelihood Methode: Man wählt denjenigen Wert $\hat{\theta}_n$, für welchen $L(X_1, \ldots, X_n; \theta)$ maximal wird.

Beispiel 1: $X_1, \ldots, X_n$ seien i.i.d. und normalverteilt $N(\mu, \sigma^2)$ mit $\theta := (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$. In diesem Fall ist

$$L(x_1, x_2, \ldots, x_n; \theta) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^2 e^{-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}}.$$

Gesucht ist der maximale Wert von L (als Funktion von θ): L maximal $\iff$ $\ln(L)$ maximal. Eine notwendige Bedingung dafür ist:

a) $\dfrac{\partial}{\partial \mu} \ln(L) = 0,$

b) $\dfrac{\partial}{\partial \sigma} \ln(L) = 0.$

Eine einfache Rechnung (siehe Übungen) liefert dann die Lösungen $\hat{\mu}_n = \bar{x}_n$, $\hat{\sigma}^2 = \frac{1}{n}\sum(x_i - \bar{x}_n)^2$. Der Maximum-Likelihood Schätzer $\hat{\theta}_n$ ist also

$$\hat{\theta}_n = \left(\bar{X}_n\,, \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right).$$

Bemerkung: Man sollte noch verifizieren, dass an der Stelle $\hat{\theta}_n$, $L(X_1, \ldots, X_n; \theta)$ den maximalen Wert annimmt. Dies ist aber trivial. (Warum?)

Beispiel 2: $X_1, \ldots, X_n$ seien i.i.d. mit der Dichte $g_\theta := \frac{1}{2}e^{-|x-\theta|}$, $\theta \in \mathbb{R}$.

$$L(x_1, \ldots, x_n; \theta) = \frac{1}{2^n}\prod_{i=1}^{n} e^{-|x_i-\theta|} = \frac{1}{2^n} e^{-\sum_{i=1}^{n}|x_i-\theta|}.$$

Der Maximum-Likelihood Schätzer ist also der Wert $\hat{\theta}_n$, für welchen die Summe $\sum_{i=1}^{n}|X_i - \theta|$ *minimal* wird, den sogenannten Zentralwert oder Median (siehe Übungen). Dieses Beispiel zeigt, dass der Maximum-Likelihood Schätzer nicht immer eindeutig bestimmt ist (n gerade!).

Bemerkungen: Wie die Methode der Momente (siehe 1.1., Beispiel 3, oben), kann die Maximum-Likelihood Methode zu unvernünftigen Schätzungen führen: $X_1, X_2, \ldots, X_n$, $Y_1, Y_2, \ldots, Y_n$ seien unabhängige reelle Zufallsgrössen, wobei X_k, Y_k, normalverteilt $N(\mu_k, \sigma^2)$ sind, $k = 1, \ldots, n$ (μ_k, σ^2, unbekannt). Als Schätzer für μ_k, σ^2 bekommen wir

$$\hat{\mu}_k = \frac{X_k + Y_k}{2}, \quad k = 1, 2, \ldots, n \quad \text{und} \quad \hat{\sigma}_n^2 = \frac{1}{4n}\sum_{k=1}^{n}(X_k - Y_k)^2.$$

$E_{\sigma_0}(X_k - Y_k)^2 = 2\sigma_0^2$, wobei σ_0^2 der wahre Wert von σ^2 ist. Wegen der Gesetze der grossen Zahlen konvergiert aber $\hat{\sigma}_n^2$ fast sicher gegen $\dfrac{\sigma_0^2}{2}$.

1.3. Die Methode der kleinsten Quadrate

Oft stellt sich das Problem, eine Gerade, Parabel oder eine andere "einfache" Funktion einer gegebenen Menge von Messwerten anzupassen. Z.B. kann in Abhängigkeit von einer Grösse x eine Grösse y gemessen worden sein, und nun liegen n Messpunkte $(x_1, y_1), \ldots, (x_n, y_n)$ vor. Wenn diese Punkte relativ gut auf einer Geraden liegen, kann man einen linearen Zusammenhang der beobachteten Grössen vermuten, der nur durch Messfehler z_i gestört ist. Dann wäre $y_i = \alpha + \beta x_i + z_i$ $(i = 1, \ldots, n)$.

In anderen Fällen könnte etwa *aus Naturgesetzen ein Ansatz* $y_i = \alpha + \beta x_i + \gamma x_i^2 + z_i$ geboten sein, in dem nur noch α, β, γ unbekannt sind.

Allgemeiner nehmen wir an, $\theta_1, \ldots, \theta_p$ seien unbekannte Parameter, und für *bekannte* Funktionen φ_i sei $\delta_i = \varphi_i(\theta_1, \ldots, \theta_p)$ $(i = 1, \ldots, n)$ der wahre zu messende Wert bei der i–ten Messung und $y_i = \delta_i + z_i$ der tatsächlich beobachtete Wert, also z_i der Messfehler. Im Beispiel der Geraden wäre $\theta_1 = \alpha$, $\theta_2 = \beta$ und $\varphi_i(\theta_1, \theta_2) = \theta_1 + \theta_2 x_i$.

Man fragt, welche Parameter am besten zu den y_i passen.

Methode der kleinsten Quadrate: Die Methode besagt, man solle die θ_k so bestimmen, dass $Q := \sum_{i=1}^{n} (y_i - \delta_i)^2$ minimal wird. Dies ist als ad hoc Ansatz ohne jede Statistik formulierbar und wird oft angewandt.

In dieser Vorlesung nehmen wir an, dass die z_i Realisierungen von Zufallsgrössen Z_i sind, wobei die Z_i unabhängig sind mit $E(Z_i) = 0$, $\forall\, i$. So ist $y = (y_1, \ldots, y_n)$ die Realisierung von $Y = (Y_1, Y_2, \ldots, Y_n)$ mit $Y_i = \delta_i + Z_i$.

Das allgemeine lineare Regressionsmodell

Definition: Das Regressionsmodell $Y_i = \varphi_i(\theta_1, \ldots, \theta_p) + Z_i$, $i = 1, \ldots, n$, heisst linear, falls sich $\varphi_i(\theta_1, \ldots, \theta_p)$ schreiben lässt als

$$\varphi_i(\theta_1, \ldots, \theta_p) = \sum_{j=1}^{p} x_{ij}\theta_j \quad \text{mit bekannten Zahlen} \quad x_{ij} \,.$$

In Matrixschreibweise lässt sich das lineare Modell so darstellen:

$$Y = X\theta + Z, \quad \text{wobei} \quad Y = (Y_1, \ldots, Y_n)^T, \quad \theta = (\theta_1, \ldots, \theta_p)^T$$

(C^T bedeudet die transponierte Matrix). $X = (x_{ij})$ ist die bekannte $n \times p$ Matrix.

Beachte: Die Methode der kleinsten Quadrate besteht darin, dass man den (einen) Wert $\hat{\theta}$ sucht, für welchen $Q(\hat{\theta}) = \min_{\theta \in \mathbb{R}^p} Q(\theta)$ mit

$$Q(\theta) := \|Y - X\theta\|^2 := \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} x_{ij}\theta_j \right)^2$$

(Euklidische Norm des Vektors $Y - X\theta$).

Satz: Wenn $p \leq n$ und $\text{Rang}(X) = p$, dann ist $\hat{\theta}$ die *einzige Lösung* des Gleichungssystems

$$(X^T X)\theta = X^T Y \quad \text{(Normalgleichungen)}.$$

Die Lösung lässt sich also explizit schreiben als

$$\hat{\theta} = (X^T X)^{-1} X^T Y.$$

Beweis. Für $i = 1, 2, \ldots, p$, bezeichne $\alpha_i (\in \mathbb{R}^n)$ den i–ten Spaltenvektor der Matrix X. Mit $\eta := X\theta$ gilt $Y = \eta + Z = \sum_{i=1}^{p} \theta_i \alpha_i + Z$. V_p sei der durch die Vektoren $\alpha_1, \ldots, \alpha_p$ gespannte Unterraum von $\mathbb{R}^n$, d.h.

$$V_p = \left\{ \sum_{i=1}^{p} \lambda_i \alpha_i : \lambda_i \in \mathbb{R}, \forall\, i \right\}.$$

Die Dimension von V_p ist gleich p.

$\hat{\eta}$ sei die orthogonale Projektion von Y auf V_p. Dann gilt $\|Y - \hat{\eta}\|^2 = \min_{\eta \in V_p} \|Y - \eta\|^2$. Da $\hat{\eta} \in V_p$, gibt es eindeutig bestimmte Zahlen $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p$ so, dass

$$\hat{\eta} = \sum_{i=1}^{p} \hat{\theta}_i \alpha_i \,.$$

Diese Zahlen sind die *einzigen Lösungen* der Normalgleichungen, denn

$\alpha_i^T (Y - \hat{\eta}) = \alpha_i^T (Y - X\hat{\theta}) = 0$ für $i = 1, 2, \ldots, p$ ($\hat{\eta}$ ist die orthogonale Projektion) $\Longleftrightarrow$

$X^T (Y - X\hat{\theta}) = 0 \Longleftrightarrow X^T Y = X^T X \hat{\theta}$ $\square$

Bemerkung: Um den Wert $\hat{\theta}$ zu bestimmen, hätten wir auch die partiellen Ableitungen von $Q(\theta)$ nach $\theta_1, \ldots, \theta_p$ Null setzen können. Die so erhaltenen Gleichungen sind die Normalgleichungen.

– Ein Beispiel wird in der Vorlesung angegeben.

§ 2 Verlustfunktion, Güte einer Schätzung, Optimalitätseigenschaften

Wie im § 1 wird in diesem Abschnitt ein Zufallsvektor $X = (X_1, \ldots, X_n)$ beobachtet, wobei die $\{X_j\}$ i.i.d. mit Verteilung $(P_\theta)_{\theta\in\Theta\subseteq\mathbb{R}^k}$. Zu schätzen ist der unbekannte Parameter θ oder eine reelle bekannte Funktion h von θ. $\big(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta}\big)$ bezeichne das zugrundeliegende statistische Modell.

2.1 Verlustfunktion, Güte einer Schätzung

Definitionen Eine *Schätzfunktion* oder kurz eine *Schätzung* für $h(\theta)$ ist eine Abbildung von $\mathcal{X}$ in $h(\Theta)$, wobei $h(\Theta) := \{h(\theta)\colon \theta \in \Theta\}$.

δ sei eine Schätzung für $h(\theta)$. Ihre *Risikofunktion* $R(\theta, \delta)$ ist definiert als $R(\theta, \delta) := E_\theta\big(\delta(X) - h(\theta)\big)^2$, $\theta \in \Theta$.

δ heisst *erwartungstreu* oder *biasfrei*, falls $E_\theta\big(\delta(X)\big) = h(\theta)$, $\forall\, \theta \in \Theta$.

$\delta(X) = \delta(X_1, \ldots, X_n) =: \delta_n(X)$ heisst *konsistent*, falls $P_\theta\big(|\delta_n(X) - h(\theta)| > \varepsilon\big) \underset{n\to\infty}{\longrightarrow} 0$, d.h. falls $\delta_n(X) \underset{n\to\infty}{\xrightarrow{P_\theta}} h(\theta)$, $\forall\, \theta \in \Theta$.

Die Funktion $L(u, t) := (u - t)^2$, $u, t \in h(\Theta)$ heisst Verlustfunktion. Die Risikofunktion ist also nichts anderes als $R(\delta, \theta) = E_\theta\big(L(\delta(X), h(\theta))\big)$, d.h. $R(\delta, \theta)$ ist der erwartete Verlust.

Spieltheoretische Interpretation von statistischen Entscheidungsproblemen

Der Spieler Nr. I sei der "Statistiker";
der Spieler Nr. II sei die "Natur".

Die Natur wählt einen Zustand $h(\theta)$ mit $\theta \in \Theta$.

Der Statistiker wählt eine *Strategie*, d.h. eine *Schätzfunktion* δ.

Wird $X = x$ beobachtet, dann wird die *Entscheidung* $\delta(x)$ getroffen. Der Statistiker verliert dann die Summe $L\big(\delta(x), h(\theta)\big)$.

Die *Risikofunktion* $R(\delta, \theta)$ ist also der erwartete Verlust, wenn δ die Strategie des ersten Spielers ist, und wenn der zweite Spieler den Zustand $\theta\big(h(\theta)\big)$ wählt.

Bemerkung: $R(\delta, \theta)$ ist ein *Mass* für die *Güte* der Schätzung δ. Je kleiner $R(\delta, \theta)$, desto besser ist die Strategie δ.

δ_1, δ_2 seien zwei Schätzer für $h(\theta)$.

Definitionen:
δ_1 ist *besser* als δ_2 an der Stelle θ, falls $R(\delta_1, \theta) < R(\delta_2, \theta)$.

δ_1 ist *überall besser* als δ_2, falls $R(\delta_1, \theta) < R(\delta_2, \theta)$ für alle $\theta \in \Theta$.

δ_1 ist *zulässig*, falls kein δ existiert, so dass $R(\delta, \theta) \le R(\delta_1, \theta)$, $\forall\, \theta$ mit $R(\delta, \theta) < R(\delta, \theta)$ für mindestens ein Element von Θ.

δ^* heisst minimax, falls $\sup_{\theta\in\Theta} R(\delta^*,\theta) = \min_{\delta}\sup_{\theta\in\Theta} R(\delta,\theta)$.

