Computerstatistik

Herwig FRIEDL

Institut für Statistik
Technische Universität Graz

März, 2005

Inhaltsverzeichnis

Kapitel 1

Einleitung: Aspekte in der Computerstatistik

1.1 Statistik Software

Die Verwaltung großer Datenmengen mit einer oft multivariaten Struktur ist sehr schwierig. Schon einfache Berechnungen können hierbei äußerst rechenintensiv sein. Der Einsatz von komplexeren Algorithmen ist daher nur noch bei der Verwendung eines Computers mit entsprechender Statistik Software möglich. Ein graphischer Output der Resultate erfordert darüberhinaus einen Laserprinter oder Plotter.
Verschiedene Statistik-Programme sind derzeit an der TU-Graz verfügbar. Dazu gehören

- `SPSS` (Campuslizenz) mit Einsatzbereichen in der Ausbildung, der technischen und klinischen Statistik, sowie der Meinungsforschung.

- `SAS` (12 Einzelplatzlizenzen) legt großen Wert auf Datensicherheit und wird deshalb vor allem auch im klinischen Bereich eingesetzt. Wegen der doch recht komplexen Bedienung ist dieses Produkt jedoch für die Ausbildung weniger gut geeignet.

- `S-Plus` (1 Einzelplatzlizenz) ist am Institut für Statistik installiert und eignet sich speziell für Wissenschaft und Forschung. Es wird immer häufiger auch in kommerziellen Bereichen eingesetzt.

Eine *free software* Version von `S-Plus` ist `R` (GNU `S`). Es verwendet denselben Befehlssatz wie `S-Plus`, bietet auch ein GUI (**G**raphical **U**ser **I**nterface) und wird von einem internationalen Entwicklungsteam laufend betreut. Hinweise zur Installation und Informationen über `R` findet man im CRAN (**C**omprehensive **R** **A**rchive **N**etwork) unter `http://www.R-project.org/`.
`R` ist ein offenes und sehr flexibles System. Das GUI ist eine direkte Schnittstelle zur objektorientierten Sprache `S`, welche von John Chambers und anderen bei den Bell Laboratories entwickelt wurde. `R` ist matrixorientiert und offeriert die Verwendung von Scripts.

1.2 Daten

Vor jeder statistischen Analyse von konkreten Daten sollte deren Messniveau und die daraus resultierenden Gegebenheiten genau betrachtet werden.

Skala	Beispiele	Kenngrößen	skaleninvariante Transformationen	Bemerkungen
Nominalskala (Klassifikatorische Skala)	Geschlecht Blutgruppe KFZ-Kennzeichen Postleitzahl Rasse, Beruf	Häufigkeiten (i) Modalwert (i)	1-1 Trafo $(m \to 0, w \to 1)$ bijektive Trafos	Klassifizierung ohne Wertung und ohne Ordnung
Ordinalskala (Rangskala)	Schulnoten sozialer Status Ranglisten Erdbebenstärke milit. Ränge	Quantile (i) (z.B. Median) Rangstatistiken (i)	monoton steigende Transformation $A \to 1$ $B \to 2$	Ordnungs-relation Reihung
Intervallskala	Temperatur $C = \frac{5}{9}F - \frac{160}{9}$ Zeitpunkte	Mittel (i) Streuung (ni) Korrelkoeff. (i)	lineare Trafos $y = ax + b \quad (a > 0)$	Nullpunkt verändert sich bei lin. Trafos $0°C \neq 0°F$
Verhältnisskala	Gewicht Länge Fläche Kosten	Mittel (i) Streuung (ni) Korrel.koeff. (i) Variat.koeff. (i)	lineare Trafos mit festem Nullpunkt $y = ax \quad (a > 0)$	
(i) ... bleibt invariant bei Transformationen (ni)... bleibt nicht invariant bei Transformationen				

Tabelle 1.1: Messniveaus von Daten

Bemerkungen zur Intervallskala:

- Die Aussage: 10°C ist doppelt so warm wie 5°C ist unsinnig!
- Der Quotient zweier Differenzen bleibt bei einer linearen Transformation $y = T(x) = ax + b$ erhalten:

$$\frac{y_1 - y_2}{y_3 - y_4} = \frac{ax_1 + b - (ax_2 + b)}{ax_3 + b - (ax_4 + b)} = \frac{x_1 - x_2}{x_3 - x_4}\,.$$

- Der direkte Quotient von Werten ändert sich bei einer allgemeinen linearen Transformation:

$$\frac{y_1}{y_2} = \frac{ax_1 + b}{ax_2 + b} \neq \frac{x_1}{x_2} \quad \text{für} \quad b \neq 0\,.$$

- Das arithmetische Mittel ist invariant im Sinne von $\overline{y} = T(\overline{x})$, da $\overline{y} = \frac{1}{n}\sum y_i = \frac{1}{n}\sum_i (ax_i + b) = a\overline{x} + b$.
- Die Streuung jedoch ist hierbei nicht invariant, da $s_y = as_x \neq as_x + b$.

Bemerkung zur Verhältnisskala (fester Nullpunkt):

- Quotienten bleiben erhalten:

$$\frac{y_1}{y_2} = \frac{ax_1}{ax_2} = \frac{x_1}{x_2}$$

  So ist es beispielsweise egal, ob man die relative Steigerung des Umsatzes in US-Dollar (x) oder in EURO (y) betrachtet.

Qualitative (diskrete) Daten werden auf einer topologischen Skala betrachtet. Zu dieser zählen die Nominalskala (S_1) und die Ordinalskala (S_2). Quantitative (stetige) Daten werden auf einer Kardinalskala gemessen zu denen die Intervallskala (S_3) und die Verhältnisskala (S_4) gehören.
Beim Übergang $S_i \to S_{i+1}$, $i = 1, 2, 3$, entsteht zwar ein Gewinn an Information aber die Empfindlichkeit gegenüber Messfehlern steigt.

1.3 Statistische Analyse

Hat man Fragen wie *„Welche Daten liegen vor?“* (Niveau, Dimension, Kenngrößen), oder *„Welche Aspekte sollen analysiert werden?“* (Konfidenzintervalle, Hypothesentests, Modelle, Ausreißer) einmal beantwortet, so kann man sich der Auswahl von geeigneten Analysemethoden widmen.
Liegt ein **eindimensionaler (univariater) Datensatz** vor, so verwendet man die Methoden der explorativen Datenanalyse (EDA) oder der graphischen Datenanalyse (GDA). Man kann auch einen Vergleich mit theoretischen Verteilungsmodellen anstellen, Konfidenzintervalle berechnen oder Hypothesentests durchführen. Eine spezielle Situation liegt beim Vergleich von eindimensionalen Stichproben vor.
Soll ein **zweidimensionaler (bivariater) Datensatz** analysiert werden, dann ist wiederum die EDA (Scatterplot mit Glättung, Korrelationsanalyse) geeignet. Ein potentieller funktionaler Zusammenhang wird mittels der Regressionsanalyse mit anschließender Betrachtung der Residuen validiert. Lokations- oder Variabiliätstests zweier abhängiger oder unabhängiger Stichproben sind Methoden, um auf Mittelwerts- oder Streuungsunterschiede zu testen.
Bei der Analyse eines **mehrdimensionalen (multivariaten) Datensatzes** helfen auch graphische Darstellungsmethoden. Die multiple lineare Regression liefert ein einfaches statistisches Modell, welches den Zusammenhang zwischen einer abhängigen und mehreren unabhängigen Variablen beschreibt. Die Diskriminanzanalyse dient zur Modellierung der Trennfunktion zwischen mehreren „als verschieden bekannte“ Datenkollektiven, während die MDS (multidimensionale Skalierung) auf Distanzmatrizen basiert und zur Dimensionsreduktion von multivariaten Daten auf eine niedrigere Dimension eingesetzt wird. Ähnliches bietet die Clusteranalyse, welche zum Auffinden von Datenbündel oder Strukturen zwischen den einzelnen Variablen verwendet wird. Durch faktorenanalytische Methoden werden hypothetische Faktoren aus einer Menge beobachteter Variablen über Kovarianzbetrachtungen berechnet (Hauptanwendung in der Psychologie – IQ-Forschung).

Kapitel 2

Univariate Stichproben

Definition 2.1 *Der n-elementige Zufallsvektor* $X = (X_1, \ldots, X_n)$ *heißt* **Stichprobe** *vom Umfang n für die Population* Y*, wenn die Zufallsvariablen* $X_1, \ldots, X_n$ *stochastisch unabhängig und so wie* Y *verteilt sind. Den reellwertigen Vektor* $x = (x_1, \ldots, x_n)$ *nennt man* **Realisation** *dieser Stichprobe.*

2.1 Parameterschätzer

Wir nehmen nun an, dass die Verteilungsfunktion $F(y|\theta)$ von Y bis auf den Parameter(vektor) θ bekannt ist. Der unbekannte Parameter θ wird nun mit Hilfe der Stichprobe geschätzt.

Definition 2.2 *Schätzt die Stichprobenfunktion* $T = T(X_1, \ldots, X_n)$ *einen unbekannten Parameter* θ*, so heißt* T **Schätzfunktion** *oder* **Schätzer** *und die entsprechende Realisation* $T(x_1, \ldots, x_n)$ **Schätzwert**.

Es gibt mehrere **Gütekriterien** für einen Schätzer T von θ. Dazu zählen

- **Erwartungstreue**:

$$\mathrm{E}(T) = \theta\,,$$

  sonst definiert $\mathrm{bias}(T, \theta) = \mathrm{E}(T) - \theta$ den **Bias** (vergleiche mit Abbildung 2.1 links).

- **Asymptotische Erwartungstreue**:

$$\lim_{n\to\infty} \mathrm{E}(T) = \theta\,.$$

- **Konsistenz**:

$$\lim_{n\to\infty} P(|T - \theta| > \epsilon) = 0\,.$$

  Dies entspricht gerade der stochastischen Konvergenz von T gegen θ.

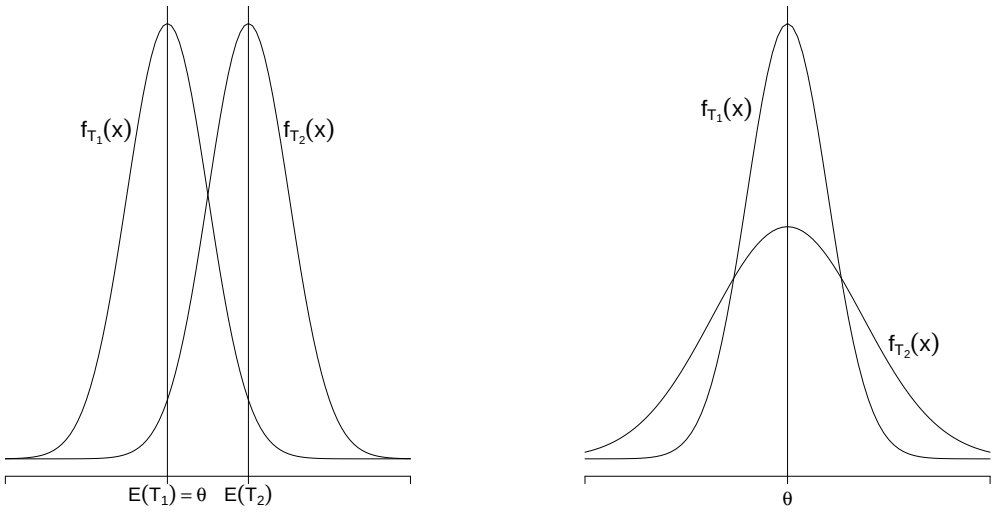


Abbildung 2.1: Vergleich von Dichten zweier Punktschätzer für den Parameter θ. Links: T_1 erwartungstreu, T_2 verzerrt; Rechts: T_1 effizienter als T_2.

- T_1 ist **wirksamster oder effektivster** Schätzer, wenn für alle anderen Schätzer T_2

$$\mathrm{E}((T_1 - \theta)^2) \leq \mathrm{E}((T_2 - \theta)^2)\,,$$

  gilt. D.h., T_1 hat den kleinsten mittleren quadratischen Fehler (MSE) (siehe Abbildung 2.1 rechts). Unter allen erwartungstreuen Schätzer hat der wirksamste Schätzer T_1 wegen $\mathrm{E}((T_1 - \mathrm{E}(T_1))^2) \leq \mathrm{E}((T_2 - \mathrm{E}(T_2))^2)$ die kleinste Varianz.

Zur Konstruktion von Schätzern:

- Die *Momentenmethode* liefert erwartungstreue und konsistente Schätzer für die Momente um Null der Ordnung k.
- Die *Maximum-Likelihood (ML) Methode* liefert keinesfalls immer erwartungstreue Schätzer. Existiert jedoch ein effektiver Schätzer, so wird dieser durch die ML Methode bestimmt.

Momente und Quantile:
1. Moment um Null:

$$\mathrm{E}(X) = \mu = \int x dF(x) dx\,.$$

Das empirische Mittel $\overline{X} = n^{-1} \sum_i X_i$ ist ein erwartungstreuer Schätzer für μ. Weiters ist $\overline{X}$ konsistent und effektiv. Für $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$ folgt $\overline{X} \sim N\,(\mu, \sigma^2/n)$.

2. zentrales Moment:

$$\mathrm{var}(X) = \sigma^2 = \mathrm{E}((X - \mu)^2)\,.$$

Die empirische Varianz $S^2 = (n-1)^{-1}\sum_i (X_i - \overline{X})^2$ ist erwartungstreuer und konsistenter Schätzer für den Varianzparameter σ^2. Der Schätzer $S_1^2 = n^{-1}\sum_i (X_i - \overline{X})^2$ ist effektivster Schätzer, hat aber bias$(S_1^2, \sigma^2) = -\sigma^2/n$.
Gilt $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$, so sind $\overline{X}$ und S^2 unabhängig und es gilt

$$\frac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1}, \quad \text{sowie} \quad \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}\,.$$

k-tes zentrales Moment:

$$\mu_k = \mathrm{E}((X-\mu)^k)\,.$$

Als standardisierte Formen der höheren zentralen Momente haben wir

- Schiefe (Skewness) $\alpha_3 = \mathrm{E}((X-\mu)/\sigma)^3 = \mu_3/\sigma^3$, welche durch

  $$\hat{\alpha}_3 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^3 / S_1^3$$

  mit var$(\hat{\alpha}_3) \approx 6/n$ geschätzt werden kann. Ist α_3 negativ (positiv) so liegt eine linksschiefe (rechtsschiefe) Verteilung vor. Für $\alpha_3 = 0$ ist die Verteilung symmetrisch.

- Kurtosis (Exzeß, Schwänzigkeit) $\alpha_4 = \mathrm{E}((X-\mu)/\sigma)^4 - 3 = \mu_4/\sigma^4 - 3$, geschätzt durch

  $$\hat{\alpha}_4 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^4 / S_1^4 - 3$$

  mit var$(\hat{\alpha}_4) \approx 24/n$. Ist α_4 negativ (positiv), so ist bei gleicher Varianz das Maximum der Dichte kleiner (größer) als bei der Normalverteilung.

In der Tabelle 2.1 sind die ersten vier Momente einiger Verteilungen angegeben. Das Verhalten der entsprechenden Dichten ist in den Abbildungen 2.2 und 2.3 dargestellt.

Verteilung	E(X)	var(X)	α_3	α_4
Normal$(0,1)$	0	1	0	0
Uniform$(-1,1)$	0	1/3	0	-1.2
Doppel-Exponential(1)	0	2	0	3
Exponential(1)	1	1	2	6
χ^2_{10}	10	20	0.894	1.2

Tabelle 2.1: Momente ausgewählter Verteilungen.

Für die Definition eines Quantils benötigt man den Begriff der geordneten Stichprobe.

Abbildung 2.2: Beispiele für die Schiefe α_3.

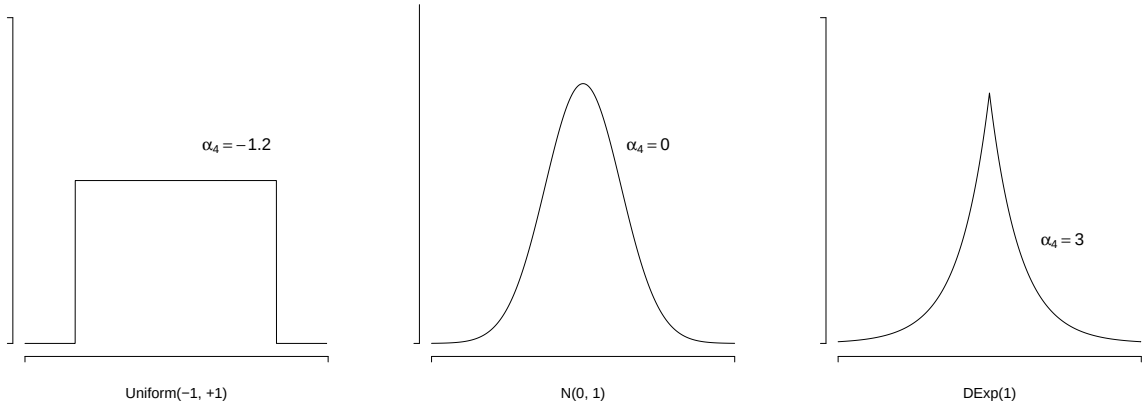


Abbildung 2.3: Beispiele für die Kurtosis α_4.

Definition 2.3 *Bezeichne* $g\colon \mathbb{R}^n \to \mathbb{R}^n$ *die Abbildung* $g(x_1, \ldots, x_n) = (x_{(1)}, \ldots, x_{(n)})$ *mit* $x_{(1)} \leq \cdots \leq x_{(n)}$. *Dann heißt* $x_{(.)} = (x_{(1)}, \ldots, x_{(n)})$ **geordnete Stichprobe** *zu* $x = (x_1, \ldots, x_n)$, $X_{(.)} = (X_{(1)}, \ldots, X_{(n)})$ *die geordnete Statistik (***Ordnungsstatistik***) und* $X_{(i)}$ *die i-te geordnete Statistik.*

Definition 2.4 *Die feste Zahl* x_p, $0 < p < 1$, *mit*

$$P(X < x_p) \leq p \leq P(X \leq x_p)$$

nennt man p-tes **theoretisches Quantil** *der Zufallsvariablen X (siehe Abbildung 2.4). Die Zufallsvariable*

$$Q(p) = \begin{cases} \frac{1}{2}(X_{(np)} + X_{(np+1)}) & \textit{für ganzzahliges } np \\ X_{([np]+1)} & \textit{sonst} \end{cases}$$

nennt man **empirisches Quantil** *oder Stichprobenquantil.*

Empirische Quantile werden als Schätzer für die (unbekannten) theoretischen Quantile verwendet. Aussagen über deren Güte liefert der folgende Satz.

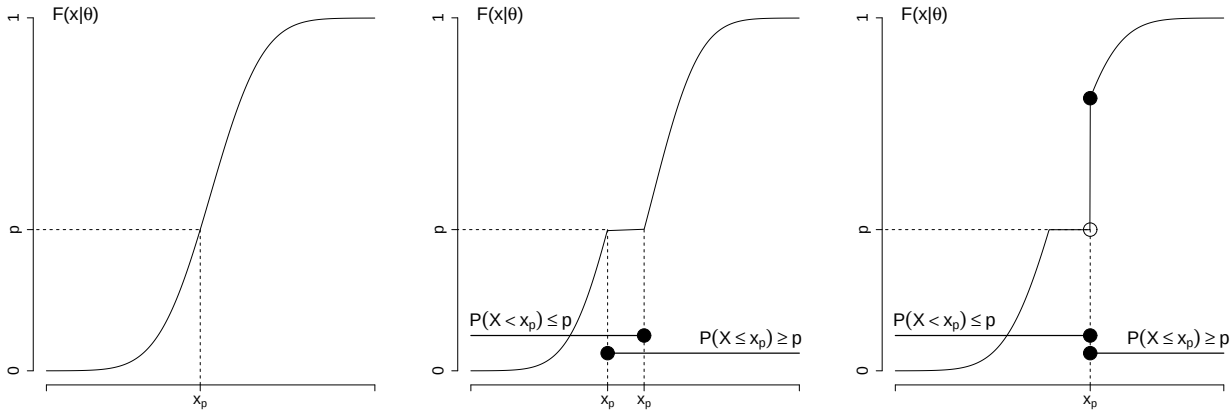


Abbildung 2.4: Zur Definition des Quantils x_p.

Satz 2.1 *Sei $X_1, \ldots, X_n$ eine Stichprobe für eine stetig verteilte Population mit Dichte $f(x|\theta)$ und Verteilungsfunktion $F(x|\theta)$. Für $0 < p < 1$ sei x_p das p-te Quantil zu $F(x|\theta)$. Ist $k = [np] + 1$ und $f(x)$ in x_p stetig und positiv, so gilt*

$$X_{(k)} \overset{as}{\sim} N\left(x_p, \frac{1}{f^2(x_p|\theta)} \frac{p(1-p)}{n}\right) .$$

Die k-te Ordnungsstatistik $X_{(k)}$ ist also ein asymptotisch erwartungstreuer Schätzer für das Quantil x_p. Darüberhinaus ist er auch konsistent.

Beispiel 2.1 *Nach Satz 2.1 hat der* **empirische Median** $\tilde{X} = Q(0.5) = X_{([n/2]+1)}$ *allgemein asymptotische Varianz* $var(\tilde{X}) = 1/(4nf^2(x_{0.5}|\theta))$.
Für die **Normalverteilung**, $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$, *gilt* $f(x_{0.5}|\mu, \sigma^2) = 1/\sqrt{2\pi\sigma^2}$ *und damit*

$$var(\tilde{X}) \approx \frac{2\pi}{4} \frac{\sigma^2}{n} = 1.5708 \frac{\sigma^2}{n} .$$

Wegen $var(\overline{X}) = \sigma^2/n < var(\tilde{X}) = 1.5708\sigma^2/n$, *ist für eine Stichprobe aus* $N(\mu, \sigma^2)$ *der Mittelwert* $\overline{X}$ *effizienter als der Median* $\tilde{X}$. *Die asymptotische relative Effizienz von* $\overline{X}$ *gegenüber* $\tilde{X}$ *ist somit*

$$are(\overline{X}, \tilde{X}) = var(\tilde{X})/var(\overline{X}) = \pi/2 = 1.5708 .$$

Um $var(\tilde{X})$ *zu schätzen benötigt man einen Schätzer für die Populationsvarianz* σ^2.

- *Der Momentenschätzer hat hierbei den Nachteil, dass* $\tilde{X}$ *auf einem ordinalen Aspekt beruht,* S^2 *aber auf einen intervallskalierten.*
- *Daher verwendet man einen auf Quantile beruhenden Schätzer für* σ^2. *Üblich ist die Verwendung des* **Inter-Quartile Range** $IQR = Q(0.75) - Q(0.25)$.

Unter $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$ *gilt*

$$iqr = x_{0.75} - x_{0.25} = (\mu + z_{0.75}\sigma) - (\mu + z_{0.25}\sigma) = 2z_{0.75}\sigma$$

mit z_p *dem p-ten Quantil der* $N(0,1)$*-Verteilung. Wegen* $z_{0.75} = 0.6745$ *folgt hierfür* $\sigma = iqr/(2 \cdot 0.6745)$*, was den robusten Varianzschätzer*

$$\hat{\sigma}^2_{IQR} = \frac{IQR^2}{1.349^2}$$

motiviert. Dieser liefert schließlich

$$\widehat{var}(\tilde{X}) = 1.5708\frac{\hat{\sigma}^2_{IQR}}{n} = 0.8639\frac{IQR^2}{n}\,.$$

Gleichverteilung*:* $X_i \overset{iid}{\sim} U(-a, a)$*,* $0 < a$*. Damit ist* $E(X) = 0$*,* $var(X) = a^2/3$ *und*

$$\begin{aligned} var(\tilde{X}) &= 4a^2/(4n) = a^2/n \\ var(\overline{X}) &= a^2/(3n)\,, \end{aligned}$$

also wiederum $var(\overline{X}) < var(\tilde{X})$*.*

Doppel-Exponential- (Laplace-) verteilung*:* $X_i \overset{iid}{\sim} DExp(\mu, \sigma^2)$ *mit Dichte*

$$f(x|\mu, \sigma^2) = 1/(2\sigma)\exp(-|x - \mu|/\sigma), \quad x, \mu \in \mathbb{R},\ \sigma > 0\,.$$

Dafür ist $E(X) = \mu$*,* $var(X) = 2\sigma^2$ *und*

$$\begin{aligned} var(\tilde{X}) &= 4\sigma^2/(4n) = \sigma^2/n \\ var(\overline{X}) &= 2\sigma^2/n\,, \end{aligned}$$

also erstmals $var(\overline{X}) > var(\tilde{X})$*. Als asymptotisch relative Effizienz folgt 1/2.*

Variationskoeffizient
Dieser ist definiert als das Momenten-Verhältnis Standardabweichung zu Erwartung, also

$$\theta = \sigma/\mu\,.$$

Somit ist er ein relatives (dimensionsloses) Streuungsmaß, dessen Einheit μ ist. Liegt eine Stichprobe X vor, so wird er geschätzt durch den empirischen Variationskoeffizienten

$$\hat{\theta} = S/\overline{X}\,.$$

Für eine **normalverteilte** Stichprobe gilt $\mathrm{var}(\hat{\theta}) = \theta^2/2n$.
Für eine **exponentialverteilte** Stichprobe mit $\mathrm{E}(X) = \lambda$ und $\mathrm{var}(X) = \lambda^2$ erhält man $\theta = \lambda/\lambda = 1$, d.h. der Variationskoeffizient ist konstant. Gerade hierbei dient θ zur Beschreibung von Größen, deren Variabilität mit dem Niveau zunimmt. Beispiele dafür sind Anzahlen oder Wartezeiten.
Liegt eine Stichprobe aus einer **Poissonverteilung** vor mit $\mathrm{E}(X) = \mathrm{var}(X) = \lambda$, so ist $\theta = \sqrt{\lambda}/\lambda = 1/\sqrt{\lambda}$.

Zusammenfassung univariater Statistiken
basierend auf Momente:

- arithmetisches Mittel: $\overline{x} = n^{-1} \sum_i x_i$.
- geometrisches Mittel: $\overline{x}_g = \sqrt[n]{\prod_i x_i}$ für $x_i > 0$.
- harmonisches Mittel: $\overline{x}_h = \frac{n}{\sum_i 1/x_i}$ für nur positive oder nur negative x_i.
  Hat man nur positive x_i, so gilt stets die Beziehung $\overline{x}_h \leq \overline{x}_g \leq \overline{x}$.
  Bei identen Beobachtungen sind die Mittel gleich.
- Varianz: $s^2 = (n-1)^{-1} \sum_i (x_i - \overline{x})^2$.
  Zur Berechnung verwende man $\sum_i (x_i - \overline{x})^2 = \sum_i x_i^2 - n\overline{x}^2$.
- Standardabweichung: $s = \sqrt{s^2}$.
- Variationskoeffizient: $s/\overline{x}$, für $\overline{x} \neq 0$.
- Schiefe: $\hat{\alpha}_3 = n^{-1} \sum_i (x_i - \overline{x})^3 / s_1^3$ mit Standardfehler $\sqrt{6/n}$.
- Kurtosis: $\hat{\alpha}_4 = n^{-1} \sum_i (x_i - \overline{x})^4 / s_1^4 - 3$ mit Standardfehler $\sqrt{24/n}$.

basierend auf Quantile:

- Median: $\tilde{x} = q(0.5)$ mit Standardfehler $0.929\text{iqr}/\sqrt{n}$.
- Minimum: $x_{(1)}$ oder Minimum Standard Score: $(x_{(1)} - \overline{x})/s$.
- Maximum: $x_{(n)}$ oder Maximum Standard Score: $(x_{(n)} - \overline{x})/s$.

  Zur Verteilung von Minimum und Maximum unabhängig ident verteilter $X_1, \ldots, X_n$ mit stetiger Dichte $f_X : (a, b) \to \mathbb{R}$ und Verteilungsfunktion F_X:

$$
\begin{aligned}
P(X_{(1)} > x) &= P(X_1 > x) \cdot \ldots \cdot P(X_n > x) = (1 - F_X(x))^n \\
F_{X_{(1)}}(x) &= \begin{cases} 0 & x < a \\ 1 - (1 - F_X(x))^n & a \leq x \leq b \\ 1 & x > b \end{cases} \\
f_{X_{(1)}}(x) &= \begin{cases} n(1 - F_X(x))^{n-1} f_X(x) & a \leq x \leq b \\ 0 & \text{sonst} \end{cases} \\
P(X_{(n)} \leq x) &= P(X_1 \leq x) \cdot \ldots \cdot P(X_n \leq x) = F_X^n(x) \\
F_{X_{(n)}}(x) &= \begin{cases} 0 & x < a \\ F_X^n(x) & a \leq x \leq b \\ 1 & x > b \end{cases} \\
f_{X_{(n)}}(x) &= \begin{cases} n F_X^{n-1}(x) f_X(x) & a \leq x \leq b \\ 0 & \text{sonst} \end{cases}
\end{aligned}
$$

- Bereich (Range): $x_{(n)} - x_{(1)}$.

- Quartile: $q(0.25)$, $q(0.75)$, sowie Interquartiler Bereich $\text{iqr} = q(0.75) - q(0.25)$.

robuste Schätzer:

- Standardabweichung: $\hat{\sigma} = \text{iqr}/1.349$.
- Tailmaß: $\hat{\tau} = \Big(q(0.975) - q(0.025)\Big) \Big/ \Big(q(0.875) - q(.125)\Big)$.
- Peakednessmaß: $\hat{\pi} = \Big(q(0.875) - q(0.125)\Big) \Big/ \Big(q(0.65) - q(0.35)\Big)$.
- Schiefe: $\hat{\gamma} = \Big(q(0.95) - \tilde{x}\Big) \Big/ \Big(\tilde{x} - q(0.05)\Big)$.

2.2 Konfidenzintervalle

Ausgangssituation in der **Parametrischen Statistik** sind n unabhängig und ident verteilte Stichprobenvariablen $X_1, \ldots, X_n \overset{iid}{\sim} F(x|\theta)$, wobei das Verteilungsmodell $F(x|\theta)$ exakt spezifiziert ist. Gesucht werden nun zwei Statistiken $L = L(X_1, \ldots, X_n)$ und $U = U(X_1, \ldots, X_n)$ (L steht für untere/lower und U für obere/upper Grenze), so dass für alle Parameter θ aus dem Parameterraum gilt

$$P_\theta(L \leq \theta \leq U) = 1 - \alpha, \quad 0 < \alpha < 1\,.$$

Das zufällige Intervall (L, U) überdeckt den wahren Parameter θ mit Wahrscheinlichkeit $1 - \alpha$. Man nennt es **zweiseitiges Konfidenzintervall** für θ zum Niveau $1 - \alpha$. Die entsprechenden **einseitigen Intervalle** erhält man, indem eine Seite offen gelassen wird, d.h. $P_\theta(L \leq \theta) = 1 - \alpha$ oder $P_\theta(\theta \leq U) = 1 - \alpha$ betrachtet.
Intuitiv bedeutet dies, dass derartige Intervalle mit großer Wahrscheinlichkeit (also α sehr klein gewählt, z.B. $\alpha = 0.05$) den unbekannten Parameter überdecken. Für eine Realisation $x_1, \ldots, x_n$ jedoch überdeckt (l, u) das wahre θ oder es überdeckt diesen auch nicht. Die Aussage, dass der Parameter mit Wahrscheinlichkeit $1-\alpha$ im Konfidenzintervall liegt ist somit unsinnig!
Beachte: Sei (L_r, U_r), $r = 1, \ldots, R$, eine Folge von iid Konfidenzintervallen für θ zum Niveau $1 - \alpha$, dann resultiert mit dem Starken Gesetz der großen Zahlen (SLLN)

$$\frac{1}{R} \sum_{r=1}^{R} I_{[L_r, U_r]}(\theta) \overset{f.s.}{\to} 1 - \alpha\,.$$

Hierbei gilt für die Indikatoren $I_{[L_r, U_r]}(\theta) \overset{iid}{\sim} Bernoulli(1 - \alpha)$ (Erwartung $= 1 - \alpha$).
Unter der Annahme $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$ resultieren folgende Intervalle:

1. Konfidenzintervall für μ bei σ bekannt:

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \Rightarrow Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

$$P(z_{\alpha/2} \leq Z \leq z_{1-\alpha/2}) = P\left(\overline{X} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \leq \mu \leq \overline{X} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}\right) = 1 - \alpha\,.$$

2. Konfidenzintervall für μ bei σ unbekannt:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2; \quad T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

$$P(t_{n-1;\alpha/2} \leq T \leq t_{n-1;1-\alpha/2}) =$$
$$P\left(\overline{X} - \frac{S}{\sqrt{n}}t_{n-1;1-\alpha/2} \leq \mu \leq \overline{X} + \frac{S}{\sqrt{n}}t_{n-1;1-\alpha/2}\right) = 1-\alpha\,.$$

3. Konfidenzintervall für σ^2 bei μ unbekannt:

$$Y = \frac{(n-1)}{\sigma^2}S^2 \sim \chi^2_{n-1}$$

$$P(\chi^2_{n-1;\alpha/2} \leq Y \leq \chi^2_{n-1;1-\alpha/2}) = P\left(\frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1;\alpha/2}}\right) = 1-\alpha\,.$$

Bei bekanntem μ verwendet man $Y = \sum_i (X_i - \mu)^2/\sigma^2 \sim \chi^2_n$. Man gewinnt daher einen Freiheitsgrad.

Für eine beliebige Verteilung F mit $\mathrm{E}(X_i) = \mu$ und $\mathrm{var}(X_i) = \sigma^2$ gilt mit dem Zentralem Grenzwertsatz (CLT), wegen $\mathrm{E}(\overline{X}) = \mu$ und $\mathrm{var}(\overline{X}) = \sigma^2/n$, dass

$$Z = \frac{\overline{X} - \mathrm{E}(\overline{X})}{\sqrt{\mathrm{var}(\overline{X})}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)\,.$$

Dies liefert Konfidenzintervalle wie oben allerdings mit einer Überdeckungswahrscheinlichkeit die nur für $n \to \infty$ gegen $1-\alpha$ strebt.

1. Approximatives Konfidenzintervall für μ bei σ^2 bekannt ($n > 30$):

$$P\left(\overline{X} - \frac{\sigma}{\sqrt{n}}z_{1-\alpha/2} \leq \mu \leq \overline{X} + \frac{\sigma}{\sqrt{n}}z_{1-\alpha/2}\right) \approx 1-\alpha\,.$$

2. Approximatives Konfidenzintervall für μ bei σ^2 unbekannt ($n > 30$):

$$P\left(\overline{X} - \frac{S}{\sqrt{n}}t_{n-1;1-\alpha/2} \leq \mu \leq \overline{X} + \frac{S}{\sqrt{n}}t_{n-1;1-\alpha/2}\right) \approx 1-\alpha\,.$$

3. Approximatives Konfidenzintervall für σ^2 bei μ unbekannt ($n > 100$):

$$P\left(\frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1;\alpha/2}}\right) \approx 1-\alpha\,.$$

Ausgangssituation in der **Nicht-Parametrischen Statistik** sind n unabhängig ident verteilte Stichprobenvariablen $X_1, \ldots, X_n \overset{iid}{\sim} F$, wobei an das Verteilungsmodell jetzt nur sehr schwache Voraussetzungen gestellt werden, wie beispielsweise die Forderung der Stetigkeit oder der Symmetrie von F. Nicht-Parametrisch ist ein Verfahren, das keine Aussagen über explizite Parameter einer Verteilung trifft (z.B. μ oder σ^2 bei der Normalverteilung). Populationsquantile werden hierbei nicht als explizite Parameter gesehen. Daher sind Verfahren nicht-parametrisch auch wenn sie sich auf implizite Parameter wie Median, Quartile, u.s.w. beziehen.
Haben die n Beobachtungen $x_1, \ldots, x_n$ zumindest ordinales Messniveau und gilt $X_i \overset{iid}{\sim} F$, wobei F streng monoton wachsend ist, so ist x_p eindeutig bestimmt, und wir definieren ein Konfidenzintervall für das unbekannte Populationsquantil x_p zum Niveau $1-\alpha$ durch

$$P(X_{(k)} < x_p < X_{(\ell)}) = 1 - \alpha$$

mit $k < \ell$. Es werden nun die Indizes k und ℓ bestimmt. Sei dazu

$$Y_i(x) = \begin{cases} 0 & \text{falls} \quad X_i > x \\ 1 & \text{falls} \quad X_i < x\,, \end{cases}$$

so ist $Y_i(x) \overset{iid}{\sim} \text{Binomial}(1, F(x))$ und $T(x) = \sum_i Y_i(x) \sim \text{Binomial}(n, F(x))$. Damit folgt

$$\begin{aligned} P(X_{(k)} < x_p < X_{(\ell)}) &= P(X_{(k)} < x_p,\ X_{(\ell)} > x_p) \\ &= P(\#(X_i < x_p) \geq k,\ \#(X_i < x_p) \leq \ell - 1) \\ &= P(k \leq T(x_p) \leq \ell - 1) = 1 - \alpha\,. \end{aligned}$$

Man bestimmt daher k und ℓ so, dass die obige Beziehung gilt. In der Regel kann aber das Niveau $1-\alpha$ nicht exakt eingehalten werden. Daher wählt man die folgende Vorgehensweise

1. Bestimme k und ℓ derart, dass $\ell - k$ minimal (Intervall kurz), und
2. $\text{Binomial}(\ell - 1|n, p) - \text{Binomial}(k - 1|n, p) \geq 1 - \alpha$ gilt (Fehler maximal α).

Für $n > 20$ ist es auch angebracht, die Binomialverteilung durch die Normalverteilung zu approximieren (Satz von DeMoivre-Laplace):

$$\begin{aligned} P(X_{(k)} < x_p < X_{(\ell)}) &= P(k \leq T(x_p) \leq \ell - 1) \\ &\approx \underbrace{\Phi\left(\frac{\ell - 1 - np + 1/2}{\sqrt{np(1-p)}}\right)}_{1-\alpha/2} - \underbrace{\Phi\left(\frac{k - np - 1/2}{\sqrt{np(1-p)}}\right)}_{\alpha/2} \approx 1 - \alpha\,. \end{aligned}$$

Also sind

$$\begin{aligned} k &= np + \frac{1}{2} + z_{\alpha/2}\sqrt{np(1-p)}\,, \\ \ell &= np + \frac{1}{2} + z_{1-\alpha/2}\sqrt{np(1-p)}\,. \end{aligned}$$

Als approximatives Konfidenzintervall ergibt sich nach Abrunden unten und Aufrunden oben

$$P\left(X_{(\lfloor k \rfloor)} < x_p < X_{(\lceil \ell \rceil)}\right) \approx 1 - \alpha\,.$$

Oft wird auch bei großem Stichprobenumfang die Verteilung des Medians durch die Normalverteilung approximiert (siehe Satz 2.1). Mit den Ergebnissen aus Beispiel 2.1 gilt

$$\widehat{\mathrm{var}}(\tilde{X}) = 0.8639\frac{\mathrm{IQR}^2}{n}$$

und als alternatives approximatives Konfidenzintervall für den theoretischen Median resultiert

$$P\left(\tilde{X} - z_{1-\alpha/2}\sqrt{\widehat{\mathrm{var}}(\tilde{X})} \le x_{0.5} \le \tilde{X} + z_{1-\alpha/2}\sqrt{\widehat{\mathrm{var}}(\tilde{X})}\right) \approx 1 - \alpha\,.$$

2.3 Hypothesentests

Zweck eines **statistischen Tests** ist es, Aussagen oder *Hypothesen* über die Verteilung einer Population Y anhand einer Stichprobe $X_1, \ldots, X_n$ zu untermauern. Zur Überprüfung der Hypothese dient eine *Teststatistik* $T = T(X_1, \ldots, X_n)$. Je nach Realisation von T entscheiden wir uns für oder gegen die vorliegende Hypothese. Das Testproblem wird in Form einer *Nullhypothese* H_0, z.B. $\theta \in \Theta_0$, und einer *Alternativhypothese* H_1, $\theta \notin \Theta_0$, formuliert. Nach bestimmten Kriterien wird eine Menge C (kritischer Bereich oder Verwerfungsbereich) gewählt, welche die Entscheidung zugunsten einer der beiden Hypothesen vorschreibt. Wir entscheiden uns für

- H_0, falls T nicht in C realisiert,
- H_1, falls T in C realisiert.

Die Entscheidung für H_0 oder H_1 kann richtig oder falsch sein. Wir unterscheiden zwischen

- *Fehler 1. Art* α: Entscheidung für H_1, d.h. H_0 ablehnen (wenn $T \in C$), obwohl H_0 zutrifft (Produzentenrisiko),
- *Fehler 2. Art* $1 - \beta$: Entscheidung für H_0, d.h. H_0 nicht ablehnen (wenn $T \notin C$), obwohl H_1 zutrifft (Konsumentenrisiko, Operationscharakteristik).

Wahrheit	Entscheidung	
	H_0	H_1
H_0	$1-\alpha$	α
H_1	$1-\beta$	β

Tabelle 2.2: Zur Definition der Fehler 1. und 2. Art bei Hypothesentests.

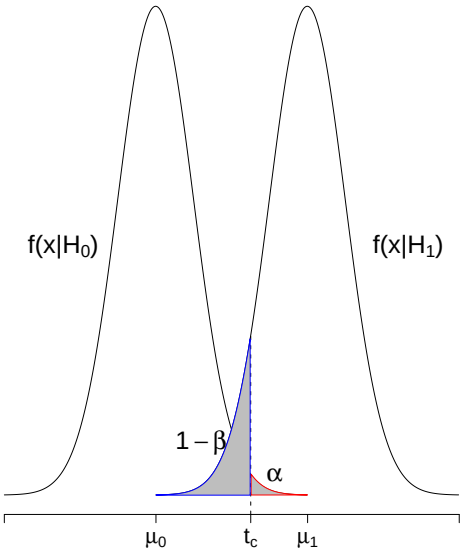


Abbildung 2.5: Type I and II Fehler beim Gaußtest H_0: $\mu = \mu_0$ gegen H_1: $\mu = \mu_1 > \mu_0$ mit Verwerfungsbereich $C = \{t : t > t_c\}$.

Natürlich möchte man die beiden Fehler α und $1 - \beta$ möglichst klein halten. Diese Forderungen widersprechen sich aber (siehe Abbildung 2.5). Üblicherweise wird α fest vorgegeben und aus dem resultierenden Annahmebereich der Fehler $1 - \beta$ bestimmt. Dabei kann es passieren, dass $1 - \beta$ sehr groß wird. Da der tatsächliche Wert des Parameters nicht bekannt ist, kann über den Fehler 2. Art auch keine genaue Auskunft gegeben werden. Nur der Fehler 1. Art ist unter Kontrolle und deshalb nur die damit verbundene Entscheidung „Hypothese H_0 **nicht** verwerfen“. Aus diesem Grund spricht man auch **nicht** von „Hypothese H_0 annehmen“ sondern von „Hypothese H_0 nicht verwerfen“.
Andererseits trifft man auch zwei Arten richtiger Entscheidungen

- $P(T \notin C | H_0 \text{ richtig}) = 1 - \alpha$ (durch Niveau bestimmt),
- $P(T \in C | H_1 \text{ richtig}) = \beta$ (Macht oder Schärfe des Tests).

Der p-Wert

Für gewöhnlich erhält man von einer Software nach Aufruf eines Tests als Ausgabe keine logische Entscheidung sondern den p-Wert. Dieser ist vereinfacht ausgedrückt die anhand der vorliegenden Stichprobe beobachtete Type I Error Rate α.
Um Eigenschaften des p-Wertes diskutieren zu können ist vorerst folgende Aussage über Eigenschaften der Verteilungsfunktion F_X ausgewertet in X notwendig.

Satz 2.2 (Probability Integral Transformation) *Habe X stetige Verteilungsfunktion $F_X(x)$ und sei die Zufallsvariable Y definiert als $Y = F_X(X)$. Dann ist Y auf $(0, 1)$ gleichverteilt, also gilt für $0 < y < 1$*

$$P(Y \leq y) = y\,.$$

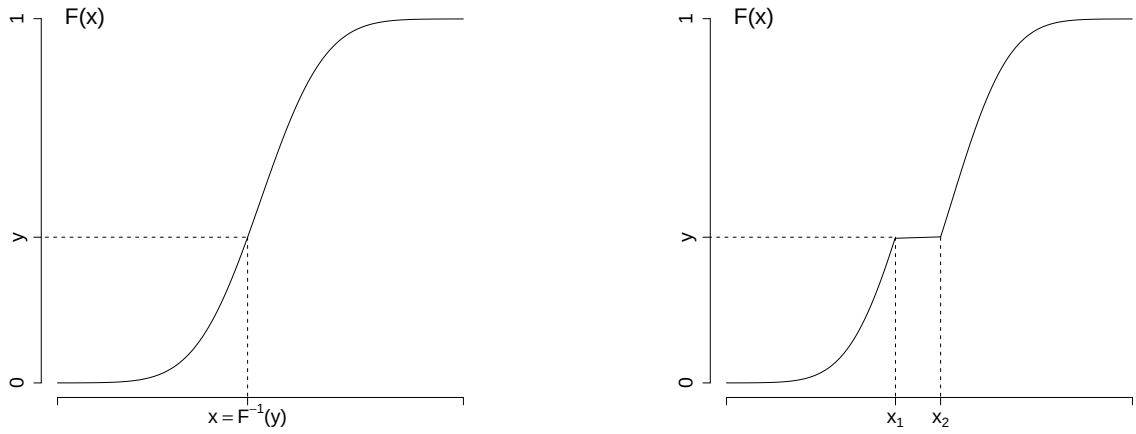


Abbildung 2.6: Links: $F(x)$ streng monoton wachsend; Rechts: $F(x)$ nicht fallend.

Bevor wir zum Beweis kommen, benötigen wir die Inverse $F_X^{-1}(y)$. Ist F_X streng monoton wachsend (siehe Abbildung 2.6 links), dann ist F_X^{-1} eindeutig definiert als

$$F_X^{-1}(y) = x \quad \Longleftrightarrow \quad F_X(x) = y\,.$$

Ist jedoch F_X über ein Intervall konstant (wie in Abbildung 2.6 rechts), dann ist $F_X(x) = y$ für jedes $x_1 \leq x \leq x_2$ und die Inverse ist dort nicht eindeutig. Wir definieren deshalb

$$F_X^{-1}(y) = \inf\{x : F_X(x) \geq y\}\,.$$

Ist F_X nicht konstant, dann stimmt diese Definition mit der obigen überein. Zusätzlich liefert sie einen eindeutigen Wert für F_X^{-1} und wir haben $F_X^{-1}(y) = x_1$. An den Endpunkten des Bereichs von y definieren wir $F_X^{-1}(1) = \infty$ und $F_X^{-1}(0) = -\infty$.
Mit dieser Definition beweisen wir den Satz 2.2. Für $Y = F_X(X)$ gilt für $0 < y < 1$

$$\begin{aligned} P(Y \leq y) &= P(F_X(X) \leq y) \\ &= P(F_X^{-1}(F_X(X)) \leq F_X^{-1}(y)) \\ &= P(X \leq F_X^{-1}(y)) \\ &= F_X(F_X^{-1}(y)) \\ &= y\,. \end{aligned}$$

An den Endpunkten haben wir $P(Y \leq y) = 1$ für $y \geq 1$, und $P(Y \leq y) = 0$ für $y \leq 0$, also ist Y gleichverteilt auf $[0, 1]$.
Die Identität

$$P(F_X^{-1}(F_X(X)) \leq F_X^{-1}(y)) = P(X \leq F_X^{-1}(y))$$

hält natürlich für streng wachsende F_X, wofür $F_X^{-1}(F_X(x)) = x$ folgt. Ist jedoch F_X für $x_1 \leq x \leq x_2$ flach, so kann dort $F_X^{-1}(F_X(x)) \neq x$ gelten. Mit der generellen Definition der Inversen von zuvor resultiert dort $F_X^{-1}(F_X(x)) \neq x_1$. Wegen $P(X \leq x) = P(X \leq x_1)$ hält aber auch dort die obige Identität. Die flache Stelle der Verteilungsfunktion bezeichnet einen Bereich mit Null-Wahrscheinlichkeit, d.h. $P(x_1 \leq X \leq x_2) = F_X(x_2) - F_X(x_1) = 0$.

Weiters werden wir noch folgenden Begriff verwenden:

Definition 2.5 *F_X ist stochastisch größer als F_Y, falls $F_X(t) \leq F_Y(t)$ für alle t gilt. Für $X \sim F_X$ und $Y \sim F_Y$ folgt $P(X \leq t) = F_X(t) \leq F_Y(t) = P(Y \leq t)$ und für alle t gilt*

$$P(X > t) \geq P(Y > t)\,.$$

Nachdem ein Hypothesentest durchgeführt ist, muss das Ergebnis mitgeteilt werden. Eine Möglichkeit ist es, α für diesen Test zu berichten und die darauf basierende Entscheidung gegen oder für H_0. Ist α klein, so ist man ziemlich überzeugt H_0 verwerfen zu können. Für großes α hingegen kann man wegen der Gefahr einer falschen Entscheidung H_0 nicht verwerfen. Alternativ können die Ergebnisse eines Tests auch mittels einer speziellen Statistik, dem p-Wert, übermittelt werden.

Definition 2.6 *Ein p-Wert $p(X)$ basierend auf die Stichprobe $X = (X_1, \ldots, X_n)$ ist eine Teststatistik mit $0 \leq p(x) \leq 1$ für jeden Stichprobenpunkt x. Kleine Werte von $p(X)$ geben einen Hinweis auf die Richtigkeit von H_1. Ein p-Wert ist gültig, falls für jedes $\theta \in \Theta_0$ und jedes $0 \leq \alpha \leq 1$ gilt*

$$P_\theta(p(X) \leq \alpha) \leq \alpha\,.$$

Ist $p(X)$ ein gültiger p-Wert, kann damit sehr einfach ein Level α Test konstruiert werden. Der Test, der H_0 genau dann verwirft wenn $p(X) \leq \alpha$ ist wegen obiger Definition ein Level α Test. Kenntnis des p-Werts ermöglicht dem Leser eine individuelle Wahl von α, die sie/er für angebracht hält. Weiters beinhaltet der p-Wert eine eher stetige Information und nicht nur die dichotome Entscheidung für oder gegen H_0.
Wie kann nun ein gültiger p-Wert definiert werden?

Satz 2.3 *Sei $W(X)$ eine Teststatistik so dass große Werte von W gegen H_0 sprechen. Definiere für einen beliebigen Stichprobenpunkt x*

$$p(x) = \sup_{\theta \in \Theta_0} P_\theta(W(X) \geq W(x))\,.$$

Damit ist $p(X)$ ein gültiger p-Wert.

Beweis: Fixiere einen Parameterwert $\theta \in \Theta_0$. Sei dafür $F_\theta(w)$ die Verteilungsfunktion von $-W(X)$. Definiere dafür weiters

$$p_\theta(x) = P_\theta(W(X) \geq W(x)) = P_\theta(-W(X) \leq -W(x)) = F_\theta(-W(x))\,.$$

Somit entspricht für diese Wahl von θ die Zufallsvariable $p_\theta(X)$ gerade $F_\theta(-W(X))$. Mit Satz 2.2 folgt, dass die Verteilung von $p_\theta(X)$ stochastisch größer oder gleich die Gleichverteilung ist (vgl. Definition 2.5 und die folgenden Bemerkungen). D.h. für jedes $0 \leq \alpha \leq 1$ gilt $P(p_\theta(X) \leq \alpha) \leq \alpha$. Nun ist der p-Wert in Satz 2.3 aber definiert über alle $\theta \in \Theta_0$, und es gilt dafür

$$p(x) = \sup_{\theta' \in \Theta_0} p_{\theta'}(x) \geq p_\theta(x)\,,$$

da der größte p-Wert für alle Elemente in Θ_0 zumindest so groß ist als der für unseren fest gehaltenen Parameterwert θ. Somit gilt auch für jedes $\theta \in \Theta_0$ und jedes $0 \leq \alpha \leq 1$

$$P_\theta(p(X) \leq \alpha) \leq P_\theta(p_\theta(X) \leq \alpha) \leq \alpha$$

und $p(X)$ ist daher ein gültiger p-Wert.

Beispiel 2.2 (p-Wert beim 2-seitigen t-Test) *Sei $X_1, \ldots, X_n$ eine Zufallsstichprobe aus $N(\mu, \sigma^2)$ und teste H_0: $\mu = \mu_0$ gegen H_1: $\mu \neq \mu_0$. Der Likelihood Quotienten Test (LRT) verwirft H_0 für große Werte von $W(X) = |\overline{X} - \mu_0|/(S/\sqrt{n})$. Für $\mu = \mu_0$ folgt $(\overline{X} - \mu_0)/(S/\sqrt{n})$ einer t_{n-1}-Verteilung, unabhängig vom Wert von σ. Deshalb gilt hierfür*

$$p(x) = P_{\theta_0}(W(X) \geq W(x)) = 2P(T_{n-1} \geq (\overline{x} - \mu_0)/(s/\sqrt{n}))\,.$$

Beispiel 2.3 (p-Wert beim 1-seitigen t-Test) *Teste H_0: $\mu \leq \mu_0$ gegen H_1: $\mu > \mu_0$. Der LRT verwirft für große Werte von $W(X) = (\overline{X} - \mu_0)/(S/\sqrt{n})$. Nun wird für diese Statistik gezeigt, dass das Supremum aus Satz 2.3 immer bei (μ_0, σ) auftritt, und dass der Wert von σ darauf keinen Einfluss hat. Betrachte ein $\mu \leq \mu_0$ und irgend ein σ:*

$$\begin{aligned} P_{\mu,\sigma}(W(X) \geq W(x)) &= P_{\mu,\sigma}\left(\frac{\overline{X} - \mu_0}{S/\sqrt{n}} \geq W(x)\right) \\ &= P_{\mu,\sigma}\left(\frac{\overline{X} - \mu}{S/\sqrt{n}} \geq W(x) + \frac{\mu_0 - \mu}{S/\sqrt{n}}\right) \\ &= P_{\mu,\sigma}\left(T_{n-1} \geq W(x) + \frac{\mu_0 - \mu}{S/\sqrt{n}}\right) \\ &\leq P\left(T_{n-1} \geq W(x)\right)\,. \end{aligned}$$

Die Ungleichung hält, weil $\mu \leq \mu_0$ und dafür $(\mu_0 - \mu)/(S/\sqrt{n})$ eine nicht-negative Zufallsvariable ist. Weiters gilt

$$P\left(T_{n-1} \geq W(x)\right) = P_{\mu_0,\sigma}\left(\frac{\overline{X} - \mu_0}{S/\sqrt{n}} \geq W(x)\right) = P_{\mu_0,\sigma}\left(W(X) \geq W(x)\right)\,,$$

was wegen $(\mu_0, \sigma) \in \Theta_0$ gerade eine der Wahrscheinlichkeiten im Supremum aus Satz 2.3 ist. Der p-Wert ist also $p(x) = P(T_{n-1} \geq W(x)) = P(T_{n-1} \geq (\overline{x} - \mu_0)/(s/\sqrt{n}))$.

2.3.1 Wichtige parametrische Tests bei Normalverteilung

1. Test auf μ bei σ^2 bekannt (Gaußtest):

H_0	H_1	Entscheidung gegen H_0, falls	kritische Werte
$\mu = \mu_0$	$\mu \neq \mu_0$	$\overline{X} < c_3$ oder $\overline{X} > c_4$	$c_3 = \mu_0 - z_{1-\alpha/2}\ \sigma/\sqrt{n}$ $c_4 = \mu_0 + z_{1-\alpha/2}\ \sigma/\sqrt{n}$
$\mu \leq \mu_0$	$\mu > \mu_0$	$\overline{X} > c_1$	$c_1 = \mu_0 + z_{1-\alpha}\ \sigma/\sqrt{n}$
$\mu \geq \mu_0$	$\mu < \mu_0$	$\overline{X} < c_2$	$c_2 = \mu_0 - z_{1-\alpha}\ \sigma/\sqrt{n}$

2. Test auf μ bei σ^2 unbekannt (t-Test):

H_0	H_1	Entscheidung gegen H_0, falls	kritische Werte
$\mu = \mu_0$	$\mu \neq \mu_0$	$\overline{X} < c_3$ oder $\overline{X} > c_4$	$c_3 = \mu_0 - t_{n-1;1-\alpha/2}\ S/\sqrt{n}$ $c_4 = \mu_0 + t_{n-1;1-\alpha/2}\ S/\sqrt{n}$
$\mu \leq \mu_0$	$\mu > \mu_0$	$\overline{X} > c_1$	$c_1 = \mu_0 + t_{n-1;1-\alpha}\ S/\sqrt{n}$
$\mu \geq \mu_0$	$\mu < \mu_0$	$\overline{X} < c_2$	$c_2 = \mu_0 - t_{n-1;1-\alpha}\ S/\sqrt{n}$

3. Test auf σ^2 bei μ unbekannt (χ^2-Test):

H_0	H_1	Entscheidung gegen H_0, falls	kritische Werte
$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$S^2 < c_3$ oder $S^2 > c_4$	$c_3 = \chi^2_{n-1;\alpha/2}\ \sigma_0^2/(n-1)$ $c_4 = \chi^2_{n-1;1-\alpha/2}\ \sigma_0^2/(n-1)$
$\sigma^2 \leq \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$S^2 > c_1$	$c_1 = \chi^2_{n-1;1-\alpha}\ \sigma_0^2/(n-1)$
$\sigma^2 \geq \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$S^2 < c_2$	$c_2 = \chi^2_{n-1;\alpha}\ \sigma_0^2/(n-1)$

All diese parametrischen Tests setzen kardinales Meßniveau (Intervalls- oder Verhältnisskala) der Beobachtungen voraus. Im folgenden Abschnitt werden nicht-parametrische Methoden behandelt. Bei einer univariaten Stichprobe gehören dazu vor allem Anpassungstests auf eine spezifizierte Verteilungsfunktion und Tests auf Quantile.