Beachte: Falls δ erwartungstreu für $h(\theta)$ ist, gilt

$$R(\delta,\theta) = E_\theta\big(\delta(X) - h(\theta)\big)^2 = \mathrm{Var}_\theta\big(\big(\delta(X)\big).$$

Beispiele:

1. Beispiel 1, II.1.1.
   Ist $h(\theta) = h(\mu,\sigma^2) = \mu$, haben wir den Schätzer $\delta(X) = \overline{X}_n := \frac{1}{n}\sum_{i=1} X_i$ vorgeschlagen. δ ist erwartungstreu und $R(\delta,\theta) = \mathrm{Var}_\theta\big(\big(\delta(X)\big) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}_\theta(X_i) = \frac{1}{n}\sigma^2$. Nach dem schwachen Gesetz der grossen Zahl ist $\delta_n(X) = \delta(X_1,\ldots,X_n)$ konsistent.

2. Beispiel 2, II.1.1.
   $\delta(X) = \overline{X}_n$ ist eine erwartungstreue Schätzung für die unbekannte Wahrscheinlichkeit θ. In diesem Fall gilt $R(\delta,\theta) = \mathrm{Var}_\theta(\overline{X}_n) = \frac{1}{n}\mathrm{Var}_\theta(X_1) = \frac{1}{n}\theta(1-\theta)$. Die Schätzung $\overline{X}_n$ ist konsistent.

3. Beobachtet wird der Zufallsvektor $X = (X_1,\ldots,X_n)$, wobei $X_1,\ldots,X_n$, i.i.d. Zufallsgrössen mit gleichförmiger Verteilung auf dem Intervall $[0,\theta]$, $\theta > 0$. Die Maximum-Likelihood Methode liefert den Schätzer $T(X) = \max(X_1, X_2,\ldots,X_n)$. Wir betrachten die folgenden erwartungstreuen Schätzer für θ: $\delta_1(X) := \frac{n+1}{n}T(X)$, $\delta_2(X) := 2\overline{X}_n$. In den Übungen wird man zeigen, dass

   $$\mathrm{Var}_\theta\big(\delta_1(X)\big) = \frac{\theta^2}{n(n+2)} \quad \text{und} \quad \mathrm{Var}_\theta\big(\delta_2(X)\big) = \frac{\theta^2}{3\cdot n}\,.$$

   δ_1 ist also immer eine bessere Strategie als δ_2.

Bemerkung 1: Man kann zeigen, dass δ_1 unter allen biasfreien Schätzungen (für θ) diejenige ist, die überall die kleinste Varianz hat.

Bemerkung 2: Wir werden später zeigen, dass $\delta(X)$ im ersten Beispiel 1 unter allen erwartungstreuen Schätzungen überall die kleinste Varianz hat.

Bemerkung 3: In den Beispielen 1 und 2 besitzt der Schätzer $\overline{X}_n$ wegen des Zentralgrenzwertsatzes die folgende Eigenschaft:

$$\text{Beispiel 1:}\quad P_\theta\left(a < \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} < b\right) \underset{n\to\infty}{\longrightarrow} \Phi(b) - \Phi(a),\ \forall\, a, b,$$

$$\text{Beispiel 2:}\quad P_\theta\left(a < \frac{\sqrt{n}(\overline{X}_n - \theta)}{\sqrt{\theta(1-\theta)}} < b\right) \underset{n\to\infty}{\longrightarrow} \Phi(b) - \Phi(a),\ \forall\, a, b \text{ und } 0 < \theta < 1.$$

Die Fisher Information

X sei eine Zufallsgrösse mit Werten in $E := \{e_1, e_2, \ldots\} \subseteq \mathbb{R}^n$ (diskreter Fall) oder ein Zufallsvektor mit einer Dichte.

Im diskreten Fall sei das statistische Modell $(E, \mathcal{P}(E), (P_\theta)_{\theta\in\Theta\subseteq\mathbb{R}})$ und wenn eine Dichte existiert $(\mathbb{R}^n, \beta_{\mathbb{R}^n}, (p_\theta)_{\theta\in\Theta\subseteq\mathbb{R}})$, wobei $P_\theta(\{e_i\}) := P_\theta(X = e_i)$, $i = 1, 2, \ldots$ und $p_\theta(x)$, $x \in \mathbb{R}^n$ die Dichte ist. Sehr oft existieren $\dfrac{dP_\theta}{d\theta}$ und $\dfrac{\partial}{\partial\theta}p_\theta(x)$. Nehmen wir an, es sei der Fall. Dann können wir die sogenannte *Fisher Information* definieren:

Definition: (Fisher Information)

1. $I(P_\theta) := \displaystyle\sum_{i=1}^{\infty} \left[\frac{\frac{dP_\theta}{d\theta}(\{e_i\})}{P_\theta(\{e_i\})}\right]^2 P_\theta(\{e_i\})$ (diskreter Fall),

2. $I(p_\theta) := \displaystyle\int \left[\frac{\frac{\partial}{\partial\theta}(p_\theta(x))}{p_\theta(x)}\right]^2 p_\theta(x)dx\,.$

Diese Grössen werden im folgenden Abschnitt eine wichtige Rolle spielen (siehe Cramer-Rao Ungleichung, unten).

Beispiele (für die Beweise siehe die Übungen)

i) X habe die Dichte $p_\theta(x) = \left(\dfrac{1}{\sqrt{2\pi}\delta}\right)^n e^{-\frac{\sum(x_i-\theta)^2}{2\sigma^2}}$ (σ^2 bekannt). Dann gilt $I(p_\theta) = \dfrac{1}{n\sigma^2}\,.$

ii) X habe die Verteilung $P_\theta\big(\{(x_1, x_2, \ldots, x_n)\}\big) = \dfrac{e^{-n\theta}\theta^{\sum_{i=1}^n x_i}}{x_1!x_2!\ldots x_n!}$, $(x_1, x_2, \ldots, x_n) \in \{0, 1, 2, \ldots\}^n$, $\theta > 0$. Dann gilt $I(P_\theta) = \dfrac{n}{\theta}$.

iii) Falls X die Verteilung $P_\theta\big(\{(x_1, \ldots, x_n)\}\big) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}$ mit $(x_1, \ldots, x_n) \in \{0, 1\}^n$ und $0 < \theta < 1$ besitzt, dann gilt $I(P_\theta) = \dfrac{n}{\theta(1-\theta)}\,.$

2.2. Die Cramer-Rao Ungleichung

X sei ein Zufallsvektor mit Werten in $\mathbb{R}^n$. Die *Dichte* von X gehöre einer Familie $\{p_\theta^{(x)}\}_{\theta\in\Theta}$ von Dichten an, wobei Θ eine *offene* Teilmenge von R^k ist. Zu schätzen ist eine *reelle Funktion* $h(\theta)$.

Satz 1: (Cramer-Rao Ungleichung) $T(X)$ sei eine biasfreie Schätzung für $h(\theta)$ mit $E_\theta\big(T^2(X)\big) < \infty$, $\forall\,\theta \in \Theta$.

Voraussetzungen:

1. $A_\theta := \{x : p_\theta(x) > 0\}$ hängt nicht von θ ab.
2. Die Dichte $p_\theta(x)$ ist für alle x nach θ differenzierbar $\left(p'_\theta(x) := \dfrac{\partial}{\partial\theta}\big(p_\theta(x)\big)\right)$ und es gilt

$$\int \left(\frac{p_{\theta+\Delta}(x) - p_\theta(x)}{\Delta p_\theta(x)} - \frac{p'_\theta(x)}{p_\theta(x)}\right)^2 p_\theta(x)dx \underset{\Delta\to 0}{\longrightarrow} 0,\ \forall\,\theta\in\Theta,$$

3. $0 < I(p_\theta) < \infty$, $\forall\,\theta\in\Theta$.
4. Man darf immer unter dem Integralzeichen ableiten.

Behauptung: $R(\theta,T) := E_\theta\left[\big(T(X) - h(\theta)\big)^2\right] = \mathrm{Var}_\theta\big(T(X)\big) \geq \dfrac{(h'(\theta))^2}{I(p_\theta)}$.

Beweis: Da $T(X)$ biasfrei ist, gilt für alle reelle Zahlen a,

(1) $\displaystyle\int \big(T(x) - a\big)p_{\theta+\Delta}(x)dx = h(\theta+\Delta) - a$ und

(2) $\displaystyle\int \big(T(x) - a\big)p_\theta(x)dx = h(\theta) - a\,.$

Indem man (2) von (1) substrahiert, erhält man

(3) $\displaystyle\int \big(T(x) - a\big)\big(p_{\theta+\Delta}(x) - p_\theta(x)\big)dx = h(\theta+\Delta) - h(\theta).$

Wegen Voraussetzung 1. gilt dann

(4) $\displaystyle\int \big(T(x) - a\big)\left(\frac{p_{\theta+\Delta}(x) - p_\theta(x)}{\Delta p_\theta(x)}\right)p_\theta(x)dx = \frac{h(\theta+\Delta) - h(\theta)}{\Delta}\,,\ \forall\,\Delta\neq 0.$

Ersetzt man in (4) a durch $h(\theta)$, dann erhält man (Schwarz'sche Ungleichung)

(5) $\displaystyle\left(\frac{h(\theta+\Delta) - h(\theta)}{\Delta}\right)^2 \leq \mathrm{Var}_\theta\big(T(X)\big)\cdot\int\left(\frac{p_{\theta+\Delta}(x) - p_\theta(x)}{\Delta p_\theta(x)}\right)^2 p_\theta(x)dx,\ \forall\,\Delta\neq 0.$

Lässt man Δ gegen 0 streben, bekommen wir (wegen Voraussetzungen 2. und 3.)

(6) $\big(h'(\theta)\big)^2 \le \mathrm{Var}_\theta\big(T(X)\big) I(p_\theta)$:

Wegen 2. gilt mit $w_\Delta(x) := \dfrac{p_{\theta+\Delta}(x) - p_\theta(x)}{\Delta p_\theta(x)}$, $\displaystyle\int \left(w_\Delta(x) - \frac{p'_\theta(x)}{p_\theta(x)}\right)^2 p_\theta(x)dx \underset{\Delta\to 0}{\longrightarrow} 0.$

Daraus folgt, dass $\{w_\Delta\}$ eine Cauchy-Folge ist, d.h. $\displaystyle\int \big(w_\Delta(x) - w_{\Delta'}(x)\big)^2 p_\theta(x)dx \underset{\Delta,\Delta'\to 0}{\longrightarrow} 0.$

Da $\displaystyle\int \big(T(x) - h(\theta)\big)\big(w_\Delta(x) - w_{\Delta'}(x)\big)p_\theta(x)dx = \frac{h(\theta+\Delta) - h(\theta)}{\Delta} - \frac{h(\theta+\Delta') - h(\theta)}{\Delta'}$,
gilt dann

$$\Big(\frac{h(\theta+\Delta) - h(\theta)}{\Delta} - \frac{h(\theta+\Delta') - h(\theta)}{\Delta'}\Big)^2 \le \mathrm{Var}_\theta\big(T(X)\big) \cdot \int \big(w_\Delta(x) - w_{\Delta'}(x)\big)^2 p_\theta(x)dx \underset{\Delta,\Delta'\to 0}{\longrightarrow} 0.$$

$\left\{\dfrac{h(\theta+\Delta) - h(\theta)}{\Delta}\right\}$ ist also eine Cauchy-Folge. Da $\mathbb{R}$ vollständig ist, konvergiert dann die Folge $\left\{\dfrac{h(\theta+\Delta) - h(\theta)}{\Delta}\right\}$. Der Limes ist natürlich $h'(\theta)$. □

Beachte: Voraussetzung 4. haben wir im Beweis nicht benützt. Wir werden sie aber brauchen, um den folgenden Satz zu beweisen:

Satz 2: $X = (X_1, \ldots, X_n)$ sei ein Zufallsvektor, wobei die $\{X_j\}$ i.i.d. sind mit Dichte $(g_\theta)_{\theta\in\Theta\subseteq\mathbb{R}}$. Wie oben, sei $h(\theta)$ zu schätzen. $T(X)$ sei ein biasfreier Schätzer mit $E_\theta\big(T^2(X)\big) < \infty$.

Behauptung: Falls die Familie $\{g_\theta\}$ die Voraussetzungen vom Satz 1 erfüllt, so ist das auch der Fall für die Dichten p_θ des Vektors X.
Ferner gilt: $I(p_\theta) = n\, I(g_\theta)$ und somit

$$\mathrm{Var}_\theta\big(T(X)\big) \ge \frac{\big(h'(\theta)\big)^2}{n\, I(g_\theta)}.$$

Beweis: Wir zeigen nur, dass $I(p_\theta) = n\, I(g_\theta)$:

$$\begin{aligned}
I(p_\theta) &= \int \left(\frac{p'_\theta(x)}{p_\theta(x)}\right)^2 p_\theta(x)dx = \int \left(\sum_{i=1}^n \frac{g'_\theta(x_i)}{g_\theta(x_i)}\right)^2 \prod_{i=1}^n g_\theta(x_i)dx_1dx_2\ldots dx_n \\
&= E\Big[\left(\sum_{i=1}^n \frac{g'_\theta(X_i)}{g_\theta(X_i)}\right)^2\Big] = \sum_{i=1}^n E\left[\left(\frac{g'_\theta(X_i)}{g_\theta(X_i)}\right)^2\right] + \sum_{i\ne j} E\left(\frac{g'_\theta(X_i)}{g_\theta(X_i)} \cdot \frac{g'_\theta(X_j)}{g_\theta(X_j)}\right) = n\, I(g_\theta) \\
&+ \sum_{i\ne j} E\left(\frac{g'_\theta(X_i)}{g_\theta(X_i)}\right) E\left(\frac{g'_\theta(X_j)}{g_\theta(X_j)}\right) \quad \text{(wegen der Unabhängigkeit)} \\
&= n\, I(g_\theta), \text{ denn } E\left(\frac{g'_\theta(X_i)}{g_\theta(X_i)}\right) = \int \frac{g'_\theta(x)}{g_\theta(x)} g_\theta(x)dx \\
&= \int g'_\theta(x)dx = \left(\int g_\theta(x)dx\right)' \quad \text{(wegen Voraussetzung 4!)} \\
&= 0\,.
\end{aligned}$$

□

Bemerkung: Unter denselben Voraussetzungen ist die Cramer-Rao Ungleichung auch im diskreten Fall gültig. Die Dichten (Integrale) werden einfach durch die Wahrscheinlichkeiten (Summen) ersetzt.

Beispiele:

1. $X := (X_1, \ldots, X_n)$ mit $X_1, \ldots, X_n$ i.i.d. und normalverteilt $N(\theta, \sigma_0^2)$, σ_0^2 bekannt. $\bar{X}_n$ ist erwartungstreu für θ mit $\mathrm{Var}_\theta(\bar{X}_n) = \frac{\sigma_0^2}{n}$. Nun gilt

$$I(p_\theta) = n\, I(g_\theta) \quad \text{mit} \quad g_\theta(x) = \frac{1}{\sqrt{2\pi}\sigma_0}\, e^{-\frac{(x-\theta)^2}{2\sigma_0^2}}.$$

$$\ln g_\theta(x) = -\frac{(x-\theta)^2}{2\sigma_0^2} - \ln(\sqrt{2\pi}\sigma_0) \quad \text{und} \quad \frac{g'_\theta(x)}{g_\theta(x)} = \frac{\partial}{\partial\theta} \ln g_\theta(x) = \frac{x-\theta}{\sigma_0^2}.$$

   Also ist $I(g_\theta) = \frac{1}{\sigma_0^4} \int (x-\theta)^2 g_\theta(x) dx = \frac{1}{\sigma_0^2}$ und deswegen gilt $\mathrm{Var}_\theta(\bar{X}_n) = \frac{1}{n\, I(g_\theta)}$, d.h. $\bar{X}_n$ *ist unter allen biasfreien Schätzungen* für θ, diejenige mit der *kleinsten* Varianz.

2. $X_1, X_2, \ldots, X_n$ seien i.i.d. Zufallsgrössen mit Werten in $\{0, 1\}$ und $Q_\theta(\{X_i = 1\}) = \theta$, $0 < \theta < 1$.
   $\bar{X}_n$ ist biasfrei mit $\mathrm{Var}_\theta(\bar{X}_n) = \frac{1}{n}\theta(1-\theta)$.
   $I(P_\theta) = n\, I(Q_\theta)$, wobei

$$\begin{aligned} I(Q_\theta) &= \left(\frac{Q'_\theta(\{X_i = 1\})}{Q_\theta(\{X_i = 1\})}\right)^2 Q_\theta(\{X_i = 1\}) + \left(\frac{Q'_\theta(\{X_i = 0\})}{Q_\theta(\{X_i = 0\})}\right)^2 \\ & \qquad Q_\theta(\{X_i = 0\}) = (\frac{1}{\theta})^2 \cdot \theta + (\frac{-1}{1-\theta})^2 (1-\theta) = \frac{1}{\theta(1-\theta)}. \end{aligned}$$

   Also ist $I(P_\theta) = \frac{n}{\theta(1-\theta)}$. Wir haben Gleichheit in der Ungleichung von Cramer-Rao, d.h. $\bar{X}_n$ ist unter allen linearen Schätzern derjenige mit der kleinsten Varianz.

3. $X_1, X_2, \ldots, X_n$ seien i.i.d. Zufallsgrössen mit gleichförmiger Verteilung auf dem Intervall $[0, \theta]$, $\theta > 0$. Sie haben gezeigt, dass $T(X_1, \ldots, X_n) := \frac{n+1}{n} \cdot \max\{X_1, \ldots, X_n\}$ biasfrei ist mit $\mathrm{Var}_\theta(T) = \frac{\theta^2}{n(n+2)}$. Ferner gilt $g_\theta(x) = \frac{1}{\theta}\, 1_{[0,\theta]}(x)$. Also ist $\frac{g'_\theta(x)}{g_\theta(x)} = -\frac{1}{\theta}\, 1_{[0,\theta]}(x)$ und somit

$$I(g_\theta) = \frac{1}{\theta} \int_0^\theta \frac{1}{\theta^2} dx = \frac{1}{\theta^2}.$$

   Daraus folgt

$$I(p_\theta) = \frac{n}{\theta^2} \quad \text{und} \quad \mathrm{Var}_\theta(T) < \frac{1}{I(p_\theta)}.$$

In diesem Fall ist die Cramer-Rao Ungleichung *nicht gültig*.
Beachte: die erste Voraussetzung vom Satz 1 ist nicht erfüllt. Man kann aber trotzdem zeigen, dass $T(X_1, \ldots, X_n)$ unter allen biasfreien Schätzungen die beste ist. Der Beweis ist nicht einfach und wird deshalb nicht in dieser Vorlesung vorgeführt.

2.3. Asymptotische Eigenschaften von Maximum-Likelihood Schätzungen

$X_1, X_2, \ldots$ sei eine Folge von i.i.d. reellen Zufallsgrössen mit Dichte g_θ, wobei $\theta \in \Theta \subseteq \mathbb{R}$. Man definiert die Funktion L_n als $L_n(x_1, \ldots, x_n) = \prod_{i=1}^{n} g_\theta(x_i)$ (= Dichte des Vektors $X = (X_1, \ldots, X_n)$). Nehmen wir an, die Gleichung $\sum_{i=1}^{n} \frac{g'_\theta(X_i)}{g_\theta(X_i)} = 0$ besitzt eine einzige Lösung $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ und dies für alle n. ($g'_\theta(X_i) := \frac{\partial}{\partial\theta}\big(g_\theta(X_i)\big)$)

$\hat{\theta}_n$ ist die sogenannte Maximum-Likelihood Schätzung für θ (aus der Stichprobe $(X_1, \ldots, X_n)$ hergeleitet).

Unter sehr schwachen Voraussetzungen über die möglichen Dichten g_θ kann man zeigen, dass $\hat{\theta}_n$ konsistent ist: $\hat{\theta}_n \xrightarrow[n\to\infty]{P_\theta} \theta$, d.h. falls θ der wahre Wert ist, dann konvergiert $\hat{\theta}_n$ in Wahrscheinlichkeit gegen θ.

Definition: $Y, Y_1, Y_2, \ldots$ seien Zufallsgrössen mit stetigen Verteilungsfunktionen. Die Folge $Y_1, Y_2, \ldots$ konvergiert *in Verteilung* gegen Y, falls

$$\lim_{n\to\infty} P(a < Y_n < b) = P(a < Y < b),\ \forall\, a, b.$$

$U(\theta)$ sei eine Zufallsgrösse mit Normalverteilung $N(0, \frac{1}{I(g_\theta)})$. Unter schwachen Bedingungen über $\{g_\theta\}$ kann man zeigen, dass die Folge $\left\{\sqrt{n}(\hat{\theta}_n - \theta)\right\}$ in Verteilung gegen $U(\theta)$ konvergiert, falls θ der wahre Wert ist. Dies bedeutet, dass für grosse Werte von n die Zufallsgrösse $\hat{\theta}_n - \theta$ angenähert $N(0, \frac{1}{n\,I(g_\theta)})$ verteilt ist. Grob gesagt: Asymptotisch ist die Cramer-Rao Schranke erreicht.

Beweisskizze für die asymptotische Normalität
Per Definition hat man

$$\sum_{i=1}^{n} h(\hat{\theta}_n, X_i) = 0, \quad \text{wobei} \quad h(\theta, x) := \frac{g'_\theta(x)}{g_\theta(x)}\,.$$

Ist die Funktion h nach θ differenzierbar, dann gilt, falls θ der wahre Wert ist,

$$0 = \frac{1}{n}\sum_{i=1}^{n} h(\hat{\theta}_n - \theta + \theta, X_i) \cong \frac{1}{n}\sum_{i=1}^{n} h(\theta, X_i) + \frac{1}{n}\sum_{i=1}^{n} h'(\theta, X_i)(\hat{\theta}_n - \theta)$$

($\hat{\theta}_n - \theta$ ist "klein"). Also ist

$$\sqrt{n}(\hat{\theta}_n - \theta) \cong \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} h(\theta, X_i)}{-\frac{1}{n}\sum_{i=1}^{n} h'(\theta, X_i)} =: \frac{I_n}{II_n}\,.$$

Asymptotisches Verhalten von II_n

$$h'(\theta, x) = \frac{\partial}{\partial\theta}\left(\frac{g'_\theta(x)}{g_\theta(x)}\right) = \frac{g''_\theta(x)g_\theta(x) - (g'_\theta(x))^2}{g^2_\theta(x)}\,.$$

Also gilt

$$\begin{aligned} E_\theta\big(h'(\theta, X_i)\big) &= \int \frac{g''_\theta(x) g_\theta(x)}{g^2_\theta(x)} g_\theta(x) dx - \int \left(\frac{g'_\theta(x)}{g_\theta(x)}\right)^2 g_\theta(x) dx \\ &= \int g''_\theta(x) dx - I(g_\theta) = \left(\int g_\theta(x) dx\right)'' - I(g_\theta) = -I(g_\theta). \end{aligned}$$

Nach dem schwachen Gesetz der grossen Zahlen konvergiert also II_n in Wahrscheinlichkeit gegen $-I(g_\theta)$.

Asymptotisches Verhalten von I_n
Es gilt

$$E_\theta\big(h(\theta, X_i)\big) = \int \frac{g'_\theta(x)}{g_\theta(x)} g_\theta(x) dx = \int g'_\theta(x) dx = \big(\int g_\theta(x) dx\big)' = 0$$

$$\text{und} \quad E_\theta\big(h^2(\theta, X_i)\big) = \mathrm{Var}_\theta\big(h(\theta, X_i)\big) = \int \left(\frac{g'_\theta(x)}{g_\theta(x)}\right)^2 g_\theta(x) dx = I(g_\theta).$$

Nach dem Zentralgrenzwertsatz Konvergiert I_n in Verteilung gegen eine Zufallsgrösse $U^*(\theta)$, die $N\big(0, I(g_\theta)\big)$ verteilt ist.

Aus den obigen Überlegungen folgt, dass $\sqrt{n}(\hat{\theta}_n - \theta)$ in Verteilung gegen die Zufallsgrösse $U(\theta) := \dfrac{U^*(\theta)}{-I(g_\theta)}$ konvergiert. Die letztere ist aber $N(0, \dfrac{1}{I(g_\theta)})$ verteilt. □

2.4. Einige Eigenschaften der Kleinsten-Quadrat-Schätzung (KQ-Schätzung)

Wie in 1.3. betrachten wir das allgemeine lineare Regressionsmodell:

$$Y = X\theta + Z, \quad \text{wobei } \theta \text{ der unbekannte Parameter ist} \quad (Y \in \mathbb{R}^n,\ \theta \in \mathbb{R}^p).$$

Definition Falls U eine zufällige Matrix ist, ist die Erwartung $E(U)$ von U definiert als die Matrix der Erwartungen der Elemente U_{ij} von U, d.h. $\big(E(U)\big)_{ij} := E(U_{ij})$.

In diesem Abschnitt machen wir die folgenden Voraussetzungen:

1. $p < n$,
2. $\text{Rang}(X) = p$,
3. $E(Z) = 0$ und $\text{Cov}(Z) := E\left[\big(Z - E(Z)\big)\big(Z - E(Z)\big)^T\right] = E[ZZ^T] = \sigma^2 I_n$ (I_n ist die $n \times n$ Identitätsmatrix.)

Beachte Falls die Komponenten $Z_1, Z_2, \ldots, Z_n$ von Z i.i.d. Zufallsgrössen sind mit $E(Z_i) = 0$ und $\text{Var}(Z_i) = \sigma^2$, ist die dritte Voraussetzung erfüllt. Die KQ-Schätzung $\hat{\theta}$ für θ ist

$$\hat{\theta} = (X^T X)^{-1} X^T Y \qquad \text{(siehe 1.3).}$$

Satz 1
Unter den gemachten Voraussetzungen gilt

a) $E_{\theta,\sigma^2}(\hat{\theta}) = \theta$, $\forall \theta, \sigma^2$, d.h. $\hat{\theta}$ ist biasfrei,

b) $\text{Cov}_{\theta,\sigma^2}(\hat{\theta}) = \big(X^T X\big)^{-1} \cdot \sigma^2$, $\forall \theta, \sigma^2$.