2.3.2 Tests auf Güte der Anpassung

Anpassungstests (Goodness-of-Fit Tests) dienen zur Überprüfung der Hypothese, dass die Population aus der die Stichprobe stammt einer hypothetischen Verteilung folgt. Dazu wird untersucht, ob sich die beobachtete Stichprobenverteilung (empirische Verteilungsfunktion) hinreichend gut dieser von uns vorgegebenen hypothetischen Verteilung anpasst.

Definition 2.7 *Sei $X_1, \ldots, X_n$ eine zumindest ordinal skalierte Zufalls-Stichprobe aus der Verteilungsfunktion F.*

$$F_n(x) = \frac{1}{n}\ \#(\ X_i \leq x), \quad x \in \mathbb{R}$$

nennt man die **empirische Verteilungsfunktion** *dieser Stichprobe.*

In der empirische Verteilungsfunktion wird jedem X_i die Wahrscheinlichkeit $1/n$ zugeordnet. Falls keine Bindungen vorliegen, d.h. falls es keine mehrfachen identen Realisationen gibt, so kann $F_n(x)$ auch über die geordnete Statistik definiert werden

$$F_n(x) = \begin{cases} 0 & \text{für } x < X_{(1)}\,, \\ i/n & \text{für } X_{(i)} \leq x < X_{(i+1)},\ i = 1, \ldots, n-1\,, \\ 1 & \text{für } X_{(n)} \leq x\,. \end{cases}$$

Allgemein hat $F_n(x)$ folgende Eigenschaften:

- F_n ist eine monoton steigende Treppenfunktion mit Sprüngen in $x_{(1)}, \ldots, x_{(n)}$.
- Bei ungebundenen Beobachtungen macht F_n in $x_{(i)}$ einen Sprung der Höhe $1/n$. Bei einer Bindung von k Beobachtungen beträgt die Höhe des Sprunges k/n.
- Für jede Realisation $x_{(1)}, \ldots, x_{(n)}$ ist F_n eine Verteilungsfunktion.
- Für jedes x ist $F_n(x)$ eine Zufallsvariable.
- $F_n(x)$ ist diskret mit den Realisationen i/n für $i = 0, \ldots, n$.

Es gilt der Zentralsatz der Statistik (Satz von Clivenko-Cantelli).

Satz 2.4 (Glivenko-Cantelli) *Seien* $X_1, \ldots, X_n \overset{iid}{\sim} F$. *Dann gilt für* $n \to \infty$

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \overset{f.s.}{\to} 0\,,$$

also die gleichmäßige fast sichere Konvergenz.

Satz 2.5 *Seien* $X_1, \ldots, X_n \overset{iid}{\sim} F$. *Dann gilt für jedes feste* $x \in \mathbb{R}$

$$P\left(F_n(x) = \frac{i}{n}\right) = \binom{n}{i} F^i(x)(1 - F(x))^{n-i}\,, \quad i = 0, \ldots, n\,.$$

Es gilt also $nF_n(x) \sim \text{Binomial}(n, F(x))$, und daher $\text{E}(F_n(x)) = F(x)$ sowie $\text{var}(F_n(x)) = F(x)(1 - F(x))/n$.

Der Kolmogorov-Smirnov Test (1933)

Seien X_i aus einer Grundgesamtheit mit uns unbekannter stetiger Verteilungsfunktion F. Zu testen ist die Hypothese, dass F eine gewisse, vollständig spezifizierte Gestalt F_0 hat. Der Zentralsatz der Statistik legt für die zweiseitige Fragestellung die Verwendung von $\sup_x |F_n(x) - F_0(x)|$ als Teststatistik nahe. Unter H_0 sollte diese Abweichung hinreichend klein sein und wir verwerfen H_0 bei einer zu großen Realisation.
Die KS-Statistik basiert auf dem Maximum von Abweichungen und setzt kardinales Meßniveau der Beobachtungen voraus. Ist die unbekannte Verteilungsfunktion F **stetig**, so ist der Test exakt. Ist F_0 diskret, oder sind die Parameter nicht vollständig spezifiziert sondern werden durch die Stichprobe geschätzt, so ist der KS-Test **konservativ**, d.h. für das wahre Testniveau gilt $P(H_0 \text{ ablehnen} | H_0 \text{ richtig}) < \alpha$. Der Test neigt in diesen Fällen also eher dazu, H_0 nicht zu verwerfen. Ersetzen wir beispielsweise in $F_0(\mu)$ den Parameter $\text{E}(X) = \mu$ durch den Stichprobenwert $\overline{x}$, dann unterscheidet sich diese geschätzte hypothetische Verteilungsfunktion $F_0(\overline{x})$ bzgl. des ersten Momentes nicht mehr von der empirischen Verteilung $F_n(x)$.
Sei F_0 vollständig spezifiziert, so können folgende Hypothesen getestet werden:
Testproblem:

- Test A: $H_0 : F(x) = F_0(x)\ \forall x \in \mathbb{R}, \quad H_1 : \exists x \in \mathbb{R} : F(x) \neq F_0(x)$
- Test B: $H_0 : F(x) \leq F_0(x)\ \forall x \in \mathbb{R}, \quad H_1 : \exists x \in \mathbb{R} : F(x) > F_0(x)$
- Test C: $H_0 : F(x) \geq F_0(x)\ \forall x \in \mathbb{R}, \quad H_1 : \exists x \in \mathbb{R} : F(x) < F_0(x)$

Die entsprechenden **KS-Teststatistiken** lauten

- Test A: $K_n = \sup_{x \in \mathbb{R}} |F_0(x) - F_n(x)|$
- Test B: $K_n^- = \sup_{x \in \mathbb{R}} (F_n(x) - F_0(x))$
- Test C: $K_n^+ = \sup_{x \in \mathbb{R}} (F_0(x) - F_n(x))$

Entscheidungsregel: H_0 wird abgelehnt, wenn

- Test A: $k_n \geq k_{n;1-\alpha}$; mit $P(K_n \geq k_{n;1-\alpha}) = \alpha$
- Test B: $k_n^- \geq k_{n;1-\alpha}^-$; mit $P(K_n^- \geq k_{n;1-\alpha}^-) = \alpha$
- Test C: $k_n^+ \geq k_{n;1-\alpha}^+$; mit $P(K_n^+ \geq k_{n;1-\alpha}^+) = \alpha$

Definition 2.8 *$X_1, \ldots, X_n \overset{iid}{\sim} F$. Die Teststatistik $T = T(X_1, \ldots, X_n)$ heißt* **verteilungsfrei***, falls die Verteilung von T nicht von F abhängt.*

Lemma 2.1 *Unter der Annahme der Stetigkeit von F_0 sind die Teststatistiken K_n, K_n^+ und K_n^- unter H_0 verteilungsfrei, d.h. unabhängig von F_0.*

Die Herleitung der Verteilung von K_n unter H_0 ist aufwendig. Die Quantile $k_{n;1-\alpha}$ sind jedoch für $n \leq 40$ **exakt** tabelliert (siehe Tabelle E). Für $n > 40$ kann auf Quantile der asymptotischen Verteilung zurückgegriffen werden.

Satz 2.6 *Ist F_0 stetig, so gilt für alle $z > 0$*

$$\begin{aligned} (1) \quad \lim_{n\to\infty} P\left(K_n \leq \frac{z}{\sqrt{n}}\right) &= L(z) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2z^2}\,, \\ (2) \quad \lim_{n\to\infty} P\left(K_n^+ \leq \frac{z}{\sqrt{n}}\right) &= L^+(z) = 1 - e^{-2z^2}\,. \end{aligned}$$

Aus Punkt (2) folgt, dass $4nK_n^{+2}$ asymptotisch χ_2^2-verteilt ist. Für $1 - \alpha = 0.95$ gilt beispielsweise $\chi^2_{2;0.95} = 5.99$, womit als Approximation $k^+_{n;0.95} \approx \sqrt{\chi^2_{2;0.95}/4n} = 1.22/\sqrt{n}$ resultiert.

Beispiel 2.4 *Zu testen ist, ob der Benzinverbrauch eines Autotyps per 100 Meilen einer $F_0 = N(12, 1)$-Verteilung entspricht. Dazu liegt eine Stichprobe vom Umfang $n = 10$ vor. Der Test sollte zweiseitig mit $\alpha = 0.05$ durchgeführt werden, d.h. man testet*

$$H_0 : F(x) = \Phi(x|12, 1) \quad \textit{gegen} \quad H_1 : F(x) \neq \Phi(x|12, 1).$$

Die Tabelle 2.3 gibt die für die Anwendung des KS-Tests benötigten Werte an. Die hypothetische Verteilung $\Phi(x|12, 1)$ und die empirische Verteilung $F_n(x)$ sind in der Abbildung 2.7 dargestellt. $F_n^+(x)$ und $F_n^-(x)$ sind der rechts- bzw. linksseitige Grenzwert von $F_n(x)$ und $d_n^+(x) = |\Phi(x|12, 1) - F_n^+(x)|$, $d_n^-(x) = |\Phi(x|12, 1) - F_n^-(x)|$. In $x_{(4)} = 12.4$ realisiert K_{10} in $k_{10} = 0.355$. Wegen $k_{10;0.95} = 0.409$ (d.h. $k_{10} < k_{10;0.95}$) kann H_0 nicht verworfen werden.
Die folgenden Zeilen zeigen beispielhaft die Verwendung entsprechender Kommandos in `R`*. Dabei werden zuerst die Daten definiert, dann der zweiseitige KS-Test angewandt (default Option), und schließlich in einer Graphik die empirische Verteilungsfunktion mit der hypothetischen Verteilung verglichen.*

i	$x_{(i)}$	$\Phi(x_{(i)})$	F_n^+	F_n^-	d_n^+	d_n^-
1	11.5	0.309	0.1	0.0	0.209	0.309
2	11.8	0.421	0.2	0.1	0.221	0.321
3	12.0	0.500	0.3	0.2	0.200	0.300
4	12.4	0.655	0.4	0.3	0.255	**0.355**
5	12.5	0.691	0.5	0.4	0.191	0.291
6	12.6	0.726	0.6	0.5	0.126	0.226
7	12.8	0.788	0.7	0.6	0.088	0.188
8	12.9	0.816	0.8	0.7	0.016	0.116
9	13.0	0.841	0.9	0.8	0.059	0.041
10	13.2	0.885	1.0	0.9	0.115	0.015

Tabelle 2.3: Kolmogorov-Smirnov-Test bei den Benzinverbrauchsdaten.

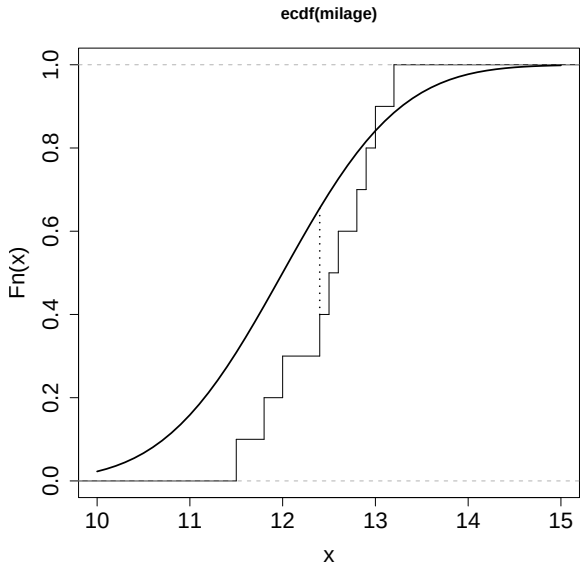


Abbildung 2.7: Vergleich der empirischen mit der hypothetischen Verteilungsfunktion.

```
> milage <- c(11.5, 11.8, 12.0, 12.4, 12.5, 12.6, 12.8, 12.9, 13.0, 13.2)
> ks.test(milage, "pnorm", 12, 1)

        One-sample Kolmogorov-Smirnov test

  data:  milage
  D = 0.3554, p-value = 0.1598
  alternative hypothesis: two.sided
> library(stepfun);  plot(ecdf(milage), do.points=FALSE, verticals=TRUE)
> x <- seq(10, 15, 0.1);  lines(x, pnorm(x, 12, 1))  # plot hypothetical c.d.f.
```

Der Pearson χ^2-Test (1900)

Dieser Test ist ein weiterer Anpassungstest, der im Gegensatz zum KS-Test, welcher Verteilungsfunktionen direkt vergleicht, die entsprechenden erwarteten Häufigkeiten zum Vergleich heranzieht. Deshalb können die Daten auch beliebiges Meßniveau haben.
Die n Beobachtungen $x_1, \ldots, x_n$ werden in k disjunkte Klassen eingeteilt. Die Teststatistik erfasst nun die Abweichungen der beobachteten Häufigkeiten n_j von den theoretischen Häufigkeiten np_j unter H_0, mit $\sum_{j=1}^{k} n_j = n$.

Klasse	K_1	K_2	$\ldots$	K_k
Anzahl d. Beobachtungen	n_1	n_2	$\ldots$	n_k

Unter der Annahme $X_i \overset{iid}{\sim} F$ unterscheidet man zwischen zwei Testproblemen:
Testproblem A: Sei F_0 eine vollständig spezifizierte Verteilungsfunktion

- $H_0 : F(x) = F_0(x) \quad \forall x \in \mathbb{R}$,
- $H_1 : \exists x \in \mathbb{R} : F(x) \neq F_0(x)$.

Teststatistik:

$$T_{\chi^2} = \sum_{j=1}^{k} \frac{(N_j - np_j)^2}{np_j} \overset{as}{\sim} \chi^2_{k-1}$$

mit $p_j = P(X_i \in K_j | H_0) = F_0(a_j) - F_0(a_{j-1})$, falls $K_j = (a_{j-1}, a_j)$.
Entscheidungsregel: Da für $n \to \infty$ die Folge der Verteilungsfunktionen von T_{χ^2} gegen die χ^2-Verteilung mit $k-1$ Freiheitsgraden strebt, wird H_0 abgelehnt, falls $t_{\chi^2} \geq \chi^2_{k-1;1-\alpha}$ gilt. Die theoretischen Quantile sind in der Tabelle C wiedergegeben.

Beispiel 2.5 *Ein Würfel wird* 120 *mal geworfen. Ist der Würfel unfair?*

$$H_0 : p_j = 1/6, \quad j = 1, \ldots, 6; \quad H_1 : \exists j : p_j \neq 1/6.$$

Klasse j	*1*	*2*	*3*	*4*	*5*	*6*	*Summe*
n_j	*20*	*30*	*20*	*25*	*15*	*10*	*120*
np_j	*20*	*20*	*20*	*20*	*20*	*20*	*120*
$\frac{(n_j - np_j)^2}{np_j}$	*0*	*5*	*0*	*5/4*	*5/4*	*5*	*12.5*

Mit $t_{\chi^2} = 12.5$ *ist für* $\alpha = 0.01$ *wegen* $\chi^2_{5,0.99} = 15.08 > t_{\chi^2}$ *der Würfel als fair zu werten. Bei* $\alpha = 0.05$ *führt dies zu* $\chi^2_{5,0.95} = 11.07 < t_{\chi^2}$ *und der Würfel kann nicht als fair eingestuft werden. Der exakte p-Wert ist hierbei etwas kleiner als 3%.*

```
> dice <- c(20, 30, 20, 25, 15, 10)
> chisq.test(dice, p = rep(1/6, 6))

        Chi-squared test for given probabilities

data:  dice
X-squared = 12.5, df = 5, p-value = 0.02854
```

Wie bereits beim KS-Test sollte auch hier noch geklärt werde, was zu tun ist wenn unbekannte Parameter in der hypothetischen Verteilungsfunktion sind und diese somit nicht vollständig spezifiziert ist. Dies führt uns zum Testproblem B.

Testproblem B: Seien r unbekannte Parameter $\theta_1, \ldots, \theta_r$ in F_0, dann testen wir

- $H_0 : F(x) = F_0(x|\theta_1, \ldots, \theta_r) \quad \forall x \in \mathbb{R}$,
- $H_1 : \exists x \in \mathbb{R} : F(x) \neq F_0(x|\theta_1, \ldots, \theta_r)$.

Seien $\hat{\theta}_1, \ldots, \hat{\theta}_r$ Schätzer für die unbekannten Parameter in F_0, dann werden in der Teststatistik diese Parameter durch deren Schätzungen ersetzt und wir erhalten dadurch die **Teststatistik:**

$$T^m_{\chi^2} = \sum_{j=1}^{k} \frac{(N_j - np_j(\hat{\theta}_1, \ldots, \hat{\theta}_r))^2}{np_j(\hat{\theta}_1, \ldots, \hat{\theta}_r)} \overset{as}{\sim} \chi^2_{k-1-r} \,.$$

Diese ist daher eine modifizierte Version der Teststatistik für das Testproblem A, in der die r unbekannten Parameter durch deren Schätzungen ersetzt werden. Dabei reduziert sich lediglich der Freiheitsgrad der χ^2-Verteilung um die Anzahl der geschätzten Parameter.
Entscheidungsregel: H_0 wird abgelehnt, falls $t^m_{\chi^2} \geq \chi^2_{k-1-r;1-\alpha}$ gilt.
Es stellt sich noch die Frage, für welche Schätzmethoden dies Gültigkeit hat. Dabei zeigt sich, dass die asymptotische χ^2-Verteilungseigenschaft nur dann gilt, wenn die Parameter $\hat{\theta}_1, \ldots, \hat{\theta}_r$ entweder mit der Maximum-Likelihood-Methode bzgl. gruppierter Daten oder mittels der Minimum-Chi-Quadrat Methode geschätzt wurden.
Bei der ML-Methode bzgl. gruppierter Daten wird die Likelihood Funktion

$$\max_{\theta_1,\ldots,\theta_r} \prod_{j=1}^{k} p_j(\theta_1, \ldots, \theta_r)^{n_j}$$

maximiert. Würde man die ML-Methode für ungruppierte Daten anwenden, d.h.

$$\max_{\theta_1,\ldots,\theta_r} \prod_{i=1}^{n} f(x_i|\theta_1, \ldots, \theta_r)$$

maximieren, so wäre $T^m_{\chi^2}$ **nicht** mehr asymptotisch χ^2-verteilt. In der Praxis verwendet man aber trotzdem fälschlicherweise die ML-Schätzer für ungruppierte Daten, da diese einfacher zu erhalten sind. Dieses Vorgehen führt dann jedoch zu einem wahren Niveau $\alpha^* > \alpha$, wodurch der Fehler 1. Art nicht mehr kontrolliert ist.

Bei der Minimum-χ^2-Methode wird einfach die Teststatistik $T^m_{\chi^2}(\theta_1, \ldots, \theta_r)$ in den Parametern maximiert. Diese Schätzer $\hat{\theta}_1, \ldots, \hat{\theta}_r$ sind daher definiert als Lösung des Gleichungssystems $\partial T^m_{\chi^2}(\theta_1, \ldots, \theta_r)/\partial\theta_j = 0$, $j = 1, \cdots, k$.

Liegen ungruppierte Daten vor, so stellt sich die Frage nach der Klasseneinteilung, d.h. für welches n und für welches p_j $(j = 1, \ldots, k)$ ist die Approximation der Verteilung von T_{χ^2} durch die χ^2-Verteilung gerechtfertigt. In der Praxis wird oft als Faustregel $np_j \geq 5$ verwendet.

Vergleich KS-Test mit χ^2-Test:

- Die exakte Verteilung von K_n liegt für kleine $n \leq 40$ vor. Der χ^2-Test kann nur für große n angewendet werden; er ist ein approximativer Test.
- Alle n Beobachtungen werden beim KS-Test unmittelbar benutzt. Beim χ^2-Test müssen sie erst zu k Klassen zusammengefasst werden (Informationsverlust). Das beinhaltet eine Willkür bei der Festlegung der Klassenanzahl und Klassenbreite.
- Der KS-Test basiert auf der Annahme einer stetigen Verteilung der Grundgesamtheit, während der χ^2-Test auch (und gerade) bei diskreten Verteilungen anwendbar ist. Der KS-Test ist in diesem Fall konservativ.
- Müssen die Parameter von $F_0(x)$ geschätzt werden, so hat $\hat{K}_n$ (Schätzer für die Parameter substituiert) nicht dieselbe Verteilung wie K_n. Der Fehler ist somit nicht unter Kontrolle. Beim χ^2-Test verringert sich in diesem Fall lediglich die Anzahl der Freiheitsgrade um die Anzahl der geschätzten Parameter.
- Der χ^2-Test ist nur zweiseitig anwendbar, während mit dem KS-Test auch Abweichungen in nur eine Richtung getestet werden können.

Der Shapiro-Wilk Test (1965)

Häufig ist man daran interessiert, Abweichungen von der Normalverteilung zu erkennen. Sowohl der KS-Test wie auch der χ^2-Test sind dafür nicht geeignet. Um den KS-Test verwenden zu können müssen μ und σ^2 in F_0 spezifiziert sein. Der χ^2-Test erlaubt nur die Verwendung von entsprechend aufwendigen Schätzungen. Shapiro und Wilk's W Statistik ist eine gute Alternative, um auf Abweichungen von der Normalverteilung zu testen. Diese ist das Verhältnis zweier Schätzungen für die Varianz einer Normalverteilung,

$$W = \frac{\left[\sum_{i=1}^{n} a_i X_{(i)}\right]^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\,.$$

Der Zähler ist proportional dem Quadrat des besten (minimale Varianz, unverzerrt) linearen Schätzers für die Standardabweichung, und der Nenner beschreibt die Quadratsumme der Abweichungen der Beobachtungen vom Mittel. Die Koeffizienten a_i liegen in Form einer Tabelle vor oder können nur approximiert werden. Es gibt jedoch Programme, welche diese Approximationen berechnen.

Beispiel 2.6 *Im KS-Test auf Normalität der Benzinverbrauchsdaten wurde $\mu = 12$ und $\sigma^2 = 1$ verwendet. Die Abbildung 2.7 zeigte aber, dass gerade diese Wahl nicht realistisch ist. Es ist zwar $\overline{x} = 12.47$, aber nur $s^2 = 0.3$.*

```
> mean(milage)
[1] 12.47
> var(milage)
[1] 0.3045556
> shapiro.test(milage)

        Shapiro-Wilk normality test

data:  milage
W = 0.9529, p-value = 0.7026
```

Haben wir beim KS-Test mit einen p-Wert von doch nur 0.16 die Nullhypothese verwerfen können, so ist dies jetzt trotz geringerem Abstand zwischen F_0 und F_n viel deutlicher. Zum Vergleich liefert der KS-Test mit geschätzter Hypothese

```
> ks.test(milage, "pnorm", mean(milage), sd(milage))

        One-sample Kolmogorov-Smirnov test

data:  milage
D = 0.1495, p-value = 0.9787
alternative hypothesis: two.sided
```

Der Binomialtest

Liegen zwei Gruppen vor, so kann man den **Binomialtest** einsetzen. Die n Beobachtungen realisieren in einer von zwei disjunkten Klassen K_1, K_2. Unter $X_i \overset{iid}{\sim} F$ sei $P(X_i \in K_1) = p$. Diese Wahrscheinlichkeit ist für alle i gleich, da die X_i ident verteilt sind.
Testproblem:

- Test A: $H_0 : p = p_0$; $\quad H_1 : p \neq p_0$
- Test B: $H_0 : p \leq p_0$; $\quad H_1 : p > p_0$
- Test C: $H_0 : p \geq p_0$; $\quad H_1 : p < p_0$

Teststatistik:

$$T = \#(X_i \in K_1)\,.$$

Da T eine Summe von Bernoulli(p)-Variablen ist, gilt offensichtlich unter H_0

$$T \sim \text{Binomial}(n, p_0)\,.$$

T ist eine diskret verteilte Zufallsvariable, weshalb es in der Regel zu einem beliebigen α kein t_α gibt, wofür $P(T \leq t_\alpha) = \alpha$ exakt eingehalten wird. Deshalb werden für diskret verteilte Teststatistiken Ungleichungen der Form $P(T \leq t_\alpha) \leq \alpha$ verwendet. Der Fehler erster Art ist dadurch maximal α. Das exakte α^* sollte angegeben werden.
Für $n \leq 20$ ergeben sich die kritischen Bereiche wie folgt:

- Test A: $t_{1-\alpha_1} = \min_k \{k | P(T \geq k) \leq \alpha_1\}$, und $t_{\alpha_2} = \max_k \{k | P(T \leq k) \leq \alpha_2\}$
  mit $\alpha_1 + \alpha_2 = \alpha$.
  H_0 wird abgelehnt, falls $t \geq t_{1-\alpha_1}$ oder $t \leq t_{\alpha_2}$ zutrifft.
- Test B: $t_{1-\alpha} = \min_k \{k | P(T \geq k) \leq \alpha\}$
  H_0 wird abgelehnt, falls $t \geq t_{1-\alpha}$ zutrifft.
- Test C: $t_\alpha = \max_k \{k | P(T \leq k) \leq \alpha\}$
  H_0 wird abgelehnt, falls $t \leq t_\alpha$ zutrifft.

Für p_0 um $1/2$ wird wegen der Symmetrie der Verteilung von T $\alpha_1 = \alpha_2 = \alpha/2$ gewählt. Für kleine (große) p_0 empfiehlt es sich manchmal α_1 größer (kleiner) als α_2 zu wählen.

Beispiel 2.7 *Von einer Maschine wird behauptet, dass diese maximal 5% defekte Geräte produziert. In einer Stichprobe vom Umfang $n = 20$ wurden 3 defekte Stücke entdeckt (15%). Kann damit die Behauptung auf $\alpha = 0.10$ aufrecht gehalten werden?*
Testproblem: *Test B*

$$H_0 : p \leq 0.05; \quad H_1 : p > 0.05\,.$$

Für den kritischen Bereich $(t_{1-\alpha}, 20)$ gilt

$$\begin{aligned} P(T \geq 2|p = 0.05) &= 1 - P(T \leq 1|p = 0.05) = 1 - 0.7358 = 0.2642 > \alpha\,, \\ P(T \geq 3|p = 0.05) &= 1 - P(T \leq 2|p = 0.05) = 1 - 0.9245 = 0.0755 < \alpha\,. \end{aligned}$$

Wegen $t_{0.90} = 3 \leq t = 3$ lehnen wir H_0 auf dem Niveau $\alpha^* = 0.0755$ ab (oder auch nicht!).

```
> binom.test(x=3, n=20, p=0.05, alternative="greater")

        Exact binomial test

 data:  3 and 20
 number of successes = 3, number of trials = 20, p-value = 0.07548
 alternative hypothesis: true probability of success is greater than 0.05
 95 percent confidence interval:
  0.04216941 1.00000000
 sample estimates:
 probability of success
                   0.15

> binom.test(x=3, n=20, p=0.05, alt="greater", conf.level=0.90)$conf.int
 [1] 0.0564179 1.0000000
 attr(,"conf.level")
 [1] 0.9
```

Gilt $n > 20$ und $10 \leq np \leq n - 10$, so liefert die Anwendung des Satzes von DeMoivre-Laplace eine gute Approximation für t_α. Damit resultiert

- Test A: $t_{1-\alpha_1} = \lceil np_0 - \frac{1}{2} + z_{1-\alpha_1}\sqrt{np_0(1-p_0)}\rceil$, und $t_{\alpha_2} = \lfloor np_0 + \frac{1}{2} + z_{\alpha_2}\sqrt{np_0(1-p_0)}\rfloor$
- Test B: $t_{1-\alpha} = \lceil np_0 - \frac{1}{2} + z_{1-\alpha}\sqrt{np_0(1-p_0)}\rceil$
- Test C: $t_\alpha = \lfloor np_0 + \frac{1}{2} + z_\alpha\sqrt{np_0(1-p_0)}\rfloor$

Hierbei wird auf- bzw. abgerundet, um das vorgelegte Niveau α möglichst gut einhalten zu können. Die Approximation ist wegen der Symmetrie der Normalverteilung umso besser, je näher p bei $1/2$ liegt. Dies ist aber bei vielen Fragestellungen gerade nicht der Fall.

Beispiel 2.8 *(Fortsetzung von Bsp. 2.4) Die Normalverteilungsapproximation liefert für $\alpha = 0.10$ als kritischen Wert*

```
> n <- 20;  p <- 0.05;  alpha <- 0.10
> ceiling(n*p -1/2 + qnorm(1-alpha)*sqrt(n*p*(1-p)))
[1] 2
```

also den kritischen Bereich $[2, 20]$ *mit einem Niveau von 0.26. Die Approximation ist in diesem Fall sehr ungenau.*

Bei binären Merkmalen ergibt sich eine Einteilung in zwei Klassen von selbst (z.B. Geschlecht (männlich, weiblich), Prüfung (bestanden, nicht bestanden)). Liegt ein quantitatives Merkmal vor, so wird ein fester Trennwert zwischen den beiden Klassen vorgegeben (z.B. Gewicht ≤ 80 kg, > 80 kg) und die Wahrscheinlichkeit dieser Gruppierung getestet.

2.3.3 Tests für Quantile

Entsprechend dem Gaußtest für normalverteilte Populationen wird nun nicht-parametrisch mittels Vorzeichentest und Wilcoxon Vorzeichen-Rangtest auf ein beliebiges Quantil und auf das Lokationszentrum getestet. Da die Wilcoxon-Statistik die Ränge der Stichprobenvariablen verwendet, werden zuerst deren Verteilungseigenschaften diskutiert.

Verteilung der Ränge

Definition 2.9 *Seien* $X_1, \ldots, X_n \overset{iid}{\sim} F$ *und* F **stetig**. *Der Rang* $R_i = R(X_i)$ *gibt die Anzahl der Variablen an, die* X_i *nicht übertreffen, also*

$$R(X_i) = \#(X_j \leq X_i), \quad j = 1, \ldots, n\,.$$

R_i *ist eine diskret verteilte Zufallsvariable auf* $1, 2, \ldots, n$.

Bemerkungen:

- Da F stetig ist, gilt $P(X_i = X_j) = 0$ für $i \neq j$. D.h. Bindungen treten nur mit Wahrscheinlichkeit Null auf.
- Der Rang von X_i legt die Position dieser Beobachtung in der geordneten Statistik fest. Der Index j von $X_{(j)}$ bezeichnet genau den Rang jenes X_i, welches dem $X_{(j)}$ entspricht. Sortiert man also die Stichprobe nach den Rängen, erhält man die geordnete Stichprobe.

```
> x <- c(3, 1, 4, 1, 5, 9, 2, 6, 5, 3, 5)
> order(x)
 [1]  2  4  7  1 10  3  5  9 11  8  6
> (r <- rank(x)) # ties are averaged
 [1] 4.5  1.5    6  1.5    8   11    3   10    8  4.5    8
> (r <- rank(x, ties.method = "first"))  # first occurrence wins
 [1]   4    1    6    2    7   11    3   10    8    5    9
> (r <- rank(x, ties.method = "random")) # ties broken at random
 [1]   4    2    6    1    8   11    3   10    9    5    7
> x[order(x)]
 [1] 1 1 2 3 3 4 5 5 5 6 9
> sort(x)
 [1] 1 1 2 3 3 4 5 5 5 6 9
```

Grundlegend für das Arbeiten mit Rängen ist der folgende Satz.

Satz 2.7 *Seien* $X_1, \ldots, X_n \overset{iid}{\sim} F$ *und* F *stetig. Dann gilt*

1. $P(R_1 = r_1, \ldots, R_n = r_n) = 1/n!$, *wobei* $r_1, \ldots, r_n$ *eine Permutation der Zahlen* $1, \ldots, n$ *ist.*
2. $P(R_i = j) = 1/n$, $i = 1, \ldots, n$, *mit* $j \in \{1, \ldots, n\}$.
3. $P(R_i = k, R_j = l) = 1/n(n-1)$, *für* $1 \leq i, j, k, l \leq n$; $i \neq j$, $k \neq l$.
4. $E(R_i) = (n+1)/2$, $i = 1, \ldots, n$.
5. $var(R_i) = (n^2 - 1)/12$, $i = 1, \ldots, n$.
6. $cov(R_i, R_j) = -(n+1)/12$, $1 \leq i, j \leq n$, $i \neq j$.
7. $cor(R_i, R_j) = -1/(n-1)$, $1 \leq i, j \leq n$, $i \neq j$.

Dieser Satz zeigt, dass die Momente der Ränge unabhängig von der Verteilung der Grundgesamtheit sind (verteilungsfrei). Mit zunehmendem Stichprobenumfang streuen die Ränge mehr, und die Korrelation zwischen zwei beliebigen Rängen nimmt ab.

Der Vorzeichentest

Wenn von einer Verteilungsfunktion F nur deren **Stetigkeit** vorausgesetzt werden kann, verwendet man als Test für Quantile x_p (d.h. $F(x_p) = p$, $0 < p < 1$) den Vorzeichentest. Dieser Test stellt eine spezielle Anwendung des Binomialtest dar, mit dem man testet, ob das unbekannte Quantil x_p der Population einem festen hypothetischen Wert x_0 entspricht.
Testproblem:

- Test A: $H_0 : x_p = x_0$; $\quad H_1 : x_p \neq x_0$,
- Test B: $H_0 : x_p \leq x_0$; $\quad H_1 : x_p > x_0$,
- Test C: $H_0 : x_p \geq x_0$; $\quad H_1 : x_p < x_0$.

Teststatistik:

$$D = \sum_{i=1}^{n} h(x_0 - X_i)\,, \qquad \text{mit} \qquad h(z) = \begin{cases} 1 & \text{für } z > 0 \\ 0 & \text{für } z < 0 \end{cases}$$

Die Statistik D beschreibt die Anzahl der Beobachtungen X_i, die kleiner als x_0 sind, d.h. die Anzahl der positiven Vorzeichen unter den n Differenzen $(x_0 - X_i)$. Wegen der vorausgesetzten Stetigkeit von F gilt $P(x_0 - X_i = 0) = 0$ und die obige Definition des Indikators h ist gerechtfertigt. Die Beobachtungen werden dadurch in zwei disjunkte Klassen K_1 und K_2 eingeteilt, welche durch x_0 getrennt sind. Falls x_0 das p-te Quantil zu F ist, dann gilt

$$P(X_i \in K_1) = P(h(x_0 - X_i) = 1) = P(X_i < x_0) = F(x_0) = p\,.$$

Da D eine Summe n unabhängiger Bernoulli(p)-Variablen ist, folgt $D \sim \text{Binomial}(n, p)$. Um über H_0 zu entscheiden, geht man also gleich vor wie beim Binomialtest.

Beispiel 2.9 *Von $n = 15$ Personen liegt die Körpergröße in cm vor. Ist der Median signifikant ($\alpha = 0.05$) von $x_0 = 180cm$ verschieden?*
Wir testen hier also den zweiseitigen Fall (Test A)

$$H_0 : x_{.50} = 180; \quad H_1 : x_{.50} \neq 180\,.$$

Zur Berechnung der Teststatistik benötigen wir

x_i	179	177	178	174	170	185	175	179	176	169	186	189	168	170	174
$180 - x_i$	1	3	2	6	10	-5	5	1	4	11	-6	-9	12	10	6

Wir können $d = 12$ positive Differenzen beobachten. Unter H_0 gilt $D \sim Binomial(15, 0.5)$. Für $\alpha_1 = \alpha_2 = \alpha/2$ folgt wegen der Symmetrie dieser Binomialverteilung

$$\begin{aligned} P(D \geq 12|p = 0.5) = P(D \leq 3|p = 0.5) &= 0.0176 < \alpha/2\,, \\ P(D \geq 11|p = 0.5) = P(D \leq 4|p = 0.5) &= 0.0592 > \alpha/2\,. \end{aligned}$$

Die Nullhypothese ist daher auf dem exakten Niveau $\alpha = 2 \cdot 0.0176 = 3.5\%$ abzulehnen. Der kritische Bereich dazu ist $\{0, 1, 2, 3\} \cup \{12, 13, 14, 15\}$.

```
> height <- c(179,177,178,174,170,185,175,179,176,169,186,189,168,170,174)
> binom.test(sum(height<180), length(height), 1/2)

        Exact binomial test

data:  sum(height < 180) and length(height)
number of successes = 12, number of trials = 15, p-value = 0.03516
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.5191089 0.9566880
sample estimates:
probability of success
                   0.8
```

Das wahre Niveau p des Quantils $x_p = 180$ scheint also zwischen 52% und 96% zu liegen.

Der Wilcoxon Vorzeichen-Rangtest für den Median

Der Vorzeichen-Rangtest von Wilcoxon ist nur als Test auf den **Median** $\tilde{x} = x_{0.5}$ geeignet und benötigt als Voraussetzung außer der **Stetigkeit** auch die **Symmetrie** der Verteilungsfunktion F. Neben der Anzahl der positiven Vorzeichen der Differenzen $D_i = X_i - \tilde{x}_0$ wird auch die Rangordnung der $|D_i|$ registriert.
Testproblem:

- Test A: $H_0 : \tilde{x} = \tilde{x}_0; \quad H_1 : \tilde{x} \neq \tilde{x}_0,$
- Test B: $H_0 : \tilde{x} \leq \tilde{x}_0; \quad H_1 : \tilde{x} > \tilde{x}_0,$
- Test C: $H_0 : \tilde{x} \geq \tilde{x}_0; \quad H_1 : \tilde{x} < \tilde{x}_0.$

Wegen der Stetigkeit von F folgt sofort $P(D_i = 0) = 0$ und $P(|D_k| = |D_l|) = 0$ für $k \neq l$. Ist nun $R(|D_i|)$ der Rang von $|D_i|$, so ist die Wilcoxon-Teststatistik definiert als

$$W^+ = \sum_{i=1}^{n} Z_i R(|D_i|), \qquad W^- = \sum_{i=1}^{n} (1 - Z_i) R(|D_i)|$$

mit

$$Z_i = \begin{cases} 1 & \text{für} \quad D_i > 0\,, \\ 0 & \text{für} \quad D_i < 0\,. \end{cases}$$

W^+ (W^-) ist die Summe der Ränge der positiven (negativen) D_i. Wegen $W^+ + W^- = n(n+1)/2$ genügt es beispielsweise nur W^+ zu betrachten.
Durch einfache Index-Transformation $j = R(|D_i|)$ kann die Teststatistik W^+ auch umgeformt werden zu

$$W^+ = \sum_{i=1}^{n} Z_i R(|D_i|) = \sum_{j=1}^{n} j Z_{(j)}, \quad \text{wobei} \quad Z_{(j)} = \begin{cases} 1 & \text{für} \quad D_i > 0\,, \\ 0 & \text{für} \quad D_i < 0\,. \end{cases}$$

Im Gegensatz zu den abhängigen $R(|D_i|)$ sind die $Z_{(j)}$ jetzt unabhängig.

Beispiel 2.10 *Kann $H_0 : \tilde{x} = 5$ auf Grund der vorliegenden Stichprobe mit Umfang $n = 8$ auf dem Niveau $\alpha = 0.05$ abgelehnt werden (Testproblem A)?*

i	1	2	3	4	5	6	7	8
x_i	3.5	4.5	4.0	0.5	2.5	7.0	8.5	8.0
$d_i = x_i - 5$	–1.5	–0.5	–1.0	–4.5	–2.5	2.0	3.5	3.0
z_i	0	0	0	0	0	1	1	1
$r(\|d_i\|)$	3	1	2	8	5	4	7	6
$z_{(i)}$	0	0	0	1	0	1	1	0

Es folgt daher

$$w^+ = \sum_{i=1}^{n} z_i r(|d_i|) = 17\,, \quad w^- = \sum_{i=1}^{n} (1 - z_i) r(|d_i|) = 19\,.$$

Um über H_0 entscheiden zu können, müssen wir die Verteilung von W^+ unter H_0 kennen.

Zur Verteilung von W^+ unter H_0:

Unter H_0 gilt wegen der Symmetrie von F um $\tilde{x}_0$:

$$P(\underbrace{(X_j - \tilde{x}_0)}_{D_j} > 0) = P(\underbrace{(X_j - \tilde{x}_0)}_{D_j} < 0) = 1/2\,,$$

d.h. $P(Z_{(i)} = 1) = P(Z_{(i)} = 0) = 1/2$. Somit folgt

$$\mathrm{E}(Z_{(i)}) = 1/2, \quad \mathrm{var}(Z_{(i)}) = \mathrm{E}(Z_{(i)}^2) - \mathrm{E}^2(Z_{(i)}) = 1/2 - 1/4 = 1/4\,.$$

Damit folgt für den Erwartungswert der Teststatistik W^+

$$\mathrm{E}(W^+) = \sum_{i=1}^{n} i\mathrm{E}(Z_{(i)}) = \frac{1}{2}\sum_{i=1}^{n} i = \frac{n(n+1)}{4}\,,$$

und wegen der Unabhängigkeit der $Z_{(i)}$ ergibt sich als Varianz

$$\mathrm{var}(W^+) = \sum_{i=1}^{n} i^2\mathrm{var}(Z_{(i)}) = \frac{1}{4}\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{24}\,.$$

Weiters gilt für die Realisationen $0 \le w \le n(n+1)/2$, wobei $w = 0$ im Falle nur negativer Differenzen resultiert und das Maximum $n(n+1)/2$ für ausschließlich positive Differenzen erreicht wird.
Der Stichprobenraum Ω besteht aus allen möglichen Tupeln $(z_{(1)}, \ldots, z_{(n)})$, d.h.

$$\Omega = \{(0,0,\ldots,0),(1,0,\ldots,0),\ldots,(1,1,\ldots,1)\}$$

mit $|\Omega| = 2^n$. Unter H_0 hat jedes Tupel Auftrittswahrscheinlichkeit $1/2^n$ (Laplace Raum). Daraus folgt

$$P(W^+ = w) = \frac{1}{2^n}\ \#\left(\text{Tupel mit } \sum iz_{(i)} = w\right) = \frac{a(w)}{2^n}\,.$$

Aufwendig aber prinzipiell leicht erweist sich die Bestimmung der Anzahl $a(w)$. Als kleine Erleichterung müssen aber nur alle $a(w)$ für $w \ge \mathrm{E}(W^+)$ berechnet werden. Wegen $P(W^+ = w) = P(W^- = w)$ und mit $W^+ + W^- = n(n+1)/2$ folgt

$$P(W^+ = w) = P(W^- = n(n+1)/2 - w) = P(W^+ = n(n+1)/2 - w)\,.$$

Dies hat zur Folge, dass die Statistik W^+ unter H_0 symmetrisch um $\mathrm{E}(W^+) = n(n+1)/4$ verteilt ist.
Wie berechnet man nun die Anzahl $a(w)$, für alle $w \ge E(W^+)$? Sei beispielsweise $n = 5$, so ergibt sich dafür $0 \le w \le 15$, $\mathrm{E}(W^+) = 7.5$, und $2^5 = 32$. Dies führt zu den folgenden Wahrscheinlichkeiten:

w	Rangtupel positiver D_i	$a(w)$	$P(W^+ = w)$
15	$(1,2,3,4,5)$	1	1/32
14	$(2,3,4,5)$	1	1/32
13	$(1,3,4,5)$	1	1/32
12	$(3,4,5)$; $(1,2,4,5)$	2	2/32
11	$(2,4,5)$; $(1,2,3,5)$	2	2/32
10	$(1,4,5)$; $(2,3,5)$; $(1,2,3,4)$	3	3/32
9	$(4,5)$; $(2,3,4)$; $(1,3,5)$	3	3/32
8	$(3,5)$; $(1,3,4)$; $(1,2,5)$	3	3/32

Damit kann man beispielsweise ablesen, dass $P(W^+ \ge 13) = 3/32 \approx 0.094$ gilt. Aus der Symmetrie folgt weiters sofort $P(W^+ = 15) = P(W^+ = 0)$ oder $P(W^+ = 8) = P(W^+ = 7)$. Für $4 \le n \le 20$ findet man die Quantile w_α von W^+ im Anhang in der Tabelle F. Bemerkenswert ist, dass schon für $n = 20$ die Betrachtung von mehr als einer Million n-Tupeln erforderlich ist.

Sind die kritischen Werte einmal berechnet, dann ist die Testentscheidung wiederum leicht und H_0 wird abgelehnt, falls

- Test A: $w^+ \geq w_{1-\alpha/2}$ oder $w^+ \leq w_{\alpha/2}$,
- Test B: $w^+ \geq w_{1-\alpha}$,
- Test C: $w^+ \leq w_{\alpha}$.

Für $n > 20$ kann die Verteilung von

$$Z = \frac{W^+ - \mathrm{E}(W^+)}{\sqrt{\mathrm{var}(W^+)}}$$

durch die $N(0,1)$-Verteilung approximiert werden (wegen der Unabhängigkeit der $Z_{(i)}$).
Beispiel 2.10 (Fortsetzung) *Für $\alpha = 0.05$ folgen $w^+_{0.025} = 3$ und $w^+_{0.975} = 33$ aus der Tabelle F. Die Daten ergaben $w^+ = 17$. Wegen $3 < 17 < 33$ kann hier H_0 nicht abgelehnt werden.*

```
> x <- c(3.5, 4.5, 4.0, 0.5, 2.5, 7.0, 8.5, 8.0)
> wilcox.test(x, mu = 5)

        Wilcoxon signed rank test

 data:  x
 V = 17, p-value = 0.9453
 alternative hypothesis: true mu is not equal to 5
```

Falls keine Bindungen vorliegen, berechnet die `R` Funktion `wilcox.test` den exakten p-Wert für $n < 50$ (wie im obigen Beispiel erfüllt). Ansonsten wird per default die Normalverteilungsapproximation mit Korrektur (`correct=TRUE`) verwendet. Liegen Bindungen vor, wird zusätzlich die Varianz von W^+ um $\sum(b_j^3 - b_j)/48$ verringert. Hierbei bezeichnet b_j die Anzahl der gebundenen absoluten Differenzen in der j-ten Bindungsgruppe.

Beispiel 2.11 *Für die bereits mit dem Vorzeichen-Test analysierten $n = 15$ Körpergrößen folgt mit dem Wilcoxon-Test auf $H_0 : \tilde{x} = 180$ gegen $H_0 : \tilde{x} \neq 180$.*

```
> wilcox.test(height, mu = 180)

        Wilcoxon signed rank test with continuity correction

 data:  height
 V = 26.5, p-value = 0.06053
 alternative hypothesis: true mu is not equal to 180

 Warning message: Cannot compute exact p-value with ties in:
 wilcox.test.default(height, mu = 180)

> sort(abs(height-180))
 [1]  1  1  2  3  4  5  5  6  6  6  9 10 10 11 12
```

Wegen der vielen Bindungen kann kein exakter p-Wert berechnet werden. Die Differenzen $|d_i| \in (1, 5, 10)$ *kommen jeweils zweimal vor, die Differenz* $|d_i| = 6$ *sogar dreimal. Dies ergibt eine Reduktion von* $var(W^+) = 310$ *um* $(3(2^3 - 2) + (3^3 - 3))/48 = 0.875$. *Der Erwartungswert* $E(W^+) = 60$ *wird dadurch nicht beeinflusst. Je nachdem, ob ohne oder mit Stetigkeitskorrektur gearbeitet wird, wird hier* $(26.5 - 60)/\sqrt{309.125} = -1.905$ *oder* $(27.0 - 60)/\sqrt{309.125} = -1.877$ *mit der Normalverteilung verglichen, was dann den p-Wert ergibt. War dieser beim Vorzeichen-Test 0.035, so ist er hier etwa doppelt so groß. Ist man auch an einem Konfidenzintervall für den Median* $\tilde{x}$ *interessiert, so ist dafür nur die zusätzliche Option* `conf.int = TRUE` *mit* `conf.level = 0.95` *notwendig.*

2.4 Dichteschätzer

2.4.1 Ein Exploratives graphisches Verfahren: der Boxplot

Fünf Kenngrößen lassen sich durch den **Box and Whisker Plot**, oder kurz **Boxplot** (Tukey, 1977), graphisch darstellen. Die Höhe des Rechtecks (der box) entspricht dem interquartilen Bereich ($q(0.25)$, $q(0.75)$) und die Unterteilung dem Median $q(0.50)$. Extremwerte liegen außerhalb der beiden *whiskers*, den um ± 1.5iqr verlängerten Quartilen. Die Abbildung 2.8 wurde erzeugt mit

```
> boxplot(VC)
```

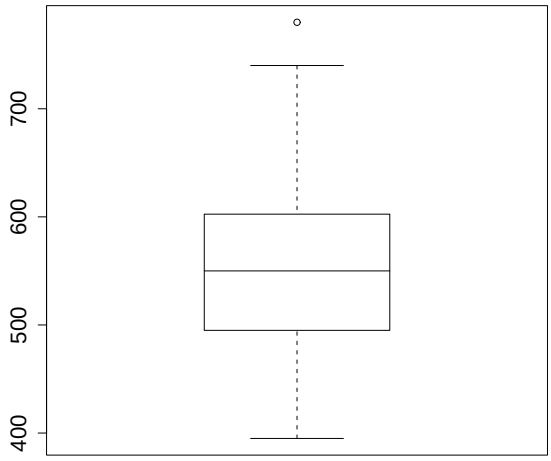


Abbildung 2.8: Darstellung der `VC` Daten mittels Boxplot.

2.4.2 Das Histogramm

Dieses Verfahren wurde bereits von Playfair 1786 in seinem politischen Atlas von London verwendet. Die Häufigkeiten von in ℓ disjunkten Klassen $([t_0, t_1), [t_1, t_2), \ldots, [t_{\ell-1}, t_\ell])$ unterteilten Daten werden hier als Stäbe dargestellt. Durch die freie Wahl der Stabbreite $h = t_j - t_{j-1}$, für $j = 1, \ldots, \ell$, können auch unterschiedliche Eindrücke erzielt werden. Alternativ zu dieser Form (absolute Häufigkeiten) können auch die normierten relativen Häufigkeiten (Dichteschätzer, siehe Abbildung 2.9) aufgetragen werden.

```
> VCmin <- min(VC)-5;  VCmax <- max(VC)+5
> hist(VC, breaks=seq(VCmin, VCmax, length = 8), freq=FALSE)
> hist(VC, breaks=seq(VCmin, VCmax, length = 9), freq=FALSE)
> hist(VC, breaks=seq(VCmin, VCmax, length =17), freq=FALSE)
```

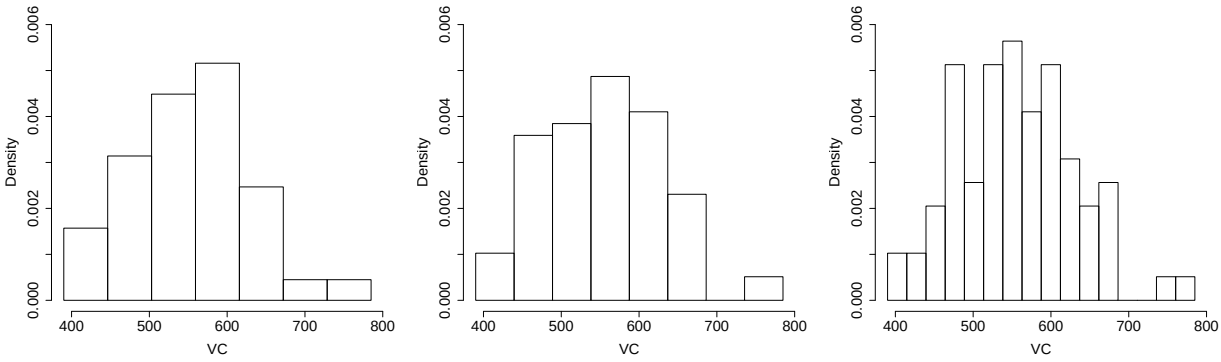


Abbildung 2.9: Histogramme von `VC` mit 7, 8 und 17 Klassen.

Bezeichnet $\hat{f}(x)$ die geschätzte Dichte der Klasse um x, so gilt

$$\hat{f}(x) = \begin{cases} \dfrac{N_j}{nh} & \text{für} \quad t_{j-1} \le x < t_j \\ 0 & \text{sonst,} \end{cases}$$

wobei N_j die Anzahl der Stichprobenelemente in der Klasse j beschreibt. Für ein Histogramm gelten somit auch die Eigenschaften von Dichten, d.h. $\hat{f}(x) \ge 0$, und

$$\int_{\mathbb{R}} \hat{f}(x)dx = \sum_{j=1}^{\ell} \frac{N_j}{nh}\, h = 1.$$

In der Literatur existieren diverse **Faustregeln** für die geeignete Wahl von ℓ, nämlich

- Sturges: $\ell_{St} = \lceil \log_2 n \rceil + 1$,
- Velleman: $\ell_V = \lceil 2\sqrt{n} \rceil$ für $n < 100$,
- Dixon: $\ell_D = \lceil 10 \log_{10} n \rceil$ für $n > 100$.

Aber auch **theoretische Kriterien** zur Bestimmung des optimalen Wertes von h gibt es. Sei dazu $f(x)$ die unbekannte Dichte der Population aus der die Zufalls-Stichprobe $X_1, \ldots, X_n$ stammt. Weiters sei $\hat{f}_h(x)$ der Dichteschätzer von $f(x)$, der vom gewählten h abhängt. Ein lokales Fehlermaß in einem festen Punkt x ist der mittlere quadratische Fehler (Mean Square Error)

$$\mathrm{MSE}_h(x) = \mathrm{E}\left(\hat{f}_h(x) - f(x)\right)^2 = \mathrm{var}\left(\hat{f}_h(x)\right) + \mathrm{bias}^2\left(\hat{f}_h(x), f(x)\right).$$

Als globale Fehlermaße über den gesamten x-Bereich bieten sich sowohl der integrierte mittlere quadratische Fehler (Integrated Mean Square Error)

$$\text{IMSE}_h = \int_{\mathbb{R}} \text{E}\left(\hat{f}_h(x) - f(x)\right)^2 dx\,,$$

als auch die maximale absolute Abweichung (Maximum Absolute Deviation)

$$\text{MAD}_h = \max_x \left|\hat{f}_h(x) - f(x)\right|$$

an. Beide Maße werden nun bzgl. h minimiert. Dieses Vorgehen liefert die theoretische Kriterien:

- Scott: $h_S = 6\left(n\int f'(x)^2 dx\right)^{-1/3}$ minimiert IMSE_h.
  Für $X_i \overset{iid}{\sim} N(0,\sigma^2)$ folgt $\hat{h}_S = 3.49\hat{s}n^{-1/3}$.
- Freedman: $h_F = c(f)\,(\log(n)/n)^{1/3}$ minimiert den MAD_h.
  Für $X_i \overset{iid}{\sim} N(0,\sigma^2)$ folgt $\hat{c}(f) = 1.66\hat{s}$ und damit $\hat{h}_F = 1.66\,(\log(n))^{1/3}\,\hat{s}n^{-1/3}$.
- Freedman/Diaconis (robust): $\hat{h}_{F^*} = 2\text{IQR}n^{-1/3}$ ist einfacher und entspricht einer robusten Version von $\hat{h}_F$.

Für `VC` mit $n = 79$ resultiert $\ell_V = \lfloor 2\sqrt{79}\rfloor = 17$, jedoch $\ell_{St} = 8$. Da $\hat{s} = 76.3$, iqr $= 107.5$ und als Bereich $r = 385$ beobachtet werden konnte, folgt $\hat{h}_S = 62.0$, $\hat{h}_F = 48.2$ und $\hat{h}_{F^*} = 50.1$. Als Klassenanzahl erhält man damit $\ell_S = \lceil r/\hat{h}_S\rceil = \lceil 6.2\rceil = 7$ und entsprechend $\ell_F = 8$ sowie $\ell_{F^*} = 8$ (vgl. Abbildung 2.9). Die Notwendigkeit von 17 Stäben (Velleman Regel) scheint hier mehr als fragwürdig. Diese Regel bezieht sich nur auf den Stichprobenumfang und nicht auf die Variabilität in den Daten. Alle übrigen Kriterien liefern hingegen meist nur acht Stäbe.

```
> nclass.Sturges(VC)
[1] 8
> nclass.scott(VC)
[1] 7
> nclass.FD(VC)
[1] 8
```

In der Abbildung 2.10 werden die Faustregeln mit den Ergebnissen aus den theoretischen Kriterien unter Annahme einer $N(0,1)$-verteilten Population verglichen. Daher wurde auch

$$\text{E(IQR)} = \Phi^{-1}(0.75) - \Phi^{-1}(0.25) = 1.349$$

für die Berechnung von $\hat{h}_{F^*}$ verwendet. Weiters muss nun die berechnete Stabbreite h zur Anzahl der Klassen ℓ umgeformt werden. Dazu verwenden wir zuerst eine Approximation für den zu erwartenden Bereichs

$$\text{E(Range)} = \text{E}\left(X_{(n)} - X_{(1)}\right) \approx 2\Phi^{-1}\left(\frac{n-3/8}{n+1/4}\right)\,,$$

womit sich $\ell = \lceil \text{E(Range)}/h\rceil$ ergibt.