Beweis

"a)": $E_{\theta,\sigma^2}(\hat{\theta}) = E_{\theta,\sigma^2}\left((X^T X)^{-1} X^T Y\right) = (X^T X)^{-1} X^T E_{\theta,\sigma^2}(Y)$ (Linearität der Erwartung!)
$= (X^T X)^{-1} X^T E_{\theta,\sigma^2}(X\theta + Z) = (X^T X)^{-1} X^T X\theta = \theta,$

"b)": $\text{Cov}_{\theta,\sigma^2}(\hat{\theta}) = E_{\theta,\sigma^2}\left[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T\right] =$

$$E_{\theta,\sigma^2}\left[\big((X^T X)^{-1} X^T Y - \theta\big)\big((X^T X)^{-1} X^T Y - \theta\big)^T\right]$$
$$= E_{\theta,\sigma^2}\left[\big((X^T X)^{-1} X^T (Y - X\theta)\big)\big((X^T X)^{-1} X^T (Y - X\theta)\big)^T\right]$$
$$= E_{\theta,\sigma_2}\left[(X^T X)^{-1} X^T Z Z^T X (X^T X)^{-1}\right]$$
$$= (X^T X)^{-1} X^T E_{\theta,\sigma^2}(ZZ^T) X (X^T X)^{-1}$$
$$= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}. \qquad \square$$

Die Diagonalelemente der Matrix $\text{Cov}_{\theta,\sigma^2}(\hat{\theta})$ geben Information über die Güte der Schätzungen $\hat{\theta}_i$, $i = 1, \ldots, p$. Es ist also notwendig, einen Schätzer für den unbekannten Parameter σ^2 zu haben. Eine Möglichkeit ist durch den folgenden Satz gegeben:

Satz 2
$\hat{\sigma}^2 := \dfrac{\|Y - X\hat{\theta}\|^2}{n-p}$ ist eine biasfreie Schätzung für σ^2, d.h. $E_{\theta,\sigma^2}(\hat{\sigma}^2) = \sigma^2$, $\forall \theta, \sigma^2$.

Beweis Führe im y–Raum (Beobachtungsraum) ein neues orthogonales Koordinatensystem ein mit den ersten p orthonormierten Basisvektoren in dem von "idealen" Messwerten $X\theta$ aufgespannten Unterraum V_p. Seien $V_1^*, \ldots, V_n^*$ die Koordinaten des Punktes Y im neuen System. Da $V^* = \Gamma Y$ mit Γ ortogonal, gilt:

1. $\delta_{\theta,\sigma^2} := E_{\theta,\sigma^2}(V^*) = \Gamma E_{\theta,\sigma^2}(Y)$ mit $(\delta_{\theta,\sigma^2})_i = 0$ für $i > p$,
2. $\mathrm{Cov}_{\theta,\sigma^2}(V^*) = E_{\theta,\sigma^2}\left[\Gamma\big(Y - E(Y)\big)\Big(\Gamma\big(Y - E(Y)\big)\Big)^T\right]$
   $= E_{\theta,\sigma^2}\left[\Gamma Z Z^T \Gamma^T\right] = \Gamma\sigma^2 I_n \Gamma^T = \sigma^2 I_n$,
3. $\|Y - X\hat{\theta}\|^2 = \|\Gamma Y - \Gamma X\hat{\theta}\|^2$ (Γ ist orthogonal!) $= \sum_{i=p+1}^{n} V_i^{*2}$.

Daraus folgt:

$$E_{\theta,\sigma^2}(\|Y - X\hat{\theta}\|^2) = \sum_{i=p+1}^{n} E_{\theta,\sigma^2}(V_i^{*2}) = \sum_{i=p+1}^{n} \mathrm{Var}_{\theta,\sigma^2}(V_i^{*2}) \quad \text{(wegen 1.)}.$$

Also gilt

$$E_{\theta,\sigma^2}(\|Y - X\hat{\theta}\|^2) = (n-p)\sigma^2 \qquad \text{(wegen 2.)}. \qquad \square$$

Sei $\psi(\theta) := \sum_{i=1}^{p} \lambda_i \theta_i$ mit $\lambda_1, \lambda_2, \ldots, \lambda_p$ bekannt.

Definition 1 Die KQ-Schätzung $\hat{\psi}$ für ψ ist definiert als $\hat{\psi}(Y) = \sum_{i=1}^{p} \lambda_i \hat{\theta}_i$.

Definition 2 Ein Schätzer $T(Y)$ für ψ heisst linear, falls T sich schreiben lässt als

$$T(Y) = \sum_{i=1}^{n} d_i Y_i\,,$$

wobei $d_1, \ldots, d_n$ Konstanten sind.

Bemerkung Die KQ-Schätzung $\hat{\psi}$ für ψ ist linear. Es gilt weiter $E_{\theta,\sigma^2}(\hat{\psi}) = \psi(\theta)$, $\forall \theta, \sigma^2$, d.h. $\hat{\psi}$ ist biasfrei.

Satz 3 (Gauss-Markov)
$\psi(\theta) := \sum_{i=1}^{n} \lambda_i \theta_i$ sei irgend eine Linearform in den unbekannten Parametern.

Behauptung Unter allen linearen biasfreien Schätzungen für $\psi(\theta)$ ist $\hat{\psi}$ diejenige mit der kleinsten Varianz.

Beweis Wenn $\tilde{\psi} = \sum_{i=1}^{n} c_i Y_i$ irgend eine lineare Schätzung von ψ ist, dann ist $\tilde{\psi}$ auch in den v^*–Koordinaten linear:

$$\tilde{\psi} = \sum_{i=1}^{n} d_i V_i^* \,.$$

Erwartungstreue ergibt

$$\psi(\theta) = E_{\theta,\sigma^2}(\tilde{\psi}) = \sum_{i=1}^{p} d_i E_{\theta,\sigma^2}(V_i^*), \text{ denn } E_{\theta,\sigma^2}(V_i^*) = 0 \text{ für } i > p.$$

Die Beobachtungsgleichungen können auch im v^*–System ausgedrückt werden; sie lauten etwa

$$V_i^* = \sum_{j=1}^{p} x'_{ij}\theta_j + Z'_i \text{ mit } x'_{ij} = 0 \text{ für } i > p.$$

Also ist $E_{\theta,\sigma^2}(V_i^*) = \sum_{j=1}^{p} x'_{ij}\theta_j$, und Einsetzen ergibt

$$\psi(\theta) = \sum_{j=1}^{p} \lambda_j \theta_j = \sum_{i=1}^{p} d_i \sum_{j=1}^{p} x'_{ij}\theta_j = \sum_{j=1}^{p} \left(\sum_{i=1}^{p} d_i x'_{ij} \right) \theta_j \,, \; \forall \theta.$$

Koeffizientenvergleich bestimmt $d_1, \ldots, d_p$ *eindeutig* (während die $d_{p+1}, \ldots, d_n$ *beliebig* sind), denn die Matrix (x'_{ij}) hat Rang p.

Wir haben $\mathrm{Var}_{\theta,\sigma^2}(\tilde{\psi}) = \sum_{i=1}^{n} d_i^2 \; \mathrm{Var}_{\theta,\sigma^2}(V_i^*) = \sigma^2 \sum_{i=1}^{n} d_i^2$; das wird minimal, wenn wir $d_{p+1} = \ldots = d_n = 0$ setzen.

Die so bestimmte lineare erwartungstreue Schätzung kleinster Varianz $\tilde{\psi} = \sum_{k=1}^{p} d_k V_k^*$ fällt aber mit der KQ-Schätzung $\hat{\psi}$ zusammen, denn auch diese ignoriert die Werte von $V_{p+1}^*, \ldots, V_n^*$, und $d_1, \ldots, d_p$ sind durch die Erwartungstreue eindeutig bestimmt. □

Bemerkung Wenn die Messungen verschiedene Varianzen $\sigma_i^2 := \mathrm{Var}(Z_i)$ besitzen, soll man

$$Q(\theta) := \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \left(Y_i - \sum_{j=1}^{p} x_{ij}\theta_j \right)^2 \quad \text{minimalisieren.}$$

Beweis Ersetze

$$Y_i = \sum_{j=1}^{p} x_{ij}\theta_j + Z_i \text{ durch } \left(\sqrt{\frac{1}{\sigma_i^2}} Y_i \right) = \sum_{j=1}^{p} \left(\sqrt{\frac{1}{\sigma_i^2}} x_{ij} \right) \theta_j + \sqrt{\frac{1}{\sigma_i^2}} Z_i, \; i = 1, \ldots, n.$$

Beispiel (siehe Übungen.)

§ 3 Vertrauensgebiete: *Vertrauensintervalle für die Erwartung*

Im Beispiel 1 (Seite 2) haben wir als Schätzer für die Erfolgswahrscheinlichkeit θ, $\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ vorgeschlagen. Ist n hinreichend gross, wissen wir (starkes Gesetz der grossen Zahlen!), dass mit grosser Wahrscheinlichkeit $|\bar{X}_n - \theta|$ klein ist. Dies legt es nahe zu versuchen, ein kleines Intervall $I(X)$ (siehe Beispiel 6, Seite 4) um den Schätzer $\bar{X}_n$ herum festzulegen, indem man θ vermuten darf. Man könnte etwa fordern, dass z.B. P ($I(X)$ enthält den wahren Wert θ) $\geq 0,95$.

Allgemein liege ein statistisches Modell $\big(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta}\big)$ vor und es sei $g(\theta)$ ($\in \mathbb{R}$) zu schätzen. Beobachtet wird also eine Zufallsgrösse X mit Werten in $\mathcal{X}$.

Definition Ein zufälliges Gebiet $C(X)$ mit der Eigenschaft

$$P_\theta\big(C(X) \text{ enthält } g(\theta)\big) \geq 1-\alpha, \quad \forall\theta\in\Theta,$$

heisst *Vertrauensgebiet* für $g(\theta)$ *zum Niveau* $1-\alpha$.

Oft ist $C(X)$ ein zufälliges Intervall. Man spricht dann von einem *Vertrauensintervall zum Niveau* $1-\alpha$.

Es ist wichtig, sich diese Definition genau anzusehen, damit die Angabe von $C(X)$ nicht *falsch interpretiert* wird: Nicht $g(\theta)$ ist zufällig, sondern X und damit $C(X)$. Wird $X = x$ beobachtet, ist dann $C(x)$ ein festes Gebiet und es gilt: entweder $g(\theta) \in C(x)$ oder nicht, aber $\{\theta : g(\theta) \in C(x)\}$ ist *kein Ereignis*. Die Aussage über das Niveau $1-\alpha$ ist vielmehr eine Aussage über die *gesamte Familie* $\{C(x) : x \in \mathcal{X}\}$, d.h. über die Vorschrift, nach der das Gebiet aus der Beobachtung bestimmt wird. Wenn wir für *jedes* x das Gebiet $C(x)$ als Vertrauensgebiet angeben, wird — was auch immer $\theta \in \Theta$ ist — das zufällige Gebiet in ca. 95 % der Fälle $g(\theta)$ enthalten (falls $\alpha = 0,05$ ist).

Beispiel 1 $X := (X_1, X_2, \ldots, X_n)$ mit $X_1, \ldots, X_n$ i.i.d. Zufallsgrössen. Nehmen wir an, X_1 sei $N(\theta, \sigma^2)$ verteilt mit σ^2 *bekannt*.

$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ ist eine biasfreie Schätzung für θ.

Falls θ der wahre Wert ist, dann besitzt $\dfrac{\sqrt{n}(\bar{X}_n - \theta)}{\sigma}$ eine $N(0,1)$ Verteilung.

$0 < \alpha < 1$ sei vorgegeben. ξ^*_α sei diejenige Zahl, für welche $\dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{-\xi^*_\alpha}^{\xi^*_\alpha} e^{-\frac{u}{2}^2} du = 1-\alpha$. (Z.B. für $\alpha = 0,05$ ist $\xi^*_\alpha \approx 1,96$.)

Dann gilt $P_\theta\left(\left|\dfrac{\sqrt{n}(\bar{X}_n - \theta)}{\sigma}\right| \leq \xi^*_\alpha\right) = 1-\alpha$, $\forall\theta$. Also ist $I(X) := \left[\bar{X}_n - \dfrac{\sigma\xi^*_\alpha}{\sqrt{n}},\ \bar{X}_n + \dfrac{\sigma\xi^*_\alpha}{\sqrt{n}}\right]$ ein Vertrauensintervall für die Erwartung θ zum Niveau $1-\alpha$.

Beispiel 2 $X_1, X_2, \ldots, X_n$ seien i.i.d. Zufallsgrössen mit Normalverteilung $N(\theta, \sigma^2)$, wobei σ^2 *unbekannt* ist. Wir definieren $V_n^2 := \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$ ($\bar{X}_n$ wie im Beispiel 1). Falls θ der wahre Wert ist, kann man zeigen, dass $T_n := \frac{\sqrt{n}(\bar{X}_n - \theta)}{V_n}$ eine Student-Verteilung mit $n-1$ Freiheitsgraden ist. f_{n-1} sei die Dichte dieser Verteilung und $0 < \alpha < 1$ sei vorgegeben. Man bestimmt dann die Zahl $t^*_{\alpha,n-1}$, für welche $\int_{t^*_{\alpha,n-1}}^{t^*_{\alpha,n-1}} f_{n-1}(x)dx = 1 - \alpha$. (Dazu benützt man eine Tabelle für die Student-Verteilung; z.B. für $n = 7$ und $\alpha = 0,05$ ist $t^*_{0,05,\,6} = 2,365$.) Mit dieser Wahl von $t^*_{\alpha,n-1}$ gilt

$$P_\theta(|T_n| \leq t^*_{\alpha,n-1}) = 1 - \alpha, \, \forall \, \theta, \sigma^2$$

und somit ist $I(X) := \left[\bar{X}_n - \frac{V_n}{\sqrt{n}} t^*_{\alpha,n-1}, \, \bar{X}_n + \frac{V_n}{\sqrt{n}} t^*_{\alpha,n-1}\right]$ ein Vertrauensintervall für θ zum Niveau $1 - \alpha$.

Beispiel 3 $X_1, X_2, \ldots, X_n$ seien i.i.d. Zufallsgrössen mit $P_\theta(X_i = 1) = \theta$ und $P_\theta(X_i = 0) = 1 - \theta$, $0 < \theta < 1$. $\bar{X}_n$ ist eine erwartungstreue Schätzung für θ. Für hinreichend grosse n besagt der Zentralgrenzwertsatz, dass $V_n := \frac{\sqrt{n}(\bar{X}_n - \theta)}{\sqrt{\theta(1-\theta)}}$ angenähert $N(0,1)$–verteilt ist (falls θ der wahre Wert ist!). ξ^*_α sei wie im Beispiel 1 definiert. Dann gilt $P_\theta(|V_n| \leq \xi^*_\alpha) \approx 1 - \alpha$, d.h.

$$P_\theta\left(\left[\bar{X}_n - \frac{\sqrt{\theta(1-\theta)}\xi^*_\alpha}{\sqrt{n}}, \, \bar{X}_n + \frac{\sqrt{\theta(1-\theta)}\xi^*_\alpha}{\sqrt{n}}\right] \text{ enthält } \theta\right) \approx 1 - \alpha.$$

Da $\theta(1-\theta) \leq \frac{1}{4}$, $\forall \theta$, gilt

$$P_\theta\left(\left[\bar{X}_n - \frac{\xi^*_\alpha}{2\sqrt{n}}, \, \bar{X}_n + \frac{\xi^*_\alpha}{2\sqrt{n}}\right] \text{ enthält } \theta\right) \tilde{\geq} 1 - \alpha, \, \forall \theta.$$

Somit ist $I(\bar{X}_n) := \left[\bar{X}_n - \frac{\xi^*_\alpha}{2\sqrt{n}}, \, \bar{X}_n + \frac{\xi^*_\alpha}{2\sqrt{n}}\right]$ ein Vertrauensintervall zum Niveau $\tilde{\geq} 1 - \alpha$.

III Testtheorie

§1 Testverfahren, Niveau, Macht.

X sei eine Zufallsgrösse und $\big(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta}\big)$ das zugrundeliegende statistische Modell. Von einem *Testproblem* spricht man, wenn man auf grund des beobachteten Wertes x der Zufallsgrösse X entscheiden soll, ob P_θ einer *bestimmten echten Teilmenge* H von Θ angehört oder nicht.

Ein *Test* ist eine Entscheidungsregel, die für jeden möglichen Wert von x festlegt, ob man sich für die *Hypothese* "$\theta \in H$" oder für die *Alternative* "$\theta \in \Theta - H$" = "$\theta \in H^c$" entscheiden soll. Man nennt auch kurz H die Hypothese und $K := H^c$ die Alternative. Die Entscheidung d_H für die Hypothese nennt man *"Annahme" der Hypothese*, und die Entscheidung d_K für die Alternative nennt man *Verwerfen der Hypothese*. Ein Test ist also (bis auf weiteres) beschrieben durch Angabe der Menge R derjenigen x, für welche die Hypothese verworfen werden soll. R wird *Verwerfungsbereich* oder *kritischer Bereich* des Tests genannt.

Innerhalb des gewählten Modells sind also zwei Arten von Fehlern möglich:
Ist $\theta \in H$ und wird die Hypothese verworfen, so spricht man von einem *Fehler erster Art*.
Ist $\theta \in K$ und wird die Hypothese "angenommen", so spricht man von einem *Fehler zweiter Art*.

Praktisch gibt man R meist mit Hilfe von einer Funktion $\varphi(x)$ an, der sogenannten Testfunktion, die $\mathcal{X}$ in $\{0,1\}$ ($[0,1]$) abbildet:

Ist $\varphi(x) = 1$, dann lehnt man die Hypothese ab.
Ist $\varphi(x) = 0$, dann wird die Hypothese "angenommen".

Bemerkung: Die Nullhypothese ist damit nicht bewiesen; sie ist bloss nicht widerlegt. Es ist möglich, dass wir einen Fehler 2. Art mit grosser Wahrscheinlichkeit begehen. Siehe unten.

Falls φ, $\mathcal{X}$ in $[0,1]$ abbildet, dann lehnt man die Hypothese mit Wahrscheinlichkeit $\varphi(x)$ ab (falls x beobachtet wurde). Der Test heisst dann randomisiert.

Bisher haben wir das Testsproblem so formuliert, dass H und K völlig symmetrische Rollen spielen. In konkreten Fragestellungen gibt es aber gewöhnlich eine Asymmetrie. Ist man z.B. daran interessiert, ob sich irgendwelche Daten innerhalb einer etablierten Theorie erklären lassen oder auf neue Effekte hindeuten, so sollte man auf neue Effekte erst schliessen, wenn wirklich deutliche Hinweise darauf vorliegen. Soll ein gebräuchliches Medikament durch ein neues ersetzt werden, so wird man bei unklaren Vergleichswerten vorerst bei den alten Medikamenten bleiben. In beiden Fällen erscheint ein vorschneller Wechsel nicht ratsam. Im Zweifel kann man sich ja gewöhnlich weitere Daten verschaffen.

In der Formulierung des Testproblems trägt man dem so Rechnung, dass man als Hypothese die Verteilung (oder die Verteilungen) wählt, die der etablierten Theorie bzw. reiner Zufälligkeit entsprechen.

Man zieht nur Verwerfungsbereiche R (d.h. Bereich der Form $\{x : \varphi(x) = 1\}$) in Betracht, für die die Wahrscheinlichkeit eines Fehlers erster Art durch eine *vorgegebene Zahl* $\alpha > 0$ begrenzt ist. Dadurch erreicht man, dass man neue Effekte oder wesentliche Vorteile des neuen Medikamentes nur dann behauptet, wenn wirklich die Daten deutlich dafür sprechen. Leider lässt sich die Wahrscheinlichkeit eines Fehlers zweiter Art (beim festen Stichprobenumfang) nicht simultan in gleicher Weise begrenzen.

Quantitative Aussagen erhält man durch Betrachtung der Gütefunktion

$\beta(\theta) := P_\theta\big(\varphi(X) = 1\big) = E_\theta\big(\varphi(X)\big)$ des Tests φ, die jedem θ die Verwerfungswahrscheinlichkeit unter P_θ zuordnet.

Definition Wir sagen, dass der Test φ das *Niveau* α hat, falls

$$\sup_{\theta \in H} \beta(\theta) \leq \alpha.$$

Beachte $\beta(\theta) \leq \alpha$, $\forall \theta \in H$ bedeutet: Die Wahrscheinlichkeit eines Fehlers erster Art ist maximal α.

Für $\theta \in K$ heisst $\beta(\theta)$ die *Macht* des Tests in θ.

Beachte Ist die Macht $\beta(\theta)$ nahe bei 1, so ist die Wahrscheinlichkeit $1 - \beta(\theta)$ eines Fehlers zweiter Art klein.

Beispiel Angenommen, jemand behauptet, er habe eine Methode gefunden, um das Zahlenverhältnis ($\sim 1 : 1$) zwischen Kuh- und Stierkälbern zugunsten der ersten zu verschieben. Eine landwirtschaftliche Organisation ist interessiert, aber skeptisch, und möchte das Verfahren zuerst ausprobieren, z.B. in $n = 20$ Fällen; dabei kommen X Kuhkälber heraus.

$X = 20$ würde sie wohl überzeugen. Nehmen wir aber an, sie entschliesst sich, das Verfahren zu empfehlen, falls $X \geq 15$. Was bedeutet das?

X ist binomial $B(20, \theta)$ verteilt, d.h.

$$P_\theta(X = x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}, \quad x = 0, 1, \ldots, 20.$$

Hypothese H: $\theta = \frac{1}{2}$ ("Behandlung wirkungslos"),

Alternative K: $\theta \geq 0,7$ ("Behandlung wirtschaftlich interessant")

$\varphi(x) = 1$, falls $x \geq 15$ und $\varphi(x) = 0$ sonst.

In diesem Beispiel ist $\beta(\frac{1}{2})$

$$= P_{1/2}(X \geq 15) = (\frac{1}{2})^{20} \left[\binom{20}{15} + \binom{20}{16} + \ldots + \binom{20}{20} \right] \approx 0,021 \quad \text{und} \quad P_{0,7}(X \geq 15) \approx 0,416\,,$$

d.h. das Niveau des Tests φ ist gleich $0,021$ und die Macht an der Stelle $\theta = 0,7$ beträgt $0,416$.

Bemerkung Würde man nach einem Test φ^* suchen, so dass $\beta^*(\frac{1}{2}) = 0,05$ und $\beta^*(0,7) = 0,90$, so müsste die Anzahl n von Versuchen grösser als 52 sein und $\varphi^*(x) = 1$, falls $x \geq 33$. Für kleinere n geht es nicht.

§2 Einige Beispiele von wichtigen Tests

2.1 Ein einfacher Test mit Hilfe des Zentralgrenzwertsatzes

Es wird immer wieder behauptet, die Wahrscheinlichkeit einer Knabengeburt sei grösser als die Wahrscheinlichkeit einer Mädchengeburt. Ist das wirklich so?

Wir versuchen die folgende Hypothese H zu testen:

Wahrscheinlichkeit p einer Knabengeburt = 0,5. Die Alternative K sei $p > 0,5$. Wir testen also einseitig (zweiseitig würde die Alternative $K^* : p \neq 0,5$ bedeuten). Die Wahl der Alternative bedeutet, dass wir praktisch sicher sind, dass $p < 0,5$ nicht in Frage kommt.

Wir benützen als Beobachtungen die Zahlen, die ein zwischen 1969 und 1972 *zufällig* ausgewählter Jahrgang des Statistischen Jahrbuches der Schweiz liefert. Dieses nennt für 1972, $n = 91'342$ Geburten mit $x = 47'179$ Knabengeburten. Diese Zahl stellt einen Wert einer *Zufallsgrösse* X dar.

Unter der Hypothese ($p = \frac{1}{2}$) ist X, $B(91'342, \frac{1}{2})$ verteilt. Es ist aber hier einfacher mit der Normal-Approximation zu arbeiten. Wir wissen, dass (unter H) $Y := \dfrac{X - \frac{n}{2}}{\frac{1}{2}\sqrt{n}}$ angenähert eine Standard Normalverteilung besitzt. Sei α (das Niveau) = 5 % und $\xi_{0,05}$ die Zahl, für welche $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\xi_{0,05}} e^{-\frac{x^2}{2}} dx = 0.95$. Es ist naheliegend, die Hypothese zu verwerfen, falls der Beobachtete Wert x von X zu gross ist, d.h. falls $y(x)$ zu gross ist. Tut man das, falls $y(x) \geq \xi_{0,05}$ (Verwerfungsbereich), dann hat unser Test das Niveau 5 %. In diesem Beispiel ist $y \approx 10$ und aus einer Tabelle der Normalverteilung liest man $\xi_{0,05} = 1,645$.

Der Test lehnt also die Hypothese ab. Die Abweichung ist sogar hochsignifikant, denn auch für das Niveau $\alpha = 1\,\%$ ($\xi_{0,01} = 2,326$), würden wir die Hypothese verwerfen.

Eine Bemerkung über den sogenannten p–Wert (p–value) Wenn man bei einem Testproblem ein Software benützt, liefert ein Computer immer im output den sogenannten p–value. Diese Zahl wollen wir im oberen Beispiel erklären.

Die Länge der Stichprobe war $n = 91342$ und der beobachtete Wert der Zufallsgrösse X gleich $x = 47179$.

Definition:
Der p–value ist die Wahrscheinlichkeit, dass die normalisierte Zufallsgrösse $Y = Y(X)$ den beobachteten Wert $y(x)$, unter H, überschreitet. Man bezeichne diese Wahrscheinlichkeit mit $\alpha_{y(x)}$.

Interpretation:
Falls für ein *vorgegebenes* Niveau α die Ungleichung $\alpha \geq \alpha_{y(x)}$ gilt, lehnt man (zum Niveau α) die Hypothese ab.

Man kann also die jetzt bei statistischen Auswertungen von den Computern berechneten p–Werte als Entscheidungsanweisungen für den Statistiker auffassen, der α fest gewählt hat. Je nach Wahl von α wird die Anweisung zu verschiedenen Entscheidungen führen.

Achtung:
Kritisch an der Verwendung von p–Werten ist vor allem, dass sie leicht fehlinterpretiert

werden. Nicht ganz so offensichtlich ist im Falle $\alpha_{y(x)} = 0,023$ darauf zu schliessen, dass H zum Niveau $0,023$ abzulehnen ist. Das Niveau *soll ja nicht* vom Ergebnis x abhängen.

Ist die Hypothese einfach ($= \frac{1}{2}$), so ist die folgende Intepretation richtig: $\alpha_{y(x)}$ ist die Wahrscheinlichkeit (unter H) dafür, dass $Y(X) \geq y(x)$ ist, also die Wahrscheinlichkeit, dass die Teststatistik $Y(X)$ einen mindestens so grossen Wert annimmt wie den aktuell beobachteten.

2.2 Beispiel 2.1 Fortsetzung

Im Beispiel 2.1 könnte man dieselbe Hypothese $p = \frac{1}{2}$ gegen die Alternative $K^* : p \neq \frac{1}{2}$ testen (zweiseitige Situation).

Wir betrachten dieselben Zufallsgrössen X und Y wie im Beispiel 2.1 und dieselben Beobachtungen aus dem Jahre 1972. In diesem Falle ist es naheliegend, die Hypothese abzulehnen, falls Y zu gross ist ($p > \frac{1}{2}$) oder zu klein ist ($p < \frac{1}{2}$), d.h. falls $|Y|$ (absoluter Betrag) zu gross ist.