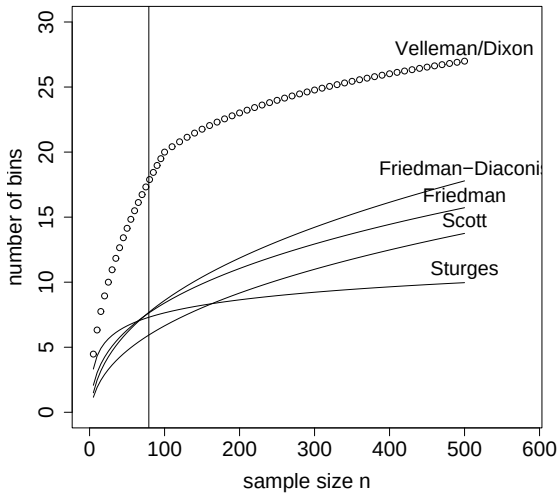


Abbildung 2.10: Anzahl von Stäben ℓ für normalverteilte Stichproben vom Umfang n mit $\sigma = 1$, basierend auf den Faustregeln sowie den theoretischen Kriterien.

2.4.3 Der naive Schätzer

Allgemein gilt für die Dichtefunktion einer stetig verteilten Zufallsvariablen X

$$f(x) = \lim_{h \to 0} \frac{1}{2h} P(x - h < X < x + h)\,.$$

Für einen festen Wert von h schätzt man nun die Wahrscheinlichkeit $P(x-h < X < x+h)$ durch die relative Häufigkeit der Stichprobenelemente $X_1, \ldots, X_n$, die in das Intervall $(x \pm h)$ fallen, also durch $N_h(x)/n$. Dies ergibt den sogenannten naiven Dichteschätzer

$$\hat{f}(x) = \frac{1}{2h} \frac{N_h(x)}{n} = \frac{1}{nh} \frac{N_h(x)}{2}$$

für $f(x)$. Um diesen Schätzer transparenter auszudrücken, definiert man Gewichte

$$w(u) = \begin{cases} 1/2 & |u| < 1 \\ 0 & \text{sonst.} \end{cases}$$

Mit diesem Indikator kann nun gezählt werden, wie viele der X_i maximal h Einheiten von x entfernt sind. Jedem dieser in der Nähe von x liegenden X_i wird das Gewicht $1/2$ zugeordnet. Damit resultiert als naiver Dichteschätzer

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} w\left(\frac{x - X_i}{h}\right)\,.$$

Der naive Schätzer $\hat{f}(x)$ kann also als Summe von Rechtecksflächen mit Breite $2h$ und Höhe $1/(2nh)$ gesehen werden, wobei diese Rechtecke über jede einzelne Beobachtung x_i symmetrisch platziert sind. Im Prinzip addiert man daher n Gleichverteilungsdichten

auf dem Intervall $(x_i \pm h)$. Dies führt direkt zur Idee des Kernschätzers, bei dem diese Gleichverteilung nur durch eine beliebige, um Null symmetrische, stetige Dichte $K(\cdot)$ ersetzt wird.
Am Beispiel `VC` zeigt die Abbildung 2.11 den naiven Schätzer für verschiedene Werte von h. Hierbei ist auch zu erkennen, dass $\hat{f}(x)$ für zu klein gewählte h einen sehr rauen Eindruck macht, und für wachsendes h immer glatter wird und dann schließlich gegen die Dichte einer Gleichverteilung strebt.

```
> h <- 33;  x <- seq(350, 800, 5);  nx <- length(x);  fhat <- 1:nx
> w <- function(data, x, h)  1/2*(abs(data-x)/h < 1)
> for (j in 1:nx)  fhat[j] <- sum(w(VC, x[j], h)/(length(VC)*h))
> plot(x, fhat, type="l")
```

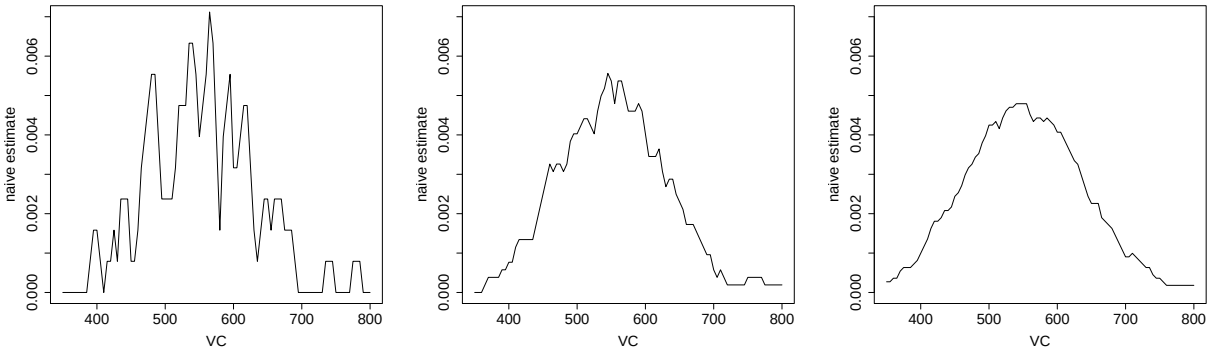


Abbildung 2.11: Naiver Dichteschätzer für `VC` mit $h = 8$, 33 und 70.

2.4.4 Der Kernschätzer

Der naive Schätzer wird verallgemeinert, indem wir das Gewicht $w(\cdot)$ durch einen stetigen Kern $K(\cdot)$ ersetzen mit

$$\int_{\mathbb{R}} K(x)dx = 1\,.$$

Für gewöhnlich ist diese Funktion eine symmetrische Dichte. So spricht man z.B. bei Wahl der Normalverteilungsdichte für $K(x)$ vom Gaußkern. Analog zum naiven Schätzer definiert man nun den Kernschätzer durch

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)\,.$$

Jetzt ist dies eine Summe von Beulen, die über den Beobachtungen x_i zentriert sind. Der Kern K bestimmt dabei die Gestalt der Beule, während die *Smoothing Bandwidth* h deren Breite definiert (Varianz der entsprechenden Zufallsvariablen mit Dichte K). Dieses Konstruktionsprinzip ist in der Abbildung 2.12 dargestellt. Man erkennt deutlich die Abhängigkeit des Kernschätzers von der verwendeten Fensterbreite h. Je größer h gewählt wird, desto glatter wird der Schätzer $\hat{f}(x)$.

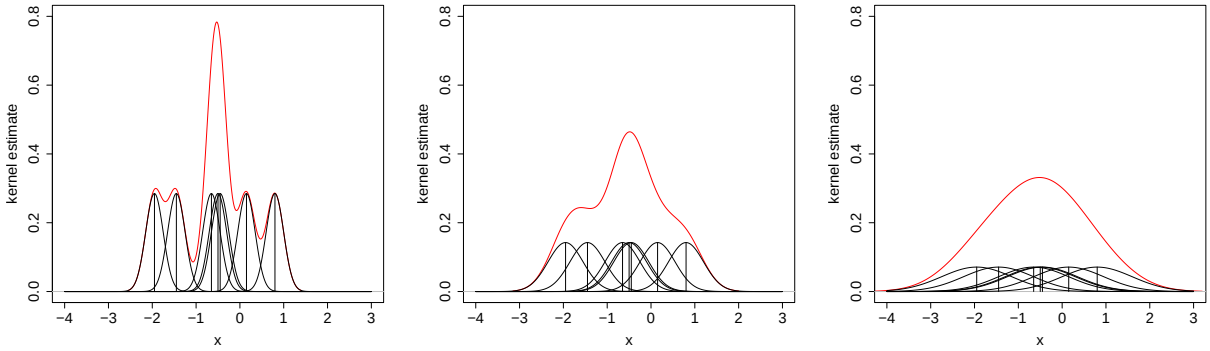


Abbildung 2.12: Kernschätzer (Gaußkern) mit $h = 0.2$, $h = 0.4$ und $h = 0.8$.

Asymptotische Betrachtung von Bias und Varianz

Unser Ziel ist es nun, die Bandbreite h und auch den Kern K derart zu wählen, dass durch diese Wahl der integrierte mittlere quadratische Fehler, d.h.

$$\mathrm{IMSE}(h, K) = \int_{\mathbb{R}} \mathrm{E}\left(\hat{f}_h(x) - f(x)\right)^2 dx$$

minimiert wird. Zunächst fordern wir an den Kern, dass er normiert ist, Erwartung Null widerspiegelt, und entsprechende Varianz σ_K^2 hat, also

$$\int_{\mathbb{R}} K(t)dt = 1, \quad \int_{\mathbb{R}} tK(t)dt = 0, \quad \int_{\mathbb{R}} t^2K(t)dt = \sigma_K^2 > 0\,.$$

Die unbekannte Populationsdichte $f(x)$, die geschätzt werden soll, besitze stetige Ableitungen aller benötigten Ordnungen.
Für eine Zufalls-Stichprobe X_i, $i = 1, \ldots, n$, aus dieser Population ergibt sich

$$\mathrm{E}\left(\hat{f}_h(x)\right) = \frac{1}{nh}\, n\mathrm{E}\left(K\left(\frac{x - X_i}{h}\right)\right) = \frac{1}{h}\int_{\mathbb{R}} K\left(\frac{x-y}{h}\right) f(y)dy\,,$$

also $\mathrm{E}(K((x - X)/h)) = h\mathrm{E}(\hat{f}_h(x))$, sowie damit auch

$$\begin{aligned}
\mathrm{var}\left(\hat{f}_h(x)\right) &= \frac{1}{n^2h^2}\, n\mathrm{var}\left(K\left(\frac{x - X_i}{h}\right)\right) \\
&= \frac{1}{nh^2}\left[\mathrm{E}\left(K^2\left(\frac{x - X_i}{h}\right)\right) - \mathrm{E}^2\left(K\left(\frac{x - X_i}{h}\right)\right)\right] \\
&= \frac{1}{nh^2}\int_{\mathbb{R}} K^2\left(\frac{x-y}{h}\right) f(y)dy - \frac{1}{n}\mathrm{E}^2\left(\hat{f}_h(x)\right)\,.
\end{aligned}$$

Der Bias an einer Stelle x hängt also nur von h und nicht von n ab und es gilt

$$\mathrm{bias}_h(x) = \mathrm{E}\left(\hat{f}_h(x)\right) - f(x) = \frac{1}{h}\int_{\mathbb{R}} K\left(\frac{x-y}{h}\right) f(y)dy - f(x)\,.$$

Setzt man $y = x - ht$, so folgt

$$\text{bias}_h(x) = \int_{\mathbb{R}} K(t) f(x-ht) dt - f(x) = \int_{\mathbb{R}} K(t) \Big(f(x-ht) - f(x) \Big) dt \,,$$

da $f(x) \int K(t) dt = f(x)$ gilt. Entwickelt man $f(x-ht)$ um $h = 0$ in eine Taylorreihe, also

$$f(x-ht) = f(x) - htf'(x) + \frac{1}{2} h^2 t^2 f''(x) - \dots \,,$$

so ergibt dies unter Verwendung der Anforderungen an K für kleine Werte von h

$$\begin{aligned} \text{bias}_h(x) &= -hf'(x) \underbrace{\int_{\mathbb{R}} tK(t) dt}_{=0} + \frac{1}{2} h^2 f''(x) \underbrace{\int_{\mathbb{R}} t^2 K(t) dt}_{=\sigma_K^2} - \dots \\ &\approx \frac{1}{2} h^2 \sigma_K^2 f''(x) \,. \end{aligned}$$

Der integrierte quadratische Bias ist somit approximativ

$$\int_{\mathbb{R}} \text{bias}_h^2(x) dx \approx \frac{1}{4} h^4 \sigma_K^4 \int_{\mathbb{R}} f''(x)^2 dx \,.$$

Als approximative Varianz des Kernschätzers folgt mit $\text{bias}_h(x) = O(h^2)$

$$\begin{aligned} \text{var}\left(\hat{f}_h(x) \right) &= \frac{1}{nh^2} \int_{\mathbb{R}} K^2 \left(\frac{x-y}{h} \right) f(y) dy - \frac{1}{n} \Big(\text{bias}_h(x) + f(x) \Big)^2 \\ &= \frac{1}{nh} \int_{\mathbb{R}} K^2(t) f(x-ht) dt - \frac{1}{n} \left(O(h^2) + f(x) \right)^2 . \end{aligned}$$

Nimmt man nun an, dass h klein und n groß wird, erhält man dafür

$$\begin{aligned} \text{var}\left(\hat{f}_h(x) \right) &= \frac{1}{nh} \int_{\mathbb{R}} K^2(t) f(x-ht) dt + O\left(h^2 n^{-1} \right) \\ &\approx \frac{1}{nh} f(x) \int_{\mathbb{R}} K^2(t) dt \,. \end{aligned}$$

Da $f(x)$ eine Dichtefunktion ist, folgt für die über alle x integrierte Varianz

$$\int_{\mathbb{R}} \text{var}\left(\hat{f}_h(x) \right) dx \approx \frac{1}{nh} \int_{\mathbb{R}} K^2(t) dt \cdot \underbrace{\int_{\mathbb{R}} f(x) dx}_{=1} = \frac{1}{nh} \int_{\mathbb{R}} K^2(t) dt \,.$$

Mit dem Ergebnis von zuvor ergibt sich approximativ

$$\begin{aligned} \text{IMSE}(h, K) &= \int_{\mathbb{R}} \text{var}\left(\hat{f}_h(x) \right) dx + \int_{\mathbb{R}} \text{bias}_h^2(x) dx \\ &\approx \frac{1}{nh} \int_{\mathbb{R}} K^2(t) dt + \frac{1}{4} h^4 \sigma_K^4 \int_{\mathbb{R}} f''(x)^2 dx \,. \end{aligned}$$

Optimale Wahl von h und K:
Vom Gesichtspunkt des minimalen IMSE, folgt nach Nullsetzen von

$$\frac{\partial}{\partial h}\text{IMSE}(h,K) \approx -\frac{1}{nh^2}\int_{\mathbb{R}} K^2(t)dt + h^3\sigma_K^4 \int_{\mathbb{R}} f''(x)^2 dx$$

als idealer Wert von h

$$h_{\text{opt}} \approx \sigma_K^{-4/5}\left\{\int K^2(t)dt\right\}^{1/5}\left\{n\int f''(x)^2dx\right\}^{-1/5}.$$

Leider hängt h_{opt} wiederum von der unbekannten Dichte f respektive von deren Rauheit (roughness) $\int f''(x)^2dx$ ab. Man sieht jedoch, dass h_{opt} mit wachsendem n langsam gegen Null konvergiert. Als minimaler IMSE resultiert damit

$$\text{IMSE}(h_{\text{opt}},K) \approx \frac{5}{4}C(K)\left\{\int f''(x)^2dx\right\}^{1/5} n^{-4/5} \quad \text{mit} \quad C(K) = \left\{\sigma_K \int K^2(t)dt\right\}^{4/5}.$$

Man sollte also Kerne verwenden, die kleine Werte von $C(K)$ erzeugen. Für den standardisierten ($\sigma_K^2 = 1$) Gaußkern

$$K(t) = \frac{1}{\sqrt{2\pi}}\exp(-t^2/2) \quad \text{für} \quad t \in \mathbb{R}$$

resultiert dafür beispielsweise

$$C(K) = \left\{\int \frac{1}{2\pi}\exp(-t^2)dt\right\}^{4/5} = \left\{\frac{1}{2\sqrt{\pi}}\right\}^{4/5} = 0.3633.$$

Die Minimierung von $C(K)$ für standardisierte Kerne ($\sigma_K^2 = 1$) reduziert sich auf die Minimierung von $\int K^2(t)dt$. Hodges und Lehmann zeigten, dass der *Epanechnikov-Kern*

$$K(t) = \frac{3}{4\sqrt{5}}\left(1 - \frac{1}{5}t^2\right) \text{für} \quad -\sqrt{5} \le t \le \sqrt{5}$$

eine Lösung dieses Minimierungsproblems darstellt.
Verwendet man den standardisierten Gaußkern unter der Annahme einer $N(0,\sigma^2)$-verteilten Stichprobe mit Dichte f als Referenz, so erhält man

$$\begin{aligned} f''(x)^2 &= \frac{1}{2\pi\sigma^{10}}(x^2-\sigma^2)^2\exp(-x^2/\sigma^2) \\ \int f''(x)^2dx &= \frac{3}{8\sqrt{\pi}\sigma^5} \approx 0.212\sigma^{-5}, \end{aligned}$$

woraus für die optimale Fensterbreite

$$h_{\text{opt}} = (4\pi)^{-1/10}\left(\frac{3}{8}\pi^{-1/2}\right)^{-1/5}\sigma n^{-1/5} = \left(\frac{4}{3}\right)^{1/5}\sigma n^{-1/5} \approx 1.06\sigma n^{-1/5}$$

folgt. Silverman zeigte, dass dieses Ergebnis für eine breite Klasse von verschieden verteilten Stichproben gute Resultate bezüglich der Minimierung von IMSE(h,K) liefert. Die Tabelle 2.4 beinhaltet die optimalen Werte von h für verschiedene Kernfunktionen in ihrer standardisierten Form mit $\sigma_K^2 = 1$ (vgl. auch Abbildung 2.13), wobei als Referenz angenommen wurde, dass die Stichprobe aus einer Normalverteilung stammt.

```
> kernels <- eval(formals(density)$kernel) # yields all available kernels
> for(i in 2:length(kernels))
    plot(density(0,bw=1,kern=kernels[i]), xlab=kernels[i])

> plot(density(VC, bw=33, kern="epanechnikov"))
```

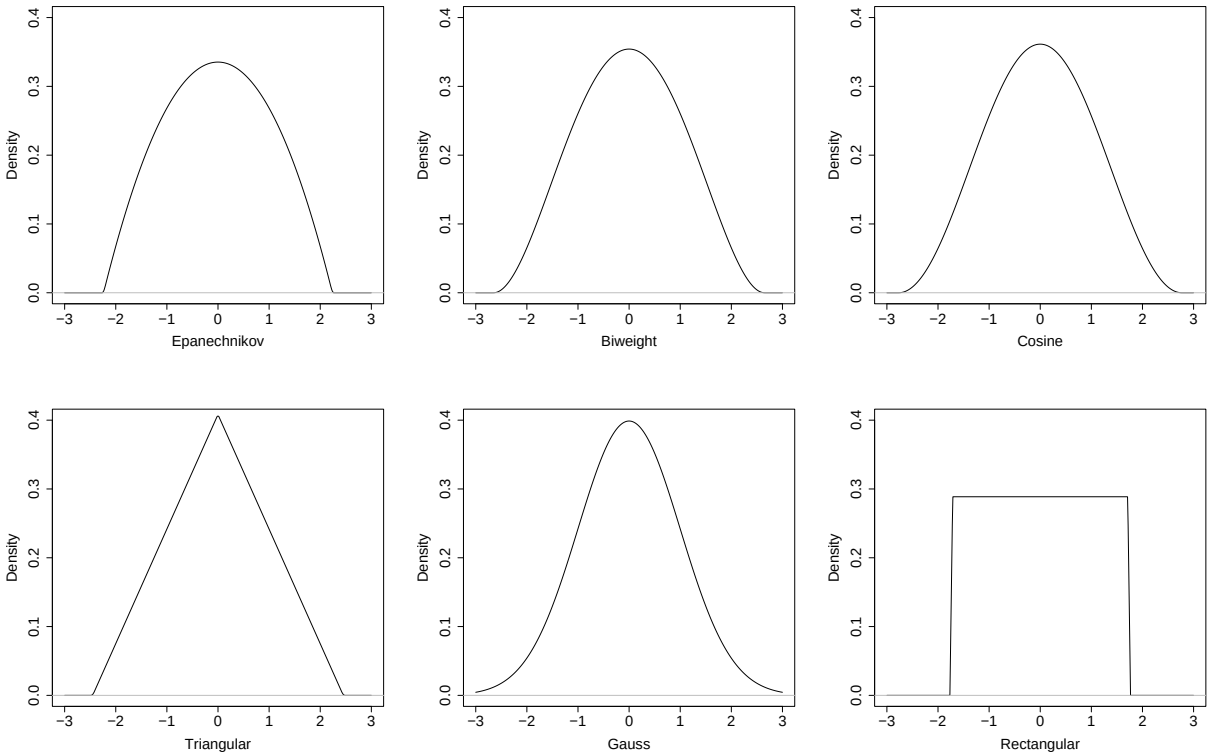


Abbildung 2.13: Vergleich der standardisierten Kerne.

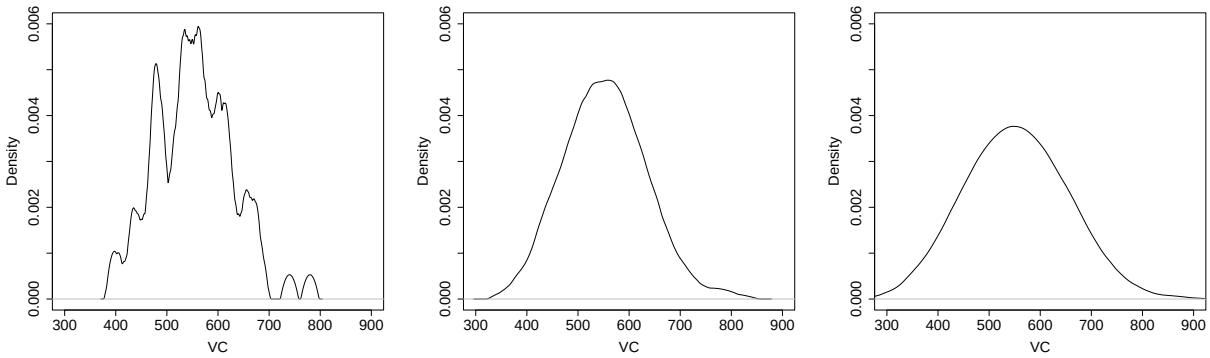


Abbildung 2.14: Kernschätzung der `VC`-Dichte mittels Epanechnikov-Kern mit Bandbreiten $h = 8$, $h = 33$ und $h = 70$.

Kern	$K(t)$	für	h_{opt}
Epanechnikov	$\frac{3}{4\sqrt{5}}\left(1-\frac{t^2}{5}\right)$	$\lvert t\rvert \le \sqrt{5}$	$1.04867\sigma n^{-1/5}$
Biweight	$\frac{15}{16}7^{-5/2}\left(7-t^2\right)^2$	$\lvert t\rvert \le \sqrt{7}$	$1.04996\sigma n^{-1/5}$
Cosinus	$\frac{1}{2\pi}\sqrt{\frac{\pi^2-6}{3}}\left(1+\cos t\sqrt{\pi^2-6}/\sqrt{3}\right)$	$\lvert t\rvert \le \pi\sqrt{\frac{3}{\pi^2-6}}$	$1.05086\sigma n^{-1/5}$
Dreieck	$\frac{1}{6}\left(\sqrt{6}-\lvert t\rvert\right)$	$\lvert t\rvert \le \sqrt{6}$	$1.05166\sigma n^{-1/5}$
Gauß	$\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$	$t \in \mathbb{R}$	$1.05922\sigma n^{-1/5}$
Rechteck	$\frac{1}{2\sqrt{3}}$	$\lvert t\rvert \le \sqrt{3}$	$1.06412\sigma n^{-1/5}$

Tabelle 2.4: Standardisierte Kernfunktionen mit optimalen Werten für die Bandbreiten h.

2.5 Graphische Darstellungen

2.5.1 Der Symmetrie-Plot

Speziell für die Überprüfung der Symmetrie in der Verteilung lässt sich eine eigene Graphik konstruieren bei der die Differenzen über dem Median $u_i = x_{(n+1-i)} - \tilde{x}$ gegen jene unter dem Median $v_i = \tilde{x} - x_{(i)}$, $i = 1, \ldots, [(n+1)/2]$, aufgetragen werden. Symmetrie um den Median $\tilde{x}$ liegt gerade dann vor, wenn diese Punkte auf der Geraden $u = v$ (Referenzlinie) liegen.

Bedeutung der Symmetrie:

- Bei Symmetrie hat die Verteilung ein eindeutiges Zentrum (Median, Erwartungswert und Modalwert sind ident).
- Eine meist einfachere Beschreibung des die Daten erzeugenden Prozesses ist möglich.
- Viele statistische Prozeduren (Wilcoxon-Test, klassische Tests) beruhen auf der Annahme einer symmetrischen Verteilung.
- Eine erkannte Asymmetrie kann dann durch Transformationen oft beseitigt werden (z.B. $\sqrt{x}$, x^2, ...).

Beispiel 2.12 *In der Abbildung 2.15 ist deutlich zu erkennen, dass die Verteilung der Variablen* `age` *nicht symmetrisch zu sein scheint, während die Annahme von Symmetrie bei den Variablen* `VC` *und* `FEV1.VC` *doch gerechtfertigt ist.*

```
> x <- VC;  n <- length(x);  i <- 1:trunc((n+1)/2)
> u <- sort(x)[n+1-i] - median(x);  v <- median(x) - sort(x)[i]
> lim <- c(0, max(u,v))
> plot(v, u, xlim=lim, ylim=lim);  abline(0,1)
```

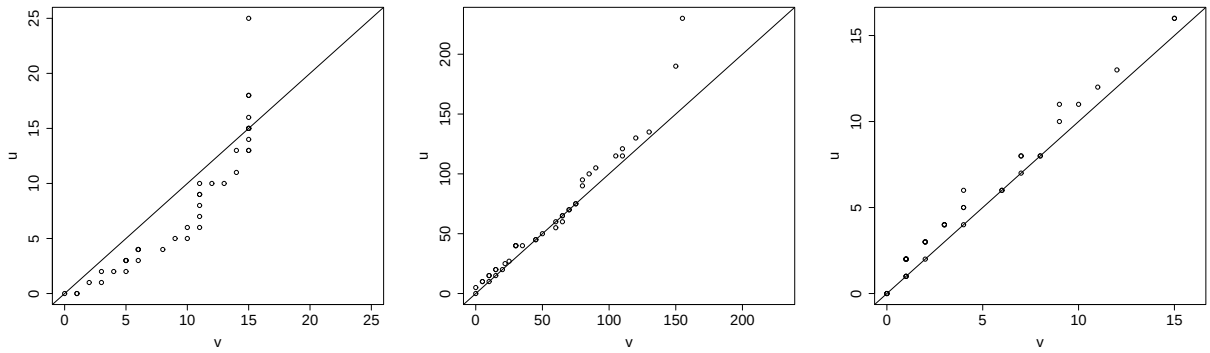


Abbildung 2.15: Symmetrie-Plot von `age` (links), `VC` (Mitte), `FEV1.VC` (rechts).

2.5.2 Die empirische Verteilungsfunktion

Hierbei werden die Wahrscheinlichkeiten $p_i = F(x_{(i)}) = i/n$ gegen die empirischen Quantile $q(p_i) = x_{(i)}$ für $i = 1, \ldots, n$ aufgetragen. Aus diesem Punktverlauf sind Quantile, IQR, Range, Extremwerte, sowie andere Charakteristiken der Verteilung ersichtlich. Häufig auftretende Datenwerte (hohe Konzentrationen) ergeben einen steilen Punktverlauf. Je steiler der Verlauf, desto größer ist dort die lokale Dichte. Die empirische Verteilungsfunktion beinhaltet die Rohdaten und somit die vollständige Information.

Beispiel 2.13 *Die empirischen Verteilungsfunktionen der Variablen* `age`*,* `VC` *und* `FEV1.VC` *sind in der Abbildung 2.16 dargestellt. Im Vergleich zu einer Normalverteilung weist die empirische Verteilung von* `age` *viel zu kurze Schwänze auf. Weiters fällt auf, dass F_n erst bei etwa 0.15 beginnt. Dies liegt darin begründet, dass fast 15% der Personen gerade 16 Jahre alt sind. Bei den beiden anderen scheinen keine besonderen Abweichungen von einem Normalverteilungsmodell ersichtlich.*

```
> library(stepfun)
> plot.ecdf(age)
```

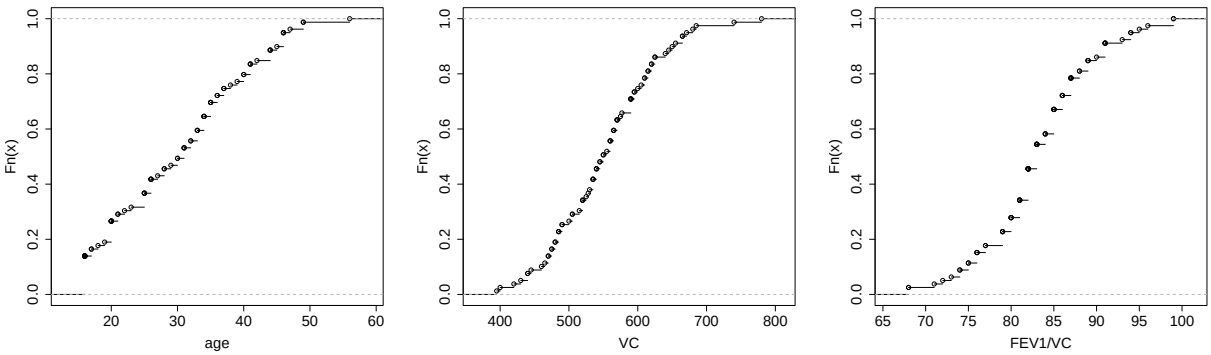


Abbildung 2.16: Empirische Verteilungen von `age` (links), `VC` (Mitte), `FEV1.VC` (rechts).

Ein **Konfidenzband** für $F(x)$ kann zusätzlich eingezeichnet werden. Als **punktweises** Band (nur im betrachteten x ist die Überdeckung $1-\alpha$) erhält man

$$F_n(x) \pm z_{1-\alpha/2}\sqrt{F_n(x)(1-F_n(x))/n}\,.$$

Hingegen soll für ein **simultanes** Band $(U(x), O(x))$ gelten, dass

$$P(\forall x \in \mathbb{R}|\ U(x) \le F(x) \le O(x)) = 1-\alpha\,.$$

Mit der Kolmogorov-Smirnov Statistik kann ein solches konstruiert werden. Wegen

$$\begin{aligned} 1-\alpha &= P\left(\sup_x |F(x)-F_n(x)| \le k_{1-\alpha}\right) \\ &= P\left(\sup_x (F(x)-F_n(x)) \le k_{1-\alpha} \text{ und } \sup_x (F_n(x)-F(x)) \le k_{1-\alpha}\right) \\ &= P\left(\forall x \in \mathbb{R}|\ F_n(x)-k_{1-\alpha} \le F(x) \le F_n(x)+k_{1-\alpha}\right) \end{aligned}$$

folgt, da $0 \le F(x) \le 1$ der natürliche Wertebereich von $F(x)$ ist, als simultanes Konfidenzband in allen $x \in \mathbb{R}$

$$P\left(\forall x \in \mathbb{R}|\ \max_x(0, F_n(x)-k_{1-\alpha}) \le F(x) \le \min_x(1, F_n(x)+k_{1-\alpha})\right) = 1-\alpha\,.$$

Beispiel 2.14 *Für die Benzinverbrauchsdaten aus Beispiel 2.4 ergibt sich für* $\alpha = 0.05$ *als KS-Quantil* $k_{0.95} = 0.409$. *Als simultanes 95%-Konfidenzband für* $F(x)$ *ergibt sich*

$$0 \le F_n(x) \pm 0.409 \le 1 \quad \forall x \in \mathbb{R}.$$

```
> lines(c(0, milage, 16), pmin(1, (0:(n+1))/n + 0.409), type="s")
> lines(c(0, milage, 16), pmax(0, (0:(n+1))/n - 0.409), type="s")
```

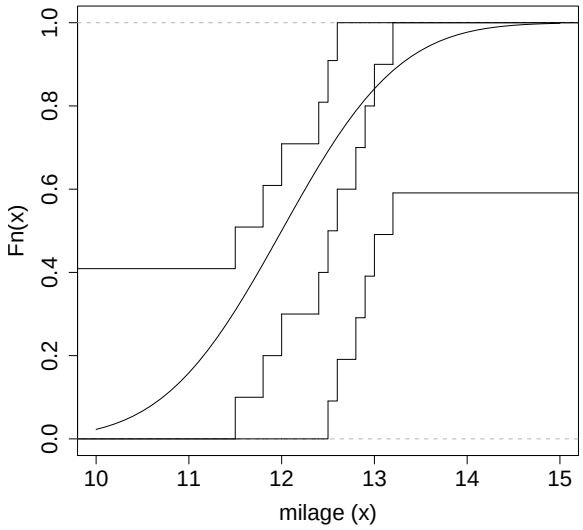


Abbildung 2.17: Empirische Verteilungsfunktion zu Beispiel 2.4 mit $N(12,1)$-Verteilung.

2.5.3 Vergleich empirische mit theoretische Verteilung

Es gibt vielfältige Grunde für eine derartige graphische Gegenüberstellung:

1. Die Daten werden als Stichprobe aus einer theoretischen Verteilung klassifiziert. Ist diese die Normalverteilung, so wäre es möglich, klassische Verfahren anzuwenden.

2. Kennt man die Verteilung der Stichprobe, so führt dies zu einem besseren Verständnis des physikalischen Mechanismuses, der die Daten generiert hat.

3. Die Transformation zu einer Normalverteilung kann leichter durchgeführt werden, wenn die theoretische Verteilung bekannt ist.

Der theoretische Quantil-Quantil-Plot

Beim theoretischen Quantil-Quantil (TQQ)-Plot werden die Quantile der empirischen Verteilung gegen entsprechende Quantile einer theoretischen Verteilung $F(x)$ aufgetragen. Wählt man beispielsweise hierbei $p_i = (i-1/2)/n$, $i = 1, \ldots, n$, so führt dies mit Definition 2.4 für np nicht ganzzahlig zum empirischen Quantil $Q(p_i) = X_{([np_i]+1)} = X_{([i-1/2]+1)} = X_{(i)}$, der i-ten geordneten Stichprobe. Im TQQ-Plot werden daher

$$q(p_i) = x_{(i)} \quad \text{gegen} \quad q_F(p_i) = F^{-1}(p_i) \qquad \text{mit} \quad p_i = \frac{i-1/2}{n} \quad \text{für} \quad i = 1, \ldots, n$$

aufgetragen.
Wird für $F(x)$ die Standard-Normalverteilung verwendet, so spricht man vom **Normal-Plot**. Gerne nimmt man auch die Halb-Normalverteilung und konstruiert den **Half-Normal-Plot**. Es können aber auch Vergleiche mit beliebigen anderen Verteilungen angestellt werden.
Falls die vorliegende Zufallsstichprobe $X_1, \ldots, X_n$ aus der uns unbekannten Populationsverteilung $G(x)$ stammt, so folgt unter Verwendung des Satzes 2.1, dass für n groß

$$X_{(i)} \approx \mathrm{E}(X_{(i)}) \approx q_G(p_i)$$

gilt. Wir betrachten nun den TQQ-Plot, in dem als theoretische Verteilung $F(x)$ verwendet wird. Wurde F darin richtig gewählt, d.h. $F = G$, dann ist $x_{(i)} \approx q_F(p_i)$ annähernd eine Gerade. Falls jedoch $G(x) = F((x-a)/b)$ gilt, also G und F unterschiedliche Lokations- und Skalenparameter aufweisen, dann gilt

$$p_i = G(q_G(p_i)) = F((q_G(p_i) - a)/b) \qquad \Rightarrow \qquad q_F(p_i) = (q_G(p_i) - a)/b\,,$$

also $q_G(p_i) = bq_F(p_i) + a$. Für $x_{(i)}$ wird daher gelten

$$x_{(i)} \approx bq_F(p_i) + a\,,$$

und der TQQ-Plot mit F wird Punkte beinhalten, die auf einer Geraden mit Intercept a und Steigung b liegen.

Folgende Aussagen können daher aus dem Verlauf der Punkte im TQQ-Plot mit F getroffen werden: Bilden die Punkte

1. die Gerade $y = x$, so ist die theoretische Verteilung F eine gute Approximation von G, d.h. $x_{(i)} \approx \mathrm{E}(X_{(i)})$,
2. eine Gerade parallel zu $y = x$, so liegt der Unterschied zwischen F und G nur im Lageparameter, d.h. $x_{(i)} + a \approx \mathrm{E}(X_{(i)})$,
3. eine Gerade, die $y = x$ schneidet, so beruht der Unterschied im Skalierungsparameter, d.h. $bx_{(i)} \approx \mathrm{E}(X_{(i)})$,
4. kein Geradenmuster, so liegt der Stichprobe eine andere Verteilung als F zu Grunde.

Das lineare Muster kann aus vielerlei Gründen verfälscht werden:

1. Sind einige wenige Punkte an den Enden der Geraden weiter entfernt, so kann dies Ausreißer hinweisen.
2. Zeigt am rechten Ende die Krümmung nach oben (oder links nach unten), so hat die empirische Verteilung rechts (bzw. links) längere Schwänze als die theoretische.
3. Vergleicht man eine unsymmetrische Verteilung gegen eine symmetrische theoretische Verteilung, so erhält man ein Kurvenmuster mit von links nach rechts steigender Krümmung (Daten sind rechtsschief) oder entsprechend umgekehrt.
4. Plateaus oder Sprünge im Plot weisen auf hohe Datenkonzentrationen an einer Stelle oder fehlende Beobachtungen über einen größeren Bereich hin.

Verteilung		Dichte $f(x)$	Bereich	Standardform
Gleich	$U(a,b)$	$\frac{1}{b-a}$	$a < x < b$	$a = 0,\ b = 1$
Normal	$N(\mu,\sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$x \in \mathbb{R},\ \sigma > 0$	$\mu = 0,\ \sigma^2 = 1$
Half-Normal	$H(\sigma^2)$	$\frac{1}{\sigma}\sqrt{\frac{2}{\pi}}\exp\left(-\frac{x^2}{2\sigma^2}\right)$	$x,\ \sigma > 0$	$\sigma^2 = 1$
Gamma	$G(a,\lambda)$	$\frac{x^{a-1}\lambda^a\exp(-\lambda x)}{\Gamma(a)}$	$x,\ a,\ \lambda > 0$	$\lambda = 1$
Exponential	$E(\lambda)$	$\lambda\exp(-\lambda x)$	$x,\ \lambda > 0$	$\lambda = 1$
χ^2	$\chi^2(\nu)$	$\frac{x^{\nu/2-1}\exp(-x/2)}{2^{\nu/2}\Gamma(\nu/2)}$	$x,\ \nu > 0$	

Tabelle 2.5: Standardformen von Verteilungen für den TQQ-Plot.

Die Verteilungen in der Tabelle 2.5 bieten sich gut als Referenzverteilung F an. Die Normalverteilung hat hierbei den Vorteil, dass sie nur von einem Lokations- und Skalenparameter abhängt. Trifft dieses Verteilungsmodell zu, so resultiert als TQQ-Plot eine Gerade die eventuell verschoben ist oder eine andere Steigung aufweist. Dies gilt auch für die Gleich- und Exponentialverteilung. Andere Verteilungen wie etwa die Gammaverteilung haben auch einen Gestaltsparameter, der vor oder mittels der graphischen Darstellung festgelegt werden muss.

Es ist vorteilhaft, die empirische Verteilung mit einer **Standardform** der theoretischen Verteilung zu vergleichen (siehe Tabelle 2.5). Aus dem dadurch entstandenen TQQ-Plot kann optisch der Lokations- und Skalenparameter geschätzt werden.

Beispiel 2.15 *Die Abbildung 2.18 vergleicht die* `VC`*-Quantile mit Quantilen diverser* $N(\mu, \sigma^2)$*-Verteilungen. Links wird dafür* $\mu = 550$ *und* $\sigma = 75$ *verwendet. Der Punktverlauf bestätigt die sehr gute Anpassung. Die beiden anderen Plots zeigen, dass* $\sigma = 50$ *(Mitte) und* $\mu = 600$ *(rechts) falsch gewählt sind. In jeder Graphik ist zusätzlich auch die erste Mediane eingezeichnet (*`abline(0, 1)`*).*

```
> n <- length(VC);  p <- (1:n - 1/2)/n  # direct calculation of the p_i's
> p <- ppoints(n)                       # the same as above but much faster
> plot(qnorm(p, 550, 75), sort(VC))  # produces N(550,75)-plot
> qqplot(qnorm(p, 550, 75), VC)      # produces N(550,75)-plot, no sorting requ.
> qqplot(qnorm(p, mean(VC), 50), VC) # produces N(mean(VC),50)-plot
> qqplot(qnorm(p, 600, sd(VC)), VC)  # produces N(600,sd(VC))-plot
```

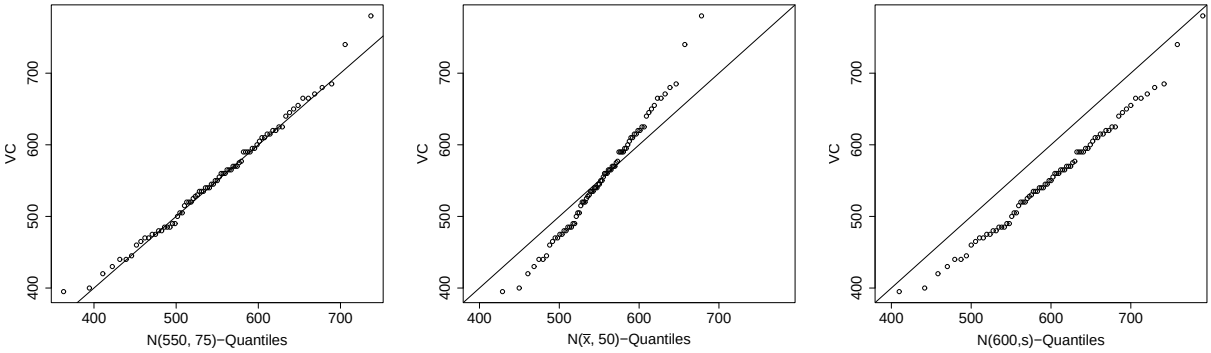


Abbildung 2.18: TQQ-Plot von `VC` verglichen mit Normalverteilungen .

In Abbildung 2.19 finden sich Vergleiche mit anderen Verteilungen. Links ist ein Vergleich der `VC`*-Quantile mit der* $N(0,1)$*-Verteilung zu sehen; dies ist ein Normal-Plot der eine optische Schätzung der beiden Momente* μ*, und* σ *erlaubt. So ist die eingezeichnete Gerade durch das empirische und theoretische erste und dritte Quartil bestimmt. Intercept und Slope (548.75, 79.69) sind ausgezeichnete Schätzer für Erwartungswert und Standardabweichung (553.49, 76.27). In der Mitte werden die Quantile von* `age` *mit einer Exponential*$(a, rate = 1/\mu)$*-Verteilung verglichen. Dazu wird zuerst die Lokation* a *auf das minimale Alter von 16 Jahren gesetzt und für* `rate` *die Schätzung* `1/mean(age-16)` *verwendet. Die empirische Verteilung von* `age` *hat aber viel kürzere Schwänze als die Exponentialverteilung. Rechts wird* `FEV1.VC` *einer um den Mittelwert verschobenen und skalierten* t_ν*-Verteilung gegenüber gestellt. Die Varianz einer* t_ν*-Verteilung ist* $\nu/(\nu - 2)$*, was etwa der empirischen Varianz entsprechen soll. Für* $\nu = 10$ *(Varianz 10/8) werden deshalb die theoretischen Quantile mit dem Faktor 5.75 gestreckt, was dann die gewünschte Standardabweichung von* 6.43 *ergibt.*

```
> qqnorm(VC)  # directly produces a N(0,1)-plot for VC
> qqline(VC)  # adds a line which passes through 1st and 3rd quartiles

> qqplot(16+qexp(p, 1/15), age); abline(0,1)  # produces Exp-plot for age
> qqplot(mean(FEV1.VC)+qt(p,10)*5.75, FEV1.VC); abline(0,1) #t-plot for FEV1.VC
```

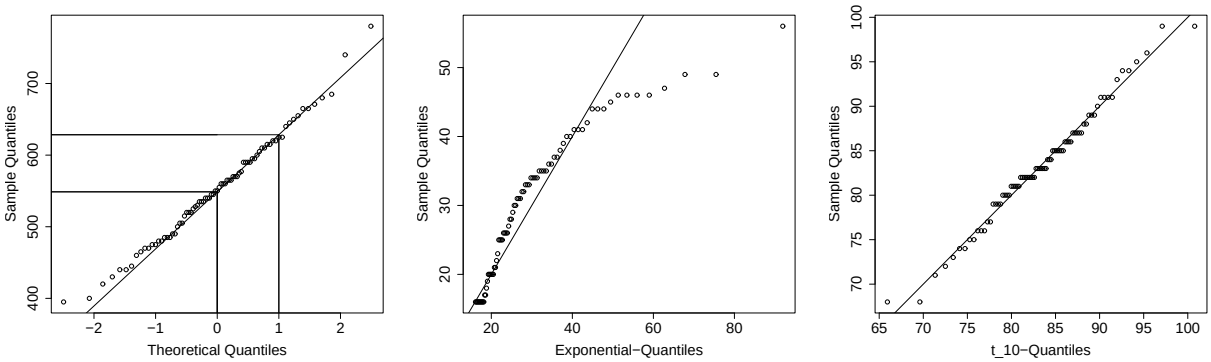


Abbildung 2.19: TQQ-Plot von `VC` (links), `age` (Mitte) und `FEV1.VC` (rechts).

Der theoretische Prozent-Prozent-Plot

Da der TQQ-Plot bei Unterschieden im Verteilungszentrum nicht sensitiv ist, wird er oft mit dem theoretischen Prozent-Prozent (TPP)-Plot kombiniert. Dieser erkennt jedoch keine Unterschiede in den Schwänzen der Verteilung, da $q(p)$ eine stark wachsende Funktion für extreme p ist, während $F(x)$ eher im Zentralbereich einen steilen Verlauf hat. Für den TPP-Plot werden die empirischen gegen die theoretischen Prozentwerte aufgetragen, also

$$p_i = \frac{i - 1/2}{n} \quad \text{gegen} \quad F\left(\frac{x_{(i)} - \mu}{\sigma}\right).$$

F bezeichnet die theoretische Verteilung mit der verglichen wird. Das Muster der Punkte ist bei Übereinstimmung linear. Bei Lokations- oder Skalenunterschieden bleibt das Referenzmuster allerdings nicht mehr linear.

Beispiel 2.16 *In der Abbildung 2.20 sieht man die gute Anpassung einer Normalverteilung für* `VC`*, während diesbezüglich deutliche Unterschiede bei* `age` *erkennbar sind.*

```
> plot(pnorm(sort(VC), mean(VC), sd(VC)), p)
> plot(pnorm(sort(age), mean(age), sd(age)), p)
```

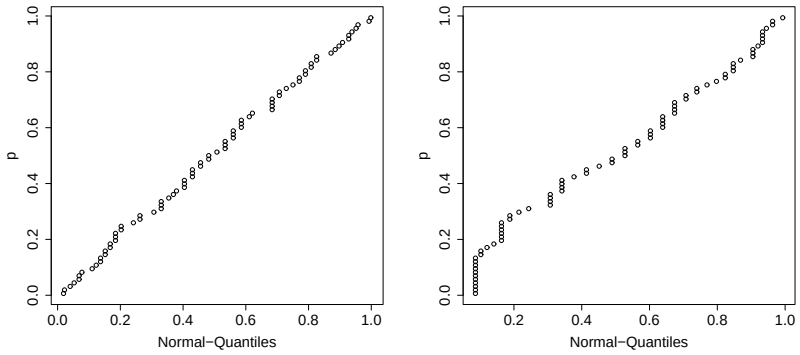


Abbildung 2.20: TPP-Plot der Variablen `VC` (links) und `age` (rechts).

Kapitel 3

Vergleich zweier eindimensionaler Stichproben

Beim *Zwei-Stichproben-Problem* unterscheidet man, ob **unabhängige** Stichproben (von zwei Populationen wird die selbe Variable beobachtet) oder **verbundene/abhängige** Stichproben (in einer Population wird eine Variable zweimal beobachtet) vorliegen. Gegeben seien Zufalls-Stichproben $X_1, \ldots, X_m$ sowie $Y_1, \ldots, Y_n$. Diese sollen beispielsweise Aufschluss über den Parameter $\theta = E(Y) - E(X)$ geben. Als Teststatistik bietet sich $T = \overline{Y} - \overline{X}$ an. Berechnet man die Varianz von T, so lässt sich ein prinzipieller Unterschied zwischen den beiden Situationen beobachten. Gilt $m = n$, so erhält man

$$\begin{aligned} \mathrm{var}(T) &= \mathrm{var}(\overline{Y}) + \mathrm{var}(\overline{X}) - 2\mathrm{cov}(\overline{Y}, \overline{X}) \\ &= \frac{1}{n}\mathrm{var}(Y) + \frac{1}{n}\mathrm{var}(X) - \frac{2}{n}\mathrm{cor}(X,Y)\sqrt{\mathrm{var}(Y)\mathrm{var}(X)}\,. \end{aligned}$$

Während für unabhängige Komponenten $\mathrm{cor}(X, Y) = 0$ gilt, ist dies bei abhängigen Stichproben nicht der Fall. Werden die Variablen (X_i, Y_i) am selben Objekt i beobachtet, so werden X_i und Y_i häufig positiv korreliert sein. Dies impliziert dann $\mathrm{cor}(X, Y) > 0$, und somit eine Verringerung der Varianz von T.
In diesem Kapitel werden Methoden zur Analyse zweier unabhängiger Zufalls-Stichproben $X_1, \ldots, X_m$ und $Y_1, \ldots, Y_n$ eingeführt. Den Verfahren für verbundene Stichproben in der Form $(X_1, Y_1), \ldots, (X_n, Y_n)$ ist dann das folgende Kapitel gewidmet.

3.1 Graphische Verfahren

Unterschiede in den Verteilungsmodellen zweier unabhängiger Stichproben $X_1, \ldots, X_m$ und $Y_1, \ldots, Y_n$ des gleichen Merkmales können häufig mittels graphischer Methoden leicht erkannt werden.
Der **empirische Quantil-Quantil-Plot** (EQQ-Plot) stellt die Punkte

$$q_Y(p) \leftrightarrow q_X(p), \quad 0 < p < 1$$

dar. Für $n = m$ wird der EQQ-Plot durch die Punkte $(x_{(i)}, y_{(i)})$ gebildet. Gilt $n \neq m$, wird für gewöhnlich die kleinere Stichprobe genommen und die Quantile der größeren

Stichprobe werden durch Interpolation bestimmt. Bei $m < n$ wird $x_{(i)} = q_X((i-1/2)/m)$ gegen das interpolierte $(i-1/2)/m$-Quantil der y-Daten aufgetragen. Dazu wird ein Wert v benötigt, für den $(v-1/2)/n = (i-1/2)/m$ gilt, woraus $v = (i-1/2)n/m+1/2$ folgt. Ist v ganzzahlig, wird $x_{(i)}$ gegen $y_{(v)}$ gezeichnet. Sonst sei j der ganzzahlige Anteil von v und θ der positive Rest. Damit folgt $q_Y((i-1/2)/m) = (1-\theta)y_{(j)} + \theta y_{(j+1)}$. Für $m = 50$ und $n = 100$ folgt beispielsweise $v = (i-1/2)100/50 + 1/2 = 2i - 1/2$, so dass $j = 2i - 1$ und $\theta = 1/2$ gilt. D.h. $x_{(1)}$ wird gegen $1/2y_{(1)} + 1/2y_{(2)}$ gezeichnet, $x_{(2)}$ gegen $1/2y_{(3)} + 1/2y_{(4)}$, usw.
Sind X und Y ident verteilt, dann ergibt der EQQ-Plot die Gerade $x = y$. Beim Vergleich von Verteilungen mit langen Schwänzen tendiert der EQQ-Plot dazu, die Unterschiede in diesen Schwänzen eher zu übertreiben, in den Zentren aber Differenzen zu verwischen.

Beispiel 3.1 *Im linken Teil der Abbildung 3.1 werden die* `VC` *Beobachtungen von jungen mit jenen von älteren Personen verglichen. Deutlich ist zu erkennen, dass sämtliche Quantile der Jungen durchwegs größere Werte haben als die entsprechenden bei den älteren Personen. Im rechten Teil ist zu sehen, dass die Verteilung von* `VC` *bei den Murauern einen kürzeren linken Schwanz zu haben scheint als bei den Aichfeldern.*

```
> qqplot(VC[age<30], VC[age>=30]);  abline(0, 1)
> qqplot(VC[region=="A"], VC[region=="M"]);  abline(0, 1)
```

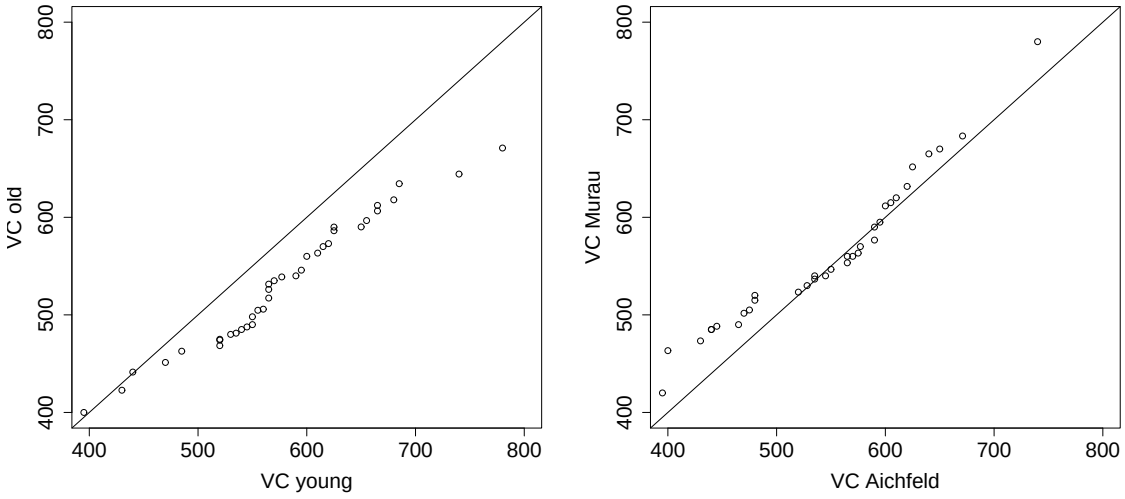


Abbildung 3.1: EQQ-Plot von `VC` für den Vergleich jung/alt (links), sowie für die Bezirke Aichfeld/Murau (rechts).

Der **empirische Prozent-Prozent-Plot** (EPP-Plot) ist ein Diagramm der Form

$$p_Y(q) \leftrightarrow p_X(q)\,.$$

Sind X und Y beide $U(0,1)$-verteilt, so sind der EQQ-Plot und der EPP-Plot ident. Bei beliebiger identischer Verteilung von X und Y ergibt der EPP-Plot die lineare Anordnung der Punkte $x = y$. Im Gegensatz zum EQQ-Plot werden nun Unterschiede in den Zentren der Verteilungen erkannt. Auf Unterschiede in den Schwänzen der Verteilungen reagiert er nicht so stark. Deswegen sollten EQQ- und EPP-Plot zusammen verwendet werden.

Beispiel 3.2 *Wie schon im Beispiel zuvor sollen die beiden* `VC`*-Stichproben von jungen und älteren Probanden, sowie von jenen aus Aichfeld und aus Murau miteinander verglichen werden. Leider bietet* `R` *für den EPP-Plot keine eigene Graphik an. Ein solcher lässt sich jedoch ganz einfach erzeugen. Im linken Teil der Abbildung 3.2 ist deutlich zu erkennen, dass die* `VC`*-Beobachtungen der jüngeren Personen Quantilen mit niedrigeren Niveaus entsprechen als dies bei den älteren der Fall ist.*

```
> s <- sort(VC);  n <- length(VC)
> sx <- VC[age< 30];  sy <- VC[age>=30]  # am Beispiel der beiden Altersgruppen
> px <- py <- 1:n
> for (i in 1:n) {
    px[i] <- (sum(sx <= s[i]) - 1/2)/lenx
    py[i] <- (sum(sy <= s[i]) - 1/2)/leny
  }
>  plot(px, py);  abline(0, 1)
```

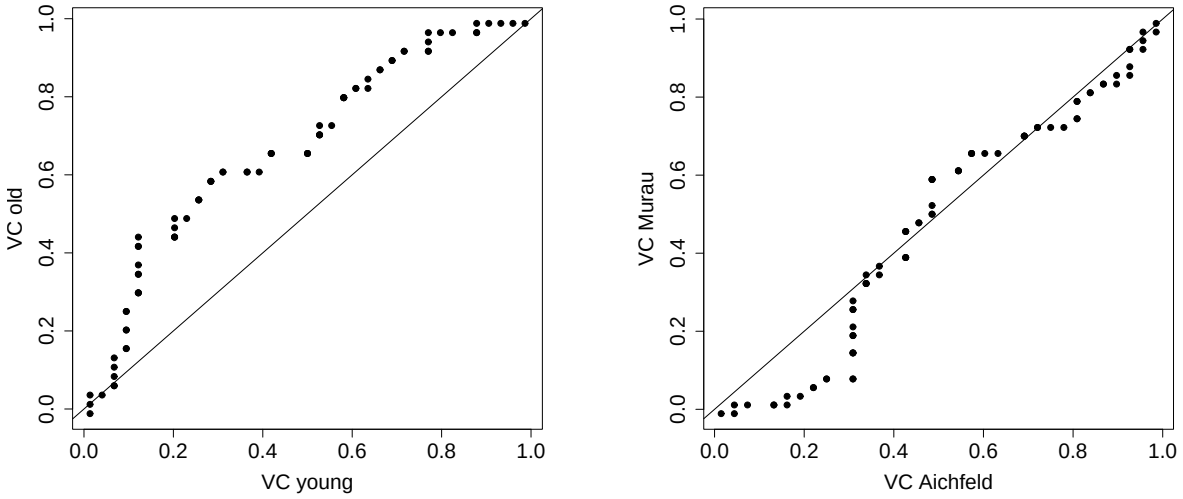


Abbildung 3.2: EPP-Plot der Variablen `VC` für die beiden Altersgruppen (links) und für die beiden der Bezirke Aichfeld/Murau (rechts).

Will man mehr als zwei Gruppen miteinander vergleichen, so eignen sich **Boxplot-Serien**. Für diesen Zweck werden zwei Modifikationen des Boxplots eingeführt.

1. Im **variable-width-box-plot** wird die Breite einer Box proportional zum Stichprobenumfang in der Gruppe gewählt.

2. Im **notched-box-plot** der Abbildung 3.3 werden Kerben (notches) der Form

$$\tilde{X} \pm cS_{\tilde{X}}$$

zusätzlich in die Box eingezeichnet.

Bemerkungen zu den Kerben im Fall zweier Gruppen:
Es soll damit ein graphischer Test auf die Gleichheit der Erwartungswerte beider Mediane $\mathrm{E}(\tilde{X}) = m_X$ und $\mathrm{E}(\tilde{Y}) = m_Y$ derart ermöglicht werden, so dass bei Überlappung der Kerben diese Gleichheitshypothese $H_0 : m_X - m_Y = 0$ auf einem Niveau α nicht verworfen werden kann.
Wir nehmen an, dass die beiden Mediane unterschiedliche Standardfehler haben können, also dass $\tilde{X} \sim N(m_X, \sigma^2)$ und $\tilde{Y} \sim N(m_Y, k^2\sigma^2)$ gilt. Unter H_0 folgt somit wegen der Unabhängigkeit der beiden Gruppen

$$\frac{(\tilde{X} - \tilde{Y}) - (m_X - m_Y)}{\sqrt{1+k^2}\sigma} \overset{H_0}{=} \frac{\tilde{X} - \tilde{Y}}{\sqrt{1+k^2}\sigma} \overset{H_0}{\sim} N(0,1)\,.$$

Wir können somit H_0 nicht verwerfen, falls die Null im Intervall

$$\left[\tilde{X} - \tilde{Y} - z_{1-\alpha/2}\sqrt{1+k^2}\sigma, \tilde{X} - \tilde{Y} + z_{1-\alpha/2}\sqrt{1+k^2}\sigma\right]$$

enthalten ist. Die Null ist aber gerade dann in diesem Intervall, wenn

$$\begin{aligned} \tilde{X} - \tilde{Y} - z_{1-\alpha/2}\sqrt{1+k^2}\sigma &\leq 0 \\ \tilde{X} - \tilde{Y} + z_{1-\alpha/2}\sqrt{1+k^2}\sigma &\geq 0\,. \end{aligned}$$

Teilt man die Konstante auf in $z_{1-\alpha/2}\sqrt{1+k^2}\sigma = c\sigma + ck\sigma = c\sigma(1+k)$, so folgt $c = z_{1-\alpha/2}\sqrt{1+k^2}/(1+k)$ und die beiden obigen Bedingungen sind äquivalent mit

$$\begin{aligned} \tilde{X} - z_{1-\alpha/2}\frac{\sqrt{1+k^2}}{1+k}\sigma &\leq \tilde{Y} + z_{1-\alpha/2}\frac{\sqrt{1+k^2}}{1+k}k\sigma \\ \tilde{X} + z_{1-\alpha/2}\frac{\sqrt{1+k^2}}{1+k}\sigma &\geq \tilde{Y} - z_{1-\alpha/2}\frac{\sqrt{1+k^2}}{1+k}k\sigma\,. \end{aligned}$$

Bei identischer Varianz ($k = 1$) muss für $\alpha = 0.05$ in jedem der beiden Boxplots $c = z_{1-\alpha/2}\sqrt{2}/2 = 1.386$ gewählt werden. Bei $k = 2$ resultiert $c = z_{1-\alpha/2}\sqrt{5}/3 = 1.461$. Wählt man $c = 1.7$, so ist man bei einer Vielzahl von Fällen über dem einzuhaltenden Niveau von $\alpha = 0.05$ und damit auf der sicheren Seite. Für die Konstruktion der Kerben folgt somit

$$\tilde{X} \pm 1.7 S_{\tilde{X}} = \tilde{X} \pm 1.7\frac{1.25\mathrm{IQR}}{1.35\sqrt{n}} = \tilde{X} \pm 1.57\frac{\mathrm{IQR}}{\sqrt{n}}\,.$$

Kommt es dabei zu einer Überlappung, so kann die Gleichheitshypothese für benachbarte Mediane nicht verworfen werden.

Beispiel 3.3 *Für die beiden Gruppierungen jung/älter sowie Aichfeld/Murau sollen die Daten mittels Boxplots dargestellt werden. Die Breite der Box soll dabei proportional zum Umfang der jeweiligen Gruppe sein, und Notches mögen einen Test auf identische Mediane erlauben. Da sich diese Kerben in der Abbildung 3.3 sowohl links (jung/älter) als auch rechts (Aichfeld/Murau) überlappen, kann kein signifikanter Unterschied in den Lokationszentren der jeweiligen Gruppen festgestellt werden.*

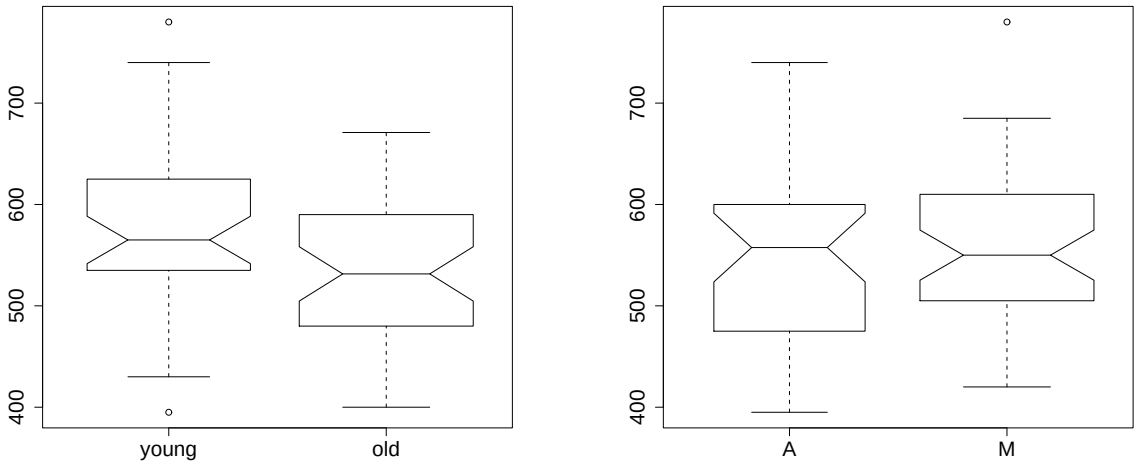


Abbildung 3.3: Boxplots der Variablen `VC` mit Notches und proportionalen Breiten für die beiden Altersgruppen (links) und für die Regionen Aichfeld/Murau (rechts).

```
> a <- as.factor(trunc(age/30));  levels(a) <- c("young", "old")
> boxplot(VC ~ a, varwidth = TRUE, notch=TRUE)
```

Um die Dichtefunktionen zweier Gruppen miteinander vergleichen zu können, werden die beiden gruppenspezifischen **Kernschätzer** in einem Plot übereinander gezeichnet. Dabei ist zu beachten, dass ein korrekter Vergleich dieser Dichteschätzungen gewährleistet ist. Dies bedeutet, dass dabei der Parameter h sowie der Kern K einheitlich gewählt werden müssen.

Beispiel 3.4 *Eine Empfehlung ist die Verwendung des Mittelwertes beider optimaler Werte für h. Dazu betrachtet man für jede Gruppe separat den optimal justierten Dichteschätzer. In der Abbildung 3.4 wurde jeweils der Gauss-Kern und das Mittel der beiden gruppenspezifischen optimalen Fensterbreiten für h verwendet. Für die beiden Altersgruppen ergaben sich* $h = 29.36$ *(junge) und* $h = 27.66$ *(ältere), also wurde* $h = 28.5$ *verwendet. Für die beiden Regionen ergaben sich* $h = 36.47$ *(Aichfeld) sowie* $h = 30.20$ *(Murau). Der rechte Teil der Abbildung 3.4 wurde mit* $h = 33.00$ *erzeugt.*

```
> plot(density(VC[a=="young"], bw=28.5)
> lines(density(VC[a=="old"], bw=28.5), lty=2)

> plot(density(VC[region=="A"], bw=33.0))
> lines(density(VC[region=="M"], bw=33.0), lty=2)
```

Boxplots sowie Kernschätzer können natürlich auch für den graphischen Vergleich von mehr als zwei Gruppen gezeichnet werden. Bei den Boxplots sind dann aber die Notches nicht mehr allgemein interpretierbar, sondern nur noch für zwei benachbarte Boxen. Die Anzahl von übereinander gelegten Kernschätzungen unterliegt ferner einer natürlichen oberen Schranke.

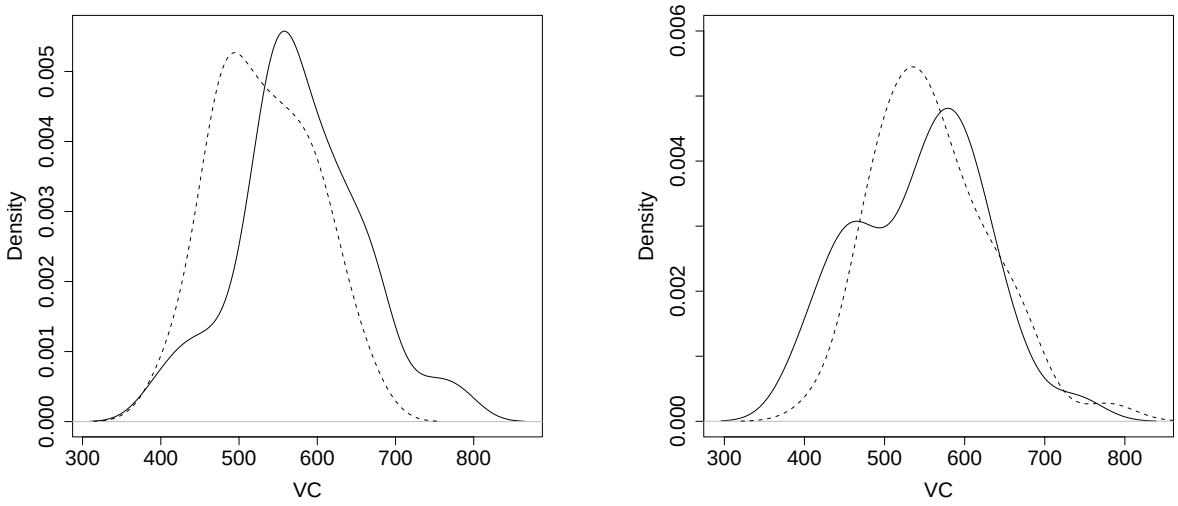


Abbildung 3.4: Kernschätzer der beiden `VC`-Gruppierungen für (links: jung/solid, älter/dashed) und (rechts: Aichfeld/solid, Murau/dashed).

3.2 Lineare Rangstatistik

Viele der nicht-parametrischen Tests, welche verschiedene Aspekte zweier unabhängiger Stichproben miteinander vergleichen, basieren auf speziellen Versionen von linearen Rangstatistiken. Wir werden uns im Folgenden mit drei Gruppen von Alternativhypothesen näher auseinander setzen. Gelte dazu allgemein, dass die beiden unabhängigen Stichproben $X_i \overset{iid}{\sim} F$, $i = 1, \ldots, m$, und $Y_j \overset{iid}{\sim} G$, $j = 1, \ldots, n$, aus stetigen Verteilungen F, und G stammen. Wir testen im Folgenden generell die Gleichheithypothese $H_0 : F(z) = G(z),\ \forall z \in \mathbb{R}$ gegen die beispielsweise hier zweiseitig formulierte

allgemeine Alternative: $H_1 : F(z) \neq G(z)$,

Lokationsalternative: $H_1 : F(z) = G(z + \theta)$, $\theta \neq 0$,

Variablitätsalternative: $H_1 : F(z) = G(z\theta)$, $\theta \neq 1$.

Zu den Tests auf eine allgemeine Alternative (entsprechend den Anpassungstests im Einstichproben-Problem) zählen der Wald-Wolfowitz-Test und der Kolmogorov-Smirnov-Test. Lineare Rangstatistiken sind dann die Tests auf Lokations- und/oder Variablitätsalternativen. Ist der t-Test der klassische parametrische Test bei Lokationsalternativen, so gehören zu den nicht-parametrischen Verfahren unter anderem der Wilcoxon-Test (Mann-Whitney-U-Test) sowie der Van der Waerden-Test. Variabilitätsunterschiede in der parametrischen Statistik werden mit dem F-Test erfasst. Hingegen verwendet man dafür nicht-parametrisch unter anderem die linearen Rangstatistiken des Siegel-Tukey-Test und des Mood-Tests.
Unter Gültigkeit der Nullhypothese stammen beide Stichproben aus der selben Verteilung und können daher zu einer kombinierten Stichprobe zusammen gefasst werden.

Definition 3.1 *In der* **kombinierten Stichprobe** $Z = (X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ *sind die Ränge* R_i *der* X_i *für* $i = 1, \ldots, m$ *bestimmt durch:*

$$R_i = R(X_i) = \sum_{j=1}^{m} T(X_i - X_j) + \sum_{k=1}^{n} T(X_i - Y_k)$$

mit

$$T(u) = \begin{cases} 0 & \text{für} \quad u < 0\,, \\ 1 & \text{für} \quad u \geq 0\,. \end{cases}$$

R_i definiert also die Anzahl der X_j und Y_k kleiner gleich X_i. Für Rangtests erweist es sich als sinnvoll, die **kombinierte, geordnete Stichprobe** $Z_{(\cdot)} = (Z_{(1)}, \ldots, Z_{(N)})$ von $X_1, \ldots, X_m$ und $Y_1, \ldots, Y_n$ mit Umfang $N = m + n$ durch den Vektor V zu charakterisieren, der die Zugehörigkeit zur Gruppe X beschreibt:

$$V_i = \begin{cases} 1 & \text{falls} \;\; Z_{(i)} \text{ eine } \; X\text{-Variable}\,, \\ 0 & \text{falls} \;\; Z_{(i)} \text{ eine } \; Y\text{-Variable}\,. \end{cases}$$

Die meisten Statistiken, die auf Ränge basieren, lassen sich in der folgenden Form (linear in den V_i) darstellen:

$$L_N = \sum_{i=1}^{N} g_i V_i$$

mit reellen Gewichtungsfaktoren g_i. L_N heißt **lineare Rangstatistik**. Generell weist diese Statistik für beliebige Gewichte g_i unter der Gleichheitshypothese der beiden Verteilungen folgende Eigenschaften auf:

Satz 3.1 *Unter* $H_0 : F = G$ *gilt für alle* $i = 1, \ldots, N$*:*

1. $E(V_i) = m/N$, $var(V_i) = mn/N^2$, $cov(V_i, V_j) = -mn/(N^2(N-1))$, $i \neq j$.
2. $E(L_N) = m/N \sum_{i=1}^{N} g_i$, $var(L_N) = mn/(N^2(N-1)) \left(N \sum_i g_i^2 - \left(\sum_i g_i\right)^2\right)$.
3. $P(V_1 = v_1, \ldots, V_N = v_N | F = G) = 1/\binom{N}{m}$.
4. $P(L_N = c | F = G) = a(c)/\binom{N}{m}$,
   wobei $a(c)$ *die Anzahl der Vektoren* $v = (v_1, \ldots, v_N)$ *ist, für die* $L_N(v) = c$ *gilt.*
5. L_N *ist symmetrisch um* $E(L_N)$ *verteilt, falls* $g_i + g_{N-i+1} = k$ *konstant für alle* i *ist.*
6. *Für* $m = n$ *ist jede lineare Rangstatistik* L_N *symmetrisch um* $E(L_N)$ *verteilt.*
7. *Für* $N = m + n \to \infty$, *wobei* $m/n \to \lambda$, $0 < \lambda < \infty$, *ist* $(L_N - E(L_N))/\sqrt{var(L_N)}$ *asymptotisch* $N(0,1)$*-verteilt.*

3.3 Tests der allgemeinen Alternative

3.3.1 Iterationstest

Die Idee des Iterationstests lässt sich für den Fall einer Stichprobe als Test auf Zufälligkeit in der Anordnung der Stichprobenelemente, und im Fall zweier Stichproben dann als Test auf Gleichheit der Verteilungsfunktionen verwenden.

Test auf Zufälligkeit

Die meisten statistischen Verfahren basieren auf der Annahme der Unabhängigkeit der Beobachtungen. Im Falle von Alternativdaten (das sind binäre Daten wie Geschlecht, Erfolg, usw.) würde das bedeuten, dass alle möglichen Reihenfolgen des Auftretens die gleiche Wahrscheinlichkeit haben. Dies ist äquivalent damit, dass eine bestimmte auftretende Reihenfolge *zufällig* ist. Falls die Annahme der Zufälligkeit nicht gesichert erscheint, sollte sie vor der Anwendung eines statistischen Verfahrens überprüft werden. Neben dieser Voruntersuchung kann ein Test auf Zufälligkeit aber auch der eigentliche Gegenstand der Untersuchung sein, wie beispielsweise die Untersuchung täglicher Kursschwankungen (steigend/fallend) bei Finanzmarktdaten. Der bekannteste Test auf Zufälligkeit ist der **Iterationstest (Runstest)**.

Definition 3.2 *Unter einer Iteration (Run) versteht man eine Folge von einem oder mehreren identischen Symbolen, denen entweder ein anderes oder kein Symbol unmittelbar vorangeht oder folgt. Die Statistik R zählt die Anzahl der Gesamtiterationen.*

Bei einer zufälligen Reihenfolge ist anzunehmen, dass sich die beiden Ausprägungen des Merkmales weder ganz regelmäßig abwechseln (viele Iterationen), noch dass zuerst mehr die eine dann mehr die andere Ausprägung auftritt (wenige Iterationen).

Beispiel 3.5 *Die $n = 20$ Schüler einer Grundschulklasse ($n_1 = 8$ Jungen und $n_2 = 12$ Mädchen) warten ungeduldig in einer Schlange vor dem Würstelstand und zwar in der folgenden Reihenfolge:*

J J	M M M M	J J J	M M M M M	J J	M M M	J
J(2)	M(4)	J(3)	M(5)	J(2)	M(3)	J(1)

In der zweiten Zeile der Tabelle sind Typ und Länge der Iterationen angegeben. Somit ergibt sich für die totale Anzahl der beobachteten Iterationen

$$r = r_J + r_M = \#(\text{Iterationen vom Typ J und M}) = 4 + 3 = 7\,.$$

Maximal könnten hierbei 17 Iterationen vorkommen und minimal 2.
Es soll nun getestet werden, ob die Reihenfolge zufällig ist, d.h.

- H_0 : Anordnung ist **zufällig**.

Lautet die Alternative

- H_1 : Anordnung ist **nicht zufällig**,

so ist ein zweiseitiger Test zu wählen. Bei der hypothetischen Annahme eines Trends ist dagegen ein einseitiger Test zu verwenden. Für unser Beispiel lauten die beiden einseitigen Alternativen

- H_1 : Anordnung ist **geschlechtshomogen** (wenig gemischt, wenige Iterationen),
- H_1 : Anordnung ist **geschlechtsinhomogen** (extrem gemischt, viele Iterationen).

Letztere sollte bei Kindern realistischerweise angenommen werden.
Im zweiseitigen Test wird H_0 abgelehnt, falls $r \leq r_{\alpha/2}$ oder $r \geq r_{1-\alpha/2}$ realisiert. Die vorher festgelegte Richtung der Abweichung von der Zufälligkeit (einseitiger Test) kann hinweisen auf

1. zu wenig Iterationen, d.h. H_0 wird abgelehnt, wenn $r \leq r_\alpha$ ist, oder
2. zu viele Iterationen, d.h. H_0 wird abgelehnt, wenn $r \geq r_{1-\alpha}$ ist.

Die Quantile r_α der Verteilung von R unter H_0 sind für $3 \leq n_1, n_2 \leq 20$ in der Tabelle G angegeben.
Bezüglich Beispiel 3.5 folgt bei Wahl von $\alpha = 0.1$ für den zweiten einseitigen Alternativfall (Geschlechtsinhomogenität) der kritische Wert $r_{0.9} = 14$, was wegen $r = 7 < r_{1-\alpha}$ nicht zur Ablehnung von H_0 führt. Bzgl. der Alternative „Geschlechtshomogenität“ resultiert $r_{0.1} = 8$ was zur Ablehnung von H_0 führen würde.

Der folgende Satz gibt Auskunft über die Verteilung der Iterationsanzahl R unter H_0. Generell gilt für den Realisationsbereich von R:

$$\begin{aligned} n_1 = n_2 &: \quad 2 \leq R \leq n\,, \\ n_1 \neq n_2 &: \quad 2 \leq R \leq 2\min(n_1, n_2) + 1\,. \end{aligned}$$

Satz 3.2 *Die Wahrscheinlichkeitsfunktion von R unter H_0 ist gegeben durch*

$$\begin{aligned} P(R = r_1 + r_2) &= P(R = 2k) = \frac{2\binom{n_1-1}{k-1}\binom{n_2-1}{k-1}}{\binom{n}{n_1}}\,, \\ P(R = r_1 + r_2) &= P(R = 2k+1) = \frac{\binom{n_1-1}{k}\binom{n_2-1}{k-1} + \binom{n_1-1}{k-1}\binom{n_2-1}{k}}{\binom{n}{n_1}}\,. \end{aligned}$$

Beweis: $n = n_1 + n_2$ Elemente können auf $n!$ Arten angeordnet werden. Da jede Permutation der n_1 Elemente vom selben Typ A und jede Permutation der n_2 Elemente vom selben Typ B die Anordnung unverändert lassen, gibt es also insgesamt $\binom{n}{n_1} = \binom{n}{n_2}$ verschiedene Anordnungen, die alle unter H_0 die gleiche Wahrscheinlichkeit haben (Anzahl der möglichen Fälle). Für die Bestimmung der Anzahl der günstigen Fälle kann folgendermaßen vorgegangen werden:

- r gerade: Sei $r = 2k$, so gibt es k Iterationen mit Elementen vom Typ A und k Iterationen mit Elementen vom Typ B ($n_1 \geq k$).

Typ A: #(günstige Fälle) = #(n_1 mal A auf k Iterationen verteilen) = #(n_1 Kugeln auf k nummerierte Zellen verteilen und keine Zelle leer)

$$\underbrace{\bullet -}_{Zelle\ 1} | - \bullet - \bullet - | - \bullet - | - \bullet - \bullet - \bullet - | \underbrace{- \bullet - \bullet}_{Zelle\ k}$$

= #((k − 1) Striche auf ($n_1 - 1$) Zwischenräume verteilen) = $\binom{n_1-1}{k-1}$.

Typ B: analoge Vorgangsweise führt zu #(günstige Fälle)= $\binom{n_2-1}{k-1}$.

Die Gesamtanzahl der verschiedenen Anordnungen, beispielsweise beginnend mit einer Iteration von Elementen des Typs A, ist somit $\binom{n_1-1}{k-1}\binom{n_2-1}{k-1}$. Das gleiche gilt für die Anordnungen, welche mit einer Iteration des Typs B beginnen. Die Gesamtanzahl aller Anordnungen ist also $2\binom{n_1-1}{k-1}\binom{n_2-1}{k-1}$.

- r ungerade: Sei $r = 2k + 1$, so gibt es $(k + 1)$ Iterationen mit Elementen vom Typ A und k Iterationen vom Typ B oder umgekehrt. Die Herleitung der Anzahl der günstigen Fälle erfolgt analog dem Fall „r gerade“.

Für $n_1 > 20$, $n_2 > 20$ approximiert man die Verteilung von

$$Z = \frac{R - \mathrm{E}(R)}{\sqrt{\mathrm{var}(R)}}$$

mit

$$\begin{aligned} \mathrm{E}(R) &= \frac{2n_1n_2}{n} + 1\,, \\ \mathrm{var}(R) &= \frac{2n_1n_2(2n_1n_2 - n)}{n^2(n-1)} \end{aligned}$$

durch die $N(0, 1)$-Verteilung.

Beispiel 3.6 *Es sei $n = 28$ mit $n_1 = n_2 = 14$, also $E(R) = 15$. Für den zweiseitigen Iterationstest und $\alpha = 0.05$ resultiert aus der Tabelle G als Annahmebereich das Intervall* $[10, 20]$. *Beispielsweise könnten die folgenden systematischen Anordnungen beobachtet werden:*

$0\ldots0\|1\ldots1\|$	$r = 2$	H_0 *ablehnen,*
$0\|1\|0\|1\|\ldots\|0\|1\|$	$r = 28$	H_0 *ablehnen,*
$00\|11\|00\|\ldots\|11\|$	$r = 14$	H_0 *fälschlicherweise nicht ablehnen,*
$000\|111\|000\|\ldots\|00\|11\|$	$r = 10$	H_0 *fälschlicherweise nicht ablehnen.*

Der Iterationstest erkennt also nur jene Arten der Abweichung von der Zufälligkeit, die zu extremen Runanzahlen führen.

Wald-Wolfowitz-Test

Im Falle zweier unabhängiger Stichproben kann man den Iterationstest als Test der allgemeinen zweiseitigen Alternative

- $H_0 : F(z) = G(z) \quad \forall z \in \mathbb{R}$
- $H_1 : F(z) \neq G(z) \quad$ für mindestens ein $z \in \mathbb{R}$

verwenden, mit $X_i \overset{iid}{\sim} F$ und $Y_j \overset{iid}{\sim} G$. Dazu wird in der kombinierten, geordneten Stichprobe die Anzahl der Iterationen von x und y beobachtet. Für die Gleichheitshypothese würde eine perfekte Mischung aus x und y sprechen (sehr viele Runs). Ist die Anzahl der Iterationen zu klein ($R < r_\alpha$), so wird H_0 zu Gunsten H_1 verworfen. Der Wald-Wolfowitz-Test is also nur zweiseitig anwendbar. Beachte aber, dass dafür die Quantile des einseitigen Iterationstest verwendet werden.

Beispiel 3.7 *Bei einer Untersuchung von Kindern soll die Körpergröße von 8 Mädchen (x) und 10 Jungen (y) verglichen werden. Folgende Beobachtungen liegen vor:*

Mädchen $x_{(i)}$	117	120	122	124	126	126	128	132		
Jungen $y_{(j)}$	110	113	114	116	116	118	119	119	123	125

Liegt dieselbe Verteilung für Mädchen und Jungen vor?
Als kombinierte, geordnete Stichprobe resultiert

110	113	114	116	116	117	118	119	119	120	122	123	124	125	126	126	128	132
y	y	y	y	y	x	y	y	y	x	x	y	x	y	x	x	x	x

Es gibt also $r = 8$ *Iterationen. Für* $\alpha = 0.05$ *folgt aus der Tabelle G* $r_{0.05} = 6$. *Wegen* $r > r_{0.05}$ *kann* H_0 *nicht verworfen werden.*

Unter H_0 ist hier zu erwarten, dass die x und y sehr gut gemischt sind (viele Runs). Die beiden Extremfälle mit minimalem Wert von r

$$\begin{array}{lll} x \ldots x | y \ldots y & \text{und} & y \ldots y | x \ldots x \\ F(z) > G(z) & & F(z) < G(z) \end{array}$$

können nicht unterschieden werden. Die Teststatistik beinhaltet daher keinerlei Informationen über Lagealternativen. Beim Übergang von quantitativen Beobachtungen zur Runanzahl entsteht daher ein sehr großer Verlust an Information und der Test ist zum Erkennen von Lageunterschieden nicht geeignet.

3.3.2 Der Kolmogorov-Smirnov-Test

Wie schon beim Vergleich einer empirischen mit einer theoretischen Verteilung kann der Kolmogorov-Smirnov-Test auch für den Vergleich zweier empirischer Verteilungen verwendet werden. Dazu wird angenommen, dass $X_i \overset{iid}{\sim} F$, $i = 1, \ldots, m$, und $Y_j \overset{iid}{\sim} G$, $j = 1, \ldots, n$, unabhängig und stetig verteilt sind. Die empirischen Verteilungsfunktionen F_m und G_n sind erwartungstreue Schätzungen für F und G.

Folgende Hypothesen können getestet werden:

- Test A: $H_0 : F(z) = G(z) \quad \forall z \in \mathbb{R}, \quad H_1 : \exists z \in \mathbb{R} : F(z) \neq G(z)$,
- Test B: $H_0 : F(z) \leq G(z) \quad \forall z \in \mathbb{R}, \quad H_1 : \exists z \in \mathbb{R} : F(z) > G(z)$,
- Test C: $H_0 : F(z) \geq G(z) \quad \forall z \in \mathbb{R}, \quad H_1 : \exists z \in \mathbb{R} : F(z) < G(z)$.

Die KS-Teststatistik ist definiert durch

- Test A: $K_{m,n} = \max_{z \in \mathbb{R}} |F_m(z) - G_n(z)|$,
- Test B: $K^+_{m,n} = \max_{z \in \mathbb{R}} (F_m(z) - G_n(z))$,
- Test C: $K^-_{m,n} = \max_{z \in \mathbb{R}} (G_n(z) - F_m(z))$.

H_0 wird abgelehnt, falls

- Test A: $k_{m,n} > k_{1-\alpha}$,
- Test B: $k^+_{m,n} > k^+_{1-\alpha}$,
- Test C: $k^-_{m,n} > k^-_{1-\alpha}$.

Beim Auftreten von Bindungen ist dieser KS-Test jedoch konservativ. Man kann zeigen, dass die Verteilung von $K_{m,n}$ unter H_0 nur von m und n und nicht von F und G abhängt. $K_{m,n}$ ist also auch hier eine verteilungsfreie Teststatistik.

Zur Verteilung von $K_{m,n}$ unter H_0: Die Herleitung der exakten Verteilung beruht auf kombinatorische Überlegungen und wird anhand eines einfachen Beispiels kurz erläutert. Der Wert von $K_{m,n}$ hängt nur von der Ordnung der x- und y- Realisationen in der kombinierten, geordneten Stichprobe ab und nicht von deren expliziten Werten.
Sei dazu $m = 2$ und $n = 3$, so gibt es mit Satz 3.1(3) insgesamt $\binom{m+n}{n} = 10$ unterscheidbare kombinierte, geordnete Stichproben, welche unter H_0 alle gleichwahrscheinlich sind. Für jede dieser Situationen wird nun die Realisierung der KS-Statistik berechnet. Die dabei resultierenden maximalen Abstände der entsprechenden empirischen Verteilungsfunktionen F_m und G_n sind in der folgenden Tabelle 3.1 zusammengefasst. Mit dieser Tabelle erhält man beispielsweise $P(K_{2,3} \leq 1/2) = 4/10$; oder $P(K_{2,3} = 1) = 0.20$. Für $m = n$ findet man $k_{1-\alpha}$, $k^+_{1-\alpha}$ und $k^-_{1-\alpha}$ in Tabelle H; für $m \neq n$ benutze man Tabelle I.

kombinierte, geordnete Stichproben	k	$P(K_{2,3} = k)$
(x,x,y,y,y), (y,y,y,x,x)	1	2/10
(y,x,x,y,y), (y,y,x,x,y), (y,y,x,y,x), (x,y,x,y,y)	2/3	4/10
(x,y,y,x,y), (x,y,y,y,x), (y,x,y,y,x)	1/2	3/10
(y,x,y,x,y)	1/3	1/10

Tabelle 3.1: Zur Verteilung von $K_{m,n}$ im Falle $m = 2$ und $n = 3$.

Beispiel 3.8 *Wie im Beispiel 3.7 zuvor, testen wir jetzt wieder auf die identische Verteilung der Körpergröße von Mädchen ($m = 8$) und Jungen ($n = 10$). Liegt dieselbe Verteilung vor?*
Es ist also der zweiseitige KS-Test (Test A) anzuwenden. Die Teststatistik realisiert in $k_{8,10} = 0.675$ für alle Körpergrößen aus dem Intervall $[119, 120)$ (siehe Tabelle 3.2). Verwenden wir $\alpha = 0.05$, so ist wegen $k_{8,10} = 0.675 > k_{0.95} = 46/80 = 0.575$ hier H_0 abzulehnen. Im Gegensatz dazu führte der Iterationstest von Wald-Wolfowitz nicht zur Ablehnung von H_0. Dies liegt darin begründet, dass der KS-Test ein etwas anderes Maß für den Unterschied zweier Verteilungen verwendet als der Wald-Wolfowitz-Test.

Intervall	$\lvert F_m(z) - G_n(z)\rvert$	Intervall	$\lvert F_m(z) - G_n(z)\rvert$
$-\infty < z < 110$	0	$120 \leq z < 122$	0.550
$110 \leq z < 113$	0.100	$122 \leq z < 123$	0.425
$113 \leq z < 114$	0.200	$123 \leq z < 124$	0.525
$114 \leq z < 116$	0.300	$124 \leq z < 125$	0.400
$116 \leq z < 117$	0.500	$125 \leq z < 126$	0.500
$117 \leq z < 118$	0.375	$126 \leq z < 128$	0.250
$118 \leq z < 119$	0.475	$128 \leq z < 132$	0.125
$119 \leq z < 120$	**0.675**	$132 \leq z < \infty$	0

Tabelle 3.2: KS-Tests für den Vergleich der Körpergrößen von Mädchen und Jungen.

```
> x <- c(117,120,122,124,126,126,128,132)          # m=8 Mädchen
> y <- c(110,113,114,116,116,118,119,119,123,125)  # n=10 Knaben
> ks.test(x, y)

        Two-sample Kolmogorov-Smirnov test

data:  x and y
D = 0.675, p-value = 0.03484
alternative hypothesis: two.sided

Warning message:
cannot compute correct p-values with ties in: ks.test(x, y)
```

Wegen vorhandener Bindungen wurde von `R` *eine Warnung ausgegeben (stetige Verteilungen würden diese nicht generieren). Diese finden sich zwar nur innerhalb der beiden Gruppen (z.B.* $126, 126$ *in* `x`*), haben jedoch trotzdem einen Einfluss auf die Verteilung der Statistik.*
Sehr einfach lassen sich auch die beiden empirischen Verteilungsfunktionen graphisch miteinander vergleichen. Den maximalen Abstand der beiden Treppenfunktionen (KS-Abstand) beobachtet man dabei im Intervall $[119, 129)$.

```
> library(stepfun)
> plot(ecdf(x)); lines(ecdf(y))
```

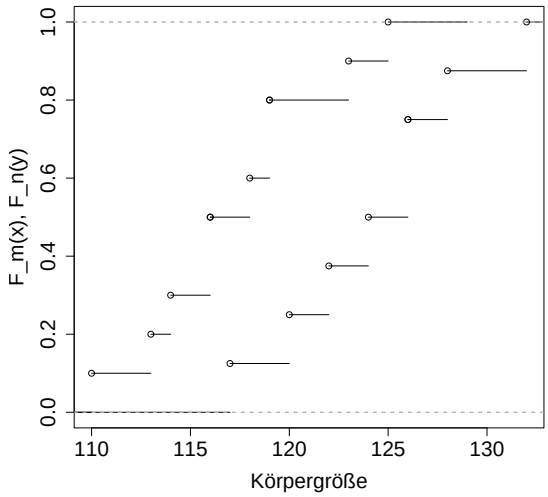


Abbildung 3.5: Graphische Interpretation des KS-Test für das Beispiel der Körpergrößen.

3.4 Tests bezüglich Lokationsalternativen

Häufig weiß man von zwei Stichproben, dass diese bis auf den Lage- (Lokations-) Parameter ident verteilt sind. Dies impliziert, dass die Varianzen beider Verteilungsfunktionen gleich sind. Man möchte in diesem Fall auf mögliche Lageunterschiede in den beiden Verteilungen F und G testen, d.h.

- $H_0 : G(z) = F(z) \quad \forall z \in \mathbb{R}$,
- $H_1 : G(z) = F(z - \theta) \quad \forall z \in \mathbb{R},\ \theta \neq 0$.

Je nachdem, welche Werte man bezüglich θ zulässt, werden die folgenden Alternativhypothesen unterschieden:

- $\theta \neq 0$ (zweiseitig); d.h. $F \neq G$,
- $\theta > 0$ (einseitig); d.h. $F > G$,
- $\theta < 0$ (einseitig); d.h. $F < G$.

Bemerke, dass unter Gültigkeit der einseitigen Alternativhypothese $\theta > 0$ gilt, dass F stochastisch größer als G ist. Dies impliziert weiters, dass beispielsweise der Median, Erwartungswert oder sämtliche Quantile von X gerade um θ kleiner sind als die entsprechenden Größen von Y. Alle höheren Momente (Varianz, Schiefe, u.s.w.) dürfen sich weder unter H_0 noch unter den Alternativen H_1 unterscheiden. Daher ist es in der Praxis auch notwendig, diese Aspekte vor der Durchführung eines Tests auf Lokationsalternativen zu prüfen.

3.4.1 Parametrische Tests bei Normalverteilung

Liegen zwei unabhängige Stichproben $X_i \overset{iid}{\sim} N(\mu_X, \sigma_X^2)$, $i = 1, \ldots, m$, und $Y_j \overset{iid}{\sim} N(\mu_Y, \sigma_Y^2)$, $j = 1, \ldots, n$, vor, so entspricht der Test auf Lokationsalternativen dem Test auf Gleichheit der beiden Erwartungswerte. Als Likelihood-Quotienten-Test resultiert dafür der so genannte Zweistichproben t-Test für unabhängige Stichproben:

H_0	H_1	Entscheidung gegen H_0, falls	kritische Werte
$\mu_Y - \mu_X = \theta = 0$	$\theta \neq 0$	$T < c_3$ oder $T > c_4$	$c_3 = t_{\alpha/2}$, $c_4 = t_{1-\alpha/2}$
$\mu_Y - \mu_X = \theta = 0$	$\theta > 0$	$T < c_1$	$c_1 = t_\alpha$
$\mu_Y - \mu_X = \theta = 0$	$\theta < 0$	$T > c_2$	$c_2 = t_{1-\alpha}$

Diesen Test kann man darüber hinaus auch für unterschiedliche Varianzen in den Gruppen verwenden. Wir unterscheiden die beiden Situationen:

1. $\sigma_X^2 = \sigma_Y^2$ (unbekannt) oder σ_X^2/σ_Y^2 bekannt. Für die Teststatistik gilt unter H_0:

$$T = \frac{\left(\overline{X} - \overline{Y}\right)\sqrt{\frac{nm}{n+m}}}{\underbrace{\sqrt{\frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}}}_{S_P}} \sim t_{n+m-2}\,.$$

   Diese Statistik vereinfacht sich im Fall gleich großer Stichprobenumfänge ($n = m$) zu

$$T = \frac{\left(\overline{X} - \overline{Y}\right)\sqrt{n}}{\sqrt{S_X^2 + S_Y^2}}\,.$$

   Die Größe S_P^2 bezeichnet man hierbei als **gepoolte Varianzschätzung**.

2. $\sigma_X^2 \neq \sigma_Y^2$. Für diese Teststatistik gilt unter H_0:

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}} \overset{ap}{\sim} t_\nu \quad \text{mit} \quad \nu = \frac{\left(\frac{S_X^2}{m} + \frac{S_Y^2}{n}\right)^2}{\frac{1}{m-1}\left(\frac{S_X^2}{m}\right)^2 + \frac{1}{n-1}\left(\frac{S_Y^2}{n}\right)^2}\,.$$

   Dies entspricht gerade dem Ergebnis der Satterthwait's Approximation für die Freiheitsgrade einer Summe gewichteter unabhängiger χ^2-Größen. Man findet diesen Test auch unter der Bezeichnung Welch-Test.

Beispiel 3.9 *Wir analysieren wiederum die beiden Stichproben mit den Körpergrößen der Kinder und testen unter Annahme zweier unabhängiger normalverteilter Populationen auf einen signifikanten Unterschied in den beiden Erwartungswerten. Auch in* `R` *erlaubt dieser Test beide obigen Varianzmodelle.*

```
> t.test(x, y, paired=FALSE, var.equal=TRUE)

         Two Sample t-test

 data:  x and y
 t = 3.2357, df = 16, p-value = 0.005174
 alternative hypothesis: true difference in means is not equal to 0
 95 percent confidence interval:
   2.439771 11.710229
 sample estimates: mean of x mean of y
   124.375   117.300

> t.test(x, y, paired=FALSE, var.equal=FALSE)

        Welch Two Sample t-test

 data:  x and y
 t = 3.2196, df = 14.837, p-value = 0.005797
 alternative hypothesis: true difference in means is not equal to 0
 95 percent confidence interval:
   2.386632 11.763368
 sample estimates: mean of x mean of y
   124.375   117.300
```

3.4.2 Der Wilcoxon-Rangsummentest (1945)

Sind $X_i \overset{iid}{\sim} F$, $Y_j \overset{iid}{\sim} G$ stetig verteilt, X, Y unabhängig, und haben $x_1, \ldots, x_m$, $y_1, \ldots, y_n$ zumindest ordinales Meßniveau, so kann der Wilcoxon-Rangsummentest auf Lokationsunterschiede verwendet werden. Folgende Hypothesen werden dabei getestet:

- Test A: $H_0 : G(z) = F(z)\ \forall z \in \mathbb{R}; \quad H_1 : G(z) = F(z-\theta)\ \forall z \in \mathbb{R},\ \theta \neq 0,$
- Test B: $H_0 : G(z) = F(z)\ \forall z \in \mathbb{R}; \quad H_1 : G(z) = F(z-\theta)\ \forall z \in \mathbb{R},\ \theta > 0,$
- Test C: $H_0 : G(z) = F(z)\ \forall z \in \mathbb{R}; \quad H_1 : G(z) = F(z-\theta)\ \forall z \in \mathbb{R},\ \theta < 0.$

Sei $N = m + n$, so ist die Teststatistik W_N definiert durch

$$W_N = \sum_{i=1}^{N} iV_i = \sum_{i=1}^{m} R_i\,,$$

wobei R_i die Ränge der X_i in der kombinierten, geordneten Stichprobe bezeichnen. Wir verwerfen H_0, falls

- Test A: $w_N \geq w_{1-\alpha/2}$ oder $w_N \leq w_{\alpha/2}$,
- Test B: $w_N \leq w_\alpha$,
- Test C: $w_N \geq w_{1-\alpha}$.

Zur Verteilung von W_N unter H_0:
Die Statistik W_N ist eine spezielle lineare Rangstatistik L_N mit den Gewichten $g_i = i$ (vgl. Abschnitt 3.2). Sie hat positive Wahrscheinlichkeiten auf dem Bereich

$$\min(W_N) = m(m+1)/2 \le W_N \le m(m+1)/2 + mn = \max(W_N)$$

sowie Momente (diese ergeben sich einfach durch Anwendung des Satzes 3.1)

$$\mathrm{E}(W_N) = \frac{m(N+1)}{2} \quad \text{und} \quad \mathrm{var}(W_N) = \frac{mn(N+1)}{12}\,.$$

Deshalb strebt die Verteilung von

$$Z = \frac{W_N - m(N+1)/2}{\sqrt{mn(N+1)/12}} \overset{as}{\sim} N(0,1)$$

für $m/n = \lambda < \infty$ (konstant) gegen die $N(0,1)$-Verteilung.
Die exakte Verteilung von W_N unter H_0 wird an einem konkreten Beispiel mit $m = 3$ und $n = 5$ erläutert. Hierfür gibt es insgesamt $\binom{N}{m} = \binom{8}{3} = 56$ verschiedene Vektoren $(v_1, \ldots, v_8)$, die alle unter $H_0 : F = G$ dieselbe Wahrscheinlichkeit 1/56 haben. Da W_N nach Satz 3.1 wegen $g_i + g_{N-i+1} = i + N - i + 1 = N + 1$ (konstant) symmetrisch um den Erwartungswert $m(N+1)/2 = 13.5$ verteilt ist, genügt es, die obere Hälfte der Verteilung als Tabelle anzuführen. Für $\alpha = 4/56 \approx 0.071$ ist also das $(1-\alpha)$-Quantil $w_{1-\alpha} = 19$.

w	Ränge der X_i (r_1, r_2, r_3)	$P(W_N = w)$
21	(6,7,8)	1/56
20	(5,7,8)	1/56
19	(4,7,8); (5,6,8)	2/56
18	(3,7,8); (4,6,8); (5,6,7)	3/56
17	(2,7,8); (3,6,8); (4,6,7); (4,5,8)	4/56
16	(1,7,8); (2,6,8); (3,5,8); (3,6,7); (4,5,7)	5/56
15	(1,6,8); (2,5,8); (2,6,7); (3,5,7); (3,4,8); (4,5,6)	6/56
14	(1,6,7); (1,5,8); (2,5,7); (2,4,8); (3,4,7); (3,5,6)	6/56

Tabelle 3.3: Zur Verteilung von W_N für den Fall $m = 3$ und $n = 5$.

In der Tabelle J sind die Quantile w_α für $m \le n$, $m, n \le 25$, angegeben (Test B). Für das Testproblem C gilt $w_{1-\alpha} = 2\mathrm{E}(W_N) - w_\alpha$. Ist $m > n$ werden die Stichproben umbenannt $(X \leftrightarrow Y)$ und damit geht der Test B in den Test C über. Für größere Stichprobenumfänge verwende man die Normalverteilungs-Approximation.

Beispiel 3.10 *Körpergrößen von $m = 8$ Mädchen und $n = 10$ Knaben.*

	y_1	y_2	y_3	y_4	y_5	x_1	y_6	y_7	y_8	x_2	x_3	y_9	x_4	y_{10}	x_5	x_6	x_7	x_8
$z_{(i)}$	110	113	114	116	116	117	118	119	119	120	122	123	124	125	126	126	128	132
v_i	0	0	0	0	0	1	0	0	0	1	1	0	1	0	1	1	1	1
g_i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

Dafür resultiert $w_{18} = \sum_i iv_i = 106$. *Für das Testproblem A mit* $\alpha = 0.05$ *entnehme man aus der Tabelle J:* $w_{\alpha/2} = 53$*;* $w_{1-\alpha/2} = 2E(W_N) - w_{\alpha/2} = 152 - 53 = 99 < w_N$*, was wie schon beim Kolmogorov-Smirnov-Test zuvor wiederum zur Ablehnung von* H_0 *führt.*
Der Wilcoxon-Test wird auch von `R` *angeboten. Interessanterweise liefert der Aufruf*

```
> wilcox.test(x, y)

        Wilcoxon rank sum test with continuity correction

 data:  x and y
 W = 70, p-value = 0.00866
 alternative hypothesis: true mu is not equal to 0

 Warning message: Cannot compute exact p-value with ties
```

den kleineren Wert 70 *als Wilcoxon-Teststatistik aber dazu den korrekten p-Wert.*

Oft bieten Statistik-Pakete statt des Wilcoxon-Rangsummentests den **Mann-Whitney-U-Test** (1947) an. Dieser ist jedoch äquivalent zum Wilcoxon-Test und basiert auf folgender Teststatistik:

$$U_N = \sum_{i=1}^{m} \sum_{j=1}^{n} W_{ij}$$

mit

$$W_{ij} = \begin{cases} 1 & \text{für} \quad Y_j < X_i, \quad i = 1, \ldots, m \\ 0 & \text{für} \quad Y_j > X_i, \quad j = 1, \ldots, n\,. \end{cases}$$

U_N gibt also an, wie oft insgesamt x-Werte den y-Werten in der kombinierten, geordneten Stichprobe folgen. Es kann gezeigt werden, dass

$$U_N = W_N - \min(W_N)$$

gilt. Alle Eigenschaften des Wilcoxon-Tests halten somit auch für den U-Test.

Beispiel 3.11 *Körpergrößen von* $m = 8$ *Mädchen und* $n = 10$ *Knaben.*

	y_1	y_2	y_3	y_4	y_5	x_1	y_6	y_7	y_8	x_2	x_3	y_9	x_4	y_{10}	x_5	x_6	x_7	x_8
w^*	8	8	8	8	8	0	7	7	7	0	0	5	0	4	0	0	0	0

Dabei gibt w^* *an, wie viele* x*-Werte den jeweiligen* y*-Werten in der kombinierten, geordneten Stichprobe folgen. Aufsummieren von* w^* *führt zu* $u = 70$*. Dies entspricht genau* $w_N - m(m+1)/2 = 106 - 36 = 70$ *(dem obigen Ergebnis mit* `R`*).*

3.4.3 Van der Waerden X_N-Test

Diese lineare Rangstatistik ist definiert als

$$X_N = \sum_{i=1}^{N} \Phi^{-1}\left(\frac{i}{N+1}\right) V_i = \sum_{i=1}^{m} \Phi^{-1}\left(\frac{R_i}{N+1}\right)$$

mit den (monoton wachsenden) Quantilen der Normalverteilung $g = \Phi^{-1}(\cdot)$ als Gewichte.

Zur Verteilung von X_N unter H_0: Unter Verwendung des Satzes 3.1 folgt für X_N:

$$g_i + g_{N-i+1} = \Phi^{-1}\left(\frac{i}{N+1}\right) + \Phi^{-1}\left(1 - \frac{i}{N+1}\right) = 0\,,$$

also Symmetrie um $\mathrm{E}(X_N)$. Wegen $\sum_{i=1}^{N} g_i = 0$ folgen als Momente

$$\mathrm{E}(X_N) = 0\,, \qquad \mathrm{var}(X_N) = \frac{mn}{N(N-1)} \sum_{i=1}^{N} \left(\Phi^{-1}\left(\frac{i}{N+1}\right)\right)^2 .$$

X_N ist daher symmetrisch um $\mathrm{E}(X_N) = 0$ verteilt.
Die Verteilung von X_n unter H_0 wird wieder für $m = 3$ und $n = 5$ hergeleitet. Wiederum gibt es $\binom{N}{m} = 56$ verschiedene Vektoren, die alle unter H_0 dieselbe Wahrscheinlichkeit $1/56$ haben. Die Gewichte g_i sind dafür gegeben durch:

i	1	2	3	4	5	6	7	8
g_i	−1.2206	−0.7647	−0.4307	−0.1397	0.1397	0.4307	0.7647	1.2206

Für die ersten 28 unterschiedlichen Rangtripel (r_1, r_2, r_3) von (X_1, X_2, X_3) in der kombinierten, geordneten Stichprobe findet man die Realisationen von X_8, und deren Auftrittswahrscheinlichkeiten $P = P(X_8 = x_8)$ in der folgenden Tabelle:

(r_1, r_2, r_3)	x_8	P	(r_1, r_2, r_3)	x_8	P	(r_1, r_2, r_3)	x_8	P
(6,7,8)	2.416	1/56	(4,6,7)	1.056	1/56	(4,5,6)	0.431	
(5,7,8)	2.125	1/56	(3,5,8)	0.930	1/56	(1,6,8)	0.431	3/56
(4,7,8)	1.846	1/56	(2,6,8)	0.887	1/56	(2,6,7)	0.431	
(5,6,8)	1.791	1/56	(1,7,8)	0.765		(2,4,8)	0.316	1/56
(3,7,8)	1.555	1/56	(3,6,7)	0.765	3/56	(3,4,7)	0.194	1/56
(4,6,8)	1.512	1/56	(4,5,7)	0.765		(1,5,8)	0.140	
(5,6,7)	1.335	1/56	(3,4,8)	0.650	1/56	(2,5,7)	0.140	3/56
(2,7,8)	1.221		(2,5,8)	0.596	1/56	(3,5,6)	0.140	
(3,6,8)	1.221	3/56	(3,5,7)	0.474	1/56	(2,3,8)	0.025	1/56
(4,5,8)	1.221							

Tabelle 3.4: Zur Verteilung von X_N für den Fall $m = 3$ und $n = 5$.

Für die restlichen 28 Möglichkeiten gilt: ist $X_N = x$ für (r_1, r_2, r_3), so ist $X_N = -x$ für $(N+1-r_1, N+1-r_2, N+1-r_3)$. Falls $\alpha = 4/56 = 0.071$ ist das $(1-\alpha)$-Quantil gegeben durch $x_{1-\alpha} = 1.79$. Weiters fällt auf, dass dieselben 4 Rangtripel wie beim Wilcoxon-Test den kritischen Bereich $C = \{(6,7,8); (5,7,8); (4,7,8); (5,6,8)\}$ für den einseitigen Test C bilden. Die kritischen Werte von X_N findet man in der Tabelle K. Ist $N > 20$, so approximiert man die Verteilung von $X_N/\sqrt{\text{var}(X_N)}$ durch die $N(0,1)$-Verteilung.

Beispiel 3.12 *Körpergrößen von $m = 8$ Mädchen und $n = 10$ Knaben. Die Gewichte sind bestimmt durch $g_i = \Phi^{-1}(i/19)$. Als Realisation von X_N erhält man nun $x_{18} = 4.945$. Die Tabelle K liefert für $\alpha = 0.05$ den kritischen Wert $x_{1-\alpha/2} = 3.616$. Da $x_{18} > x_{1-\alpha/2}$ gilt, wird auch hier H_0 abgelehnt.*
Leider beinhaltet R *diesen Test nicht standardmäßig. Er kann aber recht einfach selbst programmiert werden. Hier eine Möglichkeit:*

```
> x <- c(117,120,122,124,126,126,128,132)          # m=8 Mädchen
> y <- c(110,113,114,116,116,118,119,119,123,125)  # n=10 Knaben
> m <- length(x);  n <- length(y);  N <- m+n
> group <- c(rep("x", m), rep("y", n))
> V <- 1*(group[order(c(x, y))]=="x")  # Indikator(x) in komb-geord-StPr
> g <- qnorm((1:N)/(N+1))  # Gewichte
> X <- sum(g*V);  X        # Van der Waerden Statistik
[1] 4.944933
> var.X <- m*n/(N*(N-1))*sum(g^2)  # Varianz(X)
> var.X
[1] 3.468656
> p.value <- 2*(1 - pnorm(abs(X)/sqrt(var.X)))  # two-sided p.value
[1] 0.007928642
```

Als approximativen p-Wert liefert dies 0.008 und erlaubt somit die gleiche Aussage.

3.4.4 Weitere lineare Rangtests für Lokationsalternativen

Fisher-Yates-Terry-Hoeffding:

$$g_i = \mathrm{E}\left(Z_{(i)}\right)$$

Hier ist g_i ist der Erwartungswert der i-ten geordneten Statistik $Z_{(i)}$ einer Zufalls-Stichprobe vom Umfang N aus der $N(0,1)$-Verteilung. Diese Gewichte entnimmt man einer Tabelle.

Median-Test:

$$g_i = \begin{cases} 0 & \text{für} \quad i \le (N+1)/2 \\ 1 & \text{für} \quad i > (N+1)/2\,. \end{cases}$$

Diese Statistik summiert die Ränge der x-Werte, die größer als der Median sind.