Bemerkung: Unter der Hypothese hat Y wieder angenähert eine Standard Normalverteilung, die um 0 symmetrisch ist. Sei α (das Niveau) $= 5\,\%$ vorgegeben, und sei $\xi^*_{0,05}$ die Zahl, für welche $\frac{1}{\sqrt{2\pi}} \int\limits_{-\xi^*_{0,05}}^{\xi^*_{0,05}} e^{-\frac{x^2}{2}}\, dx = 0,95$. Man lehnt dann die Hypothese ab, falls $|y| \geq \xi^*_{0,05}$. Aus einer Tabelle der Normalverteilung liest man $\xi^*_{0,05} = 1,96$. Da $|y| \cong 10$, *lehnt der Test die Hypothese ab.* Wie vorher würde man auch zum Niveau $1\,\%$ ablehnen, da $\xi^*_{0,01} = 2,576$.

Bemerkung: Im Beispiel 2.2 (wie auch im Beispiel 2.1) ist die Alternative K^* zusammengesetzt. Alle möglichen Werte p in der Menge $(0,1) - \left\{\frac{1}{2}\right\}$ sind theoretisch möglich für K^*. Die Macht des Tests hängt also von der Alternative ab. Sie wurde definiert als

$$\beta(p) := P_p\left(|Y| \geq \xi^*_{0,05}\right),$$

wenn das Niveau $5\,\%$ beträgt und $p \in K^*$.

Wenn $p \in K^*$ nahe bei der Hypothese $p = \frac{1}{2}$ liegt, steht die Macht an dieser Stelle sehr nahe bei $5\,\%$. Die Wahrscheinlichkeit eines Fehlers 2. Art ist dann in diesem Fall sehr gross und man muss sehr vorsichtig sein:
Da α (das Niveau) frei wählbar ist, hat man eine Kontrolle über die Wahrscheinlichkeit eines Fehlers 1. Art ($\leq \alpha$), ber keine über die Wahrscheinlichkeit eines Fehlers 2. Art. Deshalb sagt man, ein Test ist *signifikant*, wenn die Hypothese abgelehnt wird. Wenn nicht, ist die Hypothese *nicht bewiesen*, sie ist *einfach nicht widerlegt.*

2.3 Vergleich von zwei Wahrscheinlichkeiten bei unabhängigen Stichproben

Auf zwei verschiedenen Anlagen wird dasselbe Objekt hergestellt. Man vermutet, dass die Wahrscheinlichkeiten an Ausschussstücken bei diesen beiden Anlagen verschieden sind und will dies durch Stichproben überprüfen. $n_1 = 200$ Objekte der ersten Anlage weisen $x = 5$ fehlerhafte Stücke auf; $n_2 = 100$ Objekte der zweiten Anlage zeigen total $y = 10$ fehlerhafte Exemplare. Sind die beiden Ausschusswahrscheinlichkeiten p_1, p_2 wirklich verschieden?

Das statistische Modell für dieses Problem: Beobachtet werden $n_1 + n_2$ *unabhängige* Zufallsgrössen $X_1, X_2, \ldots, X_{n_1}$, $Y_1, Y_2, \ldots, Y_{n_2}$, wobei $X_i = 1$ ($Y_i = 1$), falls das i–te Objekt aus der ersten (zweiten) Anlage fehlerhaft ist, $= 0$ sonst. Relevant für uns sind die Zufallsgrössen

$X := \sum_{i=1}^{n_1} X_i$, $Y := \sum_{j=1}^{n_2} Y_j$, d.h. die totalen Anzahlen von fehlerhaften Objekten in beiden Fällen.

Als Hypothese wählen wir $p_1 = p_2 =: p$.

Wir müssen zunächst p schätzen. Nach den Gesetzen der grossen Zahlen, für $n_1 + n_2$ gross, ist, *unter der Hypothese*, $\hat{p} := \frac{X+Y}{n_1+n_2}$ sehr nahe bei p. Das wird unsere Schätzung für p sein.

Nach dem Zentralgrenzwertsatz ist die Differenz $\frac{X}{n_1} - \frac{Y}{n_2}$ angenähert normalverteilt mit Erwartung $p_1 - p_2 = 0$ (unter der Hypothese) und Varianz $p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})$ (unter der Hypothese).

Somit ist, im Falle, wo $p_1 = p_2 = p$,

$$U := \frac{X/n_1 - Y/n_2}{\sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}}$$

angenähert standard normalverteilt.

Dasselbe gilt auch, wenn man für p unsern Schätzer $\hat{p}$ einführt, d.h. für

$$V := \frac{X/n_1 - Y/n_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}\,.$$

Hier ist der Test zweiseitig. So, zum Niveau 5 %, lehnt man die Hypothese ab, falls der beobachtete Wert v von V so ist, dass $|v| \geq \xi^*_{0,05} = 1.96$ (siehe Beispiel 2.2).

Für v erhalten wir mit unsern Beobachtungen ($x = 5$, $y = 10$, $\hat{p} = 0,05$) den Wert $v = -2,8$.

Der Test lehnt also die Hypothese ab.

2.4 Der Vorzeichentest für kleine gepaarte Stichproben

Bei einer Person sei der diastolische Blutdruck durch P_d bezeichnet und der systolische Blutdruck durch P_s. *Der "mittlere" Blutdruck* wird dann definiert als $\frac{2}{3}P_d + \frac{1}{3}P_s$.

An 18 zufällig ausgewählten Studenten wurde der mittlere Blutdruck *zweimal gemessen*, einmal liegend und einmal stehend. Man beobachtet also 18 Paare (x_i, y_i) von Zahlen.

Das statistische Modell für dieses Experiment besteht aus 18 i.i.d. Zufallsvektoren $Z_i := (X_i, Y_i)$.

Der Vorzeichen Test: Wir betrachten die Differenzen $Z_i := Y_i - X_i$. Wenn bei der Messung die Lage des Studenten auf seinen mittleren Blutdruck keinen Einfluss hätte, würde wohl $P(Z_i > 0) = P(Z_i < 0) = \frac{1}{2}$ gelten (Wertepaare mit Differenz 0 lassen wir zum voraus weg und zählen sie nicht mit!). Man könnte aber vermuten, dass $P(z_i > 0) > \frac{1}{2}$ ist. Deshalb wählen wir als Hypothese $H : P(Z_i > 0) = P(Z_i < 0) = \frac{1}{2}$ und als Alternative $P(Z_i > 0) > \frac{1}{2}$. (Ein Test ist signifikant, nur wenn er die Hypothese ablehnt!)

Sei V die Anzahl der positiven Z_i's, d.h. $V := \sum_{i=1}^{18} I(Z_i > 0)$, wobei I die Indikatorfunktion ist. Der (einseitige) *Vorzeichen Test* lehnt die Hypothese ab, wenn der beobachtete Wert v von V zu gross ist.

Das Verfahren:
α sei vorgegeben. Man bestimmt dann die kleinste ganze Zahl c_α so, dass $P_H(V \geq c_\alpha) \leq \alpha$.
Der Vorzeichen-Test lehnt die Hypothese ab, falls v (beobachteter Wert von V) $\geq c_\alpha$.

Tabelle (beobachtete Differenzen)

Im Jahre 1975 wurden im physiologischen Institut die folgenden *Differenzen* z_i der *mittleren Blutdrucke* beobachtet:

Student	1	2	3	4	5	6	7	8	9
Differenz	$1\frac{2}{3}$	$1\frac{2}{3}$	$2\frac{2}{3}$	$4\frac{1}{3}$	$-2\frac{2}{3}$	$-3\frac{1}{3}$	$-8\frac{1}{3}$	$-1\frac{2}{3}$	$5\frac{1}{3}$
Student	10	11	12	13	14	15	16	17	18
Differenz	5	-5	$1\frac{2}{3}$	$1\frac{2}{3}$	$3\frac{1}{3}$	5	$\frac{1}{3}$	$1\frac{2}{3}$	-5

Sei $\alpha = 5\,\%$. In einer Tabelle für die Binomialverteilung $B(n,p)$ für n klein ($n \leq 40$), die man zum Beispiel im Buch von E.L. Lehmann "Nonparametrics": Statistical Methods based on ranks, Holden Day (1975), finden kann, liest man, dass die kleinste Zahl $c_{0,05}$, für welche $P_H(V \geq c_{0,0.5}) \leq 0,05$, gleich 13 ist. Unsere Stichprobe liefert für V den Wert $v = 12$. Die Hypothese wird also *nicht abgelehnt.*

Bemerkung: Wäre die Länge n der Stichprobe (in unseremBeispiel $n = 18$) viel grösser, dann würde man die Zufallsgrösse V so normalisieren, dass die Normal Approximation anwendbar ist (etwa wie im Beispiel 2.1).

2.5 Der χ^2–Anpassungstest

Das Testproblem: Es werden n unabhängige, untereinander gleiche Teilexperimente ausgeführt. Diese haben $r \geq 2$ mögliche Ausgänge und der i–te Ausgang hat Wahrscheinlichkeit p_i. Der Parameter $\theta := (p_1, p_2, \ldots, p_r)$ ist unbekannt. Wir nehmen an, dass alle p_i positiv sind. Für einen vorgegebenen Wahrscheinlichkeitsvektor $\pi := (\pi_1, \pi_2, \ldots, \pi_r)$ ist zu testen, ob $\theta = \pi$ ist.

Das zugehörige statistische Modell: Beobachtet wird ein Zufallsvektor $X := (N_1, N_2, \ldots, N_r)$, wobei N_i die Anzahl der Auftreten des i–ten Ausganges (bei den n Wiederholungen des Experimentes) darstellt.

Beachte, dass $\sum_{i=1}^{r} N_i = n$ und dass der Vektor X eine Multinomialverteilung mit Parametern n,$p_1, \ldots, p_r$ besitzt.

Beispiel: n Würfe mit einem Würfel. Mögliche Ausgänge: $\{1\}, \{2\}, \ldots, \{6\}$. Man könnte sich die folgende Frage stellen: Ist der Würfel symmetrisch, d.h. ist $(p_1, p_2, \ldots, p_6) = (\frac{1}{6}, \frac{1}{6}, \ldots, \frac{1}{6}) =: \pi$?

Zurück zum allgemeinen Testproblem. **Der χ^2–Anpassungstest**

Falls n gross ist, ist $\frac{N_i}{n}$, unter der Hypothese, nahe bei π_i (Gesetz der grossen Zahlen!). Wenn man $N_1 = n_1, \ldots, N_r = n_r$ beobachtet hat, scheint es vernünftig, die beobachtete absolute Häufigkeit n_i mit den, unter der Hypothese $\theta = \pi$, erwarteten Häufigkeiten $n\pi_i$ zu vergleichen. Man würde also die Hypothese $\theta = \pi$ ablehnen, falls z.B. $\sum_{i=1}^{r}(n_i - n\pi_i)^2$ zu gross ist. Man benützt indessen einen besonders gut brauchbaren Wert, wenn man die Quadrate der Unterschiede noch normiert:

Definition: Die χ^2–Statistik ist definiert als

$$\chi^2 = \sum_{i=1}^{r} \frac{(N_i - n\pi_i)^2}{n\pi} .$$

Definition (χ^2–Quadrat Anpassungstest)
Man kann zeigen (aber das ist schon höhere Statistik), dass für relative grosse Werte von n, etwa $n\pi_i \geq 3, \forall\, i$, die Statistik $\chi^2(X) = \chi^2(N_1, N_2, \ldots, N_r)$, unter der Hypothese $\theta = \pi$, angenähert eine χ^2–Quadrat Verteilung mit $r - 1 =$ *Anzahl der möglichen Ausgänge* -1 Freiheitsgraden besitzt.

Der χ^2–Anpassungstest: Sei α vorgegeben und h_m die Dichte der χ^2–Quadrat Verteilung mit m Freiheitsgraden, $m = 1, 2, 3, \ldots$. Man bestimmt dann die Zahl η_α so, dass $\int_0^{\eta_\alpha} h_{r-1}(x)dx = 1 - \alpha$. Der χ^2–Quadrat Anpassungstest zum Niveau α lehnt die Hypothese $\theta = \pi$ ab, falls $\chi^2(n_1, n_2, \ldots, n_r) \geq \eta_\alpha$, wobei $n_1, n_2, \ldots, n_r$ die *beobachteten* Werte von $N_1, \ldots, N_r$ sind.

Eine Anwendung
Es wird vermutet, dass bei Pferderennen auf einer kreisförmigen Rennbahn die Startpositionen einen Einfluss auf die Gewinnchancen hat. In $n = 144$ Rennen hatten die Sieger die Startpositionen $1, 2, \ldots, 8 = r$ mit den folgenden Häufigkeiten $n_1 = 29$, $n_2 = 19$, $n_3 = 18$, $n_4 = 25$, $n_5 = 17$, $n_6 = 10$, $n_7 = 15$, $n_8 = 11$. Man teste die Hypothese, dass alle Positionen die gleiche Siegwahrscheinlichkeit besitzen zum Niveau $0,05$.

Lösung: θ_i sei die Siegwahrscheinlichkeit mit Start position i. Hier ist die Hypothese $(\theta_1, \ldots, \theta_8) = (\frac{1}{8}, \ldots, \frac{1}{8})$. Die Anzahl der Freiheitsgrade beträgt $8 - 1 = 7$. Aus einer Tabelle für die χ^2–Quadrat Verteilung liest man $\eta_{0,05} = 14.07$. Hier bekommt man $\chi^2(29, 19, 18, 25, 17, 10, 15, 11) = 16,333$. Also lehnt der Test die Hypothese ab.