3.5 Tests bezüglich Variabilitätsalternativen

In diesem Abschnitt werden Rangtests für Variabilitätsalternativen diskutiert. Die beiden Verteilungen F und G seien wiederum stetig und unter der Nullhypothese ident. Die Alternativhypothese H_1 beinhaltet jetzt die Gleichheit der Verteilungen von X und θY, mit $\theta > 0$. Damit folgt

$$F(\theta z) = P(X \leq \theta z) = P(\theta Y \leq \theta z) = G(z)\,.$$

Für die ersten Momente bedeutet dies

$$\begin{aligned} \mathrm{E}(X) &= \theta \mathrm{E}(Y) \\ \mathrm{var}(X) &= \theta^2 \mathrm{var}(Y)\,. \end{aligned}$$

Variabilitätsalternativen schließen daher sowohl Lokations- als auch Varianzunterschiede ein. Nur im Fall $\mathrm{E}(X) = \mathrm{E}(Y) = 0$ sind Tests auf Variabilität auch Tests auf Varianz. Bemerke, dass für $\theta > 1$ ($\theta < 1$) X mehr (weniger) streut als Y.
Allgemein können folgende Hypothesen getestet werden:

- $H_0 : G(z) = F(z)$, $\forall z \in \mathbb{R}$;
- $H_1 : G(z) = F(\theta z)$, $\forall z \in \mathbb{R}$ mit $\theta \neq 1$, $\theta > 1$ oder $\theta < 1$.

Während bei den Tests auf Lokationsalternativen monotone Gewichte verwendet wurden, liegt es nahe, für Tests auf Variabilität den extremen Beobachtungen in der kombinierten, geordneten Stichprobe (jene mit niedrigen oder großen Rängen) kleine Gewichte zu vergeben und die mittleren Werten hoch zu gewichten (oder genau umgekehrt). Eine derartige Gewichtung hat dann ein etwa quadratisches Aussehen (als Funktion in i gesehen).

3.5.1 Parametrischer Test bei Normalverteilung

Liegen zwei unabhängige Stichproben $X_i \overset{iid}{\sim} N(\mu_X, \sigma_X^2)$ und $Y_j \overset{iid}{\sim} N(\mu_Y, \sigma_Y^2)$ vor, mit unbekannten μ_X und μ_Y, dann entspricht der Likelihood-Quotienten-Test auf Gleichheit der beiden Varianzen dem F-Test:

H_0	H_1	Entscheidung gegen H_0, falls	kritische Werte
$\sigma_X = \sigma_Y$	$\sigma_X \neq \sigma_Y$ $(\theta \neq 1)$	$T < c_3$ oder $T > c_4$	$c_3 = f_{\alpha/2}$
			$c_4 = f_{1-\alpha/2}$
$\sigma_X = \sigma_Y$	$\sigma_X > \sigma_Y$ $(\theta > 1)$	$T > c_1$	$c_1 = f_{1-\alpha}$
$\sigma_X = \sigma_Y$	$\sigma_X < \sigma_Y$ $(\theta < 1)$	$T < c_2$	$c_2 = f_{\alpha}$

Unter H_0 gilt:

$$T = \frac{S_X^2}{S_Y^2} \sim F_{m-1,n-1}.$$

Der F-Test ist sehr empfindlich gegenüber Abweichungen von der Normalverteilung.

Beispiel 3.13 *Wir verwenden wiederum die Daten über die Körpergrößen von 8 Mädchen und 10 Knaben und testen auf Gleichheit der beiden Varianzen (unter Annahme zweier Normalverteilungen).*

```
> x <- c(117,120,122,124,126,126,128,132)          # m=8 Mädchen
> y <- c(110,113,114,116,116,118,119,119,123,125)  # n=10 Knaben
> var.test(x, y, ratio = 1)  # ratio=1 default

        F test to compare two variances

data:  x and y
F = 1.0886, num df = 7, denom df = 9, p-value = 0.8841
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
0.2593722 5.2505398
sample estimates:
ratio of variances
         1.088597
```

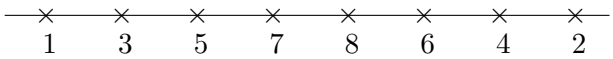
Hierbei überdeckt das angeführte Konfidenzintervall den wahren Varianzquotienten σ_X^2/σ_Y^2.

3.5.2 Siegel-Tukey-Test (1960)

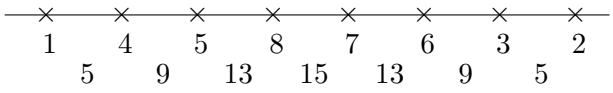
Dieser nicht-parametrische Test auf Variabilitätsalternativen stellt das Analogon zum Wilcoxon-Test auf Lokationsalternativen dar. Unter den gleichen Voraussetzungen wie für den Wilcoxon-Test können jetzt folgende Hypothesen formuliert werden:

- Test A: $H_0 : G(z) = F(z); \quad H_1 : G(z) = F(\theta z)\, \forall z \in \mathbb{R},\ \theta \neq 1,\ \theta > 0,$
- Test B: $H_0 : G(z) = F(z); \quad H_1 : G(z) = F(\theta z)\, \forall z \in \mathbb{R},\ \theta > 1,$
- Test C: $H_0 : G(z) = F(z); \quad H_1 : G(z) = F(\theta z)\, \forall z \in \mathbb{R},\ 0 < \theta < 1.$

Die Siegel-Tukey-Statistik ist eine lineare Rangstatistik und versucht symmetrische Gewichte zu verwenden, welche an den Enden der kombinierten, geordneten Stichprobe kleine Werte aufweisen. Solche Gewichte könnten mittels



vergeben werden. Während jedoch bei diesem Verfahren keine Symmetrie in den Gewichten vorhanden ist, liefert das folgende zumindest Symmetrie in der Summe benachbarter Gewichte für N gerade



Ist N ungerade, dann wird die "mittlere" Beobachtung aus der kombinierten, geordneten Stichprobe gestrichen und g_i für $N^* = N - 1$ berechnet.
Die Siegel-Tukey-Teststatistik verwendet diese Gewichtung und kann daher für gerades N wie folgt definiert werden

$$S_N = \sum_{i=1}^{N} g_i V_i \quad \text{mit} \quad g_i = \begin{cases} 2i & \text{für } i \text{ gerade und } 1 \le i \le N/2, \\ 2(N-i)+2 & \text{für } i \text{ gerade und } N/2 < i \le N, \\ 2i-1 & \text{für } i \text{ ungerade und } 1 \le i \le N/2, \\ 2(N-i)+1 & \text{für } i \text{ ungerade und } N/2 < i \le N. \end{cases}$$

Diese Gewichte bestehen wiederum aus den ganzen Zahlen $\{1, 2, \ldots, N\}$. Das sind daher die gleichen Gewichte wie schon zuvor bei der Wilcoxon-Statistik (hier nur anders nummeriert) und es folgt $a_{W_N}(c) = a_{S_N}(c)$. Daher hat unter $H_0 : F = G$ (was dann unabhängig von der Nummerierung ist) die Statistik S_N dieselbe Verteilung wie die Wilcoxon-Statistik W_N, d.h.

$$\mathrm{E}(S_N) = \frac{m(N+1)}{2}, \qquad \mathrm{var}(S_N) = \frac{mn(N+1)}{12}.$$

H_0 wird abgelehnt, falls:

- Test A: $S_N \ge w_{1-\alpha/2}$ oder $S_N \le w_{\alpha/2}$,
- Test B: $S_N \le w_\alpha$,
- Test C: $S_N \ge w_{1-\alpha}$.

Die Quantile findet man in der Tabelle J (Wilcoxon) für $m, n \le 30$. Für größere m und n verwendet man wiederum die Normalverteilungsapproximation wie bei W_N.

Beispiel 3.14 *Körpergrößen von $m = 8$ Mädchen und $n = 10$ Knaben.*

	y_1	y_2	y_3	y_4	y_5	x_1	y_6	y_7	y_8	x_2	x_3	y_9	x_4	y_{10}	x_5	x_6	x_7	x_8
$z_{(i)}$	110	113	114	116	116	117	118	119	119	120	122	123	124	125	126	126	128	132
v_i	0	0	0	0	0	1	0	0	0	1	1	0	1	0	1	1	1	1
i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
g_i	1	4	5	8	9	12	13	16	17	18	15	14	11	10	7	6	3	2

Also resultiert $s_{18} = \sum_i g_i v_i = 74$. Für $\alpha = 0.05$ entnehme man $w_{\alpha/2} = 53$ aus der Tabelle J. Damit ist $w_{1-\alpha/2} = 2E(W_N) - w_{\alpha/2} = 152 - 53 = 99$. Wegen $53 < 74 < 99$ kann H_0 nicht verworfen werden.

```
> g <- rep(1, N);  g[N] <- 2
> odd <- 1-(even <- (trunc(1:N/2)==(1:N/2)))
> for (i in 2:(N/2)) g[i] <- g[i-1] + 1*odd[i] + 3*even[i]
> for (i in (N-1):(N/2+1)) g[i] <- g[i+1] + 1*odd[i] + 3*even[i]
> S <- sum(g*V)          # [1] 74
> E.S <- m*(N+1)/2       # [1] 76
> var.S <- m*n*(N+1)/12  # [1] 126.6667
> p.value <- 2*(1-pnorm(abs(X-E.S)/sqrt(var.X)))  # two-sided
> p.value
[1] 0.858955
```

Die Anwendung des Siegel-Tukey-Tests basiert auf der Annahme, dass F und G vom selben Verteilungstyp sind und etwa den gleichen Median haben. Lageunterschiede oder Varianzunterschiede können generell damit nicht erfasst werden. Dies soll auch durch das nächste Beispiel anschaulich gemacht werden.

Beispiel 3.15 *Es liege eine kombinierte, geordnete Stichprobe der Form xxxxxyyyyy vor. Diese kann das Ergebnis zweier gänzlich unterschiedlicher Verteilungen sein mit Median(X) < Median(Y), oder der Unterschied kann nur im Lokationsparameter vorliegen. Weiters können dabei auch die x-Werte stärker streuen, als die y-Werte oder umgekehrt. Man erhält dafür* $s_{10} = 1 + 4 + 5 + 8 + 9 = 27$. *Selbst für* $\alpha = 0.50$ *führt ein zweiseitiger* S_N*-Test nicht zur Ablehnung von* H_0.

3.5.3 Mood-Test (1954)

Während die Siegel-Tukey-Statistik S_N die Variabilität der x_i- und y_j-Werte dadurch charakterisiert, dass am oberen und unteren Ende der kombinierten, geordneten Stichprobe kleine Gewichte vergeben werden, misst die Mood-Statistik unmittelbar die quadratischen Abweichungen der Ränge der x_i, also $R(X_i) = R_i$, vom mittleren Rang $\overline{R} = (N+1)/2$, d.h. die empirische Varianz der R_i in der kombinierten, geordneten Stichprobe

$$M_N = \sum_{i=1}^{N} \left(i - \frac{N+1}{2} \right)^2 V_i = \sum_{i=1}^{m} \left(R_i - \overline{R} \right)^2 .$$

Große Werte von M_N zeigen an, dass die x_i-Werte stärker variieren als die y_j-Werte.

Zur Verteilung von M_N *unter* H_0*:*
Für den Erwartungswert und die Varianz von M_N ergibt sich mit Satz 3.1

$$\mathrm{E}(M_N) = \frac{m(N^2-1)}{12} , \qquad \mathrm{var}(M_N) = \frac{mn(N+1)(N^2-4)}{180} .$$

Da $g_i + g_{N+1-i}$ nicht konstant ist, ist die Mood-Statistik im Allgemeinen nicht symmetrisch verteilt. Natürlich resultiert aber Symmetrie für den Fall $m = n$ (vgl. Satz 3.1(6)).
Für das Beispiel $m = 3$ und $n = 4$ gibt es $\binom{7}{3} = 35$ verschiedene Rangtupeln und als Gewichte resultieren $g_i = (i-4)^2$. Alle möglichen Realisationen und deren Auftrittswahrscheinlichkeiten $P = P(M_N = m_7)$ unter H_0 findet man in der Tabelle 3.5.
Die Quantile der Verteilung von M_N sind für $N \leq 20$ in der Tabelle L angegeben. Für $N > 20$ verwende man die Normalverteilungsapproximation.

Beispiel 3.16 *Für das Beispiel der Körpergrößen ist* $(N+1)/2 = 9.5$ *und damit* $m_{18} = 228$. *Für* $\alpha = 0.20$ *ist* $m_{\alpha/2} = 146$ *und* $m_{1-\alpha/2} = 284$, *d.h.* H_0 *kann nicht einmal dafür abgelehnt werden.*

```
> mood.test(x, y)
        Mood two-sample test of scale
data:  x and y
Z = 0.2341, p-value = 0.815
alternative hypothesis: two.sided
```

(r_1, r_2, r_3)	m_7	P	(r_1, r_2, r_3)	m_7	P	(r_1, r_2, r_3)	m_7	P
(1,2,7)	22	2/35	(1,2,4)	13		(2,3,4)	5	
(1,6,7)	22		(1,4,6)	13	4/35	(2,4,5)	5	4/35
(1,3,7)	19	2/35	(2,4,7)	13		(3,4,6)	5	
(1,5,7)	19		(4,6,7)	13		(4,5,6)	5	
(1,4,7)	18	1/35	(1,3,5)	11	2/35	(3,4,5)	2	1/35
(1,2,6)	17	2/35	(3,5,7)	11				
(2,6,7)	17		(1,3,4)	10				
(1,2,3)	14		(1,4,5)	10	4/35			
(1,2,5)	14		(3,4,7)	10				
(1,3,6)	14		(4,5,7)	10				
(1,5,6)	14	8/35	(2,3,6)	9	2/35			
(2,3,7)	14		(2,5,6)	9				
(2,5,7)	14		(2,4,6)	8	1/35			
(3,6,7)	14		(2,3,5)	6	2/35			
(5,6,7)	14		(3,5,6)	6				

Tabelle 3.5: Zur Verteilung von M_N im Falle $m = 3$ und $n = 4$.

```
> E.M <- m * (N^2-1)/12             # [1] 215.3333
> var.M <- m*n * (N+1)*(N^2-4)/180  # [1] 2702.222
> mood.test(x, y)$statistic*sqrt(var.M) + E.M + 1/2
    Z
 228.0
```

Die von R *ausgegebene Statistik* Z *entspricht also der standardisieren Form von M_N (mit zusätzlicher Stetigkeitskorrektur $1/2$).*

3.5.4 Weitere lineare Rangtests für Variabilitätsalternativen

Ansari-Bradley-Test (1960):

Die Ansari-Bradley-Statistik ist eine Lineare Rangstatistik mit Gewichten

$$g_i = \left(\frac{N+1}{2} - \left| i - \frac{N+1}{2} \right| \right) .$$

Ist $i \leq (N+1)/2$, so ergibt sich $g_i = i$. Für dieses i ist dann $N+1-i \geq (N+1)/2$ mit $g_{N+1-i} = i$ (symmetrische Dreiecksgewichtung). Hierbei wird also der kleinsten und größten Beobachtung der Rang 1, der zweitkleinsten und zweitgrößten der Rang 2, usw. zugeordnet. Für die Summe dieser beiden Gewichte folgt daher $g_i + g_{N+1-i} = 2i$, weshalb diese lineare Rangstatistik auch nur für $m = n$ symmetrisch ist.

Sind die Abweichungen $|i - (N+1)/2|$ groß, so wird dadurch A_N klein. Dies ist ein Hinweis für stärker streuende x_i-Werte.

Beispiel 3.17 *Für die Körpergrößen ergibt dieser Test:*

```
> ansari.test(x,y)

        Ansari-Bradley test

data:  x and y
AB = 39, p-value = 0.8574
alternative hypothesis: true ratio of scales is not equal to 1
```

Die Tests von Klotz (1962) und Capon (1961):
Klotz hat das Quadrat des Gewichts der Van der Waerden-Statistik X_N verwendet, Capon den Erwartungswert des Quadrats von $Z_{(i)}$ (vgl. mit Fisher-Yates-Terry-Hoeffding-Test für Lokationsalternativen).

$$\begin{aligned} K_N &= \sum_{i=1}^{N} \left[\Phi^{-1}\left(\frac{i}{N+1}\right)\right]^2 V_i\,, \\ C_N &= \sum_{i=1}^{N} E\left(Z_{(i)}^2\right) V_i. \end{aligned}$$

Der K_N-Test und der C_N-Test sind asymptotisch äquivalent.

Fligner-Killeen-Test (1976):

$$F_N = \sum_{i=1}^{N} \Phi^{-1}\left(\frac{1}{2} + \frac{i}{2(N+1)}\right) V_i\,.$$

Die in `R` implementierte Version dieses Tests verwendet statt dessen die Ränge der Absolutbeträge der Median-zentrierten Stichproben. Ferner wird die Verteilung dieser Teststatistik nicht durch die $N(0,1)$ sondern durch die χ^2_{k-1} approximiert. Dabei bezeichnet k die Anzahl der Stichprobengruppen (hier $k = 2$), die auf Variabilitätsunterschiede verglichen werden.

Beispiel 3.18 *Da mit* `fligner.test()` *mehrere Gruppen verglichen werden können, ist die Syntax dieser Funktion auch etwas anders, und benötigt die gesamte Stichprobe sowie den Faktor, der für die Gruppierung zuständig ist. Dieser Test ergibt für unser Beispiel:*

```
> z <- c(x, y)                          # kombinierte Stichprobe
> sex <- c(rep("F", m), rep("M", n))  # "F"=Female, "M"=Male
> fligner.test(z ~ as.factor(sex))

        Fligner-Killeen test for homogeneity of variances

data:  z by as.factor(sex)
Fligner-Killeen:med chi-squared = 0.0081, df = 1, p-value = 0.9284
```

Kapitel 4

Das Zwei-Stichproben-Problem

In diesem Kapitel werden Zufallsstichproben von n unabhängig, identisch verteilten Paaren von Zufallsvariablen diskutiert. Zuerst werden dafür einige Anwendungen des Scatter-Plots besprochen. Auch Serien von Boxplots sind geeignete graphische Darstellungen. Nur im Falle eines linearen Zusammenhangs ist dafür auch der entsprechende empirische Korrelationskoeffizient interpretierbar. Falls ein nicht-linearer Zusammenhang besteht, dann kann dieser mittels eines Glättungsverfahrens ersichtlich gemacht werden. Die konfirmatorische Statistik bietet hierzu Verfahren an, um Unterschiede in der Lokation der beiden Verteilungen zu verifizieren. Ist die Annahme einer Normalverteilung für die beiden Populationen gerechtfertigt, so wird dafür der t-Test verwendet. Nicht-parametrische Alternativen stellen der Vorzeichen-Test oder der Wilcoxon-Test dar. Abschließend werden die beiden Stichproben auf Korrelation (parametrisch wie auch nicht-parametrisch) sowie mittels Verfahren für Kontingenztafeln auf deren stochastische Unabhängigkeit untersucht.

4.1 Graphische Verfahren

Gegeben seien n Stichprobenpaare $(X_1, Y_1), \ldots, (X_n, Y_n)$ mit den entsprechenden n Realisierungen $(x_1, y_1), \ldots, (x_n, y_n)$. Dabei beschreiben (X_i, Y_i) zwei Merkmale die am i-ten Merkmalsträger (i-ten Objekt) beobachtet wurden. Der bivariate **Scatter-Plot** stellt die empirische zweidimensionale Verteilung durch die Punkte (x_i, y_i) dar. Im Gegensatz dazu lieferte der EQQ-Plot aus dem vorigen Kapitel einen Vergleich der beiden eindimensionalen empirischen Verteilungen.

Beispiel 4.1 *Während $n = 132$ Monaten wurde die Ozon-Konzentration in Yonkers (x_i) und Stamford (y_i) beobachtet. Die Abhängigkeit der Stichproben X und Y ist durch die gemeinsamen Zeitpunkte der Erhebungen i gegeben. Der Scatter-Plot (x_i, y_i) zeigt im linken Teil der Abbildung 4.1 den Zusammenhang zwischen den beiden Konzentrationen, während rechts keine spezielle Veränderung der Beobachtungen aus Stamford über die Zeit erkennbar ist.*

```
> ozon <- read.table("ozon.dat", header=TRUE)
> attach(ozon);  plot(Yonkers, Stamford);  abline(0, 1)
> plot(Monat, Stamford)
```

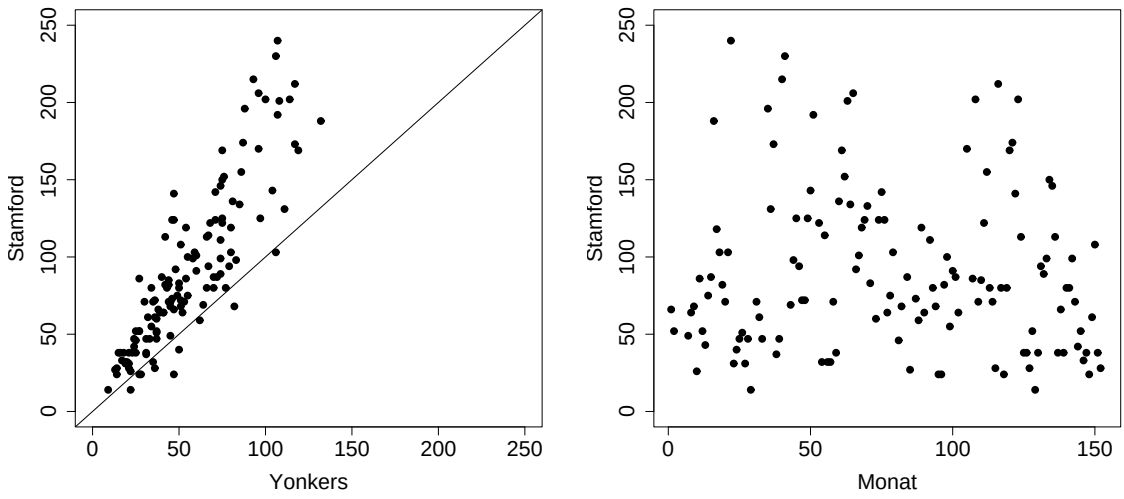


Abbildung 4.1: Scatter-Plot der Stamford und Yonkers Ozon-Daten (links) und der Stamford Ozon-Daten gegen den Zeitverlauf (rechts).

Man kann auch zwei verschiedene Merkmale durch einen Scatter-Plot darstellen und damit den möglichen funktionalen Zusammenhang optisch erkennbar machen. Diese Überlegung führt dann später zur Idee der Regression.

Beispiel 4.2 *Die Abhängigkeit der Variablen* `VC` *vom Alter und von der Größe ist in der Abbildung 4.2 dargestellt. Im linken Teil erkennt man, dass der funktionale Zusammenhang zwischen* `VC` *und* `age` *ein scheinbar annähernd quadratischer ist. Maximale Vitalkapazität erreicht man etwa bei einem Alter von 25 Jahren. Im Gegensatz dazu scheint die links dargestellte Abhängigkeit der Vitalkapazität von der Körpergröße annähernd linear zu sein. Größere Personen haben tendenziell auch eine höhere Vitalkapazität.*

```
> attach(aimu);  plot(age, VC);  plot(height, VC)
```

Der **empirische Korrelationskoeffizient** R stellt ein Maß für die **lineare** Abhängigkeit zwischen den Y_i und den X_i dar. Dieser ist definiert als Momentenschätzer der theoretischen Korrelation $\rho = \mathrm{cor}(X, Y)$ in der Population, also durch

$$R = \frac{S_{xy}^2}{\sqrt{S_x^2 S_y^2}} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2 \sum (Y_i - \overline{Y})^2}}\,.$$

Er realisiert im Intervall $[-1, 1]$. Ist $r = 1$ (bzw. $r = -1$), dann liegen alle Punkte (x_i, y_i) auf einer Geraden mit positiver (bzw. negativer) Steigung und man spricht von einem perfekten linearen Zusammenhang. Ist $r = 0$, dann besteht kein linearer Zusammenhang. Alternative nicht-parametrische Schätzer und darauf basierende Tests werden später noch in Abschnitt 4.3 diskutiert.

Beispiel 4.3 *Für die beiden in der Abbildung 4.2 dargestellten Situationen liefert der empirische Korrelationskoeffizient*

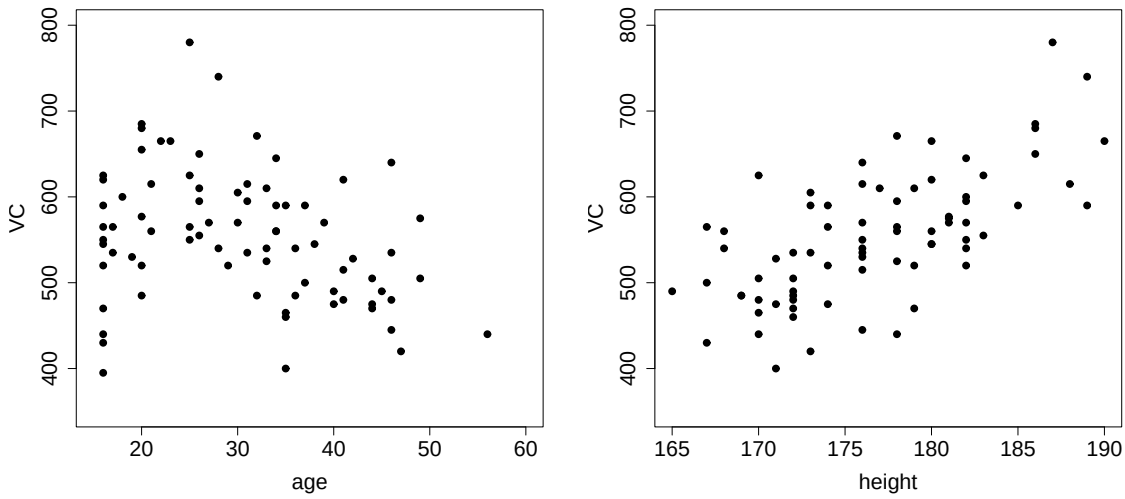


Abbildung 4.2: Scatter-Plot der `VC`-Daten gegen `age` (links) und `height` (rechts).

```
> cor(age, VC)
 [1] -0.2914085
> cor(height, VC)
[1] 0.6829789
```

Der negative Wert von -0.29 *ist sicherlich nicht adäquat als Maß für den (linearen) Zusammenhang von* `VC` *mit* `age`*. Zwar gibt es eine linear steigende Tendenz für Personen bis zu 25 Jahren, jedoch fällt diese danach wieder linear ab. Durch diese Mischung aus positiver und negativer Abhängigkeit wird der empirische Korrelationskoeffizient relativ klein. Sein Vorzeichen bestimmt die größere Gruppe der über 25-jährigen.*
Viel sinnvoller als Maß für den linearen Zusammenhang ist der Wert $+0.68$*, welcher für die Situation im rechten Teil der Abbildung 4.2 resultiert. Der Zusammenhang zwischen* `VC` *und* `height` *scheint linear sowie auch positiv zu sein. Der empirische Korrelationskoeffizient ist nur wegen der vorhandenen Streuung um die gedachte (Ausgleichs)-Gerade etwas geringer.*

Der Wert von r kann ohne zusätzliche Information über die Struktur der Daten zu falschen oder irreführenden Schlüssen führen. Die folgende Abbildung 4.3 demonstriert verschiedenartige Situationen, für die jeweils annähernd $r = 0.7$ resultiert (vergleichbar mit der Korrelation zwischen `VC` und `height`).

```
# (a): set.seed(3)
> x <- 1 + runif(20)*1.7;  y <- 1 + runif(20)*1.7
> x[21] <- y[21] <- 5.0;  cor(x, y);  plot(x, y)
 [1] 0.7159042
# (b): set.seed(1)
> x <- 1.2 + (1:9)/4;  y <- x*5/4 + (1/2-runif(9))*0.2
> x[10] <- 5;  y[10] <- 3.4;  cor(x, y);  plot(x, y)
 [1] 0.7138063
```

```
# (c) set.seed(5)
> x <- 1.1 + 3.7*runif(50);  y <- x - (x-1)*runif(50)
> cor(x, y);  plot(x, y)
 [1] 0.7147086
# (d) set.seed(7)
> x1 <- 1 + 2.5*runif(42);  y1 <- 1 + 2.5*runif(42)
> x2 <- 4 + 1.5*runif(18);  y2 <- 4 + 1.5*runif(18)
> cor(x<- c(x1, x2), y<- c(y1, y2));  plot(x, y)
 [1] 0.710447
# (e) set.seed(4)
> x1 <- (4:20)/4
> y1 <- x1   + (runif(17)-1/2)
> y2 <- 1+x1/4 + (runif(17)-1/2)
> cor(x <- c(x1, x1), y<- c(y1, y2));  plot(x, y)
 [1] 0.7105082
# (f) set.seed(1)
> x <- 3/4 + (1:17)/4
> y <- 1 + 0.7*(x-2.5)**2 + (runif(17)-1/2)/2
> cor(x, y);  plot(x, y)
[1] 0.7016779
```

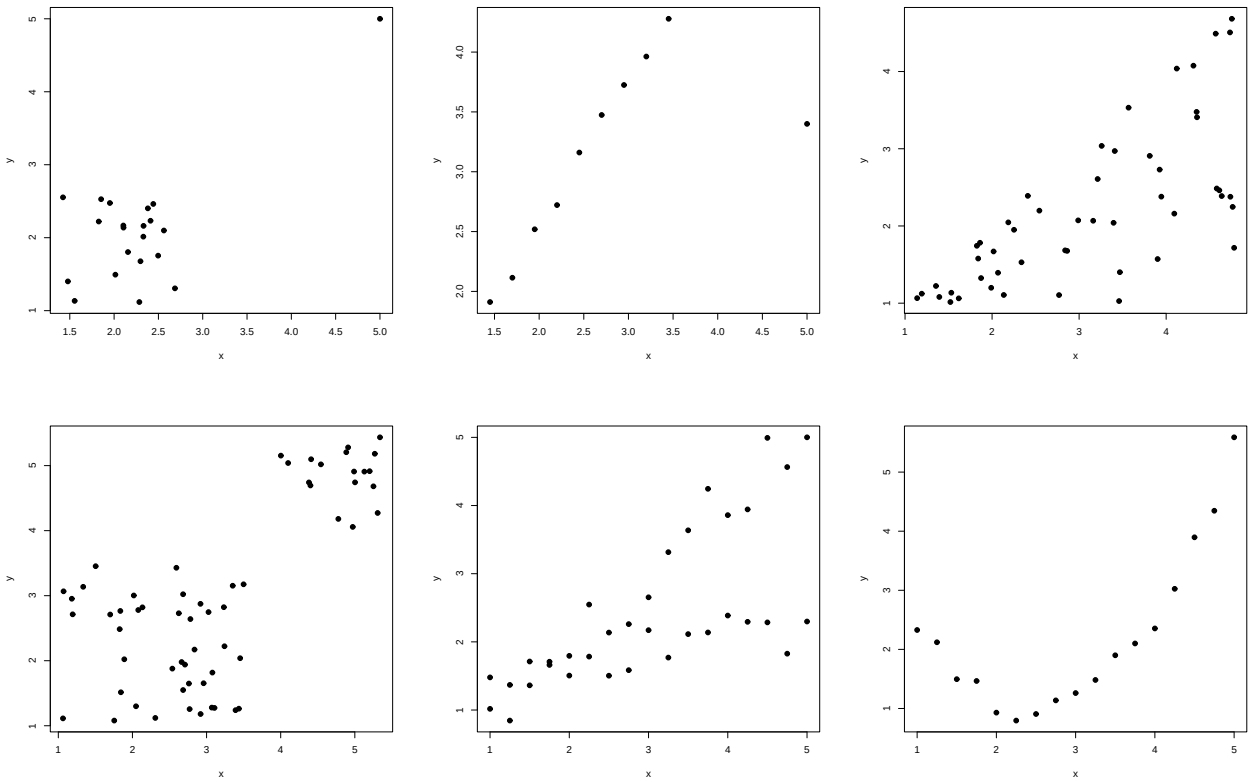


Abbildung 4.3: Diverse Scatter-Plots von Stichproben, für die jeweils etwa $r = 0.7$ gilt.

4.1.1 Analyse der Struktur der Abhängigkeit

Es liege eine Zufallsstichprobe aus der zweidimensionalen Verteilung von (X, Y) vor. Wir betrachten nun die Veränderung der empirischen lokalen Verteilung von Y in Abhängigkeit von X. Um dies zu beschreiben kann im einfachsten Fall der x-Bereich in mehrere *vertikale Streifen* unterteilt werden. Die lokale Verteilung der Y-Variable in einem Streifen wird durch einen Box-Plot dargestellt, so dass für sämtliche Streifen eine **Box-Plot-Serie** resultiert.

Beispiel 4.4 *Die Zeitreihe in der Abbildung 4.4 zeigt die mittlere Anzahl täglicher Sonnenflecken eines Monats über 33 Jahre von Jänner 1951 bis Dezember 1983. Dieser Zeitraum stellt nur einen Ausschnitt der Daten dar, denn diese Daten werden kontinuierlich von Jänner 1749 bis Dezember 1983 in Zürich erhoben (das ist ein Zeitraum von 235 Jahren). Ein Boxplot basiert somit auf das Datenmaterial der zwölf Monatsmittel dieses Jahres. Dabei fällt besonders ein 11-Jahreszyklus in der Sonnenaktivität auf.*

```
> sunspots <- scan("sunspots.dat")
> year <- gl(235, 12, label=c(1749:1983))    # 12*1749, ..., 12*1983
> plot.it <- (as.numeric(year) >= 203)
> boxplot(sunspots[plot.it] ~ year[plot.it]) # use data from 1951-1983 only
```

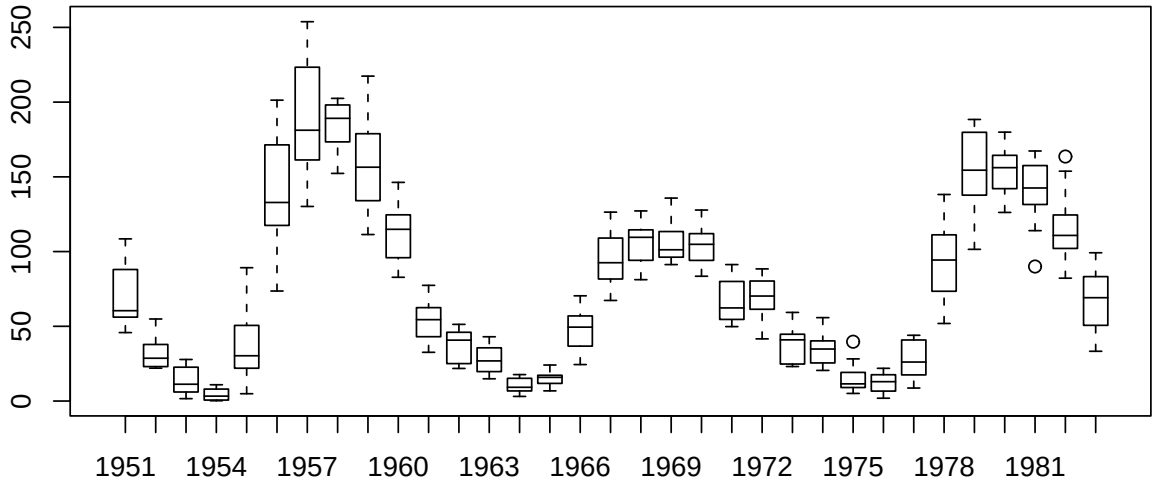


Abbildung 4.4: Box-Plot-Serie über die Aktivität der Sonne, 1951-1983.

Beispiel 4.5 *Die lokalen Verteilungen der Variablen* `VC`*, gebildet durch die Altersklassen der Probanden, sind in der Abbildung 4.5 dargestellt. Nachteil dieser Streifenbildung ist die fehlende Information über die Änderungen der lokalen Verteilung zwischen den einzelnen Streifen.*

```
> attach(aimu)
> age.class <- cut(age, breaks=seq(15,60,by=5))
> table(age.class)
 (15,20] (20,25] (25,30] (30,35] (35,40] (40,45] (45,50] (50,55] (55,60]
      21       8      10      16       8       8       7       0       1
> boxplot(VC ~ age.class)
```

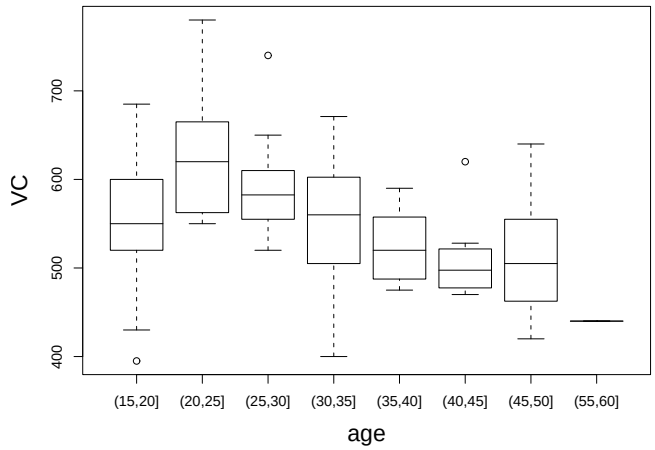


Abbildung 4.5: Box-Plots der `VC`-Werte in Abhängigkeit von der Altersgruppe.

Den Nachteil einer ausnahmslosen Betrachtung der empirischen Verteilung innerhalb disjunkter Streifen umgeht man bei den **Glättungsverfahren**, indem man den Streifen langsam und kontinuierlich von links nach rechts über den gesamten Bereich der x-Werte bewegt. Dadurch geht eine Beobachtung (x_i, y_i) öfter in die Schätzung der lokalen Verteilung ein, wodurch wiederum ein glatter (*smooth*) Verlauf gewährleistet wird. Gerade diese Vorgehensweise ist im `lowess`-Verfahren implementiert.

4.1.2 Lokal gewichtete Regression (`lowess`)

Das Verfahren *Locally Weighted Regression Scatter Plot Smoothing* wurde von Cleveland (1979) publiziert. Es besteht aus zwei Teilen: einer lokalen **Glättung** mittels Techniken der gewichtete Regression aller Daten in einem Fenster, und einer anschließenden Robustifizierung gegen etwaige Ausreißer.
Für jeden Punkt (x_i, y_i), $i = 1, \ldots, n$, wird ein geglätteter Punkt $(x_i, \hat{y}_i)$ bestimmt. Die einzelnen Schritte des Glättungsverfahrens werden nun am Beispiel von $n = 20$ künstliche generierten Punkten im Detail erläutert.

1. Um jeden Datenpunkt (x_i, y_i) wird ein vertikaler Streifen gelegt, der die $q = \lceil fn \rceil$ benachbarten Punkte enthält. Dabei zählt x_i als zu sich nächster Nachbar. Der Wert des Glättungsparameters f, mit $0 < f < 1$, soll laut Cleveland aus dem Intervall $[1/3, 2/3]$ gewählt werden. In der Abbildung 4.6 wird diese Streifenkonstruktion für $f = 1/2$ an den beiden Stellen x_6 und x_{20} (rechter Rand) demonstriert.

2. Nun definiert man für alle Punkte Nachbarschaftsgewichte mit den Eigenschaften:

   (a) Der Fenstermitte (x_i, y_i) hat das größte Gewicht.

   (b) Das Gewicht eines Punktes im Fenster nimmt mit seinem Abstand zu x_i ab.

   (c) Die Gewichtsfunktion ist symmetrisch um x_i.

   (d) Alle Punkte außerhalb des Streifens haben Gewicht Null.

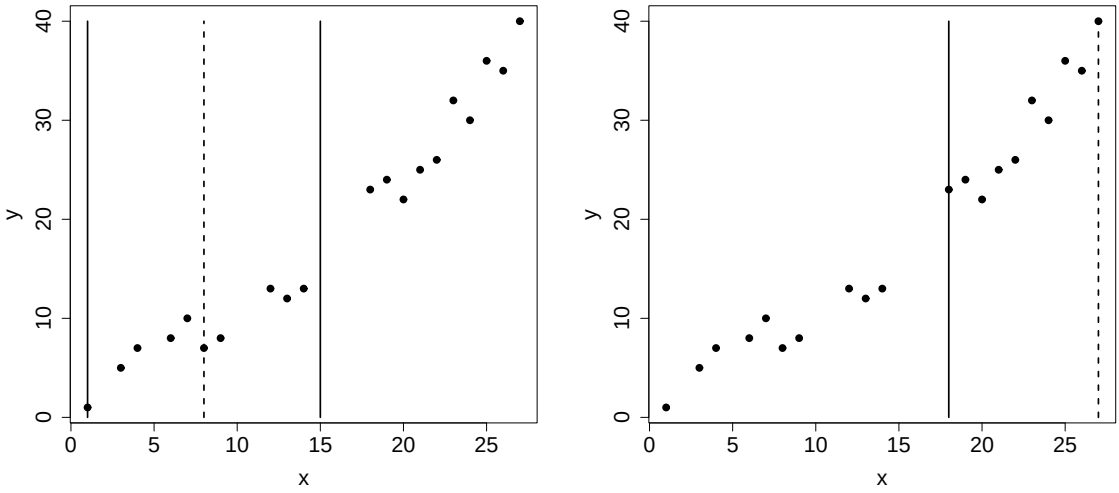


Abbildung 4.6: Streifen um $x_6 = 8$ (links) und $x_{20} = 27$ (rechts) mit Zentrum (dashed) und Rändern (solid).

Als *Nachbarschaftsgewicht* eignet sich besonders gut die Tricube-Funktion

$$T(u) = \begin{cases} (1 - |u|^3)^3 & \text{für} \quad |u| < 1 \\ 0 & \text{sonst,} \end{cases}$$

die in der Abbildung 4.7 dargestellt ist. Das Nachbarschaftsgewicht eines Punktes (x_k, y_k), $k = 1, \ldots, n$, für die Glättung im Punkt x_i ist damit definiert als

$$t_i(x_k) = T\left(\frac{x_i - x_k}{d_i}\right),$$

wobei d_i die Entfernung von x_i zum q-nächsten Nachbarn (halbe Fensterbreite um x_i) beschreibt. Falls $d_i = 0$, so haben alle q-nächsten Nachbarn gleiche Abszisse x_i. In diesem Fall bekommen all diese Punkte im Fenster das Gewicht 1.

3. Für (x_i, y_i) bestimmt man den geglätteten Wert $(x_i, \hat{y}_i)$ als jenen, der auf der Ausgleichsgeraden aller (gewichteten) Punkte im Inneren des Fensters liegt, also durch

$$\hat{y}_i = \hat{a}_i + \hat{b}_i x_i \,.$$

Die beiden Parameter $(\hat{a}_i, \hat{b}_i)$ bezeichnen die Konstante und Steigung der Ausgleichsgeraden und werden nach der gewichteten Kleinsten-Quadrate Methode bestimmt. Das bedeutet, die beiden Größen $\hat{a}_i$, $\hat{b}_i$ minimieren die gewichtete Fehlerquadratsumme (*Sum of Squared Errors*)

$$SSE_t(a_i, b_i) = \sum_{k=1}^{n} t_i(x_k)(y_k - a_i - b_i x_k)^2 \,.$$

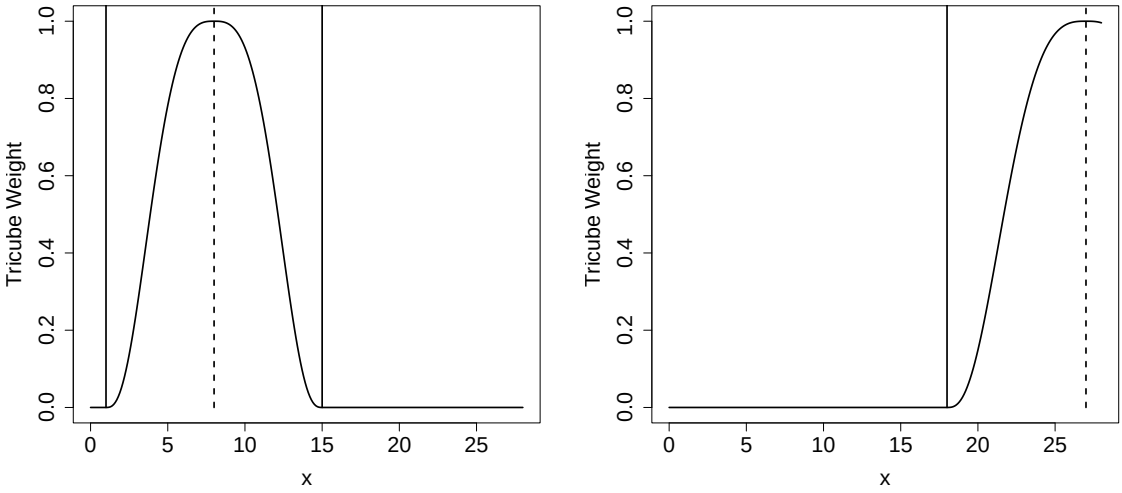


Abbildung 4.7: Nachbarschaftsgewichte um $x_6 = 8$ (links) und $x_{20} = 27$ (rechts).

Dazu werden die Lösungen der beiden Normalgleichungen $\partial SSE_t(a_i, b_i)/\partial a_i = 0$ und $\partial SSE_t(a_i, b_i)/\partial b_i = 0$, also

$$\begin{aligned} \sum_{k=1}^{n} t_i(x_k)(y_k - \hat{a}_i - \hat{b}_i x_k) &= 0\,, \\ \sum_{k=1}^{n} t_i(x_k)(y_k - \hat{a}_i - \hat{b}_i x_k)x_k &= 0 \end{aligned}$$

benötigt. Ist $d_i = 0$, so nimmt man als geglätteten Wert $\hat{y}_i$ das arithmetische Mittel aller y_k innerhalb dieses Fensters. Die Abbildung 4.8 stellt die Ergebnisse für die beiden betrachteten Fenster graphisch dar. Wendet man diese Idee der `lowess` Glättung in allen n Punkten an, so führt dies zum Ergebnis in Abbildung 4.9 (links).

Das Verfahren ist jedoch auch anfällig auf extreme y-Werte, so genannte Ausreißer. Um dies zu demonstrieren wurde der Wert von y_{11} von 23 auf 40 verändert und für diese modifizierten Daten die `lowess`-Glättung gerechnet. In der rechten Teil der Abbildung 4.9 ist dafür das Ergebnis dargestellt. Deutlich ist zu erkennen, dass dieser eine extreme y-Wert über einen großen Bereich die geglätteten Werte stark beeinflusst und nach oben zieht. Um dies zu vermeiden oder um den Einfluss einzelner extremer Punkte zu reduzieren, sollte als zusätzlicher vierter Schritt noch eine **Robustifizierung** durchgeführt werden. Als erstes berechnet man dazu in allen n Punkten die **Residuen**

$$r_i = y_i - \hat{y}_i\,.$$

Diese sind für die modifizierten Daten in der Abbildung 4.10 links dargestellt sind. Deutlich ist zu erkennen, dass die eine extreme Beobachtung auch ein extremes, positives Residuum ($r_{11} = 14$) generiert hat, und dass die Residuen in der Nachbarschaft stark

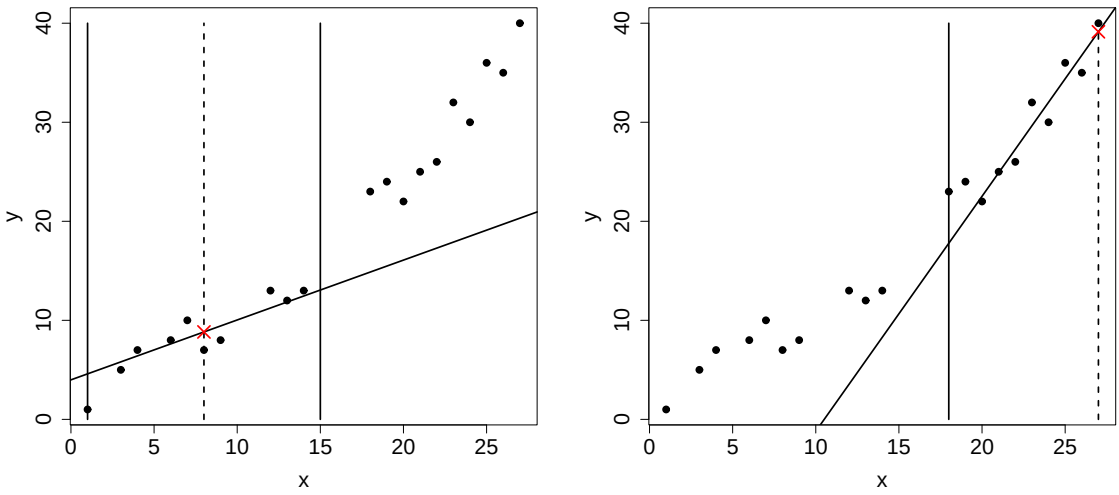


Abbildung 4.8: Ergebnisse der gewichteten linearen Regression in den Streifen um $x_6 = 8$ (links) und um $x_{20} = 27$ (rechts) mit geglätteten Punkten ($\times$).

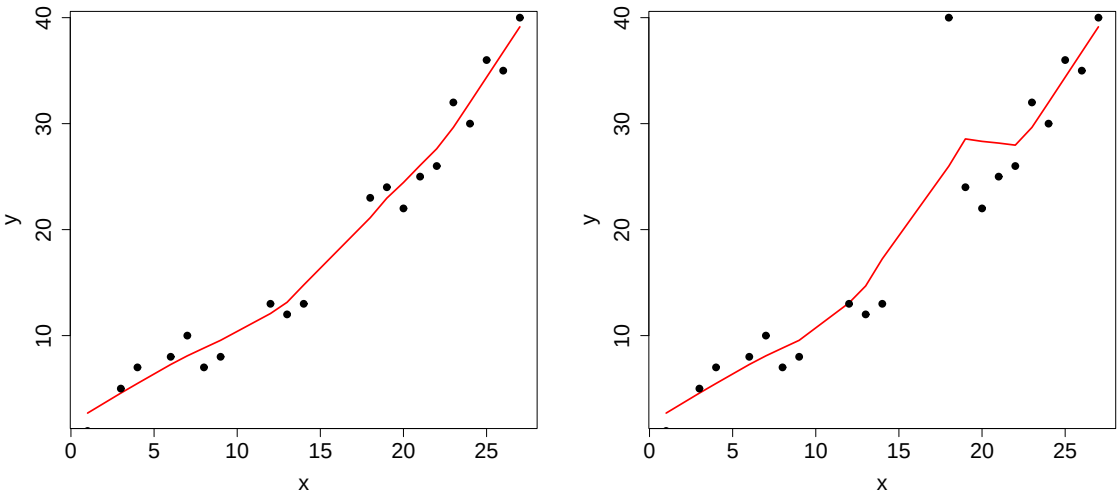


Abbildung 4.9: `lowess`-Glättung der originalen 20 Punkte (links) und der mit einem extremen Wert $y_{11} = 40$ modifizierten Daten (rechts).

negativ sind. Ausreißer haben grosse Residuen und man versucht jetzt den Einfluss von Punkten mit derartigen Residuen auf das Ergebnis der Glättung zu reduzieren. Mit der in der Abbildung 4.10 rechts dargestellten *Bisquare-Funktion*

$$B(u) = \begin{cases} (1-u^2)^2 & \text{für} \quad |u| < 1, \\ 0 & \text{sonst} \end{cases}$$

werden die Residuen relativiert, indem man zusätzlich noch Robustheitsgewichte für die

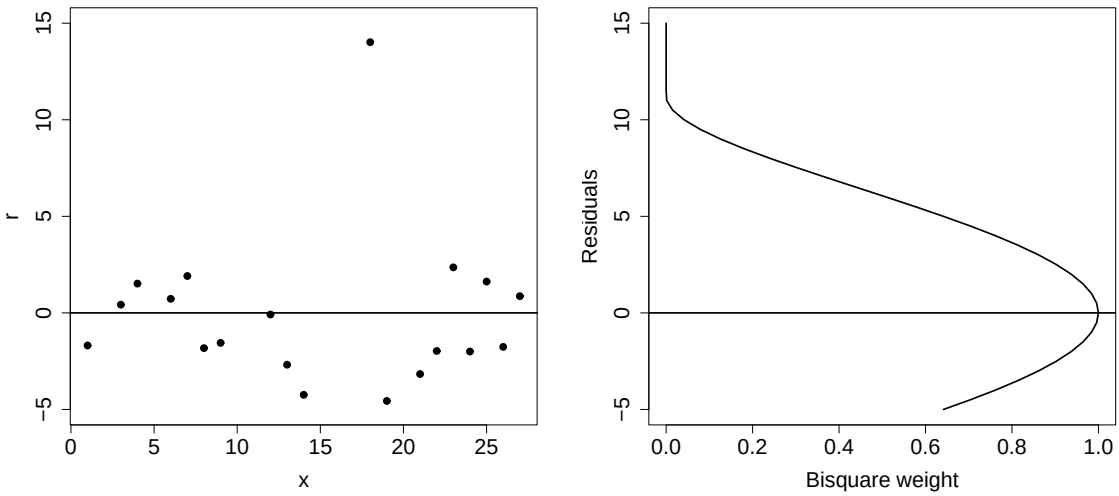


Abbildung 4.10: Scatter-Plot der Residuen r_i gegen die x_i (links) und auf diese Residuen angewandte Gewichtung (rechts).

Punkte (x_k, y_k) verwendet, die definiert sind durch

$$w(x_k) = B\left(\frac{r_k}{6m}\right) ,$$

wobei m den Median aller absoluten Residuen, $|r_k|$, bezeichnet. Große Residuen bekommen dadurch kleine Gewichte.
Bemerkung zur Skalierung durch $6m$*:* Falls eine Zufallsvariable $R \sim N(0, \sigma^2)$ vorliegt, dann folgt für deren Betrag, dass dieser halbnormal-verteilt ist, also $|R| \sim H(\sigma^2)$ gilt. Die Halbnormal-Verteilungsfunktion ist gerade $2\Phi\left(|r|/\sigma\right) - 1$. Daher gilt für den theoretischen Median m dieser Verteilung $2\Phi\left(m/\sigma\right) - 1 = 1/2$. Somit entspricht m/σ dem Quantil $z_{3/4}$ der Standard-Normalverteilung. Mit dieser Überlegung ergibt sich $m = 0.675\sigma \approx 2/3\sigma$, also $6m \approx 4\sigma$. Wir skalieren hier daher etwa mit der vier-fachen Standardabweichung der absoluten Residuen.
Man geht jetzt zurück zu Schritt 2. und verwendet wiederum die Methode der gewichteten Kleinsten-Quadrate, wobei sich nun die Gewichte aus dem Produkt der Nachbarschafts- mit den Robustheitsgewichten zusammensetzen. Es werden nun jene Schätzer $\hat{a}_i^*$ und $\hat{b}_i^*$, mit $i = 1, \ldots, n$, bestimmt, welche die derart gewichteten Fehlerquadratsummen

$$SSE_{tw}(a_i, b_i) = \sum_{k=1}^{n} w(x_k) t_i(x_k)(y_k - a_i - b_i x_k)^2$$

minimieren. Dadurch resultieren die geglätteten Werte

$$\hat{y}_i^* = \hat{a}_i^* + \hat{b}_i^* x_i$$

mit den entsprechenden Residuen

$$r_i^* = y_i - \hat{y}_i^* .$$

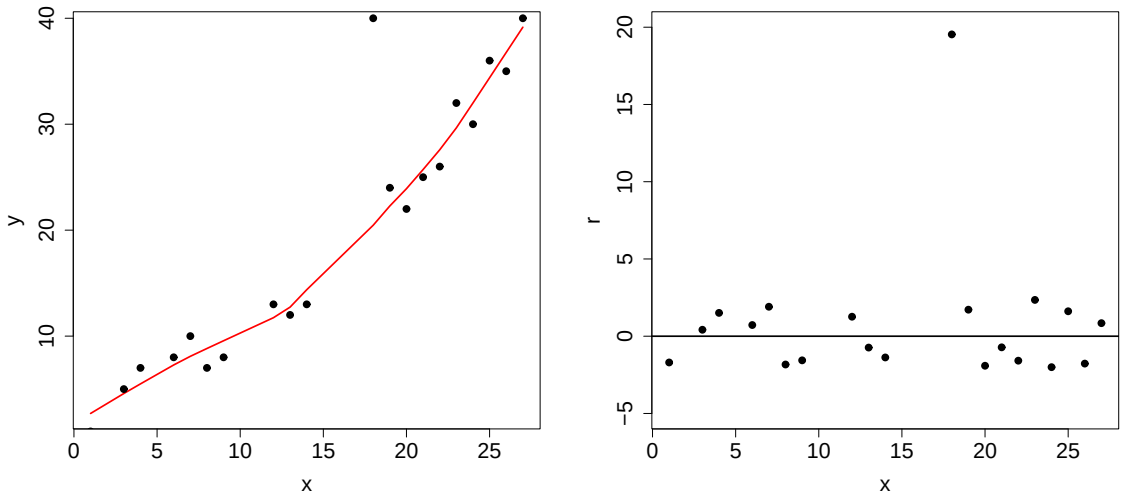


Abbildung 4.11: Geglättete Werte nach dem 1. Robustifizierungsschritt (links) mit Scatter-Plot der Residuen r_i^* (rechts).

Dieser Schritt einer Robustifizierung sollte zumindest zweimal durchgeführt werden. In der Abbildung 4.11 ist links das Ergebnis der `lowess` Glättung (wiederum mit $f = 1/2$) nach der ersten Robustifizierung dargestellt. Die entsprechenden Residuen im rechten Teil dieser Abbildung zeigen, dass zwar jetzt $r_{11} = 19.5$ gilt, aber dafür die benachbarten Residuen kleiner geworden sind und keine auffällige Struktur mehr haben.