2.6 Der χ^2–Anpassungstest in einem komplizierteren Falle

Jemand hat 100 Messungen einer chemischen Grösse gemacht. Die Resultate seien $x_1, x_2, \ldots, x_{100}$. Da bei jeder Messung ein *zufälliger* Fehler auftritt, können die Zahlen $x_1, x_2, \ldots, x_{100}$ als n_{100} Beobachtungen einer Zufallsgrösse X betrachtet werden. Wegen des Zentralgrenzwertsatzes könnte man sich fragen, ob X *eine Normalverteilung* besitzt. *Dies wird unsere Hypothese H sein.*

Ein mögliches Verfahren, um H zu testen:

1. Die Parameter μ und σ^2, unter H, sind unbekannt. Als Schätzer für μ wählen wir $\bar{x}_{100} := \frac{1}{100}\sum_{i=1}^{100} x_i$ und für σ^2, $s^2_{100} := \frac{1}{100}\sum_{i=1}^{100}(x_i - \bar{x}_{100})^2$.

2. Nehmen wir an, $\bar{x}_{100} = 37,54$, $s_{100} = 2,81$.

3. Man wählt dann z.B. 5 Intervalle I_1, I_2, I_3, I_4, I_5 um $\bar{x}_{100}$ aus und bezeichnet mit n_k die Anzahl der x_i, die im k–ten Intervall fallen.
   Die Situation sei die folgende:

Intervalle (Klassen)	beobachtete Häufigkeiten
$I_1 = [29,5,\ 32,5]$	$n_1 = \ 4$
$I_2 = [32,5,\ 35,5]$	$n_2 = 17$
$I_3 = [35,5,\ 38,5]$	$n_3 = 43$
$I_4 = [38,5,\ 41,5]$	$n_4 = 29$
$I_5 = [41,5,\ 44,5]$	$n_5 = \ 7$

4. Y sei $N(\bar{x}_{100}, s^2_{100}) = N(37,54; (2,81)^2)$ verteilt und sei $p_i := P(Y \in I_i)$, $i = 1,2,\ldots,5$. Die, unter der Hypothese, *erwarteten* Häufigkeiten sind dann durch $100p_i$, $i = 1,\ldots,5$, gegeben.
   Wir bekommen also die folgende Tabelle (siehe die Übungen für die Bestimmung der p_i!):

Intervalle	beobachtete Häufigkeiten	p_i	erwartete Häufigkeiten
I_1	4	$0,035$	$3,5$
I_2	17	$0,196$	$19,6$
I_3	43	$0,400$	$40,0$
I_4	29	$0,288$	$28,8$
I_5	7	$0,072$	$7,2$
Totale	100	1	100

5. *Man lehnt die Hypothese ab*, falls die Chiquadrat-Statistik $X^2 := \sum_{i=1}^{5} \frac{(n_i - 100p_i)^2}{100p_i} = 0,648$ zu gross ist:

   Das vorgegebene Niveau sei α. Man bestimmt dann die Zahl η_α so, dass $\int_0^{\eta_\alpha} h_2(x)dx = 1-\alpha$ und lehnt die Hypothese ab, falls $0,648 \geq \eta_\alpha$. Wenn $\alpha = 5\,\%$, dann ist z.B. $\eta_{0,05} = 5,99$ und die Hypothese wird nicht abgelehnt.

   Bemerkung: In diesem Beispiel ist die Anzahl der Freiheitsgrade $= 5 - 1 - 2 =$ Anzahl der Intervalle (Klassen) $-1 -$ Anzahl der geschätzten Parameter (μ, σ^2!).

2.7 Der χ^2–Test als Unabhängigkeitstest

Wenn man am Montag die Zeitungen liest, so hat man oft den Eindruck, am Wochenende (Sa, So) sei der Anteil der Verkehrsunfälle mit tödlichem Ausgang, *bezogen auf die Gesamtzahl der Verkehrsunfälle*, grösser als während der Woche.

Als Hypothese nehmen wir an, der Anteil mit tödlichem Ausgang sei vom Wochentag *unabhängig*. Zum Testen ziehen wir eine Verkehrsstatistik mit $n = 135'876$ Unfällen heran.

	Anzahl Verkehrsunfälle mit tödlichem Ausgang A	Anzahl Verkehrsunfälle ohne tödlichen Ausgang A^c	Totale
Wochenende B	$n_{11} = 2'808$	$n_{12} = 45'708$	$n_{1\cdot} = 48'516$
Woche (Mo-Fr) B^c	$n_{21} = 4'680$	$n_{22} = 82'680$	$n_{2\cdot} = 87'360$
Totale	$n_{\cdot 1} = 7'488$	$n_{\cdot 2} = 128'388$	$n = 135'876$

Es liegen hier also *vier* Klassen vor, die wir in einer sogenannten Vierfeldertafel (oder 2×2 Kontingenz-Tafel) dargestellt haben.

Das zugehörige statistische Modell

Beobachtet wurde ein Zufallsvektor $(N_{11}, N_{12}, N_{21}, N_{22})$, wobei die Zufallsgrössen $N_{11}(N_{12}, N_{21}, N_{22})$ die totale Anzahl der Auftreten des Ereignisses $A \cap B(A^c \cap B, A \cap B^c, A^c \cap B^c)$ darstellt (A^c bedeutet das Komplement von A!).

Nun seien $\theta_{11} = P(A \cap B)$, $\theta_{12} = P(B \cap A^c)$, $\theta_{21} = P(A \cap B^c)$, $\theta_{22} = P(A^c \cap B^c)$, $p_1 = P(B)$, $q_1 = P(B^c)$, $p_2 = P(A)$, $q_2 = P(A^c)$. Alle diese Zahlen sind natürlich a priori unbekannt. Mann könnte sie aber mit Hilfe der Kontingenz-Tafel schätzen.

Der χ^2–Test für Unabhängigkeit

Wäre die Hypothese richtig, dann würden die Ereignisse A, A^c, B, B^c unabhängig sein. In diesem Fall würde dann das folgende gelten:

$$\theta_{11} = p_1 p_2, \quad \theta_{12} = p_1 q_2, \quad \theta_{21} = q_1 p_2, \quad \theta_{22} = q_1 q_2,$$

(C und D sind unabhängig, falls $P(C \cap D) = P(C)P(D)$!).

Beachte, dass $p_1 + q_1 = 1$, $p_2 + q_2 = 1$ gilt.

Statt vier Paramter zu schätzen, bleiben, unter der Hypothese, nur 2 zu schätzen, etwa p_1 und p_2. Nach dem schwachen Gesetz der grossen Zahlen kann p_1 (p_2) durch die relative Häufigkeit $\hat{p}_1 : \frac{n_{1\cdot}}{n} = \frac{n_{11}+N_{12}}{n}$ ($\hat{p}_2 : \frac{n_{\cdot 1}}{n} = \frac{n_{11}+n_{21}}{n}$) geschätzt werden.

Die Idee ist jetzt die folgende: Man vergleicht die beobachteten Häufigkeiten (siehe Tafel) n_{ij} mit den, unter der Hypothese, erwarteten Häufigkeiten

$$\hat{n}_{11} := \tfrac{n_{1\cdot}}{n} \cdot \tfrac{n_{\cdot 1}}{n} \cdot n, \quad \hat{n}_{12} := \tfrac{n_{1\cdot}}{n} \left(1 - \tfrac{n_{1\cdot}}{n}\right) n, \quad \hat{n}_{21} := \tfrac{n_{\cdot 1}}{n} \left(1 - \tfrac{n_{1\cdot}}{n}\right) n,$$

$$\hat{n}_{22} := \left(1 - \tfrac{n_{1\cdot}}{n}\right) \left(1 - \tfrac{n_{\cdot 1}}{n}\right) n.$$

Der χ^2–Test für Unabhängigkeit lehnt die Hypothese ab, falls

$$\chi^2(n_{11}, n_{12}, n_{21}, n_{22}) := \frac{(n_{11} - \hat{n}_{11})^2}{\hat{n}_{11}} + \frac{(n_{12} - \hat{n}_{12})^2}{\hat{n}_{12}} + \frac{(n_{21} - \hat{n}_{21})^2}{\hat{n}_{21}} + \frac{(n_{22} - \hat{n}_{22})^2}{\hat{n}_{22}}$$

zu gross ist.

Bestimmung des Ablehnungsbereichs:
Ersetzt man in der Definition der $\hat{n}_{ij}$ die Grössen $n_{1\cdot}, n_{\cdot 1}$ durch die Zufallsvariablen $N_{1\cdot}$, $N_{\cdot 2}$, dann bekommt man *Zufallsgrössen* $\hat{N}_{ij}$ für die erwarteten Häufigkeiten.

Man kann dann zeigen, dass

$$\chi^2(N_{11}, N_{12}, N_{21}, N_{22}) := \frac{(N_{11} - \hat{N}_{11})^2}{\hat{N}_{11}} + \frac{(N_{12} - \hat{N}_{12})^2}{\hat{N}_{12}} + \frac{(N_{21} - \hat{N}_{21})^2}{\hat{N}_{21}} + \frac{(N_{22} - \hat{N}_{22})^2}{\hat{N}_{22}},$$

unter der Hypothese, angenähert eine χ^2–Verteilung mit $\nu = 4 - 1 - 2 = 1 =$ Anzahl von Klassen $-1-$ Anzahl der geschätzten Parameter (p_1 und p_2!) besitzt. (Der Beweis ist nicht so einfach!)

Zurück zum Beispiel
Die boebachteten Fälle liefern

$$\chi^2(2808,\, 45708,\, 4680,\, 82680) = 10.43.$$

Als Niveau wähle man 5 %.

Analog wie im Beispiel 2.5 bestimmt man mit Hilfe einer Tabelle die Zahl $\eta_{0,05}$ so, dass $\int_0^{\eta_{0,05}} h_1(x)dx = 0,95$. Man bekommt in diesem Fall $3,84$.

Der χ^2–Test für Unabhängigkeit lehnt also zum Niveau 5 % die Hypothese ab, da

$$\chi^2(2808,\, 45708,\, 4680,\, 82680) = 10,43 > 3,84.$$

2.7 Testen eines Mittelwertes bei unbekannter Varianz: Der einseitige Student-Test

An einer Frauenklinik hat man während längerer Zeit das Geburtsgewicht der lebend und reif geborenen Mädchen bestimmt und gemittelt. Das Resultat, 3200 g $:= \mu_0$, betrachtet man als Erwartung.

Einige Jahre später, führen weitere Beobachtungen zur Vermutung, dass die Erwartung μ nicht mehr 3200 g betrage, dass aber $\mu > \mu_0$.

Die Hypothese sei $\mu = \mu_0$ und die Alternative $\mu > \mu_0$.

Um die Hypothese zu testen, will man bei den 25 nächsten Geburten von lebenden, reif geborenen Mädchen das Gewicht messen.

Das zugehörige statistische Model:
Beobachtet wird der Zufallsvektor $X = (X_1, \ldots, X_{25})$, wobei X_i das Gewicht bei i–ter Geburt ist. Man kann hier annehmen, dass die Zufallsgrössen X_i's, i.i.d. normal-verteilt $N(\mu, \sigma^2)$ sind mit unbekannten Parametern μ, σ^2.

Aus der Schätztheorie wissen wir, dass $\bar{X}_{25} := \frac{1}{25} \sum_{i=1}^{25} X_i$ und $V_{25}^2 := \frac{1}{24} \sum_{i=1}^{25} (X_i - \bar{X}_{25})^2$ sehr gute Schätzungen für μ und σ^2 sind.

Bemerkung Die emprische Varianz ist $S_n^2 := \frac{1}{25} \sum_{i=1}^{25} (X_i - \bar{X}_{25})^2$. Für V_{25}^2 hat man die Summe der Quadrate durch 24 dividiert. Der Grund dafür ist die folgende

Behauptung Unter der Hypothese $\mu = \mu_0$ hat die Statistik $T = \dfrac{\bar{X}_{25} - \mu_0}{V_{25}/5}$ genau eine Student-Verteilung mit 24 Freiheitsgraden.

(Darüber werden wir in den Übungen sprechen, aber nur für Mathematiker und Physiker!)

Wir bezeichnen mit f_m die Dichte der Student-Verteilung mit m Freiheitsgraden (siehe "Einführung in die Wahrscheinlichkeitstheorie") und, für $0 < \alpha < 1$, mit $t_{\alpha,m}$ die Zahl, für welche $\int_{-\infty}^{t_{\alpha,m}} f_m(x)dx = 1 - \alpha$.

Die Idee: Der Test von Student (einseitig) lehnt die Hypothese ab, falls der beobachtete Wert t von T zu gross ist.

Zurück zum Beispiel
Nehmen wir an, wir haben $X_1 = x_1, \ldots, X_{25} = x_{25}$ beobachtet, und das folgende erhalten: $\bar{x}_{25} = \frac{1}{25} \sum_{i=1}^{25} x_2 = 3470\,\text{g}$, $v_{25} = 408\,\text{g}$. Dann bekommen wir $t = \dfrac{\bar{x}_{25} - 3200}{408/5} = 3,31$.
α sei 5 %.
Aus einer Tabelle für die Student-Verteilung liest man $t_{0,05,\,24} = 1,711$.

Folgerung: *Der Student-Test lehnt die Hypothese ab*, da $3,31 > 1,711$.

Der Test ist sogar hoch signifikant, weil er auch zum Niveau 1 % ablehnt: $t_{0,01,24} = 2,492$.

2.8 Beispiel 2.7: Fortsetzung. Der zweiseitige Student-Test

Die Bezeichnungen sind dieselben, wie unter 2.7.
Die Hypothese ist wie oben, d.h. $\mu = \mu_0 = 3200\,\text{g}$. Wir betrachten aber jetzt als Alternative $K^* : \mu \neq \mu_0$. Für $0 < \alpha < 1$ vorgegeben, sei $t^*_{\alpha,m}$ die Zahl für welche $\int_{-t^*_{\alpha,m}}^{t^*_{\alpha,m}} f_m(x)dx = 1 - \alpha$.