Beispiel 4.6 *Für die Variable* `VC` *resultiert als geglättete Abhängigkeit von* `age` *der in der Abbildung 4.12 links dargestellte Verlauf. Dabei wurde $f = 1/2$ gewählt und zweimal robustifiziert. Bemerke, dass das Ergebnis von* `lowess` *nur aus den x-Werten (aufsteigend sortiert) und den dazugehörenden geglätteten y-Werten (in entsprechender Reihenfolge) besteht. Bei der korrekten Berechnung von Residuen muss also auf die Sortierung aufgepasst werden. Diese sind im rechten Teil der Abbildung 4.12 gezeigt und weisen keinerlei auffällige Struktur auf.*

```
> o <- order(age);  o.age <- age[o];  o.VC  <- VC[o]
> lowess.fit <- lowess(o.VC ~ o.age, f=1/2, iter=2)
> plot(lowess.fit);  points(age, VC)

> yhat <- lowess.fit$y;  r <- o.VC - yhat
> plot(o.age, r);  abline(h=0)
```

4.2 Lokationstests bei abhängigen Stichproben

Gegeben sind unabhängige, zweidimensionale Stichprobenvektoren $(X_1, Y_1), \ldots, (X_n, Y_n)$. Sei $D_i = X_i - Y_i \overset{iid}{\sim} F$ mit $\mathrm{E}(D_i) = \mathrm{E}(X_i) - \mathrm{E}(Y_i) = \theta$. Das Mittel der Differenzen $\overline{D} = \overline{X} - \overline{Y}$ ist ein unverzerrter Schätzer für θ. Beschreibt X die Messung vor einer Behandlung und Y den Wert danach, so kann ein möglicher *Behandlungseffekt* durch Testen der Hypothese $H_0 : \theta = 0$ gegen $H_1 : \theta \neq 0$ erkannt werden.

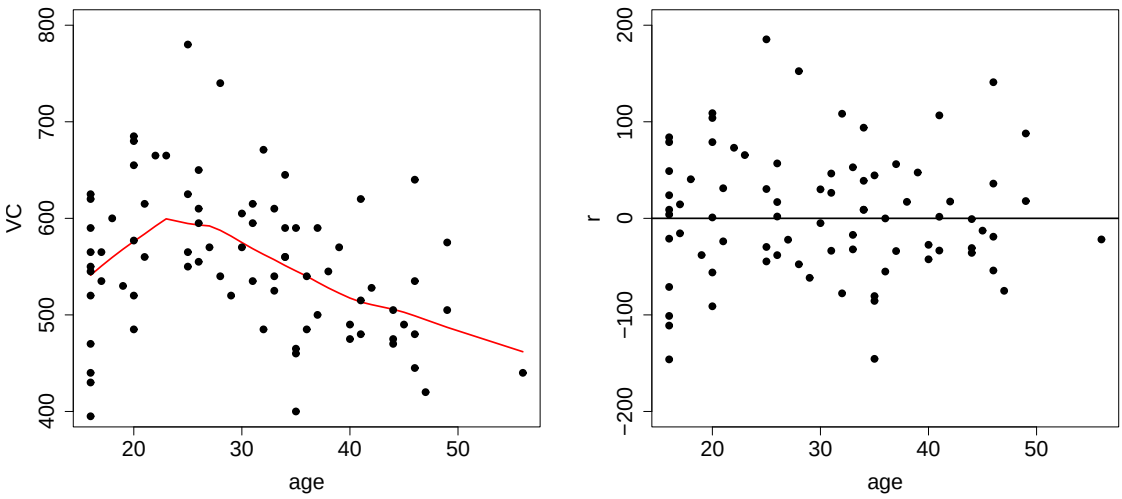


Abbildung 4.12: Ergebnis von `lowess` für `VC` in Abhängigkeit von `age` mit $f = 1/2$ und zweimaliger Robustifizierung (links) sowie entsprechende Residuen (rechts).

4.2.1 Parametrischer Test bei Normalverteilung

Unter den Annahmen $X_i \sim N(\mu_X, \sigma_X^2)$, $Y_i \sim N(\mu_Y, \sigma_Y^2)$, und $\sigma_{XY}^2 = \mathrm{cov}(X_i, Y_i) \neq 0$ folgt

$$D_i = X_i - Y_i \overset{iid}{\sim} N\left(\theta, \sigma_D^2\right)$$

mit $\theta = \mu_X - \mu_Y$ und $\sigma_D^2 = \sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}^2$, sowie

$$\overline{D} = \overline{X} - \overline{Y} \sim N(\theta, \sigma_D^2/n)\,.$$

Folgende Hypothesen können getestet werden:

H_0	H_1	Entscheidung gegen H_0, falls	kritische Werte
$\theta = 0$	$\theta \neq 0$	$T < c_3$ oder $T > c_4$	$c_3 = t_{\alpha/2}$
			$c_4 = t_{1-\alpha/2}$
$\theta \leq 0$	$\theta > 0$	$T > c_1$	$c_1 = t_{1-\alpha}$
$\theta \geq 0$	$\theta < 0$	$T < c_2$	$c_2 = t_\alpha$

mit

$$T = \frac{\overline{D}}{S_D}\sqrt{n}, \quad \text{und} \quad S_D = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2}.$$

Unter $H_0 : \theta = 0$ gilt $\mathrm{E}(\overline{D}) = 0$ und $T \sim t_{n-1}$.

Beispiel 4.7 *An $n = 10$ PKW's wird die Leistung von zwei Arten von Kraftstoff A und B getestet. Dabei ergaben sich die folgenden Fahrleistungen in km:*

PKW	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
A	*89*	*110*	*105*	*101*	*90*	*92*	*104*	*100*	*101*	*98*
B	*95*	*109*	*111*	*110*	*91*	*95*	*106*	*99*	*104*	*101*
d_i	*-6*	*1*	*-6*	*-9*	*-1*	*-3*	*-2*	*1*	*-3*	*-3*

Wegen $\overline{d} = -3.1$ *und* $s_D = 3.18$ *folgt* $t = -3.08$ *als Realisierung der Teststatistik. Für* $\alpha = 0.05$ *ist beim zweiseitigen Test* $t_{1-\alpha/2;n-1} = t_{0.975;9} = 2.26$. *Da* $t > t_{1-\alpha/2}$ *wird die Nullhypothese, dass die durchschnittliche Fahrleistung bei A und B gleich ist, verworfen.*

```
> A <- c(89, 110, 105, 101, 90, 92, 104, 100, 101,  98)
> B <- c(95, 109, 111, 110, 91, 95, 106,  99, 104, 101)
> t.test(A, B, paired = TRUE)

        Paired t-test

data:  A and B
t = -3.0846, df = 9, p-value = 0.01304
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.3734398 -0.8265602
sample estimates:
mean of the differences
                  -3.1
```

Wir vermuten, dass Kraftstoff B besser als A ist, und führen auch den einseitigen Test der Nullhypothese „Kraftstoff A bringt mehr Fahrleistung als B“ durch. Hierfür ist $t_{1-\alpha} = 1.83$ *und* H_0 *kann natürlich auch verworfen werden.*

```
> t.test(A, B, alt="less", paired = TRUE)

        Paired t-test

data:  A and B
t = -3.0846, df = 9, p-value = 0.006521
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -1.257744
sample estimates:
mean of the differences
                  -3.1
```

Der einseitige p-Wert entspricht natürlich genau dem halben p-Wert des zweiseitigen Test. Eigentlich vermuten wir sogar, dass Treibstoff B sogar im Mittel um mindestens 2 Liter besser als A ist. Die entsprechende Nullhypothese kann jetzt nicht mehr verworfen werden.

```
> t.test(A, B, alt="less", paired = TRUE, mu=-2)

        Paired t-test

data:  A and B
t = -1.0945, df = 9, p-value = 0.1511
alternative hypothesis: true difference in means is less than -2
```

4.2.2 Der Vorzeichentest

Für den **Vorzeichentest** benutzt man als Teststatistik entweder die Anzahl der Differenzen $X_i - Y_i$ mit positiven Vorzeichen (falls kardinales Meßniveau), oder die Anzahl der Paare mit $X_i > Y_i$ (falls ordinales Meßniveau, z.B. Vergleich Mathematiknote – Deutschnote beim i-ten Schüler). Unter der Annahme, dass die entsprechenden Indikatoren unabhängig ident verteilte Stichprobenvariablen sind und dass $P(X_i = Y_i) = 0$ (stetig) gilt, können die folgenden Hypothesen getestet werden:

- Test A: $H_0 : P(X < Y) = P(X > Y), \quad H_1 : P(X < Y) \neq P(X > Y),$
- Test B: $H_0 : P(X < Y) \leq P(X > Y), \quad H_1 : P(X < Y) > P(X > Y),$
- Test C: $H_0 : P(X < Y) \geq P(X > Y), \quad H_1 : P(X < Y) < P(X > Y).$

Die Teststatistik ist hierbei definiert durch

$$T = \sum_{i=1}^{n} Z_i \quad \text{mit} \quad Z_i = \begin{cases} 1 & \text{falls} \quad X_i > Y_i, \\ 0 & \text{falls} \quad X_i < Y_i. \end{cases}$$

Bei kardinalem Meßniveau gibt T die Anzahl der positiven Differenzen $X_i - Y_i$ an. Bei ordinalem Niveau wird T als Zählvariable interpretiert, welche die Anzahl der Paare mit $X_i > Y_i$ angibt. Aufgrund der Annahmen genügt die Teststatistik T unter H_0 einer Binomialverteilung mit Parameter n und $p = P(X > Y) = 1/2$.
Entscheidungsregel: H_0 wird abgelehnt, wenn

- Test A: $t \leq t_{\alpha/2}$ oder $t \geq n - t_{\alpha/2}$,
- Test B: $t \geq n - t_{\alpha}$,
- Test C: $t \leq t_{\alpha}$,

wobei t_α das α-Quantil einer $Binomial(n, 1/2)$-Verteilung ist.
Der Vorzeichen-Test lässt sich bei kardinalem Meßniveau auch zur Prüfung der Hypothese

$$H_0' : \text{'Der Median von } X - Y \text{ ist } m_0\text{'}$$

verwenden. Die Teststatistik ist dann

$$T' = \sum_{i=1}^{n} Z_i' \quad \text{mit} \quad Z_i' = \begin{cases} 1 & \text{falls} \quad m_0 < X_i - Y_i\,, \\ 0 & \text{falls} \quad m_0 > X_i - Y_i\,, \end{cases}$$

und T' ist wiederum unter der Nullhypothese $Binomial(n, 1/2)$-verteilt.
Für $n \geq 20$ approximiert man die Verteilung von $Z = (T - n/2)/\sqrt{n/4}$ durch die $N(0, 1)$-Verteilung.
Ist die Variable $X - Y$ symmetrisch um ihren Median verteilt und liegt dafür zumindest kardinales Meßniveau vor, so ist der Wilcoxon-Test dem Vorzeichen-Test vorzuziehen, da dieser mehr Information aus der Stichprobe nutzt.

Beispiel 4.8 *Treibstoffe A und B werden auf unterschiedliche Fahrleistung getestet.*

PKW	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
A	*89*	*110*	*105*	*101*	*90*	*92*	*104*	*100*	*101*	*98*
B	*95*	*109*	*111*	*110*	*91*	*95*	*106*	*99*	*104*	*101*
d_i	*-6*	*1*	*-6*	*-9*	*-1*	*-3*	*-2*	*1*	*-3*	*-3*
z_i	*1*	*0*	*1*	*1*	*1*	*1*	*1*	*0*	*1*	*1*

Wir beobachten also $t = 2$. *Für* $\alpha = 0.055$ *ist gerade* $t_\alpha = 2$ *(d.h.* $P(T \leq 2) = \alpha$*) und damit* $t_{1-\alpha} = n - t_\alpha = 10 - 2 = 8$ *(d.h.* $P(T \geq 8) = \alpha$*). Es wird somit* H_0*: „Treibstoff A ist im Mittel besser als B“ gerade noch abgelehnt.*

```
> binom.test(sum(A>B), length(A), p=1/2, alt="two.sided")

        Exact binomial test

data:  sum(A > B) and length(A)
number of successes = 2, number of trials = 10, p-value = 0.1094
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.02521073 0.55609546
sample estimates:
probability of success
                   0.2

> binom.test(sum(A>B), length(A), p=1/2, alt="less")

        Exact binomial test

data:  sum(A > B) and length(A)
number of successes = 2, number of trials = 10, p-value = 0.05469
alternative hypothesis: true probability of success is less than 0.5
95 percent confidence interval:
 0.0000000 0.5069013
probability of success
                   0.2
```

4.2.3 Wilcoxon-Test

Der Wilcoxon-Test bei verbundenen Stichproben entspricht dem Wilcoxon-Vorzeichen-Rang-Test für den Median beim Einstichproben-Problem. Seien die $D_i = X_i - Y_i \overset{iid}{\sim} F$, wobei F stetig und **symmetrisch** um den Median der Differenzen m_D verteilt ist. Wegen der Symmetrie von F gilt $\mathrm{E}(D) = m_D$ und folgende Hypothesen werden getestet:

- Test A: $H_0 : m_D = 0; \quad H_1 : m_D \neq 0,$
- Test B: $H_0 : m_D \leq 0; \quad H_1 : m_D > 0,$
- Test C: $H_0 : m_D \geq 0; \quad H_1 : m_D < 0.$

Die Teststatistik ist definiert als

$$W^+ = \sum_{i=1}^{n} Z_i R(|D_i|) \quad \text{mit} \quad Z_i = \begin{cases} 1 & \text{für} \quad D_i > 0, \\ 0 & \text{für} \quad D_i < 0, \end{cases}$$

wobei $R(|D_i|)$ den Rang von $|D_i|$ beschreibt. Obwohl Wilcoxon's Vorzeichen-Rangtest für den Median und der Wilcoxon-Test für verbundene Stichproben auf verschiedene Probleme angewandt werden, haben sie die Teststatistik gemeinsam.
H_0 wird abgelehnt, falls

- Test A: $w^+ \leq w_{\alpha/2}$ oder $w^+ \geq w_{1-\alpha/2}$,
- Test B: $w^+ \geq w_{1-\alpha}$,
- Test C: $w^+ \leq w_{\alpha}$.

Die Quantile w_α^+ werden wieder aus der Tabelle F entnommen. Für $n > 20$ verwende man die Normalverteilungsapproximation.

Beispiel 4.9 *Benzinverbrauch zweier Kraftstoffarten A und B.*

PKW	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
A	*89*	*110*	*105*	*101*	*90*	*92*	*104*	*100*	*101*	*98*
B	*95*	*109*	*111*	*110*	*91*	*95*	*106*	*99*	*104*	*101*
d_i	*-6*	*1*	*-6*	*-9*	*-1*	*-3*	*-2*	*1*	*-3*	*-3*
$r(\lvert d_i \rvert)$	*8.5*	*2*	*8.5*	*10*	*2*	*6*	*4*	*2*	*6*	*6*

Für die Vergabe der Ränge der $|d_i|$ wurden wegen der Bindungen der Realisierungen 1, 3 und 6 die Methode der Durchschnittsränge verwendet. Es kann also $w^+ = 4$ beobachtet werden. Für $\alpha = 0.05$ ist $w_\alpha = 10$, d.h. für das einseitige Testproblem (Test C) wird H_0 wegen $2 < 10$ abgelehnt.

```
> wilcox.test(A, B, paired = TRUE, alt="less")

        Wilcoxon signed rank test with continuity correction

data:  A and B
V = 4, p-value = 0.009182
alternative hypothesis: true mu is less than 0
```

Es ist zu beachten, dass die Hypothese $H_0 : m_D = 0$ im Allgemeinen nicht äquivalent ist mit $H_0 : m_X = m_Y$. Sind jedoch X bzw. Y symmetrisch um ihre Mediane m_X bzw. m_Y verteilt, so sind die folgenden Aussagen äquivalent:

- $m_D = 0$,
- $m_X = m_Y$,
- $E(X) = E(Y)$.

Der Wilcoxon Vorzeichen-Rangsummen-Test für den Median kann als Sonderfall des Wilcoxon-Tests für verbundene Stichproben angesehen werden, indem man $X_1, \ldots, X_n$ als Paare $(X_1, m_X), \ldots, (X_n, m_X)$ auffasst.

4.3 Korrelation und Unabhängigkeit

Die beiden Begriffe „Korrelation“ und „Kontingenz“ werden beide als Synonyme für die stochastische Unabhängigkeit zweier Zufallsvariablen verwendet. Es hat sich jedoch eingebürgert von Kontingenz dann zu sprechen, wenn sowohl X als auch Y nominal skaliert sind, und von Korrelation, wenn X und Y zumindest ordinal skaliert vorliegen.
Der Korrelationskoeffizient der Grundgesamtheit

$$\rho(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}}$$

beschreibt den Grad der linearen Abhängigkeit zweier Variablen X und Y.
Bei Kontingenztafeln wird häufig das „Odds Ratio“ verwendet. Liegt eine 2×2 Tafel vor mit Wahrscheinlichkeiten

	$Y=1$	$Y=2$	$P(X)$
$X=1$	π_{11}	π_{12}	π_{1+}
$X=2$	π_{21}	π_{22}	π_{2+}
$P(Y)$	π_{+1}	π_{+2}	

so beschreibt $\Omega_i = \pi_{i1}/\pi_{i2}$ die Chance („odds“), dass bei $X = i$, $i = 1, 2$, die Variable Y in Spalte 1 statt in Spalte 2 ist. Jedes Ω_i ist nicht-negativ mit Werten größer 1, falls für Y die Spalte 1 wahrscheinlicher ist als die Spalte 2. Den Quotienten der beiden Chancen Ω_1 und Ω_2 nennt man Chancenverhältnis („odds ratio“)

$$\theta = \Omega_1/\Omega_2 = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}\,.$$

Sind X und Y unabhängig, so gilt $\pi_{ij} = \pi_{i+}\pi_{+j}$, und es folgt $\theta = 1$.

4.3.1 Betrachtung von ρ bei bivariater Normalverteilung

Definition 4.1 *Die Dichte der bivariaten Normalverteilung $N_2(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ ist definiert als*

$$f(x,y)=\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2-2\rho\frac{x-\mu_X}{\sigma_X}\frac{y-\mu_Y}{\sigma_Y}+\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right)\right).$$

Eigenschaften von ρ:

1. $|\rho| \leq 1$,
2. $|\rho| = 1 \Leftrightarrow P(Y = aX + b) = 1$, wobei $a \neq 0$ und b Konstanten sind.
3. X, Y stochastisch unabhängig $\Rightarrow \rho = 0$. Die Umkehrung gilt im Allgemeinen nicht,
4. Sind $(X, Y) \sim N_2(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ mit $\rho = 0$, so gilt X und Y sind stochastisch unabhängig. (Umkehrung von 3. also nur bei Normalverteilung).

Unter der Normalverteilungsannahme können zur Prüfung der Unabhängigkeitshypothese die folgenden drei Situationen betrachtet werden:

H_0	H_1	Entscheidung gegen H_0, falls	kritische Werte
$\rho = 0$	$\rho \neq 0$	$T < c_3$ oder $T > c_4$	$c_3 = t_{\alpha/2}$
			$c_4 = t_{1-\alpha/2}$
$\rho = 0$	$\rho > 0$	$T > c_1$	$c_1 = t_{1-\alpha}$
$\rho = 0$	$\rho < 0$	$T < c_2$	$c_2 = t_{\alpha}$

Für die Teststatistik T gilt nur unter $H_0 : \rho = 0$

$$T = R\sqrt{\frac{n-2}{1-R^2}} \sim t_{n-2}$$

da X und Y unter H_0 unabhängig sind. Den Schätzer

$$R = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2 \sum_{i=1}^n (Y_i - \overline{Y})^2}}$$

nennt man empirischen Korrelationskoeffizient nach Pearson. Stammen die (X_i, Y_i) aus einer bivariaten Normalverteilung und sind die Paare (X_i, Y_i) voneinander unabhängig, so ist R der Maximum-Likelihood Schätzer für ρ.
Zur Konstruktion eines Konfidenzintervalls für ρ ist es notwendig, über die Verteilung von R unter beliebigen Werten von ρ Bescheid zu wissen. Nur im Fall $\rho = 0$ gilt $T \sim t_{n-2}$. Bei $\rho \neq 0$ ist es zweckmäßig, R passend zu transformieren. Für **Fisher's Z-Transformation** gilt:

$$Z = \frac{1}{2}\log\frac{1+R}{1-R} \overset{as}{\sim} N\left(\frac{1}{2}\log\frac{1+\rho}{1-\rho},\ \frac{1}{n-3}\right).$$

Bezeichne

$$\mu = \frac{1}{2}\log\frac{1+\rho}{1-\rho},$$

so folgt für die standardisierte Größe

$$U = (Z - \mu)\sqrt{n-3} \overset{as}{\sim} N(0,1).$$

Aus der Definition von μ folgt als inverse Funktion

$$\rho = \frac{e^{\mu} - e^{-\mu}}{e^{\mu} + e^{-\mu}} = \tanh(\mu).$$

Der Erwartungswert der Fisher Z-Transformaierten bildet also $\rho \in (-1, +1)$ auf $\mu \in \mathbb{R}$ ab. Als Konfidenzintervall für ein beliebiges ρ ergibt sich somit

$$P\left(z_{\alpha/2} \leq U \leq z_{1-\alpha/2}\right) = P\left(-\frac{z_{1-\alpha/2}}{\sqrt{n-3}} \leq Z - \mu \leq \frac{z_{1-\alpha/2}}{\sqrt{n-3}}\right) = 1 - \alpha.$$

Bezeichne $a = z_{1-\alpha/2}/\sqrt{n-3}$, so folgt wegen der strengen Monotonie von tanh

$$P\left(\tanh(Z-a) \leq \tanh(\mu) = \rho \leq \tanh(Z+a)\right) = 1 - \alpha.$$

Beispiel 4.10 *$n = 9$ Personen bewerben sich für eine freie Stelle. Für die Selektion werden zwei Kommissionen A und B eingesetzt, um die Bewerber zu testen. Nach einigen durchgeführten Tests werden die folgenden Punkte vergeben:*

Bewerber i	1	2	3	4	5	6	7	8	9		
$x_i \ldots$ *Punkte von A*	75	62	87	76	73	66	81	74	77	$\overline{x} = 74.56$	$s_X^2 = 54.78$
$y_i \ldots$ *Punkte von B*	82	69	89	84	80	68	79	70	74	$\overline{y} = 77.22$	$s_Y^2 = 54.19$

Wie groß ist die Übereinstimmung im Urteil der beiden Kommissionen? Wir vermuten ferner, dass die Abhängigkeit der beiden Beurteilungen positiv ist. Unter Annahme einer bivariaten Normalverteilung ist daher die Hypothese $H_0 : \rho = 0$ gegen $H_1 : \rho > 0$ zu testen. Abschließend soll noch ein zweiseitiges Konfidenzintervall für ρ mit Überdeckungswahrscheinlichkeit von 95% bestimmt werden.
Mit der empirischen Kovarianz $s_{XY}^2 = 42.99$ folgt $r = 0.789$. Wegen $t = 3.40 > t_{7,1-0.05} = 1.89$ kann H_0 verworfen werden, d.h. beide Kommissionen erzielten positiv übereinstimmende Resultate. Mit $z_{0.975} = 1.96$ und $a = 1.96/\sqrt{6} = 0.80$ ergibt sich das asymmetrische Konfidenzintervall $\tanh(z - a) = \tanh(0.269) = 0.26$ bzw. $\tanh(z + a) = \tanh(1.869) = 0.95$, also $(0.26, 0.95)$.

```
> A <- c(75,62,87,76,73,66,81,74,77)
> B <- c(82,69,89,84,80,68,79,70,74)
> cor.test(A, B)

        Pearson's product-moment correlation

data:  A and B
t = 3.3971, df = 7, p-value = 0.01149
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2622175 0.9534845
sample estimates:
      cor
0.7889483

> cor.test(A, B, alt="greater")

        Pearson's product-moment correlation

data:  A and B
t = 3.3971, df = 7, p-value = 0.005744
alternative hypothesis: true correlation is greater than 0
95 percent confidence interval:
0.3774919 1.0000000
sample estimates:
      cor
0.7889483
```

4.3.2 Rangkorrelationskoeffizient von Spearman (1904)

Es soll nun eine alternative nicht-parametrische Maßzahl, die zudem auch als Teststatistik für Tests auf Unabhängigkeit benutzt werden kann. Wir nehmen an, dass (X, Y) aus einer beliebigen stetigen zweidimensionalen Verteilung stammt. Liegen die Daten zumindest ordinal skaliert vor, so können jeweils die Ränge der X_i: $R_1, \ldots, R_n$ bzw. der Y_i: $S_1, \ldots, S_n$ bestimmt werden. Nun wird statt mit den originalen Beobachtungen mit diesen beiden Rängen der Korrelationskoeffizient nach Pearson gebildet, d.h.

$$R_S = \frac{\sum(R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum(R_i - \overline{R})^2 \sum(S_i - \overline{S})^2}}\,.$$

Dieser Ausdruck kann vereinfacht werden, denn es gilt:

$$\overline{R} = \frac{1}{n}\sum_{i=1}^{n} R_i = \frac{1}{n}\sum_{i=1}^{n} i = \frac{n+1}{2} = \overline{S}$$

und

$$\sum_{i=1}^{n}\left(R_i - \frac{n+1}{2}\right)^2 = \sum_{i=1}^{n}\left(i - \frac{n+1}{2}\right)^2 = \frac{(n-1)n(n+1)}{12} = \sum_{i=1}^{n}\left(S_i - \frac{n+1}{2}\right)^2\,.$$

Damit ergibt sich

$$R_S = \frac{12}{(n-1)n(n+1)}\sum_{i=1}^{n}\left(R_i - \frac{n+1}{2}\right)\left(S_i - \frac{n+1}{2}\right)\,.$$

Setzt man nun noch $D_i = R_i - S_i$, oder noch besser

$$D_i = \left(R_i - \frac{n+1}{2}\right) - \left(S_i - \frac{n+1}{2}\right)\,,$$

so folgt dafür als Quadratsumme

$$\begin{aligned} D &= \sum_{i=1}^{n} D_i^2 \\ &= \sum_{i=1}^{n}\left(R_i - \frac{n+1}{2}\right)^2 + \sum_{i=1}^{n}\left(S_i - \frac{n+1}{2}\right)^2 - 2\sum_{i=1}^{n}\left(R_i - \frac{n+1}{2}\right)\left(S_i - \frac{n+1}{2}\right) \\ &= 2\frac{(n-1)n(n+1)}{12} - 2\frac{(n-1)n(n+1)}{12}R_S \\ &= \frac{(n-1)n(n+1)}{6}(1 - R_S)\,. \end{aligned}$$

Es resultiert damit die einfache Darstellung

$$R_S = 1 - \frac{6D}{(n-1)n(n+1)}\,.$$

Eigenschaften der Realisierungen von r_S:

1. $-1 \leq r_S \leq +1$,
2. $r_S = +1 \Leftrightarrow r_i = s_i \quad \forall i = 1, \ldots, n$,
3. $r_S = -1 \Leftrightarrow r_i = n + 1 - s_i \quad \forall i = 1, \ldots, n$,
4. r_S ist invariant gegenüber monotonen Transformationen der Daten.

Interpretation von r_S:

1. Ist r_S nahe bei $+1$, so deutet dies auf eine positive Korrelation zwischen X und Y hin. Hat x_i einen hohen (niedrigen) Rang, so hat auch y_i einen hohen (niedrigen) Rang.
2. Ist r_S nahe bei -1, so deutet dies auf eine negative Korrelation zwischen X und Y hin. Hat x_i einen hohen (niedrigen) Rang, so hat auch y_i einen niedrigen (hohen) Rang.
3. Ist r_S nahe Null, so besteht kein Zusammenhang zwischen X und Y.

Bewertung von r_S:

1. r_S ist bei ordinal skalierten Daten anwendbar, jedoch entsteht ein Verlust an Information, falls die Daten kardinal skaliert vorliegen.
2. r_S ist gut interpretierbar, falls es in der Nähe von -1, 0 oder $+1$ liegt. Ansonsten ist es jedoch nicht zu interpretieren.
3. r_S ist nicht als Schätzung für ρ geeignet.

Für einen **Test auf Unabhängigkeit**, d.h.

- Test A: H_0: X und Y sind unabhängig, H_1: X und Y sind korreliert
- Test B: H_0: X und Y sind unabhängig, H_1: X und Y sind positiv korreliert
- Test C: H_0: X und Y sind unabhängig, H_1: X und Y sind negativ korreliert

verwendet man die **Hotelling-Pabst-Statistik** D. Sie ist eine lineare Funktion in R_S. Sortiert man die Stichprobe (X_i, Y_i) nach den X_i in aufsteigender Reihenfolge, so wird dadurch $R_i = i$ und wegen $\sum S_i^2 = \sum i^2$ ergibt sich

$$D = \sum_{i=1}^{n} (i - S_i)^2 = \sum_{i=1}^{n} i^2 - 2\sum_{i=1}^{n} iS_i + \sum_{i=1}^{n} S_i^2 = \frac{n(n+1)(2n+1)}{3} - 2\sum_{i=1}^{n} iS_i \,.$$

Die Verteilung von D und damit auch die von R_S hängt somit nur von der Verteilung von $\sum iS_i$ ab. Unter der Unabhängigkeitshypothese nimmt der Rangvektor $(S_1, \ldots, S_n)$ die $n!$ Permutationen von $(1, \ldots, n)$ mit gleicher Wahrscheinlichkeit an. Für $n = 3$ ist in der Tabelle 4.1 die Berechnung der Verteilung illustriert.

(s_1, s_2, s_3)	$\sum is_i$	d	r_S
(1, 2, 3)	14	0	1
(1, 3, 2)	13	2	1/2
(2, 1, 3)	13	2	1/2
(2, 3, 1)	11	6	-1/2
(3, 1, 2)	11	6	-1/2
(3, 2, 1)	10	8	-1

Tabelle 4.1: Zur Verteilung von R_S und D im Falle $n = 3$.

R_S und D sind diskret verteilt und haben folgende Wahrscheinlichkeitsfunktionen:

r_S	-1	$-1/2$	$1/2$	1
$P(R_S = r_S)$	1/6	2/6	2/6	1/6

d	8	6	2	0
$P(D = d)$	1/6	2/6	2/6	1/6

Es ist zu beachten, dass im Gegensatz zu R_S kleine (große) Werte von D für eine positive (negative) Korrelation sprechen.
Weiters gilt nach Satz 2.3 über die Verteilung von Rängen

$$\mathrm{E}(S_i) = \frac{n+1}{2} \qquad \mathrm{var}(S_i) = \frac{n^2-1}{12}\,, \qquad \mathrm{cov}(S_i, S_j) = -\frac{n+1}{12} \quad \forall i \neq j\,.$$

Daraus folgt sofort

$$\begin{aligned} \mathrm{E}\left(\sum_{i=1}^{n} iS_i\right) &= \sum_{i=1}^{n} i\mathrm{E}(S_i) = \frac{n(n+1)^2}{4}, \\ \mathrm{var}\left(\sum_{i=1}^{n} iS_i\right) &= \sum_{i=1}^{n} i^2\mathrm{var}(S_i) + \sum_{i\neq j} ij\mathrm{cov}(S_i, S_j) = \frac{(n-1)n^2(n+1)^2}{144} \end{aligned}$$

und schließlich

$$\mathrm{E}(D) = \frac{(n-1)n(n+1)}{6}\,, \qquad \mathrm{var}(D) = \frac{(n-1)n^2(n+1)^2}{36}\,,$$

sowie

$$\mathrm{E}(R_S) = 0\,, \qquad \mathrm{var}(R_S) = \frac{1}{n-1}\,.$$

Für $n \leq 11$ wird H_0 abgelehnt, falls

- Test A: $d \leq d_{\alpha/2}$ oder $d \geq d_{1-\alpha/2}$,
- Test B: $d \leq d_{\alpha}$,
- Test C: $d \geq d_{1-\alpha}$.

Die exakten Quantile d_α werden für $n \leq 11$ aus der Tabelle M entnommen. Für $12 \leq n \leq 20$ approximiert man $T = R_S\sqrt{(n-2)/(1-R_S^2)}$ durch die t_{n-2}-Verteilung. Für $n > 20$ verwende die Approximation $Z = R_S\sqrt{n-1} \overset{as}{\sim} N(0,1)$.

Bemerkung: R_S ist nur unter der Unabhängigkeitshypothese eine verteilungsfreie Statistik und deshalb auch nicht für die Konstruktion eines Konfidenzintervalls für ρ geeignet. Ferner wird ρ durch R_S oft überschätzt, weshalb R_S auch keine brauchbare Schätzung für ρ darstellt.

Beispiel 4.11 *Ergebnisse der beiden Kommissionen A und B:*

Bewerber i	1	2	3	4	5	6	7	8	9
$x_i \ldots$ *Punkte von A*	75	62	87	76	73	66	81	74	77
$y_i \ldots$ *Punkte von B*	82	69	89	84	80	68	79	70	74
Rang r_i *von A*	5	1	9	6	3	2	8	4	7
Rang s_i *von B*	7	2	9	8	6	1	5	3	4
d_i^2	4	1	0	4	9	1	9	1	9

Es resultiert $d = \sum d_i^2 = 38$ *und somit* $r_S = 1 - 6 \cdot 38/720 = 0.683$. *Der Pearson Koeffizient ergab hingegen* $r = 0.789$. *Mit* $\alpha = 0.05$ *liefert die Tabelle M für Test B:* $d = 38 < d_{0.05} = 48$, *was zur Ablehnung der Unabhängigkeitshypothese führt.*

```
> cor.test(A, B, alt="greater", method="spearman")

        Spearman's rank correlation rho

data:  A and B
S = 38, p-value = 0.02516
alternative hypothesis: true rho is greater than 0
sample estimates:
      rho
0.6833333
```

4.3.3 Kendall's τ

Die Abhängigkeit der beiden Variablen kann auch durch die Anzahl der *konkordanten* und *diskordanten* Paare beschrieben werden.

Definition 4.2 *Das Paar* $[(x_i, y_i), (x_j, y_j)]$ *heißt konkordant (übereinstimmend), falls*

1. $x_i < x_j \Rightarrow y_i < y_j$, *oder*
2. $x_i > x_j \Rightarrow y_i > y_j$

gilt. Andernfalls heißt das Paar diskordant.

Aus der Stichprobe können $\binom{n}{2}$ Paare $[(x_i, y_i), (x_j, y_j)]$, $i < j$, ausgewählt werden. Sei n_k die Anzahl konkordanter Paare und n_d die Anzahl diskordanter Paare, so muss $n_k + n_d = \binom{n}{2}$ gelten. Als Maß für die Korrelation zwischen X und Y verwendet man

$$\tau = \frac{n_k - n_d}{\binom{n}{2}}.$$

Offensichtlich ist $-1 \leq \tau \leq +1$ und es gelten die folgenden Beziehungen:

- $\tau = +1 \Leftrightarrow n_k = \binom{n}{2} \Leftrightarrow$ perfekte positive Korrelation,
- $\tau = 0 \Leftrightarrow n_k = n_d \Leftrightarrow$ keine Korrelation,
- $\tau = -1 \Leftrightarrow n_d = \binom{n}{2} \Leftrightarrow$ perfekte negative Korrelation.

Ebenso wie R_S kann auch mit der Teststatistik τ auf Unabhängigkeit der Variablen X und Y getestet werden. Einfacher jedoch ist Kendall's S

$$S = N_k - N_d \,.$$

Die Quantile s_α findet man für $n \leq 10$ in der Tabelle N. Die Testprozedur ist äquivalent mit der Spearman-Prozedur. Unter H_0 kann gezeigt werden, dass schon für $n \geq 8$ gilt

$$\tau \overset{as}{\sim} N\left(0, \frac{2(2n+5)}{9n(n-1)}\right) .$$

Beispiel 4.12 *Im Vergleich der beiden Testergebnis-Reihen können die folgenden konkordanten bzw. diskordanten Paare beobachtet werden:*

Konkordante Beurteilungen	*[1,2],*	*[1,3],*	*[1,4],*	*[1,5],*	*[1,6],*	*[1,8],*	*[2,3],*	*[2,4],*	*[2,5]*
	[2,7],	*[2,8],*	*[2,9],*	*[3,4],*	*[3,5],*	*[3,6],*	*[3,7],*	*[3,8],*	*[3,9]*
	[4,5],	*[4,6],*	*[4,8],*	*[5,6],*	*[6,7],*	*[6,8],*	*[6,9],*	*[7,8],*	*[7,9]*
	[8,9]								
Diskordante Beurteilungen	*[1,7],*	*[1,9],*	*[2,6],*	*[4,7],*	*[4,9],*	*[5,7],*	*[5,8],*	*[5,9]*	

Wir finden $n_k = 28$ konkordante und $n_d = 8$ diskordante aus den insgesamt $\binom{n}{2} = 36$ Paaren. Somit ist $\tau = (28-8)/36 = 0.556$ (noch kleiner als r_S). Kendall's S realisiert in $s = 20$. Für $n = 9$ und $s = 20$ liefert die Tabelle N (Test B) $\alpha = 0.025$. Da $\alpha < 0.05$ gilt, wird auch hier H_0 verworfen.

```
> cor.test(A, B, alt="greater", method="kendall")

        Kendall's rank correlation tau

data:  A and B
T = 28, p-value = 0.02231
alternative hypothesis: true tau is greater than 0
sample estimates:
      tau
0.5555556
```

Hierbei bezeichnet T nicht Kendall's S sondern es gibt die Anzahl konkordanter Paare an. Alle drei vorgestellten Teststatistiken führten zum Verwerfen von H_0 zu Gunsten der positiven Zusammenhangs-Hypothese. Der kleinste p-Wert wurde mit der Pearson-Statistik erzielt (jedoch unter sehr starken Annahmen).

4.4 Kontingenztafeln

Wie bereits erwähnt, verwenden wir den Begriff *Kontingenz* für den Zusammenhang (generelle stochastische Abhängigkeit) von nominal skalierten Variablen (X, Y). Derartige Variablen nennt man auch **Faktoren** und all ihre möglichen Realisierungen bezeichnet man als **Stufen** dieses Faktors. Der nun vorgestellte Test auf die stochastische Unabhängigkeit zweier Faktoren kann unmittelbar auch auf mehr als zwei Faktoren verallgemeinert werden.

4.4.1 Der χ^2-Test auf Unabhängigkeit

Sei $(X_1, Y_1), \ldots, (X_n, Y_n)$ eine Zufallsstichprobe aus einer zweidimensionalen Population. Die Daten können hierbei beliebiges Meßniveau aufweisen. Meist betrachtet man die Situation, in der sowohl X als auch Y Faktoren sind mit k bzw. ℓ Faktorstufen, welche durch A und B bezeichnet sind. D.h., A und B können jeweils nur die möglichen Realisierungen $\{a_1, \ldots, a_k\}$ und $\{b_1, \ldots, b_\ell\}$ haben.
Unter der Annahme, dass diese n Paare unabhängig sind, kann man das Testproblem

- H_0 : A und B sind unabhängig,
- H_1 : A und B sind abhängig

betrachten. Bezeichne dazu

$$P(A = a_i, B = b_j) = \pi_{ij}, \qquad P(A = a_i) = \pi_{i+}, \quad P(B = b_j) = \pi_{+j},$$

so gilt für stochastisch unabhängige Variablen A und B

$$P(A = a_i, B = b_j) = P(A = a_i)P(B = b_j)\,,$$

für $i = 1, \ldots, k$ und $j = 1, \ldots \ell$, wodurch sich das obige Testproblem auch alternativ formulieren lässt durch

- H_0^* : $\pi_{ij} = \pi_{i+}\pi_{+j}$ (für alle Paare (i, j)),
- H_1^* : $\pi_{ij} \neq \pi_{i+}p_{+j}$ (für zumindest ein Paar (i, j)).

Die am weitesten verbreitete Statistik zum Testen dieser Hypothese ist definiert durch

$$X^2 = \sum_{i=1}^{k}\sum_{j=1}^{\ell} \frac{(N_{ij} - \hat{\mathrm{E}}(N_{ij}))^2}{\hat{\mathrm{E}}(N_{ij})}.$$

Der Berechnung dieser Teststatistik X^2 liegt die folgende Kontingenztafel zugrunde:

$A\backslash B$	b_1	$\ldots$	b_j	$\ldots$	b_ℓ	n_{i+}
a_1	n_{11}	$\ldots$	n_{1j}	$\ldots$	$n_{1\ell}$	n_{1+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_i	n_{i1}	$\ldots$	n_{ij}	$\ldots$	$n_{i\ell}$	n_{i+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_k	n_{k1}	$\ldots$	n_{kj}	$\ldots$	$n_{k\ell}$	n_{k+}
n_{+j}	n_{+1}	$\ldots$	n_{+j}	$\ldots$	$n_{+\ell}$	n

Hierbei bezeichnet n_{ij} die Häufigkeit des Paares $(A = A_i, B = B_j)$. Entsprechend steht $n_{i+} = \sum_j n_{ij}$ für die Anzahl von Stichprobenpaaren mit $A = A_i$, und $n_{+j} = \sum_i n_{ij}$ für die Anzahl mit $B = B_j$. Natürlich gilt $\sum_i n_{i+} = \sum_j n_{+j} = n$. Die Statistik X^2 nennt man häufig auch *Pearson-Statistik*. Diese basiert auf den quadrierten Abständen zwischen den beobachteten und erwarteten Zellhäufigkeiten.
Die Erwartungswerte der Zellhäufigkeiten werden unter der Annahme der Unabhängigkeit der beiden Faktoren berechnet. Dazu nehmen wir an, dass die Kontingenztafel

$$N = \begin{pmatrix} N_{11} & \dots & N_{1\ell} \\ \vdots & & \vdots \\ N_{k1} & \dots & N_{k\ell} \end{pmatrix} \quad \text{mit} \quad \mathrm{E}(N) = n \begin{pmatrix} \pi_{11} & \dots & \pi_{1\ell} \\ \vdots & & \vdots \\ \pi_{k1} & \dots & \pi_{k\ell} \end{pmatrix}$$

eine Zufallsmatrix ist, deren Elemente N_{ij} die Anzahl jener Beobachtungen angibt, die in die Klasse (a_i, b_j) fallen. Generell ist N multinomial-verteilt, d.h.

$$P(N_{11} = n_{11}, \dots, N_{k\ell} = n_{k\ell}) = \frac{n!}{n_{11}! \cdot \ldots \cdot n_{k\ell}!} \pi_{11}^{n_{11}} \cdot \ldots \cdot \pi_{k\ell}^{n_{k\ell}}$$

mit $k\ell$ unbekannten Parametern π_{ij}, wovon jedoch wegen $\sum_i \sum_j \pi_{ij} = 1$ nur $k\ell - 1$ frei wählbar sind.
Für dieses Wahrscheinlichkeitsmass folgt unter H_0 die Vereinfachung

$$P(N_{11} = n_{11}, \dots, N_{k\ell} = n_{k\ell}) = \frac{n!}{n_{11}! \cdot \ldots \cdot n_{k\ell}!} (\pi_{1+}\pi_{+1})^{n_{11}} \cdot \ldots \cdot (\pi_{k+}\pi_{+\ell})^{n_{k\ell}} .$$

Wegen $\sum_i \pi_{i+} = \sum_j \pi_{+j} = 1$ müssen jetzt nur noch die $k + \ell - 2$ unbekannten Parameter π_{i+} und π_{+j} geschätzt werden. Als relevanter Teil der Likelihood Funktion resultiert dafür unter H_0

$$\begin{aligned} L(\pi|N) &= L(\pi_{1+}, \dots, \pi_{k+}, \pi_{+1}, \dots, \pi_{+\ell}|N) \\ &\propto \prod_{i=1}^{k} \prod_{j=1}^{\ell} (\pi_{i+}\pi_{+j})^{n_{ij}} \\ &= \prod_{i=1}^{k} \pi_{i+}^{\sum_j n_{ij}} \prod_{j=1}^{\ell} \pi_{+j}^{\sum_i n_{ij}} \\ &= \prod_{i=1}^{k} \pi_{i+}^{n_{i+}} \prod_{j=1}^{\ell} \pi_{+j}^{n_{+j}} . \end{aligned}$$

Erinnern wir uns, dass beispielsweise $\pi_{k+} = 1 - \sum_{i=1}^{k-1} \pi_{i+}$ gilt, so folgt damit

$$L(\pi|N) = \left(1 - \sum_{i=1}^{k-1} \pi_{i+}\right)^{n_{k+}} \prod_{i=1}^{k-1} \pi_{i+}^{n_{i+}} \prod_{j=1}^{\ell} \pi_{+j}^{n_{+j}} ,$$

beziehungsweise

$$\log L(\pi|N) = n_{k+} \log \left(1 - \sum_{i=1}^{k-1} \pi_{i+}\right) + \sum_{i=1}^{k-1} n_{i+} \log \pi_{i+} + \sum_{j=1}^{\ell} n_{+j} \log \pi_{+j} .$$

Setzt man die partiellen Ableitungen (das sind die π_{i+} Scores)

$$\frac{\partial \log L}{\partial \pi_{i+}} = -\frac{n_{k+}}{\pi_{k+}} + \frac{n_{i+}}{\pi_{i+}}$$

gleich Null, so resultieren die Maximum-Likelihood Schätzer der Randwahrscheinlichkeiten

$$\hat{\pi}_{i+} = \hat{\pi}_{k+}\frac{N_{i+}}{N_{k+}} \quad \forall i = 1, \ldots, k-1\,.$$

Da deren Summe wiederum Eins ergeben muss, d.h. $1 = \sum_i \hat{\pi}_{i+} = \frac{\hat{\pi}_{k+}}{n_{k+}} \sum_i n_{i+} = \frac{n\hat{\pi}_{k+}}{n_{k+}}$ gilt, folgt für den Maximum-Likelihood Schätzer

$$\hat{\pi}_{i+} = \frac{N_{i+}}{n}\,.$$

Analog geht man für π_{+j} vor und erhält

$$\hat{\pi}_{+j} = \frac{N_{+j}}{n}\,.$$

Damit wird nun der Schätzer für $\mathrm{E}(N_{ij}) = n\pi_{i+}\pi_{+j}$ konstruiert und wir erhalten

$$\hat{\mathrm{E}}(N_{ij}) = n\hat{\pi}_{i+}\hat{\pi}_{+j} = \frac{1}{n}N_{i+}N_{+j}\,.$$

Somit handelt es sich bei X^2 um einen modifizierten χ^2-Test auf Anpassung mit $r = (k-1)+(\ell-1)$ Maximum-Likelihood geschätzten Parametern π_{i+} und π_{+j} und es gilt

$$X^2 = \sum_{i=1}^{k}\sum_{j=1}^{\ell}\frac{(N_{ij} - n\hat{\pi}_{i+}\hat{\pi}_{+j})^2}{n\hat{\pi}_{i+}\hat{\pi}_{+j}} \overset{as}{\sim} \chi^2_{(k-1)(\ell-1)}\,.$$

Wir verwerfen H_0, falls $X^2 \geq \chi^2_{1-\alpha;(k-1)(\ell-1)}$ gilt.
Testen wir damit auf die Unabhängigkeit zweier Faktoren mit sehr vielen Stufen, dann muss auf die Güte der Approximation geachtet werden. Diese ist gut, falls die folgenden Faustregeln erfüllt sind

- nach Cochran (1954)
  - kein $\hat{\mathrm{E}}(N_{ij}) < 1$ und
  - für maximal 20% der Felder gilt: $\hat{\mathrm{E}}(N_{ij}) < 5$;
- nach Conover (1971)
  - fast alle $\hat{\mathrm{E}}(N_{ij})$ von derselben Größenordnung sind,
  - alle $\hat{\mathrm{E}}(N_{ij}) > 1$,
  - die Anzahl der Klassen klein ist.

Beispiel 4.13 *(Agresti, p. 80) In einer Studie wurde die religiöse Einstellung (Fundamentalismus) einer Person mit deren höchsten Ausbildungsgrad gegenübergestellt. Dazu wurden $n = 2726$ zufällig ausgewählte Person befragt.*

	Religiöser Glaube			
Ausbildung	*Fundamentalist*	*gemäßigt*	*liberal*	*Total*
Weniger als high school	*178*	*138*	*108*	*424*
High school oder junior college	*570*	*648*	*442*	*1660*
Bachelor oder Graduate	*138*	*252*	*252*	*642*
Total	*886*	*1038*	*802*	*2726*

```
> n <- matrix(c(178, 138, 108, 570, 648, 442, 138, 252, 252), 3, 3, byrow=TRUE)
> n
     [,1] [,2] [,3]
 [1,]  178  138  108
 [2,]  570  648  442
 [3,]  138  252  252
> chisq.test(n)

        Pearson's Chi-squared test

 data:  n
 X-squared = 69.1568, df = 4, p-value = 3.42e-14
```

Es gibt also ein extrem starkes Anzeichen, dass hierbei eine Assoziation vorliegt.
Die Funktion `chisq.test` *liefert darüberhinaus auch noch die geschätzten Erwartungswerte $\hat{E}(N_{ij})$ sowie die Residuen $(N_{ij} - \hat{E}(N_{ij}))/\sqrt{\hat{E}(N_{ij})}$.*

```
> chisq.test(n)$expected
          [,1]     [,2]     [,3]
 [1,] 137.8078 161.4497 124.7425
 [2,] 539.5304 632.0910 488.3786
 [3,] 208.6618 244.4593 188.8789

> chisq.test(n)$residuals
           [,1]       [,2]      [,3]
 [1,]  3.423775 -1.8455228 -1.499038
 [2,]  1.311771  0.6327815 -2.098646
 [3,] -4.891737  0.4822914  4.592852
```

Man sieht, dass die Zelle Bachelor oder Graduate / Fundamentalist den größten Beitrag zur Teststatistik ausmacht. Dieser ist mit $-4.89^2 = 23.93$ etwa ein Drittel vom gesamten X^2. Unter der Null-Hypothese wird für diese Zelle eine bedeutend größere Anzahl erwartet ($\hat{E}(N_{ij}) = 208.66$) als dies beobachtet werden konnte ($n_{31} = 138$). Interessanterweise waren in dieser Gruppe viel mehr liberale als erwartet. Ohne Bachelors oder Graduierte erhält man

```
> n <- matrix(c(178,138,108,570,648,442), 2, 3, byrow=TRUE)
```

```
> chisq.test(n)

        Pearson's Chi-squared test

data:  n
X-squared = 9.439, df = 2, p-value = 0.00892

> chisq.test(n)$residuals
          [,1]       [,2]       [,3]
[1,]  2.092665 -1.7330336 -0.3686978
[2,] -1.057617  0.8758623  0.1863372
```

Dies ist wiederum ein starker Hinweis dafür, dass für diesen Personenkreis der Ausbildungsgrad nicht unabhängig vom Fundamentalismus zu sein scheint. Die Residuen sind jetzt auch viel gleichmässiger als zuvor.

4.4.2 Der exakte Test von Fisher

Für den Spezialfall einer 2×2-Tafel (Vierfelder-Tafel) ist für kleine Stichproben Fisher's exakter Test anwendbar. Dies steht somit im Gegensatz zum χ^2-Test, bei dem die Verteilung der Teststatistik gerade für große Stichproben durch die χ^2-Verteilung approximiert werden kann.
Zuvor wurde bereits gezeigt, dass die Verteilung einer 2×2 Kontingenztafel unter H_0 (die beiden Faktoren A und B sind unabhängig) geschrieben werden kann als

$$P(N_{11} = n_{11}, \cdots, N_{22} = n_{22}) = \frac{n!}{n_{11}! \cdot \ldots \cdot n_{22}!} (\pi_{1+}\pi_{+1})^{n_{11}} \cdot \ldots \cdot (\pi_{2+}\pi_{+2})^{n_{22}} .$$

Das bedeutet, dass die Verteilung von N nur von den marginalen Häufigkeiten abhängt, welche somit eine suffiziente Statistik für die Parameter darstellen. Somit ist die konditionale Verteilung von N gegeben diese suffiziente Statistik von den Parametern unabhängig. Um dies zu erreichen, müssen wir also die beiden Ränder festhalten.
Wir ermitteln also alle Tafeln mit der gleichen Randhäufigkeit wie die Stichprobentafel. Andere mögliche Tafeln mit gleicher Randhäufigkeiten sind ($0 \leq x \leq \min(a+b, a+c)$:

	b_1	b_2	n_{i+}
a_1	a	b	$a+b$
a_2	c	d	$c+d$
n_{+j}	$a+c$	$b+d$	n

	b_1	b_2	n_{i+}
a_1	x	$a+b-x$	$a+b$
a_2	$a+c-x$	$d-a+x$	$c+d$
n_{+j}	$a+c$	$b+d$	n

Welche Auftrittswahrscheinlichkeit hat die Stichprobentafel unter der Unabhängigkeitshypothese? Diese ergibt sich durch das hypergeometrische Modell. In einer Urne mit n Kugeln befinden sich $a+c$ rote und $b+d$ blaue Kugeln. Es werden $a+b$ Kugeln ohne Zurücklegen gezogen. X ist die Anzahl der gezogenen roten Kugeln mit

$$P(X = x) = \frac{\binom{a+c}{x}\binom{b+d}{a+b-x}}{\binom{n}{a+b}} \quad \text{mit} \quad \mathrm{E}(X) = \frac{(a+b)(a+c)}{n} .$$

Die Zufallsvariable X wird beim Fisher-Test als Teststatistik verwendet. Wir lehnen H_0 ab, falls $x \leq c_{\alpha/2}$ oder $x \geq c_{1-\alpha/2}$ ist. Dabei bezeichnet c_α das α-Quantil der Hypergeometrisch$(n, a+c, a+b)$-Verteilung. Diese ist genau dann symmetrisch, wenn $(a+c)/n = 1/2$ oder $(a+b)/n = 1/2$ gilt. Die Wahrscheinlichkeitsverteilung von X kann mit der folgenden Rekursionsbeziehung bestimmt werden:

$$\begin{aligned} P(X=0) &= \frac{(b+d)!\,(c+d)!}{(d-a)!\,n!} \\ P(X=x+1) &= P(X=x)\frac{(a+b-x)(a+c-x)}{(x+1)(d-a+x+1)}\,. \end{aligned}$$

Beispiel 4.14 *(Fisher, 1935, Agresti, p.92) In diesem originalen Beispiel (Fisher's Tea Drinker) beschreibt Sir R. A. Fisher seine Tage in der Rothamsted Experiment Station. Eine Kollegin behauptete dort unterscheiden zu können, ob zuerst Tee oder Milch in die Tasse gegeben wurde. Sie testete acht Tassen, in die jeweils viermal Milch und Tee zuerst gegeben wurde, und kam dabei auf folgendes Ergebnis:*

	Vermutung		
Zuerst in der Tasse	*Milch*	*Tee*	*Total*
Milch	*3*	*1*	*4*
Tee	*1*	*3*	*4*
Total	*4*	*4*	*8*

```
> n <- matrix(c(3, 1, 1, 3), 2, 2, byrow=TRUE)
> fisher.test(n)

        Fisher's Exact Test for Count Data

 data:  n
 p-value = 0.4857
 alternative hypothesis: true odds ratio is not equal to 1
 95 percent confidence interval:
   0.2117329 621.9337505
 sample estimates:
 odds ratio
  6.408309

> chisq.test(n)

        Pearson's Chi-squared test with Yates' continuity correction

 data:  n
 X-squared = 0.5, df = 1, p-value = 0.4795
```

Obwohl der Ausgang des Experiments beeindruckte, war seine Meinung über seine Kollegin somit bekräftigt. Man bemerke, dass dafür nur bei perfekter Übereinstimmung ($x = 4$) die Hypothese H_0 mit einem p-Wert von 0.029 verworfen werden kann. Andererseits sollte ihn jedoch die Kollegin laut Fisher's Tochter doch von ihrer Fähigkeit überzeugt haben.

Kapitel 5

Lineare Regression

In der Korrelationsanalyse erhält man ein quantitatives Maß für den Grad des linearen Zusammenhangs der beiden Zufallsvariablen (X, Y). In der Regressionsanalyse hingegen wird ein funktionaler (meist linearer) Zusammenhang spezifiziert und es wird angenommen, dass an gegebenen Stellen x die Zufallsvariable Y eine von x abhängige Verteilung (Erwartung) hat. Es wird also das Paar (x, Y) betrachtet, wobei x als gewöhnliche Variable (**keine** Zufallsvariable) und Y als eine von x abhängige Zufallsvariable aufgefasst wird. Diese Abhängigkeit der Zufallsvariablen Y von x nennt man **Regression**. Häufig bezeichnet man dabei Y als interessierende (abhängige) Variable oder **Responsevariable**, während x erklärende (unabhängige) Variable oder **Prädiktorvariable** genannt wird. Regressionsmodelle sind relevant um zu verstehen, wie sich der Erwartungswert der Response Y mit x ändert.