Definition *Der zweiseitige Student-Test* für H gegen K^* lehnt die Hypothese zum Niveau α ab, falls

$$|t| = \frac{\bar{x}_{25} - 3200}{408/5} \geq t^*_{\alpha,\,24} \text{ ist .}$$

Wäre z.B. $\alpha = 5\,\%$, dann würde man in einer Tabelle für die Student-Verteilung mit 24 Freiheitsgraden $t^*_{0{,}05,\,24} = 2{,}06$ finden. Da $|t| = 3{,}31$, lehnt also der Student-Test die Hypothese ab. Wie vorher würde der Test die Hypothese auch zum Niveau $1\,\%$ ($t^*_{0{,}01,\,24} = 2{,}80$!) ablehnen.

2.9 Testen von zwei Mittelwerten bei unbekannter Varianz

Der einseitige (zweiseitige) Student-Test für zwei unabhängige Stichproben
Häufig stellt sich das Problem des qualitativen Vergleiches von zwei Methoden, z.B. des Vergleiches von zwei Behandlungsmethoden A und B. Man hat dann zwei Reihen von Zufallsgrössen (Messungen) $X_1, \ldots, X_{n_1}$ (Methode A) und $Y_1, \ldots Y_{n_2}$ (Methode B), die alle *unabhängig* sind. Weiter nimmt man häufig an, $X_1, \ldots, X_{n_1}$ seien $N(\mu_1, \sigma_1^2)$–verteilt und $Y_1, \ldots, Y_{n_2}$ seien $N(\mu_2, \sigma_2^2)$–verteilt.

Wir wollen hier die Hypothese $\mu_1 = \mu_2$ gegen die Alternative $K : \mu_2 > \mu_1$ (einseitiger Fall) oder die Hypothese $\mu_1 = \mu_2$ gegen die Alternative $K^* : \mu_1 \neq \mu_2$ (zweiseitiger Fall) testen.

Im folgenden werden wir annehmen, dass $\sigma_1^2 = \sigma_2^2 =: \sigma^2$ ist.
(Den Fall, wo die Varianzen verschieden sind, werden wir in der Vorlesung kurz besprechen.)

Man definiere

$$\bar{X}_{n_1} := \tfrac{1}{n_1} \sum_{i=1}^{n_1} X_i, \quad \bar{Y}_{n_2} := \tfrac{1}{n_2} \sum_{j=1}^{n_2} Y_j \quad \text{und}$$

$$V^2_{n_1+n_2} = V^2_n := \tfrac{1}{n_1+n_2-2} \left\{ \sum_{i=1}^{n_1} (X_i - \bar{X}_{n_1})^2 + \sum_{j=1}^{n_2} (Y_j - \bar{Y}_{n_2})^2 \right\}.$$

Dann kann man den folgenden Satz beweisen:

Satz (ohne Beweis)
Unter der Hypothese $\mu_1 = \mu_2$ besitzt die Statistik

$$T(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}) : \frac{\bar{X}_{n_1} - \bar{Y}_{n_2}}{V_n \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

eine Student-Verteilung mit $n_1 + n_2 - 2$ Graden.

$t_{\alpha,n}$ und $t^*_{\alpha,m}$ seien wie unter 2.7 und 2.8 definiert.

Definition Wenn man $X_1 = x_1, \ldots, X_{n_1} = x_{n_1}, Y_1 = y_1, \ldots, Y_{n_2} = y_{n_2}$ beobachtet hat, lehnt, zum Niveau α, der einseitige Student-Test (zweiseitige Student-Test) die Hypothese ab, falls

$$T(x_1, \ldots, x_{n_1}, y_1, \ldots, y_{n_2}) \geq t_{\alpha, n_1+n_2-2} \left(|T(x_1, \ldots, x_{n_2}, y_1, \ldots, y_{n_2})| \geq t^*_{\alpha, n_1+n_2-2} \right)$$

Beispiel Schweinemast mit zwei verschiedenen Futtermitteln A und B. Beobachtet wurden bei 14 zufällig ausgewählten Schweinen die Gewichtszunahme (in kg) während einer bestimmten Periode. Dabei waren 7 Schweine mit A gefüttert worden, die anderen mit B.

Hier sind die Resultate:

	Gruppe A x	Gruppe B y
1	33,17	53,77
2	66,25	53,13
3	26,08	37,75
4	43,79	73,45
5	46,22	58,25
6	55,81	61,14
7	54,50	38,80

Dann bekommen wir

$$T(x_1, \ldots, x_7, y_1, \ldots, y_7) = 1,023\,.$$

Wir haben hier 12 Freiheitsgrade für die Student-Verteilung. Aus einer Tabelle liest man, für $\alpha = 5\,\%$, $t_{0,05,\,12} = 1,782$. Also wird die Hypothese $\mu_1 = \mu_2$ gegen $\mu_2 > \mu_1$ *nicht* abgelehnt. Für den zweiseitigen Fall ($\mu_1 \neq \mu_2$) hat man $t^*_{0,05,\,12} = 2,179$. Also wird hier auch die Hypothese *nicht* abgelehnt.

2.10 Ein anderer Test zum Vergleich von zwei Mittelwerten: Der Wilcoxon-Test oder Mann-Whitney U–Test

Der Einfachheit halber betrachten wir dasselbe Problem und dasselbe Beispiel wie unter 2.9. (Der Wilcoxon-Test ist für sehr allgemeine Situationen anwendbar. Man braucht z.B. nicht wie beim Student-Test eine Normalverteilung für die Zufallsgrössen vorauszusetzen.)

Es werden also $n = n_1 + n_2$ *unabhängige* Zufallsgrössen $X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}$ mit den X_i's i.i.d. $N(\mu_1, \sigma^2)$ verteilt und den Y_j's i.i.d. $N(\mu_2, \sigma^2)$ verteilt, beobachtet.

Als Hypothese nehmen wir wie vorher $\mu_1 = \mu_2$ (es gibt also keinen Unterschied zwischen den Futtermitteln A und B!) und als Alternativen, einmal $K : \mu_2 > \mu_1$ (einseitig) und einmal $K^* : \mu_1 \neq \mu_2$ (zweiseitig).

Das Verfahren
Man ordnet alle X_i, Y_j gemeinsam der Grösse nach an. Jeder Zufallsgrösse ordnet man dann ihren Rang in der gesamten Stichprobe zu.

R_i sei der Rang von X_i, $i = 1, \ldots, n_1$.

Q_j sei der Rang von Y_j, $j = 1, \ldots, n_2$,.

Beachte: Die Ränge sind Zufallsgrössen.

U_1 (U_2) sei die Summe der Ränge der X_i (Y_j), also $U_1 := \sum_{i=1}^{n_1} R_i$, $U_2 := \sum_{j=1}^{n_2} Q_j$.

Die Idee im einseitigen Fall (zweiseitigen Fall): Man lehnt die Hypothese ab, falls der beobachtete Wert u_2 von U_2 zu gross ist (falls u_2 zu gross oder zu klein ist).

Illustration anhand des Beispiels von 2.9.

Die geordnete Stichprobe sieht so aus:

x	x	y	y	x	x	y
26,08	33,17	37,71	38,80	43,79	46,22	53,13

y	x	x	y	y	x	y
53,77	54,50	55,81	58,25	61,14	66,25	73,45

Die Ränge der y_j sind $\{3, 4, 7, 8, 11, 12, 14\}$.

Die Summe u_2 dieser Ränge ist also $u_2 = 59$.

Aus einer Tabelle für die Wilcoxon Statistik liest man, für den einseitigen Fall, dass, *unter der Hypothese*, $P_H\left(U_2 := \sum_{j=1}^{7} Q_j \geq 59\right) = 0,22789$ (siehe z.B. Lehmann "Nonparametrics": Statistical methods based on ranks).

Der Wert $\alpha_{59} = 0,22789$ ist der sogenannte p–Wert, der im Abschnitt 2.1 erklärt wurde.

Wählt man $\alpha = 5\,\%$, dann gilt $\alpha < \alpha_{59}$. Deshalb *lehnt der Wilcoxon-Test die Hypothese* $\mu_1 = \mu_2$ *nicht ab* (siehe 2.1).

Auch im zweiseitigen Fall wird die Hypothese nicht abgelehnt.

Bemerkung Für grosse Werte von n_1 und n_2 (siehe oben: Das Verfahren), normiert man die Statistik $U_2 = \sum_{j=1}^{n_2} Q_j$ so, dass, unter der Hypothese, eine Approximation durch die Standard-Normal-Verteilung möglich ist. (Siehe z.B. Lehmann "Nonparametrics": Statistical methods based on ranks.)

2.11 Vergleich zweier unabhängiger binomial-verteilter Zufallsgrössen (siehe Abschnitt 2.3)
Wir betrachten dieselbe Situation wie unter 2.3. Wie dort seien X, Y zwei unabhängige binomial-verteilte Zufallsgrössen mit Parametern n_1, p_1 bzw. n_2, p_2. Wie vorher sei die Hypothese H: $p_1 = p_2 = p$, wobei p unbekannt ist.

Der χ^2–Test für die Hypothese H
Nehmen wir an, wir haben $X = x$ und $Y = y$ beobachtet. Die Resultate können wir in einer Tafel zusammenfassen:

Anlage 1	Anlage 2	Totale
x	y	$x + y$
$n_1 - x$	$n_2 - y$	$n_1 + n_2 - x - y$
Total $= n_1$	Total $= n_2$	$n = n_1 + n_2$

Das Verfahren

1. Unter der Hypothese schätzt man p durch $\frac{x+y}{n_1+n_2} =: \hat{p}$.

2. Die erwartete Häufigkeit der fehlerhaften Stücke bei der Anlage 1 (Anlage 2) ist durch $\hat{x} := n_1\hat{p}$ ($\hat{y} := n_2\hat{p}$) gegeben.

3. Die χ^2–Statistik ist dann

$$\chi^2 := \frac{(x-\hat{x})^2}{\hat{x}} + \frac{(y-\hat{y})^2}{\hat{y}} + \frac{(n_1 - x - (n_1-\hat{x}))^2}{n_1-\hat{x}} + \frac{(n_2 - y - (n_2-\hat{y}))^2}{n_2-\hat{y}}\,.$$

4. Die Anzahl von Freiheitsgraden ist gleich $v := 4-1-1 = 2 =$ Dimension der Tafel $-1-$Anzahl der geschätzten Parameter (p!).

5. Sei $0 < \alpha < 1$ vorgegeben. Sei η_α die Zahl, für welche $\int_0^{\eta_\alpha} h_2(x)dx = 1-\alpha$.

6. Der χ^2–Test lehnt die Hypothese ab, falls $\chi^2 \geq \eta_\alpha$.

Beispiel (siehe 2.3)
$n_1 = 200$, $x = 5$, $n_2 = 100$, $y = 10$.

χ^2 ist dann gleich 7.85 und $\eta_{0,05} = 5,9991$.

Der χ^2–Test lehnt also die Hypothese ab.

Bemerkung Auf dieselbe Weise kann man den χ^2–Test benützen, um zwei unabhängige multinomial-verteilte Zufallsgrössen zu vergleichen:

Beispiel Man würfelt mit einem Würfel A, n_1–mal und mit einem Würfel B, n_2–mal. Sei $P_A(\{i\}) =: p_i =$ Wahrscheinlichkeit bei einem Wurf mit A, i zu bekommen, $i = 1,2,\ldots,6$. $q_i := P_B(\{i\})$ sei analog definiert.

Frage Gilt $p_i = q_i =: w_i$, $i = 1,2,\ldots,6$, wobei die w_i unbekannt sind? D.h. besitzen die beiden Würfel dieselben probabilistischen Eigenschaften?

Das Verfahren, um die Hypothese $H : p_i = q_i$, $\forall\, i$, zu testen:

Man definiere $n_{k1} :=$ Anzahl von k bei den n_1 Würfen mit A und $n_{k2} :=$ Anzahl von k bei den n_2 Würfen mit B. Das sind die beobachteten Häufigkeiten, $k = 1,2,\ldots,6$.

Man schätzt, unter der Hypothese, w_i durch $\hat{w}_i : \frac{n_{i1}+n_{i2}}{n_1+n_2}$, $i = 1,\ldots,6$. Die erwarteten Häufigkeiten sind dann durch $\hat{n}_{k1} := n_1\hat{w}_k$ und $\hat{n}_{k2} := n_2\hat{w}_k$ gegeben, $k = 1,2,\ldots,6$.

Die χ^2–Statistik ist dann definiert als

$$\chi^2 = \sum_{i=1}^{6} \frac{(n_{i1}-\hat{n}_{i1})^2}{\hat{n}_{i1}} + \sum_{i=1}^{6} \frac{(n_{i2}-\hat{n}_{i2})^2}{\hat{n}_{i2}}\,.$$

Die Anzahl von Freiheitsgraden ist gleich $12-1-5 = 6 =$ Dimension der Tafel (der n_{ik}, $(i = 1,\ldots,6,\ k = 1,2)$ $-1-$ Anzahl der geschätzten Parameter $(w_1, w_2, \ldots, w_5!)$.

Der χ^2–Test lehnt die Hypothese zum Niveau α ab, falls $\chi^2 \geq \eta_\alpha$, wo η_α so ist, dass $\int_0^{\eta_\alpha} h_6(x)dx = 1-\alpha$.