5.1 Einfache lineare Regression

Bei der einfachen linearen Regression wird angenommen, dass es in der Population Y eine Regressionsgerade gibt mit

$$\mathrm{E}(Y) = \mu(x) = \beta_0 + \beta_1 x\,.$$

Das Modell ist linear in den Parametern β_0 und β_1. Die Varianz von Y sei konstant, d.h. unabhängig von x.
Eine Stichprobe vom Umfang n unter diesem Modell kommt zustande, indem man n Werte $x_1, \ldots, x_n$ wählt und dort unabhängige Zufallsvariablen Y_i, $i = 1, \ldots, n$, beobachtet. Die Stichprobenversion des Regressionsmodells kann auch geschrieben werden als

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i\,, \quad i = 1, \ldots, n\,.$$

Dabei bezeichnen die Zufallsvariablen ϵ_i die **statistischen Fehler**, die nicht beobachtbar sind und für die $\mathrm{E}(\epsilon_i) = 0$, $\mathrm{var}(\epsilon_i) = \sigma^2$ (unbekannt und konstant), $\mathrm{cov}(\epsilon_i, \epsilon_j) = 0$ für $i \neq j$, gilt.
Unter der zusätzlichen Annahme $Y_i \overset{ind}{\sim} N\left(\mu(x_i), \sigma^2\right)$ folgt für den statistischen Fehler

$$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)\,.$$

Diese Situation ist auch in der Abbildung 5.1 schematisch dargestellt.

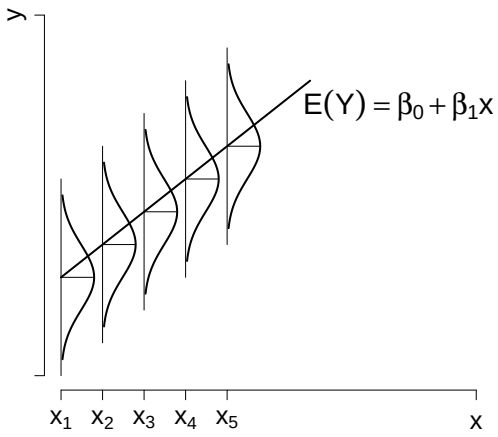


Abbildung 5.1: Verteilungsannahmen beim einfachen linearen Regressionsmodell.

5.1.1 Schätzen der Parameter

Liegt die Stichprobe $(x_1, Y_1), \ldots, (x_n, Y_n)$ vor, so können die Regressionsparameter (β_0, β_1) durch die Methode der **Kleinsten-Quadrate** (least squares) geschätzt werden. Dabei wird die Fehlerquadratsumme (sum of squared errors)

$$\mathrm{SSE}(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \mu(x_i))^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2 \,,$$

also die Quadratsumme aller vertikalen Abweichungen zwischen den beobachteten Werten (x_i, Y_i) und den erwarteten Werten auf der Regressionsgeraden $(x_i, \mu(x_i))$ in (β_0, β_1) minimiert. Als erste partielle Ableitungen erhält man die sogenannten **Normalgleichungen**

$$\begin{aligned} \frac{\partial}{\partial \beta_0} \mathrm{SSE}(\beta_0, \beta_1) &= -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i) \,, \\ \frac{\partial}{\partial \beta_1} \mathrm{SSE}(\beta_0, \beta_1) &= -2 \sum_{i=1}^{n} x_i (Y_i - \beta_0 - \beta_1 x_i) \,. \end{aligned}$$

Setzt man die beiden Gleichungen Null, so ergibt sich für die Lösung $(\hat{\beta}_0, \hat{\beta}_1)$

$$\begin{aligned} \hat{\beta}_0 + \hat{\beta}_1 \overline{x} &= \overline{Y} \\ \hat{\beta}_0 n \overline{x} + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i Y_i \,. \end{aligned}$$

Als Kleinste-Quadrate Schätzer (LSE) resultieren

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i Y_i - n\overline{x}\,\overline{Y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{s_{xY}^2}{s_x^2}\,.$$

Die minimale Fehlerquadratsumme ist

$$\begin{aligned} \mathrm{SSE}(\hat{\beta}_0, \hat{\beta}_1) &= \sum_{i=1}^{n} \left(Y_i - \overline{Y} - \hat{\beta}_1(x_i - \overline{x})\right)^2 \\ &= (n-1)(s_Y^2 - \hat{\beta}_1^2 s_x^2) \\ &= (n-1)(s_Y^2 - s_{xY}^4/s_x^2)\,. \end{aligned}$$

Alle Punkte (x_i, y_i) liegen genau dann auf der Geraden $\mu(x)$, wenn $\mathrm{SSE}(\hat{\beta}_0, \hat{\beta}_1) = 0$ und somit $s_{xY}^4 = s_x^2 s_Y^2$ gilt, also wenn

$$\left(\frac{s_{xY}^2}{s_x s_Y}\right)^2 = \widehat{\mathrm{cor}}^2(x, Y) = 1$$

und somit perfekte Korrelation zwischen den x Werten und der Response Y vorliegt.
Nachdem $\hat{\beta}_0$ und $\hat{\beta}_1$ gefunden sind ist es möglich, den **Prognose**- oder **Vorhersagewert** (den Schätzer für den Erwartungswert unter dem Modell) anzugeben. Für diesen ergibt sich

$$\hat{\mu}(x) = \hat{\beta}_0 + \hat{\beta}_1 x = \overline{Y} + \hat{\beta}_1(x - \overline{x})\,.$$

Die Differenz zwischen einer Beobachtung und dem geschätzten Erwartungswert

$$r = Y - \hat{\mu}(x)$$

nennt man **Residuum** oder beobachtbaren Fehler. Damit lässt sich die minimale Fehlerquadratsumme darstellen als

$$\mathrm{SSE}(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n}(Y_i - \hat{\mu}(x_i))^2 = \sum_{i=1}^{n} r_i^2\,.$$

Weiters folgt aus der ersten Normalgleichung die Identität

$$\sum_{i=1}^{n}(Y_i - \hat{\mu}(x_i)) = \sum_{i=1}^{n} r_i = 0\,.$$

Die Summe der Residuen ist somit immer Null und es gilt, dass die Summe der Beobachtungen der Summe der Vorhersagen entspricht.
Die Minimierung von $\mathrm{SSE}(\beta_0, \beta_1)$ liefert jedoch keinen Schätzer für die Varianz σ^2. Für gewöhnlich wird deshalb zusätzlich angenommen, dass Y normalverteilt ist, d.h.

$$Y_i \overset{ind}{\sim} N(\mu(x_i), \sigma^2)\,.$$

Diese Annahme erlaubt die Herleitung der Maximum-Likelihood Schätzer für die beiden Regressionsparameter sowie für den Varianzparameter. Unter diesem Modell folgt als Log-Likelihood Funktion der Stichprobe

$$\log L(\beta_0, \beta_1, \sigma^2|y) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

und damit das Score-System

$$\begin{aligned}\frac{\partial}{\partial\beta_0}\log L(\beta_0, \beta_1, \sigma^2|y) &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)\\ \frac{\partial}{\partial\beta_1}\log L(\beta_0, \beta_1, \sigma^2|y) &= \frac{1}{\sigma^2}\sum_{i=1}^{n}x_i(y_i - \beta_0 - \beta_1 x_i)\,.\end{aligned}$$

Diese beiden Gleichungen entsprechen gerade den Normalgleichungen zu den Kleinsten-Quadrate Schätzungen. Daher sind die gefundenen LSE's identisch den MLE's unter der Annahme normalverteilter Responses.
Als zusätzliche Scoregleichung für den Varianzparameter σ^2 erhält man

$$\frac{\partial}{\partial\sigma^2}\log L(\beta_0, \beta_1, \sigma^2|y) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

mit Lösung

$$\hat\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat\beta_0 - \hat\beta_1 x_i)^2 = \frac{1}{n}\mathrm{SSE}(\hat\beta_0, \hat\beta_1)\,.$$

Dieser MLE für den Varianzparameter unter dem Regressionsmodell entspricht von der Idee der Konstruktion des MLE's bei Vorliegen einer Zufallsstichprobe. Der Unterschied liegt dabei nur in der Schätzung der Erwartungswerte.

5.1.2 Verteilungseigenschaften der Schätzer

Wir zeigen zuerst, dass $\hat\beta_0$ und $\hat\beta_1$ Linearkombinationen der Responses $(Y_1, \ldots, Y_n)$ sind. Dazu schreiben wir die Schätzer um zu

$$\begin{aligned}\hat\beta_1 &= \frac{1}{(n-1)s_x^2}\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y}) = \sum_{i=1}^{n}\frac{x_i - \overline{x}}{(n-1)s_x^2}Y_i = \sum_{i=1}^{n}a_i Y_i\\ \hat\beta_0 &= \frac{1}{n}\sum_{i=1}^{n}Y_i - \overline{x}\sum_{i=1}^{n}a_i Y_i = \sum_{i=1}^{n}\left(\frac{1}{n} - a_i\overline{x}\right)Y_i = \sum_{i=1}^{n}b_i Y_i\end{aligned}$$

mit den Konstanten

$$a_i = \frac{x_i - \overline{x}}{(n-1)s_x^2}\,, \qquad b_i = \frac{1}{n} - \overline{x}\frac{x_i - \overline{x}}{(n-1)s_x^2}\,.$$

Nun sind Linearkombinationen normalverteilter Variablen selbst normalverteilt.

Die Verwendung dieser Schreibweisen erlaubt auch eine einfache Herleitung der Momente beider Schätzer. Mit den Ergebnissen

$$\begin{aligned}
\sum_{i=1}^n a_i &= \frac{1}{(n-1)s_x^2}\sum_{i=1}^n (x_i-\overline{x}) = 0\\
\sum_{i=1}^n a_i x_i &= \frac{1}{(n-1)s_x^2}\sum_{i=1}^n (x_i-\overline{x})x_i = 1\\
\sum_{i=1}^n a_i^2 &= \frac{1}{(n-1)^2 s_x^4}\sum_{i=1}^n (x_i-\overline{x})^2 = \frac{1}{(n-1)s_x^2}\\
\sum_{i=1}^n b_i &= 1 - \frac{\overline{x}}{(n-1)s_x^2}\sum_{i=1}^n (x_i-\overline{x}) = 1\\
\sum_{i=1}^n b_i x_i &= \overline{x} - \frac{\overline{x}}{(n-1)s_x^2}\sum_{i=1}^n (x_i-\overline{x})x_i = 0\\
\sum_{i=1}^n b_i^2 &= \frac{1}{n} + \overline{x}^2\sum_{i=1}^n \frac{(x_i-\overline{x})^2}{(n-1)^2 s_x^4} - 2\cdot 0 = \frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}
\end{aligned}$$

folgt

$$\begin{aligned}
\mathrm{E}(\hat{\beta}_1) &= \sum_{i=1}^n a_i \mathrm{E}(Y_i) = \sum_{i=1}^n a_i(\beta_0+\beta_1 x_i) = \beta_1\\
\mathrm{E}(\hat{\beta}_0) &= \sum_{i=1}^n b_i \mathrm{E}(Y_i) = \sum_{i=1}^n b_i(\beta_0+\beta_1 x_i) = \beta_0\,,
\end{aligned}$$

sowie

$$\begin{aligned}
\mathrm{var}(\hat{\beta}_1) &= \sum_{i=1}^n a_i^2 \mathrm{var}(Y_i) = \frac{\sigma^2}{(n-1)s_x^2} = \frac{\sigma^2}{\sum\limits_{i=1}^n (x_i-\overline{x})^2}\\
\mathrm{var}(\hat{\beta}_0) &= \sum_{i=1}^n b_i^2 \mathrm{var}(Y_i) = \sigma^2\frac{1}{n} + \sigma^2\frac{\overline{x}^2}{\sum\limits_{i=1}^n (x_i-\overline{x})^2}\,.
\end{aligned}$$

Für zwei beliebige lineare Formen $a^t y$ und $b^t y$ in $y \sim N(\mu, \sigma^2 I_n)$ gilt

$$\begin{aligned}
\mathrm{cov}(a^t y, b^t y) &= \mathrm{E}\left(a^t(y-\mu)b^t(y-\mu)\right) = \mathrm{E}\left(a^t(y-\mu)(y-\mu)^t b\right)\\
&= \sigma^2 a^t b\,.
\end{aligned}$$

Somit sind $a^t y$ und $b^t y$ genau dann unabhängig wenn

$$a^t b = 0\,.$$

Betrachte $\overline{y}$ und $\hat{\beta}_1$. Beides sind lineare Formen und mit

$$\begin{aligned} a &= \frac{1}{n}(1,\ldots,1)^t \\ b &= \frac{1}{\sum_{i=1}^n (x_i-\overline{x})^2}(x_1-\overline{x},\ldots,x_n-\overline{x})^t \end{aligned}$$

schreiben wir

$$\overline{y} = a^t y\,, \qquad \hat{\beta}_1 = b^t y\,.$$

Wegen

$$\begin{aligned} a^t b &= \frac{1}{n}(1,\ldots,1)\frac{1}{\sum_{i=1}^n (x_i-\overline{x})^2}\begin{pmatrix} x_1-\overline{x} \\ \vdots \\ x_n-\overline{x}\end{pmatrix} \\ &= \frac{1}{n\sum_{i=1}^n (x_i-\overline{x})^2}\sum_{i=1}^n (x_i-\overline{x}) = 0 \end{aligned}$$

folgt die Unabhängigkeit des Responsemittels und des Steigungsschätzers. Bemerke jedoch, dass mit $\hat{\beta}_0$ keine Unabhängigkeit besteht.

5.1.3 Quadratsummen-Zerlegung

Eine wesentliche Rolle in der Regressionsanalyse spielt die Zerlegung der sogenannten **totalen Quadratsumme**

$$\mathrm{SST} = \sum_{i=1}^n (Y_i-\overline{Y})^2\,.$$

Aus dem rechten Teil der Abbildung 5.2 erkennt man, dass die Summe der unquadrierten Differenzen linear zerlegt werden kann in

$$\sum_{i=1}^n (Y_i-\overline{Y}) = \sum_{i=1}^n (\hat{\mu}_i-\overline{Y}) + \sum_{i=1}^n (Y_i-\hat{\mu}_i)\,.$$

Interessanterweise hält diese Zerlegung auch im quadratischen Sinne, denn

$$\begin{aligned} \sum_{i=1}^n (Y_i-\overline{Y})^2 &= \sum_{i=1}^n \Big((Y_i-\hat{\mu}_i)+(\hat{\mu}_i-\overline{Y})\Big)^2 \\ &= \sum_{i=1}^n (\hat{\mu}_i-\overline{Y})^2 + \sum_{i=1}^n (Y_i-\hat{\mu}_i)^2\,. \end{aligned}$$

Die Summe der gemischten Terme verschwindet, weil

$$\begin{aligned} \sum_{i=1}^n (Y_i-\hat{\mu}_i)(\hat{\mu}_i-\overline{Y}) &= \sum_{i=1}^n (Y_i-\hat{\mu}_i)(\hat{\beta}_0+\hat{\beta}_1 x_i-\overline{Y}) \\ &= \hat{\beta}_0\sum_{i=1}^n (Y_i-\hat{\mu}_i) + \hat{\beta}_1\sum_{i=1}^n x_i(Y_i-\hat{\mu}_i) - \overline{Y}\sum_{i=1}^n (Y_i-\hat{\mu}_i)\,. \end{aligned}$$

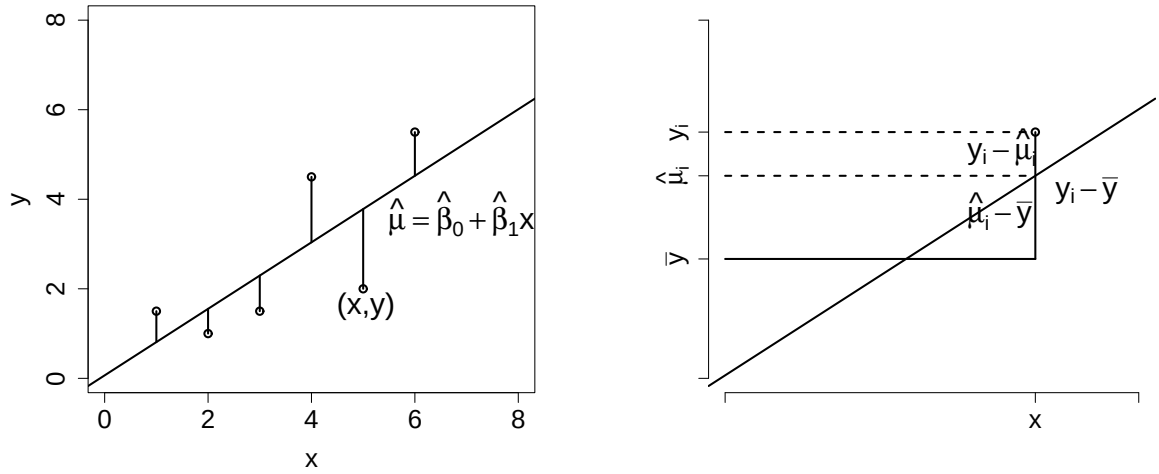


Abbildung 5.2: Stichprobensituation bei der einfachen linearen Regression (links), und zur Quadratsummenzerlegung (rechts).

Die erste und dritte Summe verschwindet, da die Summe der Residuen Null ist. Die zweite Summe definiert den Score und ist daher an der Stelle des Schätzers Null.
Die Zerlegung liefert also

$$\begin{aligned} \sum_{i=1}^{n}(Y_i - \overline{Y})^2 &= \sum_{i=1}^{n}(\hat{\mu}_i - \overline{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{\mu}_i)^2 \\ \text{SST} &= \text{SSR}(\hat{\beta}_0, \hat{\beta}_1) + \text{SSE}(\hat{\beta}_0, \hat{\beta}_1)\,. \end{aligned}$$

Hier ist zu bemerken, dass die totale Quadratsumme (SST) gar nicht vom Regressionsmodell abhängt. Die **Regressions-Quadratsumme**

$$\text{SSR}(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n}(\hat{\mu}_i - \overline{Y})^2$$

beschreibt den Unterschied zwischen den n geschätzten Erwartungswerten unter Annahme eines Regressionsmodells, $\hat{\mu}_i$, und der einen geschätzten Erwartung bei Annahme einer Zufallsstichprobe, $\overline{Y}$. Je größer diese Quadratsumme ist, desto wesentlicher ist das Regressionsmodell. Die Summe

$$\text{SSE}(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n}(Y_i - \hat{\mu}_i)^2$$

ist die bereits definierte **Fehler-Quadratsumme**. Sie ist minimal in den LSE.
Diese Partitionierung kann auch als Variationszerlegung interpretiert werden

$$\begin{pmatrix} \text{totale Variabilität} \\ \text{in } Y \end{pmatrix} = \begin{pmatrix} \text{durch das Modell} \\ \text{erklärte Variabilität} \end{pmatrix} + \begin{pmatrix} \text{durch das Modell } \mathbf{nicht} \\ \text{erklärte Variabilität} \end{pmatrix}.$$

Da das arithmetische Mittel aller $\hat{\mu}_i$ dem Mittelwert $\overline{y}$ entspricht, beschreibt SSR die Variation in den angepassten Werten. Da die Summe der Residuen Null ist, beschreiben die beiden Variationskomponenten die Variation bezüglich der Regressionsgeraden und die Variation um die Regressionsgerade.

5.1.4 Bestimmtheitsmass

Natürlich sollte $\mathrm{SSE}(\hat{\beta}_0, \hat{\beta}_1)$ im Vergleich zu $\mathrm{SSR}(\hat{\beta}_0, \hat{\beta}_1)$ möglichst klein sein. Die Schätzer $(\hat{\beta}_0, \hat{\beta}_1)$ minimieren die Fehler-Quadratsumme unter dem Regressionsmodell. Die Größe SST hingegen hängt nur von den Responses Y_i ab. In dieser Quadratsumme ist keine Information über das Modell enthalten. Somit haben wir

$$\underbrace{\mathrm{SST}}_{(fest)} = \underbrace{\mathrm{SSR}(\hat{\beta}_0, \hat{\beta}_1)}_{(maximal)} + \underbrace{\mathrm{SSE}(\hat{\beta}_0, \hat{\beta}_1)}_{(minimal)} .$$

Zur Beurteilung der Güte der Anpassung wird häufig das Bestimmtheitsmass

$$R^2 = \frac{\mathrm{SSR}(\hat{\beta}_0, \hat{\beta}_1)}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}(\hat{\beta}_0, \hat{\beta}_1)}{\mathrm{SST}}, \qquad 0 \leq R^2 \leq 1$$

verwendet. R^2 gibt den relativen Varianzanteil an, der durch das Modell erklärt wird. Für das einfache lineare Modell gilt

$$\begin{aligned} R^2 &= \frac{\sum(\hat{\mu}_i - \overline{Y})^2}{\sum(Y_i - \overline{Y})^2} = \frac{\sum(\hat{\beta}_1(x_i - \overline{x}))^2}{\sum(Y_i - \overline{Y})^2} \\ &= \hat{\beta}_1^2 \frac{s_x^2}{s_Y^2} = \frac{s_{xY}^4}{s_x^4} \frac{s_x^2}{s_Y^2} \\ &= \left(\frac{s_{xY}^2}{s_x s_Y} \right)^2 = \widehat{\mathrm{cor}}^2(x, Y) \end{aligned}$$

und R^2 stimmt mit dem Quadrat des empirischen Korrelationskoeffizienten zwischen Y_i und x_i überein.

5.2 Multiple lineare Regression

Nun betrachten wir ein lineares Regressionsmodell mit mehreren erklärenden Variablen. Das multiple lineare Modell lautet somit

$$y = X\beta + \epsilon$$

mit dem Responsevektor $y = (y_1, \ldots, y_n)^t$, dem Parametervektor $\beta = (\beta_0, \ldots, \beta_{p-1})^t$ und dem Fehlervektor $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^t$, sowie der Designmatrix

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \ldots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{n,p-1} \end{pmatrix} .$$

Die erste Spalte in der Designmatrix X ist mit 1 belegt, um für ein Modell mit Intercept (Modellkonstante) verwendbar zu sein. Wie schon zuvor nehmen wir auch jetzt wiederum an, dass

$$y \sim N(X\beta, \sigma^2 I_n)$$

oder äquivalent dazu, dass

$$\epsilon \sim N(0, \sigma^2 I_n)\,.$$

5.2.1 Schätzen der linearen Parameter

Um den MLE $\hat{\beta}$ zu finden muss wiederum

$$\begin{aligned} \mathrm{SSE}(\beta) &= (y - X\beta)^t (y - X\beta) \\ &= y^t y - 2\beta^t X^t y + \beta^t X^t X \beta \end{aligned}$$

bzgl. β minimiert werden. Als Scorevektor erhalten wir

$$\frac{\partial}{\partial \beta} \mathrm{SSE}(\beta) = -2X^t y + 2X^t X\beta\,.$$

Das Minimum ist somit definiert durch

$$X^t X\hat{\beta} = X^t y\,.$$

Falls $X^t X$ regulär (von vollem Rang p), dann existiert die Inverse und wir erhalten als MLE

$$\hat{\beta} = (X^t X)^{-1} X^t y\,.$$

Ist dies überhaupt ein Minimum (ist die Matrix der zweiten Ableitung positiv semidefinit)?

$$\frac{\partial^2}{\partial \beta \partial \beta^t} \mathrm{SSE}(\beta) = 2X^t X > 0\,.$$

Wir haben somit den MLE von β gefunden. Ganz einfach ist jetzt zu sehen, dass jede Komponente des Vektors $\hat{\beta}$ jeweils eine Linearkombination der normalverteilten Responses ist, und dass deshalb

$$\hat{\beta} \sim N\left(\mathrm{E}(\hat{\beta}), \mathrm{var}(\hat{\beta})\right)\,,$$

gilt mit

$$\mathrm{E}(\hat{\beta}) = \mathrm{E}((X^t X)^{-1} X^t y) = (X^t X)^{-1} X^t \mathrm{E}(y) = (X^t X)^{-1} X^t X\beta = \beta$$

und

$$\mathrm{var}(\hat{\beta}) = \mathrm{var}((X^t X)^{-1} X^t y) = (X^t X)^{-1} X^t \mathrm{var}(y) X (X^t X)^{-1} = \sigma^2 (X^t X)^{-1}\,.$$

Als **Prognosevektor** erhalten wir

$$\hat{\mu} = X\hat{\beta} = X(X^t X)^{-1} X^t y = Hy$$

mit der Matrix

$$H = X(X^tX)^{-1}X^t\,.$$

H ist symmetrisch und wurde von Tukey als **Hat Matrix** bezeichnet. H angewandt auf y erzeugt den Prognosevektor $\hat{\mu}$ (setzt dem μ das Dach auf). H ist auch idempotent, da

$$\begin{aligned} HH^t &= X(X^tX)^{-1}X^t(X(X^tX)^{-1}X^t)^t \\ &= X(X^tX)^{-1}X^t = H\,. \end{aligned}$$

Der Vektor der **Residuen** ist wiederum definiert als

$$r = y - \hat{\mu} = y - Hy = (I - H)y\,.$$

Auch $I - H$ ist symmetrisch und idempotent, da

$$(I - H)(I - H)^t = I - 2H + H = I - H\,.$$

Man bemerke, dass die Elemente sowohl in $\hat{\mu}$ als auch in r Linearkombinationen der normalverteilten Responses sind und damit

$$\hat{\mu} \sim N\left(\mathrm{E}(\hat{\mu}), \mathrm{var}(\hat{\mu})\right)\,, \qquad r \sim N\left(\mathrm{E}(r), \mathrm{var}(r)\right)$$

folgt. Für die Momente erhält man

$$\begin{aligned} \mathrm{E}(\hat{\mu}) &= X\mathrm{E}(\hat{\beta}) = X\beta = \mu \\ \mathrm{var}(\hat{\mu}) &= X\mathrm{var}(\hat{\beta})X^t = \sigma^2 H \end{aligned}$$

sowie

$$\begin{aligned} \mathrm{E}(r) &= (I - H)\mathrm{E}(y) = (I - H)X\beta = 0 \\ \mathrm{var}(r) &= (I - H)\mathrm{var}(y)(I - H) = \sigma^2(I - H) \end{aligned}$$

5.2.2 Schätzen der Varianz

In diesem Abschnitt werden wir zuerst zeigen, dass der MLE $\hat{\beta}$ und $\mathrm{SSE}(\hat{\beta}) = r^tr$ unabhängig sind. Wir betrachten dazu die Kovarianz zwischen $\hat{\beta}$ und r, schreiben diese Variablen jedoch vorerst um in Termen des statistischen Fehlers ϵ, d.h.

$$\hat{\beta} = (X^tX)^{-1}X^ty = (X^tX)^{-1}X^t\left(X\beta + \epsilon\right) = \beta + (X^tX)^{-1}X^t\epsilon$$

und

$$r = (I - H)y = (I - H)(X\beta + \epsilon) = X\beta - X(X^tX)^{-1}X^tX\beta + (I - H)\epsilon = (I - H)\epsilon\,.$$

Somit folgt das interessante Resultat

$$\begin{aligned} \mathrm{cov}(\hat{\beta}, r) &= \mathrm{cov}\left(\beta + (X^tX)^{-1}X^t\epsilon, (I - H)\epsilon\right) \\ &= (X^tX)^{-1}X^t \underbrace{\mathrm{cov}\left(\epsilon, \epsilon\right)}_{=\mathrm{var}(\epsilon)=\sigma^2 I} (I - H) \\ &= \sigma^2(X^tX)^{-1}X^t(I - H) \\ &= \sigma^2(X^tX)^{-1}X^t - \sigma^2(X^tX)^{-1}X^tX(X^tX)^{-1}X^t \\ &= 0\,, \end{aligned}$$

und daher die Unabhängigkeit von $\hat{\beta}$ und $\mathrm{SSE}(\hat{\beta}) = r^t r$ unter Normalverteilungsannahme. Im nächsten Schritt wird nun gezeigt, dass $\mathrm{SSE}(\hat{\beta})/\sigma^2$ einer χ^2-Verteilung genügt mit $n-p$ Freiheitsgraden. Dazu betrachten wir die Summe aller quadrierten Fehlerterme

$$\begin{aligned}\epsilon^t\epsilon &= (y-X\beta)^t(y-X\beta) = \left(y-X\hat{\beta}+X\hat{\beta}-X\beta\right)^t\left(y-X\hat{\beta}+X\hat{\beta}-X\beta\right)\\ &= \left(r+X\hat{\beta}-X\beta\right)^t\left(r+X\hat{\beta}-X\beta\right) = \left(r+X(\hat{\beta}-\beta)\right)^t\left(r+X(\hat{\beta}-\beta)\right).\end{aligned}$$

Dies entspricht der Summe von n quadrierten unabhängigen normalverteilten Größen mit Varianz σ^2. Um dafür eine χ^2_n-Verteilung zu erhalten, betrachten wir

$$\epsilon^t\epsilon/\sigma^2 = r^t r/\sigma^2 + (\hat{\beta}-\beta)^t X^t X(\hat{\beta}-\beta)/\sigma^2\,.$$

Wegen $HX = X$ verschwindet der gemischte Term, denn damit ist

$$r^t X = ((I-H)y)^t X = y^t(I-H)X = 0\,.$$

Weiters wissen wir, dass

$$\epsilon^t\epsilon/\sigma^2 = \sum_{i=1}^{n}(\epsilon_i/\sigma)^2 \sim \chi^2_n\,.$$

Wegen

$$\hat{\beta} \sim N_p(\beta, \sigma^2(X^tX)^{-1})$$

gilt weiters

$$(\hat{\beta}-\beta)(X^tX)^{1/2}/\sigma \sim N_p(0, I)$$

und für den letzten Term folgt, dass

$$(\hat{\beta}-\beta)^t X^t X(\hat{\beta}-\beta)/\sigma^2 \sim \chi^2_p\,.$$

Zur Erinnerung ist die momentenerzeugende Funktion einer χ^2_n-verteilten Zufallsvariablen $M(t) = (1-2t)^{-n/2}$, $t < 1/2$. Wendet man die Momentenerzeugende auf die Zerlegung an, so liefert dies

$$(1-2t)^{-n/2} = \mathrm{E}(\exp(tr^t r/\sigma^2))(1-2t)^{-p/2}\,.$$

Daraus folgt nun direkt, dass die uns interessierende momentenerzeugende Funktion der Fehlerquadratsumme

$$\mathrm{E}(\exp(tr^t r/\sigma^2)) = (1-2t)^{-(n-p)/2}\,,$$

die einer χ^2_{n-p} Verteilung entspricht. Also gilt

$$r^t r/\sigma^2 = \frac{1}{\sigma^2}\mathrm{SSE}(\hat{\beta}) \sim \chi^2_{n-p}\,,$$

und somit

$$\mathrm{E}(\mathrm{SSE}(\hat{\beta})) = \sigma^2(n-p)$$

wie auch

$$\mathrm{var}(\mathrm{SSE}(\hat{\beta})) = 2\sigma^4(n-p)\,.$$

Jetzt ist es verständlich, dass der MLE modifiziert werden muss um unverzerrt zu sein, und wir verwenden als erwartungstreuen Schätzer für die Varianz σ^2 unter dem multiplen linearen Regressionsmodell

$$S^2 = \frac{1}{n-p}\text{SSE}(\hat{\beta})\,.$$

Gerade haben wir die Verteilungseigenschaft vom minimalen $\text{SSE}(\hat{\beta})$ für normalverteilte Responsevariablen diskutiert. Interessanterweise resultiert $\text{E}(\text{SSE}(\hat{\beta})) = \sigma^2(n-p)$ auch ohne Annahme der Normalverteilung. Wir schreiben

$$\text{SSE}(\hat{\beta}) = r^t r = y^t(I-H)^t(I-H)y = y^t(I-H)y\,.$$

und berechnen unter der Annahme $\text{E}(y) = X\beta = \mu$ und $\text{var}(y) = \sigma^2 I$

$$\begin{aligned}\text{E}(\text{SSE}(\hat{\beta})) &= \text{E}(y^t(I-H)y)\\ &= \text{E}\Big((y-\mu)^t(I-H)(y-\mu) + \mu^t(I-H)y + y^t(I-H)\mu - \mu^t(I-H)\mu\Big)\end{aligned}$$

mit den Skalaren $y^t(I-H)\mu = \mu^t(I-H)y$ wofür $\text{E}(y^t(I-H)\mu) = \mu^t(I-H)\mu$ gilt. Wir erhalten somit

$$\text{E}(\text{SSE}(\hat{\beta})) = \text{E}\Big((y-\mu)^t(I-H)(y-\mu)\Big) + \mu^t(I-H)\mu\,.$$

Für das erste Skalar folgt

$$\begin{aligned}\text{E}\Big((y-\mu)^t(I-H)(y-\mu)\Big) &= \text{E}\Big(\text{trace}\Big((y-\mu)^t(I-H)(y-\mu)\Big)\Big)\\ &= \text{trace}\Big(\text{E}\Big((I-H)(y-\mu)(y-\mu)^t\Big)\Big)\\ &= \text{trace}\Big((I-H)\sigma^2 I\Big)\\ &= \sigma^2\text{trace}(I-H)\\ &= \sigma^2(n-p)\end{aligned}$$

während das zweite Skalar generell verschwindet, denn

$$\begin{aligned}\mu^t(I-H)\mu &= (X\beta)^t(I-H)(X\beta)\\ &= \beta^t X^t X\beta - \beta^t X^t X(X^tX)^{-1}X^tX\beta\\ &= 0\,.\end{aligned}$$

Daher gilt ganz allgemein unter $\text{E}(y) = X\beta = \mu$ und $\text{var}(y) = \sigma^2 I$

$$\text{E}(\text{SSE}(\hat{\beta})) = \sigma^2(n-p)\,.$$

5.2.3 Likelihood Terme

Beobachtete und erwartete Informationsmatrizen (Fisher Information) sind auch zentrale Größen bei linearen Regressionsmodellen. Sie liefern Varianzschranken für die MLE's (Cramér-Rao Schranke multivariat).

Wir beginnen mit den $(p+1)$ Scoregleichungen

$$\begin{aligned}\frac{\partial}{\partial\beta_j}\log f(y|\beta,\sigma^2) &= \frac{1}{\sigma^2}\sum_{i=1}^{n} x_{ij}(y_i-\mu_i)\,, \quad j=0,\ldots,p-1\\ \frac{\partial}{\partial\sigma^2}\log f(y|\beta,\sigma^2) &= -\frac{n}{2}\frac{1}{\sigma^2}+\frac{1}{2}\frac{1}{\sigma^4}\sum_{i=1}^{n}(y_i-\mu_i)^2\,,\end{aligned}$$

und den negativen zweiten Ableitungen

$$\begin{aligned}-\frac{\partial}{\partial\beta_j\partial\beta_k}\log f(y|\beta,\sigma^2) &= \frac{1}{\sigma^2}\sum_{i=1}^{n} x_{ij}x_{ik}\,, \quad j,k=0,\ldots,p-1\\ -\frac{\partial}{\partial\beta_j\partial\sigma^2}\log f(y|\beta,\sigma^2) &= \frac{1}{\sigma^4}\sum_{i=1}^{n} x_{ij}(y_i-\mu_i)\,, \quad j=0,\ldots,p-1\\ -\frac{\partial}{\partial(\sigma^2)^2}\log f(y|\beta,\sigma^2) &= \frac{1}{2}\left(-\frac{n}{\sigma^4}+\frac{2}{\sigma^6}\sum_{i=1}^{n} x_{ij}(y_i-\mu_i)^2\right)\,.\end{aligned}$$

Die Fisherinformation beinhaltet nun den Erwartungswert dieser Größen

$$\begin{aligned}\mathrm{E}\left(-\frac{\partial}{\partial\beta_j\partial\beta_k}\log f(y|\beta,\sigma^2)\right) &= \frac{1}{\sigma^2}x_j^t x_k\,, \quad j,k=0,\ldots,p-1\\ \mathrm{E}\left(-\frac{\partial}{\partial\beta_j\partial\sigma^2}\log f(y|\beta,\sigma^2)\right) &= 0\,, \quad j=0,\ldots,p-1\\ \mathrm{E}\left(-\frac{\partial}{\partial(\sigma^2)^2}\log f(y|\beta,\sigma^2)\right) &= \frac{n}{2}\frac{1}{\sigma^2}\,.\end{aligned}$$

Die erwartete Informationsmatrix ist somit

$$I(\beta,\sigma^2) = \begin{pmatrix}\sigma^{-2}X^tX & 0\\ 0 & \frac{1}{2}n\sigma^{-4}\end{pmatrix}$$

mit Inverser

$$I^{-1}(\beta,\sigma^2) = \begin{pmatrix}\sigma^2(X^tX)^{-1} & 0\\ 0 & 2\sigma^4/n\end{pmatrix}$$

Die Null in der Nebendiagonalen weist darauf hin, dass die entsprechenden MLE's $\hat\beta$ und $\hat\sigma^2$ asymptotisch unkorreliert sind. Wir wissen auch, dass der Eintrag $\sigma^2(X^tX)^{-1}$ gerade die Varianz/Kovarianzmatrix des MLE $\hat\beta$ darstellt, weshalb der für β unverzerrte Schätzer $\hat\beta$ die Varianzschranke erreicht. Generell könnte dies gezeigt werden, indem man eine passende Faktorisierung der Scorefunktion sucht. Nun ist

$$\begin{aligned}\frac{\partial}{\partial\beta}\log f(y|\beta,\sigma^2) &= \frac{1}{\sigma^2}X^t(y-X\beta) = \frac{1}{\sigma^2}\left(X^ty - X^tX\beta\right)\\ &= \frac{1}{\sigma^2}(X^tX)\left((X^tX)^{-1}X^ty-\beta\right)\\ &= \frac{1}{\sigma^2}(X^tX)\left(\hat\beta-\beta\right)\end{aligned}$$

mit $\mathrm{E}(\hat{\beta}) = \beta$. Also erreicht $\hat{\beta}$ die Cramér-Rao Schranke. Wie sieht dies für einen Varianzschätzer aus? Die Scorefunktion lässt sich schreiben als

$$\begin{aligned}\frac{\partial}{\partial\sigma^2}\log f(y|\beta,\sigma^2) &= -\frac{n}{2}\frac{1}{\sigma^2}+\frac{1}{2}\frac{1}{\sigma^4}\sum_{i=1}^{n}(y_i-\mu_i)^2 \\ &= \frac{n}{2\sigma^4}\left(\frac{1}{n}\sum_{i=1}^{n}(y_i-\mu_i)^2-\sigma^2\right)\end{aligned}$$

wobei der resultierende Schätzer $n^{-1}\mathrm{SSE}(\beta)$ erwartungstreu für σ^2 ist. Also erreicht der Schätzer $n^{-1}\mathrm{SSE}(\beta)$, mit β bekannt, die Cramér-Rao Schranke. Wir haben schon gezeigt, dass auch $S^2 = (n-p)^{-1}\sum_i(y_i-\hat{\mu}_i)^2$ erwartungstreu für σ^2 ist und dass $(n-p)S^2/\sigma^2 \sim \chi^2_{n-p}$ gilt. Deshalb resultiert als Varianz dieses Schätzers

$$\mathrm{var}(S^2) = 2\frac{\sigma^4}{n-p}$$

was natürlich größer ist als $2\sigma^4/n$.

5.2.4 Konfidenz- und Vorhersageintervalle

Wir haben nun Schätzer für die unbekannten Parameter β und σ^2 hergeleitet und deren Verteilungseigenschaften beschrieben. In diesem Abschnitt wollen wir prüfen, ob unter dem postulierten Regressionsmodell auch wirklich jede einzelne erklärende Komponente x_j relevant ist. Wir betrachten also Hypothesen der Form

$$H_0 : \beta_j = 0, \quad j = 1,\ldots,p-1\,.$$

Kann H_0 für ein spezielles $1 < j < p-1$ verworfen werden, so scheint der Einfluss der j-ten Variablen im Erwartungswertmodell sehr gering zu sein. Deswegen sollten auch derartige erklärenden Größen nicht in das Modell aufgenommen werden.
Generell gilt für jede Komponente $\hat{\beta}_j$

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2 v_{jj})$$

wobei v_{jj} das j-te Diagonalelement von $(X^tX)^{-1}$ bezeichnet. Somit gilt

$$\frac{\hat{\beta}_j-\beta_j}{\sqrt{\sigma^2 v_{jj}}} \sim N(0,1)\,.$$

Es wurde bereits gezeigt, dass $\hat{\beta}$ und $\mathrm{SSE}(\hat{\beta})$ unabhängig sind und dass

$$\frac{n-p}{\sigma^2}S^2 \sim \chi^2_{n-p}\,.$$

Daher definiert man

$$T = \frac{\frac{\hat{\beta}_j-\beta_j}{\sqrt{\sigma^2 v_{jj}}}}{\sqrt{\frac{n-p}{\sigma^2}S^2/(n-p)}} = \frac{\hat{\beta}_j-\beta_j}{\sqrt{S^2 v_{jj}}} \sim t_{n-p}\,.$$

Möchte man also die Relevanz der j-ten erklärenden Größe testen, so verwendet man dazu die Statistik

$$T = \frac{\hat{\beta}_j}{\sqrt{S^2 v_{jj}}}$$

und verwirft H_0, falls $|T| > t_{n-p,1-\alpha/2}$.
Ein zweiseitiges Konfidenzintervall für die Parameterkomponente β_j ist

$$KIV(\beta_j) = \hat{\beta}_j \pm s v_{jj}^{1/2} t_{n-p,1-\alpha/2}\,.$$

Eine gänzlich andere Fragestellung richtet sich auf neue, noch nicht beobachtete Vektoren $x_+ = (1, x_{+1}, \ldots, x_{+p-1})^t$ und auf den Erwartungswert dafür, also auf $\mu_+ = x_+^t\beta$. Nun ist $\hat{\mu}_+ = x_+^t\hat{\beta}$ der MLE für μ_+ und es gilt dafür

$$\begin{aligned} \mathrm{E}(\hat{\mu}_+) &= x_+^t\beta \\ \mathrm{var}(\hat{\mu}_+) &= x_+^t \mathrm{var}(\hat{\beta}) x_+ = \sigma^2 x_+^t (X^tX)^{-1} x_+\,. \end{aligned}$$

Nun ist $\hat{\mu}_+$ eine Linearkombination der normalverteilten Komponenten $\hat{\beta}$ und somit selbst normalverteilt mit den oben angegebenen Momenten. Da S^2 unabhängig von $\hat{\beta}$ folgt

$$\frac{x_+^t\hat{\beta} - x_+^t\beta}{\sqrt{S^2 x_+^t (X^tX)^{-1} x_+}} \sim t_{n-p}\,.$$

Bemerke, dass die Varianz eines geschätzten Erwartungswertes unter dem Modell, $\hat{\mu}_i = x_i^t\hat{\beta}$, geschrieben werden kann als

$$\mathrm{var}(\hat{\mu}_i) = \sigma^2 x_i^t (X^tX)^{-1} x_i = \sigma^2 h_{ii}$$

mit h_{ii} dem i-ten Diagonalelement der Hat Matrix $H = X(X^tX)^{-1}X^t$.
Ein **Vorhersageintervall** für eine neue Beobachtung $y_+ = x_+^t\beta + \epsilon_+$, mit $\epsilon_+ \sim N(0, \sigma^2)$ unabhängig von allen y mit denen $\hat{\beta}$ berechnet wurde, ist ein Intervall, welches mit Wahrscheinlichkeit $(1-\alpha)$ die Realisierung dieser zusätzlichen Response y_+ beinhalten wird. Die Momente des MLE von y_+ sind gerade

$$\begin{aligned} \mathrm{E}(x_+^t\hat{\beta} + \epsilon_+) &= x_+\beta \\ \mathrm{var}(x_+^t\hat{\beta} + \epsilon_+) &= \mathrm{var}(x_+^t\hat{\beta}) + \mathrm{var}(\epsilon_+) = \sigma^2(1 + x_+^t(X^tX)^{-1}x_+)\,. \end{aligned}$$

Weiters gilt

$$\begin{aligned} \mathrm{E}(y_+ - x_+^t\hat{\beta}) &= x_+\beta - x_+\beta = 0 \\ \mathrm{var}(y_+ - x_+^t\hat{\beta}) &= \sigma^2(1 + x_+^t(X^tX)^{-1}x_+)\,. \end{aligned}$$

Also basieren wir die Inferenz bezüglich einer neuen Response unter einem multiplen Regresssionsmodell auf die Statistik

$$\frac{y_+ - x_+^t\hat{\beta}}{\sqrt{S^2(1 + x_+^t(X^tX)^{-1}x_+)}} \sim t_{n-p}\,.$$

Ein graphischer Vergleich eines Vorhersageintervalls für y_+ und eines Konfidenzintervalls für μ_+ findet man in der Abbildung 5.3.

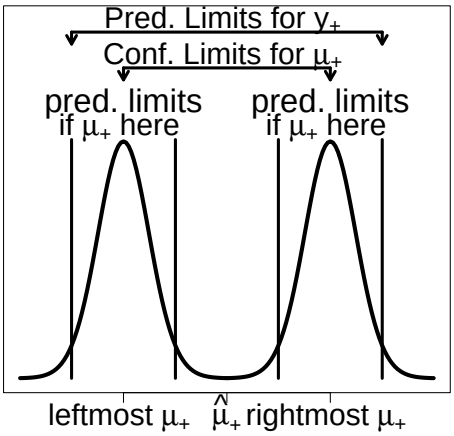


Abbildung 5.3: Zur Konstruktion von Konfidenzintervall und Vorhersageintervall bei der linearen Regression.

Beispiel 5.1 *Wir wollen nun zwei Gruppen normalverteilter Responses mit einem linearen Regressionsmodell analysieren. Habe dazu die erste Gruppe Erwartung β_0, d.h. für Responses aus dieser Gruppe gelte*

$$y_{0i} = \beta_0 + \epsilon_{0i}\,, \qquad i = 1, \ldots, n_0$$

und die andere Gruppe habe Erwartung $\beta_0 + \beta_1$, also gilt für Responses aus der zweiten Gruppe

$$y_{1i} = \beta_0 + \beta_1 + \epsilon_{1i}\,, \qquad i = 1, \ldots, n_1\,.$$

Hierbei sind die statistischen Fehlerterme ϵ_{gi}, $g \in \{1, 2\}$, alle unabhängig mit Varianz σ^2. Die Matrixform dieses Modells lautet somit

$$\begin{pmatrix} y_{01} \\ \vdots \\ y_{0n_0} \\ y_{11} \\ \vdots \\ y_{1n_1} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_{01} \\ \vdots \\ \epsilon_{0n_0} \\ \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \end{pmatrix}$$

Der MLE $\hat{\beta} = (X^t X)^{-1} X^t y$ ist für dieses Modell gleich

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} n_0 + n_1 & n_1 \\ n_1 & n_1 \end{pmatrix}^{-1} \begin{pmatrix} n_0 \overline{y}_0 + n_1 \overline{y}_1 \\ n_1 \overline{y}_1 \end{pmatrix}$$

mit $\overline{y}_g = n_g^{-1} \sum_i y_{gi}$ die jeweiligen Gruppenmittel bezeichnet. Nun gilt

$$\begin{pmatrix} n_0 + n_1 & n_1 \\ n_1 & n_1 \end{pmatrix}^{-1} = \frac{1}{(n_0 + n_1) n_1 - n_1^2} \begin{pmatrix} n_1 & -n_1 \\ -n_1 & n_0 + n_1 \end{pmatrix}$$

$$= \frac{1}{n_0 n_1}\begin{pmatrix} n_1 & -n_1 \\ -n_1 & n_0+n_1 \end{pmatrix}$$
$$= \begin{pmatrix} n_0^{-1} & -n_0^{-1} \\ -n_0^{-1} & n_0^{-1}+n_1^{-1} \end{pmatrix}$$

womit folgt

$$\begin{pmatrix} \hat\beta_0 \\ \hat\beta_1 \end{pmatrix} = \begin{pmatrix} n_0^{-1} & -n_0^{-1} \\ -n_0^{-1} & n_0^{-1}+n_1^{-1} \end{pmatrix}\begin{pmatrix} n_0\overline{y}_0 + n_1\overline{y}_1 \\ n_1\overline{y}_1 \end{pmatrix}$$
$$= \begin{pmatrix} \overline{y}_0 \\ \overline{y}_1 - \overline{y}_0 \end{pmatrix}$$

Bemerke, dass die Elemente von $\sigma^2(X^tX)^{-1}$ gerade die Varianzen und die Kovarianz der MLE's $\hat\beta_0$ und $\hat\beta_1$ beinhalten. So ist beispielsweise $var(\hat\beta_0) = var(\overline{y}_0) = \sigma^2/n_0$ oder $var(\hat\beta_1) = var(\overline{y}_1 - \overline{y}_0) = \sigma^2(1/n_1 + 1/n_0)$.
Als geschätzte Erwartungen unter diesem Modell erhält man $\widehat{E}(y_{0i}) = \hat\beta_0 = \overline{y}_0$ sowie $\widehat{E}(y_{1i}) = \hat\beta_0 + \hat\beta_1 = \overline{y}_1$.
Die Hypothese $H_0 : E(y_{0i}) = E(y_{1i})$ ist somit äquivalent zur Hypothese $H_0 : \beta_1 = 0$. Als Teststatistik resultiert dafür

$$T = \frac{\hat\beta_1 - 0}{\sqrt{\widehat{var}(\hat\beta_1)}} = \frac{\overline{y}_1 - \overline{y}_0}{\sqrt{S^2(1/n_0 + 1/n_1)}} \sim t_{n-2}\,.$$

Dies ist gerade die uns bekannte t-Test Statistik, falls die Varianzen in den beiden Gruppen gleich sind.

5.2.5 Ein angewandtes Beispiel

Beispiel 5.2 *Die $n = 79$ Vitalkapazitäten sollen nun als Responses betrachtet werden und ihr Erwartungswert soll in Termen der Körpergröße und/oder des Alters modelliert werden. Einfache lineare Modelle mit Intercept und einer erklärenden Variablen schätzt man durch*

```
> lm(vc ~ height)
  Coefficients:
  (Intercept)       height
      -812.31         7.72

> lm(vc ~ age)
  Coefficients:
  (Intercept)          age
      617.741       -2.122
```

Hierbei werden jeweils nur die Parameterschätzer von Konstante und Steigung ausgegeben. Ausführlichere Resultate erhält man hingegen durch

```
> summary(lm(vc ~ height))
  Coefficients:
              Estimate Std. Error t value Pr(>|t|)
  (Intercept) -812.3122   166.5833  -4.876 5.69e-06 ***
  height         7.7197     0.9409   8.205 4.10e-12 ***
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

  Residual standard error: 56.07 on 77 degrees of freedom
  Multiple R-Squared: 0.4665,     Adjusted R-squared: 0.4595
  F-statistic: 67.32 on 1 and 77 DF,  p-value: 4.099e-12

> summary(lm(vc ~ age))
  Coefficients:
              Estimate Std. Error t value Pr(>|t|)
  (Intercept) 617.7413    25.4151  24.306  < 2e-16 ***
  age          -2.1219     0.7938  -2.673  0.00917 **
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

  Residual standard error: 73.43 on 77 degrees of freedom
  Multiple R-Squared: 0.08492,    Adjusted R-squared: 0.07303
  F-statistic: 7.146 on 1 and 77 DF,  p-value: 0.00917
```

Falls die Vitalkapazität nur durch die Körpergröße erklärt ist (erstes Modell), so ist diese dafür signifikant notwendig. Jeder einzelne Zentimeter mehr an Körpergröße bedeutet im Mittel 7.7 Zentiliter mehr an Vitalkapazität. Im anderen Modell wird die Vitalkapazität erfolgreich durch das Alter erklärt. Jedes zusätzliche Jahr im Alter eines Probanden lässt im Mittel die Vitalkapazität um 2.1 Zentiliter abnehmen.
Kombiniert man diese beiden einfachen Regressionsmodelle, so liefert dies

```
> summary(lm(vc ~ height + age))
  Coefficients:
              Estimate Std. Error t value Pr(>|t|)
  (Intercept) -717.3919   175.2551  -4.093 0.000105 ***
  height         7.3525     0.9593   7.664 4.82e-11 ***
  age           -0.9892     0.6180  -1.601 0.113584
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

  Residual standard error: 55.51 on 76 degrees of freedom
  Multiple R-Squared: 0.4839,     Adjusted R-squared: 0.4703
  F-statistic: 35.62 on 2 and 76 DF,  p-value: 1.216e-11
```

Hierbei ist deutlich zu erkennen, dass die zusätzliche Altersinformation zu einem Modell, das bereits die Körpergröße inkludiert, keine signifikante Notwendigkeit aufweist (p-Wert 11%). Natürlich ist uns auch schon bekannt, dass kein linearer Zusammenhang zwischen `VC` *und* `age` *besteht sondern eher ein quadratischer oder sogar kubischer (vergleiche mit den LOWESS-Abbildungen). Die Schätzung eines Modells mit Größe, sowie linearem, quadratischem und kubischen Alterseffekt*

```
> summary(lm(vc ~ height + age + I(age^2) + I(age^3)))
  Coefficients:
               Estimate Std. Error t value Pr(>|t|)
  (Intercept) -1.106e+03  2.055e+02  -5.381 8.33e-07 ***
  height       6.660e+00  9.106e-01   7.314 2.55e-10 ***
  age          4.888e+01  1.575e+01   3.105  0.00270 **
  I(age^2)    -1.462e+00  4.979e-01  -2.935  0.00443 **
  I(age^3)     1.318e-02  4.945e-03   2.665  0.00944 **
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

  Residual standard error: 51.59 on 74 degrees of freedom
  Multiple R-Squared: 0.566,      Adjusted R-squared: 0.5426
  F-statistic: 24.13 on 4 and 74 DF,  p-value: 8.468e-13
```

zeigt, dass all diese Effekte signifikant im Modell berechtigt sind. Wie sieht nun die Designamatrix zu diesem doch schon recht komplexen Modell aus? Diese erhält man sehr einfach durch

```
> mod <- lm(vc ~ height + age + I(age^2) + I(age^3), x=TRUE)
> mod$x
  (Intercept) height age I(age^2) I(age^3)
 1           1    171  42     1764    74088
 2           1    178  32     1024    32768
 3           1    176  46     2116    97336
 :           :      :   :        :        :
```

Damit könnte man jetzt die Wurzel der Diagonalelemente in $s^2(X^tX)^{-1}$ berechnen, die oben angeführten Standardfehler der Parameterschätzer:

```
> X <- mod$x
> sqrt(diag(solve(t(X) %*% X))) * 51.59
  (Intercept)       height          age     I(age^2)     I(age^3)
 2.054779e+02 9.106415e-01 1.574705e+01 4.979678e-01 4.945972e-03
```

Wir entscheiden uns für dieses multiple lineare Modell und definieren eine neue Person x_+ mit Alter 45 Jahre und einer Größe von 184 cm. Welchen vorhergesagten (geschätzten) mittleren VC-Wert hat diese Person?

```
> height <- 184; age <- 45
> x.plus <- matrix(c(1, height, age, age^2, age^3))  # column vector
> hat.beta <- matrix(coef(mod))
> hatmu.plus <- as.numeric(t(x.plus) %*% hat.beta);  hatmu.plus
 [1] 560.766
```

Gebe ein 95% Konfidenzintervall für diesen Parameter μ_+ an:

```
> alpha <- 0.05;  s <- summary(mod)$sigma;  df <- summary(mod)$df[2]
> var.hatmu.plus <- as.numeric(s^2 * t(x.plus) %*% solve(t(X) %*% X) %*% x.plus)
> hatmu.plus - sqrt(var.hatmu.plus)* qt(1-alpha/2, df)
[1] 530.158
> hatmu.plus + sqrt(var.hatmu.plus)* qt(1-alpha/2, df)
[1] 591.374
```

Sehr viel einfacher erhält man dieses Resultat mittels

```
> predict(mod, new=data.frame(age=age, height=height), interval="confidence")
         fit     lwr     upr
[1,] 560.766 530.158 591.374
```

Ein Vorhersageintervall für eine neue Responsebeobachtung in x_+ ergibt

```
> var.haty.plus <- as.numeric(s**2 * (1 + t(x.plus) %*% solve(t(X) %*% X) %*% x.plus))
> hatmu.plus - sqrt(var.haty.plus)* qt(1-alpha/2, df)
[1] 453.52
> hatmu.plus + sqrt(var.haty.plus)* qt(1-alpha/2, df)
[1] 668.012
```

oder vereinfacht berechnet

```
> predict(mod, new=data.frame(age=age, height=height), interval="predict")
         fit    lwr     upr
[1,] 560.766 453.52 668.012
```

Wir wollen nun den Verlauf des geschätzten Modells zusammen mit punktweisen Vorhersageintervalle zeichnen.

```
> a <- seq(15, 50); g <- rep(175, length(a))
> p <- predict(mod, new=data.frame(age=a, height=g), interval="predict")
> plot (a, p[ , "fit"], ylim=c(400,700))
> lines(a, p[ , "lwr"]); lines(a, p[ , "upr"])
```

In der Abbildung 5.4 ist deutlich der gut passende kubische Altersverlauf zu sehen.

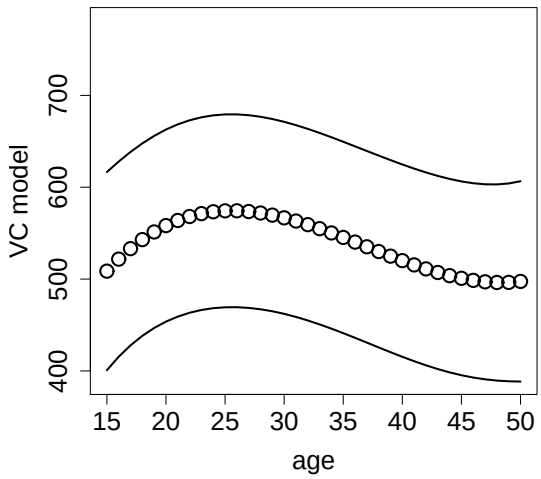


Abbildung 5.4: Geschätzte Erwartungswerte unter dem Modell mit kubischem Alterseinfluss und 95% punktweise Vorhersageintervalle für neue Responsebeobachtungen.

5.2.6 Multiples Bestimmtheitsmass

Auch für multiple lineare Regressionsmodelle ist das Bestimmtheitsmass definiert als

$$R^2 = \frac{\mathrm{SSR}(\hat{\beta})}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}(\hat{\beta})}{\mathrm{SST}}\,, \quad 0 \le R^2 \le 1\,.$$

Extreme Situationen liefern:
• $r^2 = 1$: perfekte Anpassung $y_i = \hat{\mu}_i \Rightarrow \mathrm{SSE}(\hat{\beta}) = 0$
• $r^2 = 0$: keine Abhängigkeit von $x_1, \ldots, x_{p-1} \Rightarrow \beta_1 = \cdots = \beta_{p-1} = 0 \Rightarrow \hat{\mu}_i = \overline{y} \Rightarrow \mathrm{SSR}(\hat{\beta}) = 0$.
Nachteil dieses Masses ist, dass es mit zunehmender Anzahl von erklärender Variablen im Modell wächst und damit gegen 1 geht. Aus diesem Grund wird es adjustiert und wir betrachten

$$R^2_{adj} = 1 - \frac{\mathrm{SSE}(\hat{\beta})/(n-p)}{\mathrm{SST}/(n-1)}\,.$$

Dieses Mass kann auch bei sehr schlecht passenden Modellen negative Werte haben.

5.3 Varianzanalyse – ANOVA

5.3.1 Geometrische Interpretation der Schätzer

Der LSE $\hat{\beta}$ minimiert $\mathrm{SSE}(\beta) = (y - X\beta)^t(y - X\beta)$, die quadrierte Distanz zwischen y und $\mu = X\beta$. Diese Distanz wird minimal wenn der Vektor $(y - X\beta)$ orthogonal zu dem von den Spalten in X aufgespannten Raum ist. Damit haben wir für jede Spalte x aus X

$$x^t(y - X\hat{\beta}) = 0\,,$$

und die Score- bzw. Normalgleichungen $X^t(y - X\hat{\beta}) = 0$ halten.
Der Punkt $\hat{\mu} = X\hat{\beta} = X(X^tX)^{-1}X^ty = Hy$ ist die orthogonale Projektion von y auf die durch die Spalten von X aufgespannte Ebene. Die Matrix H repräsentiert gerade diese Projektion. Bemerke, dass $\hat{\mu}$ eindeutig ist unabhängig davon ob X^tX invertierbar ist oder nicht.
Abbildung 5.5 zeigt, dass der Residuenvektor $r = y - \hat{\mu} = (I - H)y$ und der Vektor der angepassten Werte $\hat{\mu} = Hy$ orthogonal sind. Um dies algebraisch zu erkennen, prüfen wir

$$\hat{\mu}^t r = y^tH^t(I - H)y = y^t(H - H)y = 0\,.$$

Es gibt also eine unmittelbare Beziehung zwischen der Orthogonalität und der Unabhängigkeit normalverteilter Vektoren. Wir haben bereits gezeigt, dass zwei lineare Formen $u = a^ty$ und $v = b^ty$, $y \sim N(\mu, \Sigma)$, unabhängig sind falls $\mathrm{cov}(u, v) = a^t\Sigma b = 0$. Für zwei Vektoren von linearen Formen $u = A^ty$ und $v = B^ty$ impliziert dies Unabhängigkeit, falls $\mathrm{cov}(u, v) = A^t\Sigma B = 0$. In unserem Regressionskontext gilt $\Sigma = \sigma^2 I_n$, $\sigma^2 > 0$. Das Kriterium für Unabhängigkeit wird damit zu $\mathrm{cov}(u, v) = \sigma^2 A^tB = 0$. Dies ist gerade dann erfüllt, wenn die beiden Vektoren orthogonal sind. Orthogonalität ist somit äquivalent zur Unabhängigkeit bei Normalverteilung!

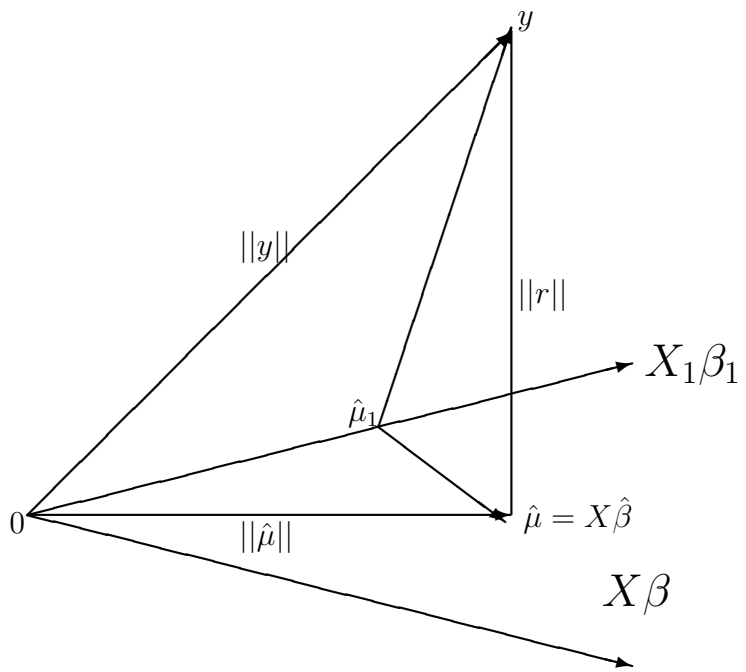


Abbildung 5.5: Zur Geometrie der Kleinsten Quadrate.

So gilt auch

$$y^t y = (y - \hat{\mu} + \hat{\mu})^t (y - \hat{\mu} + \hat{\mu}) = (r + \hat{\mu})^t (r + \hat{\mu}) = r^t r + \hat{\mu}^t \hat{\mu} + 2 r^t \hat{\mu} \,.$$

Wie bereits gezeigt sind r und $\hat{\mu}$ orthogonal und somit $r^t \hat{\mu} = 0$. Wir könne daher die overall sum of squares schreiben als

$$y^t y = r^t r + \hat{\mu}^t \hat{\mu} \,.$$

Dies resultiert auch unmittelbar aus dem Satz von Pythagoras und ist auch in der Abbildung 5.5 dargestellt.

5.3.2 F Statistiken

In den meisten Regressionsmodellen besteht die Schlüsselfrage darin, ob die erklärenden Variablen die Responsevariable beeinflussen oder nicht. Dies wirft auch die Frage auf, welche erklärenden Variablen im Modell benötigt werden.
Um konkret zu werden nehmen wir an, dass unser Modell lautet

$$y = X\beta + \epsilon = (X_1, X_2) \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \epsilon = X_1 \beta_1 + X_2 \beta_2 + \epsilon \,.$$

Hierbei ist X_1 eine $n \times q$ Matrix, X_2 eine $n \times (p - q)$ Matrix mit $q < p$,und β_1 sowie β_2 sind Parametervektoren mit entsprechenden Längen. Die erklärenden Variablen in X_2

sind nicht notwendig wenn $\beta_2 = 0$. In diesem Fall hält das einfachere Modell $y = X_1\beta_1 + \epsilon$. Wie kann man diese Situation erkennen?
In der Abbildung 5.5 repräsentiert die Gerade $x = 0$ in der horizontalen Ebene durch den Ursprung den linearen Unterraum der durch die Spalten in X_1 aufgespannt ist. Die Schätzung $\hat{\mu}_1 = X_1(X_1^tX_1)^{-1}X_1^ty$ ist die orthogonale Projektion von y auf diesen Unterraum. der Residuenvektor $y - \hat{\mu}_1 = (I - X_1(X_1^tX_1)^{-1}X_1^t)y$ zerfällt in die beiden orthogonalen Vektoren $y - \hat{\mu}$ und $\hat{\mu} - \hat{\mu}_1$, also

$$y - \hat{\mu}_1 = (y - \hat{\mu}) + (\hat{\mu} - \hat{\mu}_1)\,,$$

mit $(y - \hat{\mu})(\hat{\mu} - \hat{\mu}_1) = 0$. Diese Vektoren sind das Residuum vom komplexeren Modell $y - \hat{\mu}$, und die Änderung in den geschätzten Werten, falls X_2 in die Designmatrix aufgenommen wird, $\hat{\mu} - \hat{\mu}_1$. Da diese beiden Vektoren orthogonale lineare Funktionen der normalverteilten Responses y sind, Sind sie auch unabhängig. Der Satz von Pythagoras impliziert, dass

$$(y - \hat{\mu}_1)^t(y - \hat{\mu}_1) = (y - \hat{\mu})^t(y - \hat{\mu}) + (\hat{\mu} - \hat{\mu}_1)^t(\hat{\mu} - \hat{\mu}_1)$$

oder äquivalent

$$\mathrm{SSE}(\hat{\beta}_1) = \mathrm{SSE}(\hat{\beta}) + \Big(\mathrm{SSE}(\hat{\beta}_1) - \mathrm{SSE}(\hat{\beta})\Big)\,.$$

Die Quadratsumme des einfacheren Modells ist die Summe zweier unabhängiger Teile: der Quadratsumme des komplexeren Modells, $\mathrm{SSE}(\hat{\beta})$, und der Reduktion in der Quadratsumme wenn die Spalten von X_2 ins Modell aufgenommen werden, $\mathrm{SSE}(\hat{\beta}_1) - \mathrm{SSE}(\hat{\beta})$.
Ist das einfachere Untermodell korrekt, dann ist auch das komplexere Modell richtig, denn dort muss nur $\beta_2 = 0$ gesetzt werden. In diesem Fall folgt $\mathrm{SSE}(\hat{\beta}_1)$ einer $\sigma^2\chi^2_{n-q}$-Verteilung und $\mathrm{SSE}(\hat{\beta})$ hat eine $\sigma^2\chi^2_{n-p}$-Verteilung. Weiters ist nun $\mathrm{SSE}(\hat{\beta}_1) - \mathrm{SSE}(\hat{\beta})$ unabhängig von $\mathrm{SSE}(\hat{\beta})$. Falls $\beta_2 = 0$ so folgt $\mathrm{SSE}(\hat{\beta}_1) - \mathrm{SSE}(\hat{\beta})$ einer $\sigma^2\chi^2_{p-q}$-Verteilung (Argument mittels momentenerzeugender Funktion wie zuvor in Abschnitt 5.2.2). Somit gilt unter $\beta_2 = 0$

$$F = \frac{(\mathrm{SSE}(\hat{\beta}_1) - \mathrm{SSE}(\hat{\beta}))/(p-q)}{\mathrm{SSE}(\hat{\beta})/(n-p)} \sim F_{p-q,n-p}\,.$$

Ist jedoch β_2 nicht Null, so wird die durchschnittliche Reduktion in der Quadratsumme größer sein als unter $\beta_2 = 0$ erwartet. F wird daher im Vergleich zur $F_{p-q,n-p}$-Verteilung groß sein. Wir testen $H_0 : \beta_2 = 0$ mit F und verwerfen H_0, falls $F > F_{p-q,n-p;1-\alpha}$.
Besteht der Teil X_2 nur aus einer erklärenden Größe x_2, dann ist β_2 ein Skalar. Wir modellieren $y = X_1\beta_1 + x_2\beta_2 + \epsilon$ und berechnen daraus

$$T = \frac{\hat{\beta}_2 - \beta_2}{\sqrt{S^2 v_{rr}}}\,,$$

wobei v_{rr} das zu β_2 gehörende Diagonalelement von $(X^tX)^{-1}$ ist, mit $X = (X_1, x_2)$. Der Varianzschätzer $S^2 = (n-p)^{-1}\mathrm{SSE}(\hat{\beta})$ bezieht sich auch auf das komplexe Modell. Wir haben bereits gezeigt, dass unter $H_0 : \beta_2 = 0$ gilt: $T \sim t_{n-p}$. Hier besteht eine ganz einfache Beziehung zu F:

$$F = T^2 = \frac{\hat{\beta}_2^2}{S^2 v_{rr}}\,.$$

Beinhaltet andererseits X_1 nur den Intercept, dann wird mit F die Hypothese getestet, dass alle $p-1$ Slopeparameter im Modell Null sind, d.h. dass

$$H_0 : \beta_1 = \cdots = \beta_{p-1} = 0\,.$$

Unter dieser Hypothese ist keine der erklärenden Variablen relevant. Daher gilt $\mathrm{SSE}(\hat{\beta}_1) = \mathrm{SSE}(\overline{y}) = \mathrm{SST}$, und $\mathrm{SST} - \mathrm{SSE}(\hat{\beta}) = \mathrm{SSR}(\hat{\beta})$ ist unabhängig von $\mathrm{SSE}(\hat{\beta})$. Dieser Test wird häufig **globaler F-Test** genannt und wird von vielen Programmen gerechnet.

5.3.3 Quadratsummen

Die Interpretation von Quadratsummen ist dann besonders hilfreich, falls diese zerlegbar ist in die Reduktionen die durch sukzessives Aufnehmen einzelner erklärender Variablen in die Designmatrix entstehen.
Angenommen wir haben ein lineares Normalverteilungsmodell

$$y = 1_n\beta_0 + X_1\beta_1 + X_2\beta_2 + \cdots + X_m\beta_m + \epsilon\,,$$

worin die Matrizen $1_n, X_1, \ldots, X_m$ Terme genannt werden. Für gewöhnlich lautet das einfachste in Betracht kommende Modell

$$y = 1_n\beta_0 + \epsilon\,,$$

was die Schätzungen $\hat{\mu} = 1_n\overline{y}$ zu Folge hat und wofür als Fehlerquadratsumme $\mathrm{SSE}_0 = \mathrm{SSE}(\hat{\beta}_0) = \sum_i (y_i - \overline{y})^2$ mit $\nu_0 = n-1$ Freiheitsgraden resultiert.
Nun reduzieren wir sukzessive das SSE indem wir weitere Terme in die Designmatrix aufnehmen. Bezeichne $\hat{\mu}_r$ die Schätzung wenn die Terme $X_1, \ldots, X_r$ inkludiert sind, und schreibe

$$y - \hat{\mu}_0 = (y - \hat{\mu}_m) + (\hat{\mu}_m - \hat{\mu}_{m-1}) + \cdots + (\hat{\mu}_1 - \hat{\mu}_0)\,.$$

Dies ist eine Verallgemeinerung der Überlegung von zuvor. Die Geometrie der LSE ergibt wiederum, dass die Klammerausdrücke alle orthogonal sind. Somit folgt mit Pythagoras

$$(y-\hat{\mu}_0)^t(y-\hat{\mu}_0) = (y-\hat{\mu}_m)^t(y-\hat{\mu}_m) + (\hat{\mu}_m-\hat{\mu}_{m-1})^t(\hat{\mu}_m-\hat{\mu}_{m-1}) + \cdots + (\hat{\mu}_1-\hat{\mu}_0)^t(\hat{\mu}_1-\hat{\mu}_0)$$

oder äquivalent dazu

$$\mathrm{SSE}_0 = \mathrm{SSE}_m + \Big(\mathrm{SSE}_{m-1} - \mathrm{SSE}_m\Big) + \cdots + \Big(\mathrm{SSE}_0 - \mathrm{SSE}_1\Big)\,.$$

Alle Klammerausdrücke auf der rechten Seite sind unabhängige Zufallsvariablen. Die Differenz $(\mathrm{SSE}_{r-1} - \mathrm{SSE}_r)$ ist die Reduktion in der Fehlerquadratsumme durch Hinzunahme von X_r in das Modell, das jedoch die Terme $1_n, X_1, X_2, \ldots, X_{r-1}$ bereits beinhaltet. Gibt man immer mehr Terme in das Modell, so reduzieren sich dadurch die Freiheitsgrade und es gilt $\nu_0 \geq \nu_1 \geq \cdots \geq \nu_m$.
Die Situation $\nu_r = \nu_{r+1}$ tritt gerade dann auf wenn die Spalten von X_{r+1} eine Linearkombination der Spalten von $1_n, X_1, X_2, \ldots, X_r$ sind. In diesem Fall ist X_{r+1} redundant und es gilt $\nu_r = \nu_{r+1}$, $\hat{\mu}_r = \hat{\mu}_{r+1}$, $\mathrm{SSE}_r = \mathrm{SSE}_{r+1}$.

Die Quadratsummenkomponenten werden zusammengefasst in einer ANOVA Tabelle. Ihr Prototyp ist

Terme	*df*	Resid. QS	Term zugeben	*df*	Reduktion in QS	mittlere QS
1_n	$n-1$	SSE_0				
$1_n, X_1$	ν_1	SSE_1	X_1	$n-1-\nu_1$	$\text{SSE}_0 - \text{SSE}_1$	$\frac{\text{SSE}_0 - \text{SSE}_1}{n-1-\nu_1}$
$1_n, X_1, X_2$	ν_2	SSE_2	X_2	$\nu_1 - \nu_2$	$\text{SSE}_1 - \text{SSE}_2$	$\frac{\text{SSE}_1 - \text{SSE}_2}{\nu_1 - \nu_2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$1_n, X_1, \ldots, X_m$	ν_m	SSE_m	X_m	$\nu_{m-1} - \nu_m$	$\text{SSE}_{m-1} - \text{SSE}_m$	$\frac{\text{SSE}_{m-1} - \text{SSE}_m}{\nu_{m-1} - \nu_m}$

Beispiel 5.3 *Das multiple lineare Regressionsmodell für die $n = 79$ Vitalkapazitäten beinhaltete die Terme Körpergröße und einen linearen, quadratischen, sowie kubischen Alterseffekt. Sequentielle F Test werden durchgeführt mittels*

```
> anova(mod)
  Analysis of Variance Table

  Response: vc
            Df Sum Sq Mean Sq F value    Pr(>F)
  height     1 211652  211652 79.5378 2.372e-13 ***
  age        1   7896    7896  2.9672  0.089144 .
  I(age^2)   1  18376   18376  6.9057  0.010441 *
  I(age^3)   1  18901   18901  7.1028  0.009443 **
  Residuals 74 196915    2661
  ---
  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpretation: Gibt man in ein Intercept Modell die Größeninformation dazu, so stellt dies eine signifikante Verbesserung (***) dar. Beinhaltet das Modell die Größe, so ist die zusätzliche Altersinformation nur mit einem p-Wert von 8.9% relevant.
Man bemerke, dass dies gerade eine spezielle Sequenz darstellt. Interessant wäre es auch, zuerst die Altersterme aufzunehmen und dann erst die Körpergröße. Dazu muss jedoch das Modell neu geschätzt werden.

```
> modh <- lm(vc ~ age + I(age^2) + I(age^3) + height)
> anova(modh)
  Analysis of Variance Table

  Response: vc
            Df Sum Sq Mean Sq F value    Pr(>F)
  age        1  38531   38531  14.480 0.0002896 ***
  I(age^2)   1  39624   39624  14.890 0.0002415 ***
  I(age^3)   1  36331   36331  13.653 0.0004188 ***
  height     1 142339  142339  53.490 2.546e-10 ***
  Residuals 74 196915    2661
  ---
  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

5.4 Residuenanalyse

Bei der Schätzung des Parametervektors ist darauf zu achten, dass die Regression von Ausreißern (in x- und in y-Richtung) beeinflusst werden kann. Die verwendete Least-Squares Methode ist leider sehr anfällig gegenüber Ausreißern. Man verwendet daher Verfahren, um solche Ausreißer zu lokalisieren, beziehungsweise um das Modell zu verifizieren.
Die **Hat-Matrix** H erweist sich hierbei als wichtiges Werkzeug und ist definiert durch

$$H = X(X^tX)^{-1}X^t\,.$$

Geometrisch kann man H auf y anwenden und erhält

$$\hat{\mu} = Hy\,.$$

Für die einzelnen Elemente gilt

$$h_{ij} = x_i^t(X^tX)^{-1}x_j\,, \qquad i,j = 1,\ldots,n\,.$$

Weiters folgt für den i-ten Vorhersagewert

$$\hat{\mu}_i = \sum_{j=1}^{n} h_{ij}y_i = h_{ii}y_i + \sum_{j\neq i} h_{ij}y_j\,, \quad i = 1,\ldots,n\,.$$

Daraus resultiert

$$\frac{\partial\hat{\mu}_i}{\partial y_i} = h_{ii}\,, \quad i = 1,\ldots,n\,.$$

Die Diagonalelemente h_{ii} können somit als Maß für den Einfluss von y_i auf den Vorhersagewert $\hat{\mu}_i$ interpretiert werden.
Da H idempotent und symmetrisch ist, gilt für die Diagonalelemente zusätzlich

$$h_{ii} = \sum_{j=1}^{n} h_{ij}h_{ji} = \sum_{j=1}^{n} h_{ij}^2 = h_{ii}^2 + \sum_{j\neq i} h_{ij}^2\,, \quad i = 1,\ldots,n\,.$$

Dies bedeutet aber gerade, dass

1. $0 \leq h_{ii} \leq 1$,
2. falls $h_{ii} = 0 \Rightarrow h_{ij} = 0, \quad \forall i,j$,
3. falls $h_{ii} = 1 \Rightarrow h_{ij} = 0, \quad \forall i,j,\ i \neq j$,
4. falls $h_{ii} = 0 \Rightarrow \hat{\mu}_i$ wird nicht von y_i beeinflusst,
5. falls $h_{ii} = 1 \Rightarrow \hat{\mu}_i = y_i$, d.h. das Modell liefert eine exakte Schätzung für y_i.

Da trace$(H) = p$, folgt für das arithmetische Mittel der Diagonalelemente von H

$$\overline{h} = \frac{1}{n}\sum_{i=1}^{n} h_{ii} = \frac{p}{n}\,.$$

Falls alle h_{ii} gleich groß sind, d.h. falls $h_{ii} = \overline{h}$, $i = 1, \ldots, n$, spricht man von einem **D-optimalen Design**.
Für das einfache lineare Modell $\mu(x) = \beta_0 + \beta_1 x$ hält für h_{ii} die interessante Darstellung

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum\limits_{k=1}^{n}(x_k - \overline{x})^2}\,.$$

Für ein multiples lineares Modell mit Intercept resultiert die verallgemeinerte Darstellung

$$h_{ii} = \frac{1}{n} + \frac{1}{n-1}(x_i^1 - \overline{x}_1)^t S^{-1}(x_i^1 - \overline{x}_1)\,,$$

wobei $x_i^1 = (x_{i1}, \ldots, x_{ip-1})^t$ die i-te Designzeile ohne Intercept bezeichnet. S ist die empirische Varianz-Kovarianzmatrix dieser x_i^1.
Somit gilt für ein Modell mit Intercept, dass $h_{ii} \geq 1/n$. Aus den obigen Darstellungen von h_{ii} ist außerdem zu erkennen, dass h_{ii} wächst, falls die Distanz zwischen x_i und $\overline{x}$ beziehungsweise x_i^1 und $\overline{x}_1$ größer wird. Die Werte h_{ii} können somit direkt als Distanz der x_i^1 zum Zentrum $\overline{x}_1$ interpretiert werden. Je größer h_{ii} ist, desto extremer ist der i-te Designpunkt. Hoaglin und Welsch bezeichnen die i-te Beobachtung als **high-leverage Punkt** (Punkte mit großer Hebelwirkung), falls

$$h_{ii} > 2\overline{h} = \frac{2(p+1)}{n}\,.$$

High-leverage Punkte sind daher extreme Designpunkte, deren Vorhersagewerte $\hat{\mu}_i$ sehr stark von den Beobachtungen y_i abhängen. Zur Identifizierung solcher Punkte stehen außer den h_{ii} noch weitere Distanzen, wie die euklidische- oder die **Mahalanobisdistanz** zur Verfügung. Die Mahalanobisdistanz für die i-te Beobachtung ist definiert durch

$$MD_i^2 = (n-1)\left(h_{ii} - \frac{1}{n}\right)\,.$$

Eine Beobachtung, die als high-leverage identifiziert ist, muss nicht unbedingt auch eine **einflussreiche Beobachtung** sein. Beobachtungen werden einflussreich genannt, wenn ihre Elimination eine starke Änderung in der Schätzung der Parameter ergeben. Dies kann von Ausreißern in y- und/oder in x-Richtung verursacht sein. Das Auffinden solcher Beobachtungen wird aufgrund der Tatsache, dass diese sowohl einzeln, als auch gemeinsam einflussreich sein können, erschwert (Masking-Effekt).
Nun wird auf die Eigenschaften verschiedener Typen von Residuen eingegangen.

5.4.1 Gewöhnliche Residuen

Die Residuen r_i (beobachtbare Fehler) sind bei Least-Squares Schätzung gegeben durch

$$r = y - \hat{\mu} = (I - H)y\,.$$

Sie stellen den Anteil dar, der durch das Modell nicht erklärt werden konnte. Dadurch ist es möglich, getroffene Modellannahmen auf ihre Richtigkeit zu verifizieren. Zwischen Residuenvektor r und nicht beobachtbaren Fehlervektor ϵ gilt folgende Beziehung

$$r = (I - H)(X\beta + \epsilon) = (I - H)\epsilon\,,$$

d.h.

$$r_i = \epsilon_i - \sum_{j=1}^{n} h_{ij}\epsilon_j\,, \quad i = 1, \ldots, n\,.$$

Diese Darstellung zeigt, dass die Beziehung zwischen r und ϵ einzig von H abhängt. Sind die h_{ij} hinreichend klein, so kann r als vernünftiger Ersatz für ϵ angenommen werden. Die skalenabhängigen Residuen r_i haben zwar denselben Erwartungswert wie die ϵ_i, jedoch andere Varianz. Vergleiche

$$\begin{aligned} \mathrm{E}(\epsilon) &= 0 \qquad \mathrm{var}(\epsilon) = \sigma^2 I \\ \mathrm{E}(r) &= 0 \qquad \mathrm{var}(r) = \sigma^2(I - H)\,. \end{aligned}$$

Dies bedeutet, dass die r_i im Gegensatz zu den ϵ_i unterschiedliche Varianz haben. Deshalb sollten auch keine gewöhnliche Residuen miteinander verglichen werden. Genügt ϵ einer Normalverteilung, so sind auch die Residuen r_i normalverteilt mit

$$r_i \sim N\left(0, \sigma^2(I - H)\right)\,.$$

Für die nicht unabhängigen Residuen gilt

$$\begin{aligned} \mathrm{cov}(r_i, r_j) &= -\sigma^2 h_{ij} \\ \mathrm{cor}(r_i, r_j) &= -\frac{h_{ij}}{\sqrt{(1 - h_{ii})(1 - h_{jj})}}\,. \end{aligned}$$

Sie sind also negativ korreliert.
Falls man Least-Squares Schätzungen verwendet hat, folgt für die Summe der Residuen

$$\sum_{i=1}^{n} r_i = 0\,.$$

5.4.2 Standardisierte Residuen

Gewöhnliche Residuen mit großen h_{ii} haben kleine Varianz und umgekehrt. Um konstante Varianz zu bekommen, werden die

$$r_i = y_i - \hat{\mu}_i \quad \text{mit} \quad \mathrm{var}(r_i) = \sigma^2(1 - h_{ii})$$

zu

$$r_i^* = \frac{y_i - \hat{\mu}_i}{s\sqrt{1-h_{ii}}}$$

transformiert. Nun gilt für alle $i = 1, \ldots, n$

$$\mathrm{E}(r_i^*) = 0 \quad \text{und} \quad \mathrm{var}(r_i^*) = 1\,.$$

Jedoch ist die Summe der standardisierten Residuen nicht mehr Null.
Weiters kann gezeigt werden, dass die Quadrate der r_i^* einer Betaverteilung genügen mit

$$\frac{r_i^{*2}}{n-p} \sim \beta\left(\frac{1}{2}, \frac{n-p-1}{2}\right)\,.$$

Daraus folgt, dass die Verteilung der r_i^* eine monotone Transformation einer t-Verteilung ist, deren Teststatistik zum Erkennen von Ausreißern Verwendung findet. Weisberg definierte y_i als *potentiellen Ausreißer*, falls gilt

$$|r_i^*| > 2\sqrt{\mathrm{var}(r_i^*)} = 2\,.$$

Diese standardisierten Residuen werden in der Literatur leider auch oft als *studentisierte Residuen* bezeichnet. Da aber r_i und s^2 *nicht unabhängig* sind, stammen die r_i^* nicht aus einer t-Verteilung.

5.4.3 Deletion (Jackknife) Residuen

Es drängt sich die interessante Frage auf, wie sich das Residuum von y_i verändert, wenn die i-te Beobachtung für die Schätzung gar nicht verwendet wird. Daraus können Schlüsse über den Einfluss der i-ten Beobachtung auf die Schätzungen gezogen werden.
Bevor wir die Deletion-Residuen definieren, untersuchen wir den Effekt auf die Matrix $(X^tX)^{-1}$, auf $\hat{\beta}$ und auf die Summe der quadratischen Residuen (SSE), falls die i-te Beobachtung nicht berücksichtigt wird.
• Effekt auf $(X^tX)^{-1}$:
Sei $X_{(i)}$ die Designmatrix ohne i-ter Beobachtung (Zeile), so kann man zeigen

$$X_{(i)}^t X_{(i)} = X^tX - x_i x_i^t.$$

Unter der Voraussetzung, dass $X_{(i)}^t X_{(i)}$ eine reguläre Matrix ist, gilt für die Inverse dieser quadratischen Matrix

$$\left(X_{(i)}^t X_{(i)}\right)^{-1} = (X^tX)^{-1} + \frac{1}{1-h_{ii}}(X^tX)^{-1}x_i x_i^t (X^tX)^{-1}.$$

• Effekt auf den Parameterschätzer $\hat{\beta}$:
Entfernen wir die i-te Beobachtung, so gilt für den Parameterschätzer der $n-1$ verbleibenden Beobachtungen

$$\hat{\beta}_{(i)} = \left(X_{(i)}^t X_{(i)}\right)^{-1} X_{(i)}^t y_{(i)} = \left(X_{(i)}^t X_{(i)}\right)^{-1} (X^ty - x_i y_i),$$

wobei $y_{(i)}$ der Responsevektor ohne der i-ten Komponente ist. Substituieren wir darin die Inverse, so folgt

$$\begin{aligned}\hat{\beta}_{(i)} &= \hat{\beta} - (X^tX)^{-1}x_iy_i + \frac{\hat{\mu}_i}{1-h_{ii}}(X^tX)^{-1}x_i - \frac{y_ih_{ii}}{1-h_{ii}}(X^tX)^{-1}x_i \\ &= \hat{\beta} - \frac{r_i}{1-h_{ii}}(X^tX)^{-1}x_i.\end{aligned}$$

• Effekt auf die Summe der quadratischen Residuen (SSE):

$$\text{SSE} = (n-p)s^2 = y^ty - \hat{\beta}^tX^ty\,.$$

Äquivalent dazu gilt für die Summe der quadrierten Residuen ohne der i-ten Beobachtung

$$\text{SSE}_{(i)} = (n-1-p)s^2_{(i)} = y^ty - y_i^2 - \hat{\beta}^t_{(i)}(X^ty - x_iy_i)\,.$$

Setzen wir den Ausdruck für $\hat{\beta}_{(i)}$ ein, so folgt

$$\begin{aligned}\text{SSE}_{(i)} &= y^ty - y_i^2 - \left(\hat{\beta}^t - \frac{r_i}{1-h_{ii}}x_i^t(X^tX)^{-1}\right)(X^ty - x_iy_i) \\ &= \text{SSE} - y_i^2 + y_i\hat{\beta}^tx_i + \frac{r_i}{1-h_{ii}}x_i^t(X^tX)^{-1}X^ty - \frac{r_iy_i}{1-h_{ii}}x_i^t(X^tX)^{-1}x_i \\ &= \text{SSE} - y_i^2 + y_i\hat{\mu}_i + \frac{r_i\hat{\mu}_i}{1-h_{ii}} - \frac{r_iy_ih_{ii}}{1-h_{ii}} \\ &= \text{SSE} - \frac{r_i^2}{1-h_{ii}} \\ &= \text{SSE} - r_i^{*2}s^2 = s^2(n-p-r_i^{*2})\,.\end{aligned}$$

Die **Deletion-Residuen** sind daher definiert als

$$t_i = \frac{y_i - x_i^t\hat{\beta}_{(i)}}{s_{(i)}\sqrt{1 + x_i^t\left(X_{(i)}^tX_{(i)}\right)^{-1}x_i}} = \frac{r_i}{s_{(i)}\sqrt{1-h_{ii}}} = \frac{sr_i^*}{s_{(i)}}\,,$$

weil $x_i^t\left(X_{(i)}^tX_{(i)}\right)^{-1}x_i = h_{ii}/(1-h_{ii})$ gilt. Dies ist eine sinnvolle Definition, denn

$$\text{var}(y_i - \hat{\mu}_{i(i)}) = \sigma^2\left(1 + x_i^t\left(X_{(i)}^tX_{(i)}\right)^{-1}x_i\right) = \frac{\sigma^2}{1-h_{ii}}\,.$$

Jetzt wird nur noch das unbekannte σ^2 durch $s^2_{(i)}$ geschätzt. Der Unterschied zwischen den standardisierten und den Deletion-Residuen besteht somit in der Schätzung von σ. Betrachten wir nun die Quadratsummenzerlegung

$$\text{SSE} = \text{SSE}_{(i)} + \frac{r_i^2}{1-h_{ii}}$$

so folgt die Unabhängigkeit von $\text{SSE}_{(i)}$ und r_i^2. Bei Normalverteilungsannahme ist weiters $\text{SSE}_{(i)} \sim \sigma^2\chi^2_{n-1-p}$. Da $\text{E}\left(\frac{r_i}{\sigma\sqrt{1-h_{ii}}}\right) = 0$ und $\text{var}\left(\frac{r_i}{\sigma\sqrt{1-h_{ii}}}\right) = 1$ folgt $\frac{r_i}{\sigma\sqrt{1-h_{ii}}} \sim N(0,1)$ und damit $\frac{r_i^2}{\sigma^2(1-h_{ii})} \sim \chi^2_1$.
Mit diesen beiden χ^2-verteilten Größen wird das Deletion-Residuum gebildet, denn

$$\frac{\frac{r_i^2}{\sigma^2(1-h_{ii})}/1}{\frac{SSE_{(i)}}{\sigma^2}/(n-1-p)} = \frac{r_i^2}{s^2_{(i)}(1-h_{ii})} = t_i^2\,.$$

Weiters gilt daher

$$t_i^2 \sim F_{1,n-1-p}, \quad \text{bzw.} \quad t_i \sim t_{n-1-p}\,.$$

Für die Deletion-Residuen folgt aus der Varianzeigenschaft einer t-verteilten Zufallsvariablen

$$\text{E}(t_i) = 0 \quad \text{und} \quad \text{var}(t_i) = \frac{n-1-p}{n-3-p}\,.$$

Die Varianz ist also konstant (unabhängig von i). Allerdings sind die Zufallsvariablen t_i und t_j mit $i \neq j$ *nicht unabhängig.*
Eine andere Schreibweise der t_i liefert Informationen über die Beziehung zwischen standardisierten und Deletion-Residuen

$$t_i = r_i^*\sqrt{\frac{s^2}{s^2_{(i)}}} = r_i^*\sqrt{\frac{s^2(n-1-p)}{s^2(n-p-r_i^{*2})}} = r_i^*\sqrt{\frac{n-1-p}{n-p-r_i^{*2}}}\,.$$

Daraus ist ersichtlich, dass t_i^2 eine monotone, nichtlineare Transformation der r_i^{*2} ist. Weisberg bezeichnet y_i als Ausreißer zum Niveau α, falls

$$|t_i| \geq t_{n-1-p;1-\alpha/2n}\,.$$

Dieses auf die Bonferroni-Ungleichung basierende Kriterium ist zwar sicherlich sehr konservativ, berücksichtigt aber das theoretisch durchzuführende *n-fache Testen.*

5.5 Distanzanalyse

Die gewöhnlichen, standardisierten und Deletion-Residuen dienen primär zur Anzeige von auffälligen y-Werten (bzw. zur Validierung des Modells). Als Erweiterung dazu sollen nun Maße für den Einfluss einer einzelnen Beobachtung (x_i, y_i) auf die Schätzung von β konstruiert werden.

DFBETA

$$DFBETA_i = \hat{\beta} - \hat{\beta}_{(i)} = \frac{r_i}{1-h_{ii}}(X^tX)^{-1}x_i$$

DFSIGMA

$$\text{SSE}_{(i)} = \text{SSE} - \frac{r_i^2}{1-h_{ii}}$$

DFFIT

$$DFFIT_i = x_i^t\hat{\beta} - x_i^t\hat{\beta}_{(i)} = r_i\frac{h_{ii}}{1-h_{ii}}$$

DFBETAS

$$DFBETAS_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{S_{(i)}\sqrt{(X^tX)^{-1}_{jj}}}$$

DFFITS

$$DFFITS_i = \frac{r_i}{S_{(i)}\sqrt{1-h_{ii}}}\left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2}$$

COVRATIO

$$COVRATIO_i = \frac{\det(S_{(i)}(X^t_{(i)}X_{(i)})^{-1})}{\det(S\ (X^tX)^{-1})}$$

Cook-Distanz
Um den Einfluss der i-ten Beobachtung auf die Schätzung von β zu messen, hat Cook (1977) die normierte quadratische Distanz von $\hat{\beta}$ zu $\hat{\beta}_{(i)}$ vorgeschlagen, nämlich

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^t X^t X(\hat{\beta}_{(i)} - \hat{\beta})}{ps^2}, \quad i = 1, \ldots, n.$$

Ein großer Wert von D_i signalisiert einen starken Einfluss der i-ten Beobachtung auf die Schätzung von β. Da $D_i \sim F_{p,n-p}$, vergleiche man D_i mit den Quantilen der entsprechenden F-Verteilung, um eine Größeneinschätzung der Distanz von $\hat{\beta}$ zu $\hat{\beta}_{(i)}$ durchzuführen. Die Darstellung von D_i kann noch vereinfacht werden. Substituiert man $\hat{\beta}_{(i)}$ in D_i, so resultiert

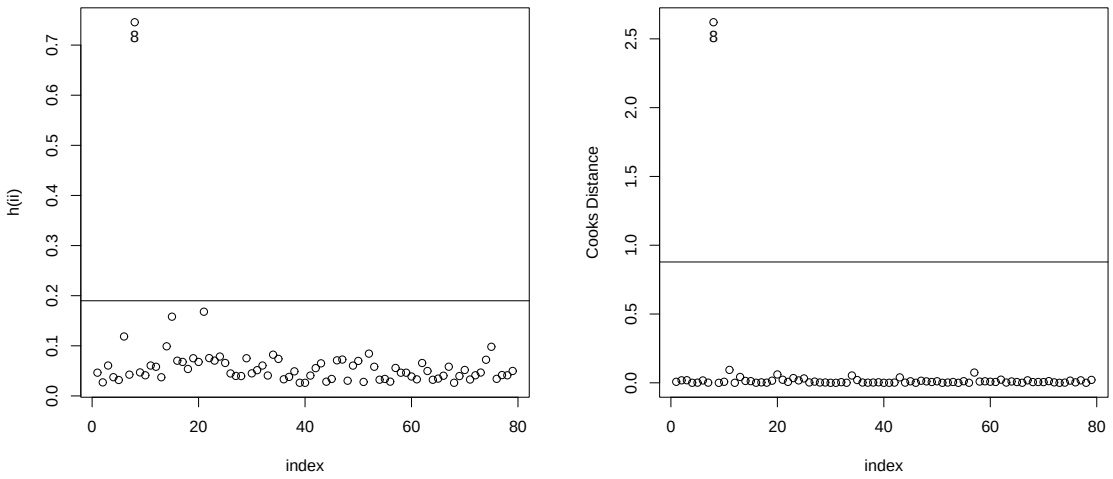
$$D_i = \frac{r_i^2 x_i^t(X^tX)^{-1}x_i}{s^2(1-h_{ii})^2p} = \frac{r_i^2 h_{ii}}{s^2(1-h_{ii})^2p} = \frac{r_i^{*2}}{p}\left(\frac{h_{ii}}{1-h_{ii}}\right).$$

Die Größe der **Cook-Distanz** hängt also nur vom standardisierten Residuum r_i^* und von h_{ii} ab, welches ein Maß für den Abstand von x_i^1 zum Zentrum $\overline{x}_1$ ist. D_i ist groß , falls die i-te Beobachtung ein high-leverage point ist, oder das entsprechende standardisierte Residuum groß ist, oder beides vorliegt.

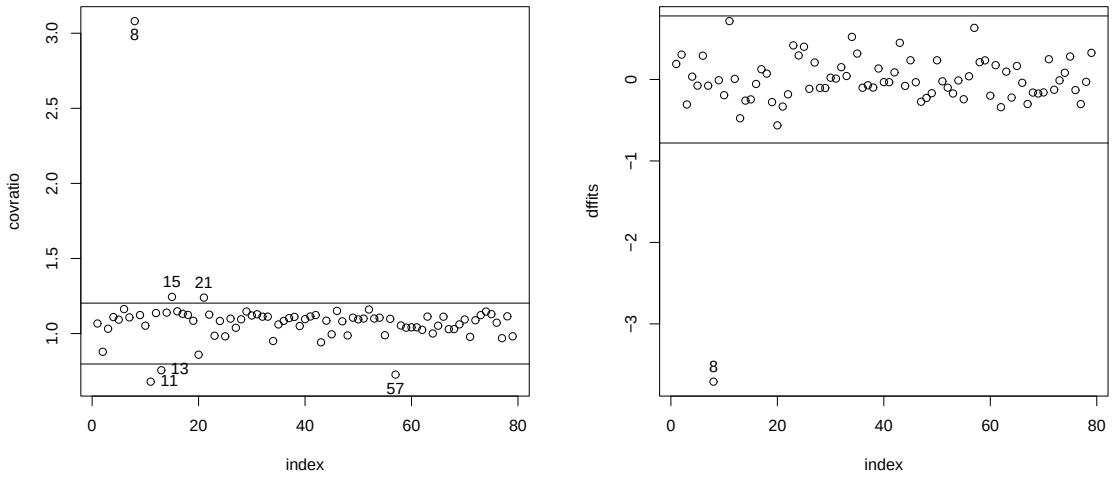
5.6 Angewandte Diagnostics

```
> summary(inflm.mod <- influence.measures(mod))
  Potentially influential observations of
          lm(formula = vc ~ height + age + I(age^2) + I(age^3)) :

    dfb.1_  dfb.hght dfb.age dfb.I(^2 dfb.I(^3 dffit   cov.r   cook.d  hat
  8  1.29_*  0.25    -2.21_*  2.45_*  -2.70_*  -3.71_*  3.08_* 2.62_*  0.75_*
 11  0.09    0.15    -0.28    0.27    -0.24     0.72    0.68_* 0.09    0.06
```

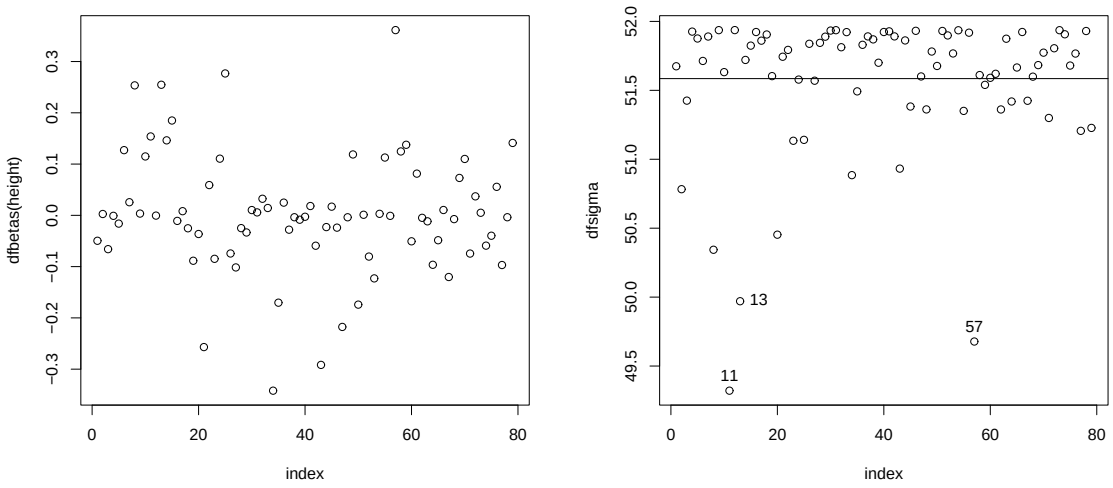
```
 13 -0.17    0.25    -0.02   -0.01     0.04    -0.48    0.76_* 0.04    0.04
 15 -0.20    0.18     0.06   -0.05     0.04    -0.24    1.24_* 0.01    0.16
 21  0.07   -0.26     0.16   -0.14     0.12    -0.33    1.24_* 0.02    0.17
 57 -0.51    0.36     0.31   -0.31     0.30     0.63    0.73_* 0.07    0.06

> which(apply(inflm.mod$is.inf, 1, any))
 8 11 13 15 21 57
 8 11 13 15 21 57

> h <- lm.influence(mod)$hat
> plot(1:n, h, xlab="index", ylab="h(ii)")
> abline(h = 3*5/79);  identify(1:n, h)

> c <- cooks.distance(mod) # pf(c, p, n - p) > 0.5
> pf(c, 5, 74) > 0.5       # obs nr 8 only
> plot(1:n, c, xlab="index", ylab="Cooks Distance")
> abline(h = qf(0.5, 5, 74)); identify(1:n, c)
```



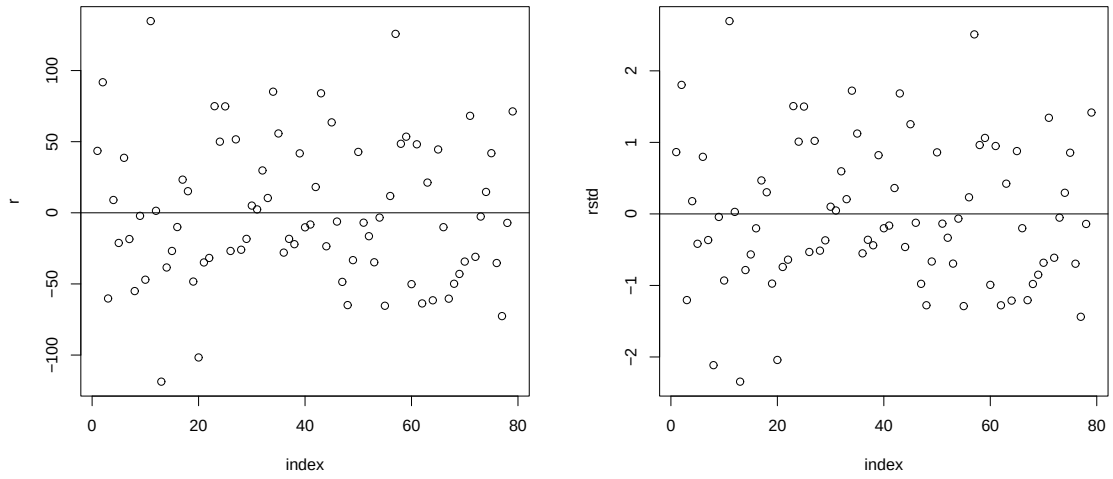

```
> cov.r <- covratio(mod) # 3p/(n - p)
> plot(1:n, cov.r, xlab="index", ylab="covratio")
> abline(h = c(1 - 3*5/74, 1 + 3*5/74)); identify(1:n, cov.r)

> dffits <- dffits(mod) # > 3*sqrt(p/(n - p))
> plot(1:n, dffits, xlab="index", ylab="dffits")
> abline(h = c(+3*sqrt(5/74), -3*sqrt(5/74))); identify(1:n, dffits)
```

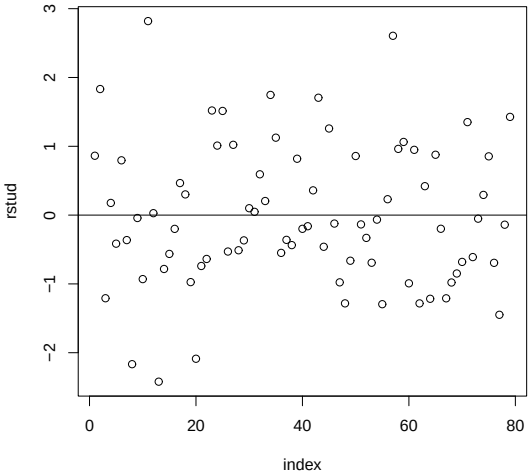



```
> dfbs <- dfbetas(mod) # > 1
> dfbs
     (Intercept)        height           age      I(age^2)      I(age^3)
 1   0.101258665 -0.0495899609 -0.090844379  0.0971073825 -0.0987564488
 2  -0.089469667  0.0026368428  0.104212527 -0.0815388812  0.0593050671
 3  -0.036833659 -0.0659490671  0.120175549 -0.1152957392  0.1011026054
 4  -0.017335527 -0.0009253076  0.024847735 -0.0239820293  0.0225516747
 :
> plot(1:n, dfbs[ , 2], xlab="index", ylab="dfbetas(height)")
> abline(h = c(1,-1)); identify(1:n, dfbs[,2])

> dfsigma <- lm.influence(mod)$sigma  # MSE_(i)
       1        2        3        4        5        6        7
51.67445 50.78282 51.42537 51.92606 51.87597 51.71320 51.89032
> plot(1:n, dfsigma, xlab="index", ylab="dfsigma")
> abline(h = summary(mod)$sigma); identify(1:n, dfsigma)
```




```
> r <- residuals(mod)    # raw residuals
> plot(1:n, r, xlab="index", ylab="r"); abline(h = 0)

> rstd <- rstandard(mod) # stand. residuals
> plot(1:n, rstd, xlab="index", ylab="rstd")
> abline(h = c(-3, 0, +3)); identify(1:n, rstd)

> rstud <- rstudent(mod) # deletion resididuals
> plot(1:n, rstud, xlab="index", ylab="rstud")
> abline(h = c(0, qt(1-0.05/158, 73))); identify(1:n, rstud)
```




rstud
3
2
1
0
−1
−2
0
20
40
60
80
index

Anhang A

Der Datensatz 'Vitalkapazität'

Statistik	total		Aichfeld Plz $<$ 8800		Murau Plz $\geq$ 8800		jung Alter $<$ 30		alt Alter $\geq$ 30	
Stichprobenumfang	79		34		45		37		42	
Arithmetisches Mittel	5.53	(0.09)	5.45	(0.14)	5.60	(0.11)	5.78	(0.13)	5.32	(0.10)
Varianz	0.58		0.67		0.52		0.66		0.42	
Standardabweichung	0.76		0.82		0.72		0.82		0.65	
Variationskoeffizient	0.14		0.15		0.13		0.14		0.12	
Geometrisches Mittel	5.48		5.39		5.56		5.72		5.28	
Harmonisches Mittel	5.43		5.32		5.52		5.67		5.24	
Skewness	0.29	(0.28)	0.03	(0.42)	0.67	(0.37)	0.11	(0.40)	0.13	(0.38)
Kurtosis	0.02	(0.55)	-0.64	(0.84)	0.35	(0.73)	0.07	(0.81)	-0.84	(0.76)
Median	5.50	(0.12)	5.58	(0.20)	5.50	(0.14)	5.65	(0.14)	5.32	(0.15)
Minimum	3.95		3.95		4.20		3.95		4.00	
Minimum Std. Score	-2.03		-1.98		-1.81		-2.09		-2.03	
Maximum	7.80		7.40		7.80		7.80		6.71	
Maximum Std.Score	3.02		2.22		3.20		2.64		2.15	
Bereich	3.85		3.45		3.60		3.85		2.71	
1.Quartil	4.90		4.73		5.05		5.35		4.80	
3.Quartil	6.05		5.98		6.10		6.25		5.83	
Interquartiler Bereich	1.15		1.25		1.05		0.90		1.03	
Standarddeviation	0.85		0.93		0.78		0.67		0.76	
1.Dezil	4.60		4.35		4.85		4.70		4.53	
9.Dezil	6.55		6.33		6.65		6.80		6.13	
Interdeziler Bereich	1.95		1.98		1.80		2.10		1.60	
Tailness	1.94		2.78		1.32		2.14		1.60	
Peakedness	3.37		2.19		3.78		3.00		2.07	
Skewness	1.08		0.64		1.63		1.30		0.97	

Tabelle A.1: Univariate Grundstatistiken in den Gruppen.

Aichfeld					Murau				
Alter	Groesse	Gewicht	VC	FEV1	Alter	Groesse	Gewicht	VC	FEV1
41	170	70	480	365	34	178	73	560	465
16	182	63	550	400	44	174	78	475	335
16	183	71	625	495	40	172	80	490	400
16	178	73	565	435	35	189	94	590	505
16	180	74	545	515	38	180	64	545	470
35	170	66	465	395	33	177	82	610	495
46	176	74	535	400	31	172	75	535	450
16	182	85	520	500	33	182	91	540	425
16	195	80	620	510	44	172	75	505	385
39	181	94	570	485	37	167	73	500	400
49	181	88	575	435	21	176	67	615	545
17	167	54	565	455	25	170	70	625	545
32	178	72	671	550	20	179	72	520	515
16	172	54	470	410	47	173	85	420	340
17	173	81	535	500	34	182	88	645	530
30	173	68	605	450	20	169	70	485	420
18	182	79	600	540	23	190	75	665	635
26	179	86	610	500	34	180	80	560	510
40	171	72	475	375	22	180	87	665	595
34	185	81	590	500	21	168	100	560	475
26	182	90	595	495	36	176	90	540	470
46	176	82	640	510	20	192	76	655	550
46	176	78	445	341	30	176	85	570	420
28	189	85	740	500	35	172	94	460	375
16	174	64	590	490	31	188	100	615	445
16	178	58	440	415	33	178	88	525	460
20	181	81	577	484	44	179	83	470	415
16	167	69	430	390	29	174	63	520	420
16	160	60	395	335	36	172	76	485	390
26	186	80	650	490	49	170	68	505	415
42	171	71	528	456	19	176	72	530	525
56	170	79	440	390	45	165	61	490	420
46	172	75	480	435	26	183	85	555	440
35	171	67	400	330	31	178	90	595	515
					20	186	92	680	575
					20	186	80	685	570
					25	187	102	780	580
					41	176	70	515	425
					41	180	80	620	510
					27	182	85	570	470
					28	168	64	540	430
					37	173	78	590	400
					25	174	80	565	470
					25	176	76	550	500
					32	169	68	485	410

Tabelle A.2: Datensatz AIMU.DAT.

Anhang B

Tabellen

Hier steht die Tabelle für die Normalverteilung

Hier steht die Tabelle für die t–Verteilung

Hier steht die Tabelle für die Chi...

Hier steht die Tabelle für die F...

Hier steht die Tabelle für die F...

Hier steht die Tabelle für die F...

Hier steht die Tabelle für die F...

Hier steht die Tabelle für die F...

Hier steht die Tabelle für die Kolmogorov–Smirnov...

Hier steht die Tabelle für die Wilcoxon...

Hier steht die Tabelle für die Wald–Wolfowitz..

Hier steht die Tabelle für die Wald–Wolfowitz..

Hier steht die Tabelle für die Wald–Wolfowitz..

Hier steht die Tabelle für die Wald–Wolfowitz..

Hier steht die Tabelle für die Kolmogorov–Smirnov...

Hier steht die Tabelle für die Kolmogorov–Smirnov...

Hier steht die Tabelle für die Kolmogorov–Smirnov...

Hier steht die Tabelle für die Kolmogorov–Smirnov...

Hier steht die Tabelle für die Kolmogorov–Smirnov...

Hier steht die Tabelle für die Wilcoxon...

Hier steht die Tabelle für die Wilcoxon...

Hier steht die Tabelle für die Wilcoxon...

Hier steht die Tabelle für die Wilcoxon...

Hier steht die Tabelle für die Van der Waerden...

Hier steht die Tabelle für die Van der Waerden...

Hier steht die Tabelle für die Mood...

Hier steht die Tabelle für die Mood...

Hier steht die Tabelle für die Hotelling-Pabst...
Hier steht die Tabelle für die Kendall...