

Statistik für Maschinenbauer (Teil II)

Die Statistik versucht auf Grund einer relativ kleinen **Stichprobe** von Daten Aussagen über eine große Menge von Daten (die **Gesamtpopulation**) zu machen. Zum Beispiel soll auf Grund einer Umfrage von 500 Personen eine Vorhersage über das Wahlverhalten der Gesamtbevölkerung gemacht werden. In Abhängigkeit von den verwendeten Methoden unterteilt man die Statistik in die **Beschreibende Statistik** und die **Schließende Statistik**.

Beschreibende (Deskriptive) Statistik: Methoden zur numerischen und graphischen Darstellung von Daten mit folgenden Zielen:

- Beschreibung der oft großen Datenmengen mit wenigen Kenngrößen und Graphiken.
- Überblick über die Zusammenhänge in den Daten.
- Durch Aufbereitung der Daten (z.B. Erkennen von fehlerhaften Daten (Tippfehler!), Transformation der Daten) soll die Analyse der Daten mit Standardmethoden der **Schließenden Statistik** vorbereitet werden (Schlagwort: EDA – **Explorative Daten Analyse**).

Schließende Statistik: Auf Grund der Explorativen Daten Analyse (oder im Idealfall auf Grund theoretischen Wissens über die Herkunft der Daten) wird ein wahrscheinlichkeitstheoretisches Modell des zugrundeliegenden zufälligen Geschehens erstellt. Anschließend kann mit den Methoden der Schließenden Statistik überprüft werden, ob die Daten zum unterstellten Modell passen oder nicht. Das wahrscheinlichkeitstheoretische Modell erlaubt dann Aussagen über die Gesamtpopulation (**Vorhersagen**). Zu den Methoden der Schließenden Statistik gehören:

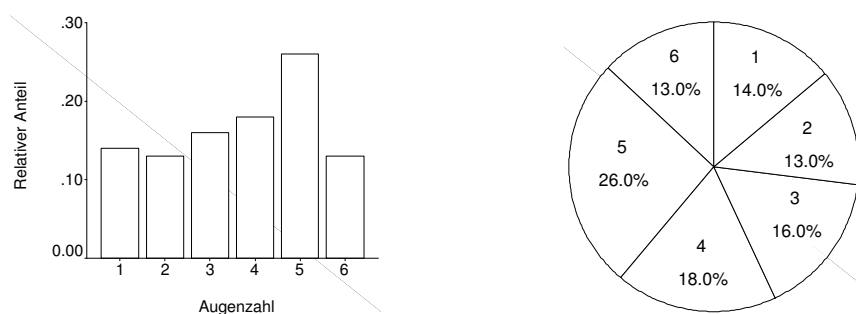
- Schätzmethoden
- Konfidenzintervalle
- Hypothesentests

Kapitel 1

Deskriptive Statistik

Man unterscheidet zwischen **qualitativen (kategoriellen) Merkmalen** (z.B. Familienstand, Haarfarbe, etc.) und **quantitativen Merkmalen** (z.B. Temperatur, Körpergröße, etc.). Qualitative Merkmale können nach einer geeigneten Kodierung (z.B. rot=1, blau=2) natürlich auch durch Zahlen beschrieben werden. Allerdings kann die Lage dieser Zahlen auf der Zahlengerade nicht interpretiert werden (z.B. rot|blau ?). Wir beschäftigen uns hauptsächlich mit den für die Technik bedeutenderen quantitativen Merkmalen. Ein quantitatives Merkmal wird durch eine reellwertige **Zufallsvariable** modelliert. Bei quantitativen Merkmalen unterscheidet man zwischen **diskreten** Merkmalen und **stetigen** Merkmalen.

Diskrete Merkmale: Nehmen nur bestimmte auf der Zahlengerade getrennt liegende Werte an (z.B. nur ganze Zahlen). Die graphische Darstellung kann etwa mittels **Balkendiagrammen** oder **Kreisdiagrammen** erfolgen.



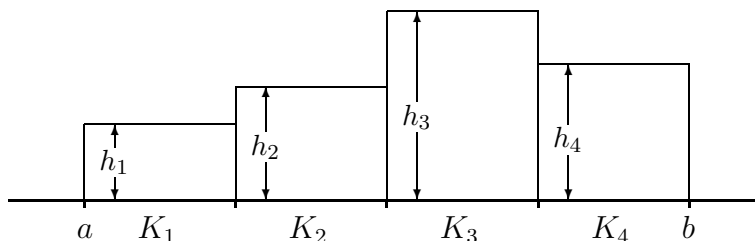
Balkendiagramm und Kreisdiagramm von 100 Würfelwürfen. Diese Daten werden in Kapitel 3 analysiert.

Stetige Merkmale: Jeder Wert eines Intervalls der Zahlengerade ist (theoretisch) möglich. Die graphische Darstellung kann etwa mittels **Histogrammen** oder **Boxplots** erfolgen.

Graphische Darstellung mittels Histogramm

Es sei x_1, \dots, x_n eine Stichprobe bestehend aus n reellen Meßwerten. Man wählt zuerst ein Intervall $[a, b]$, das alle Stichprobenwerte enthält. Dieses Intervall wird dann in k gleich

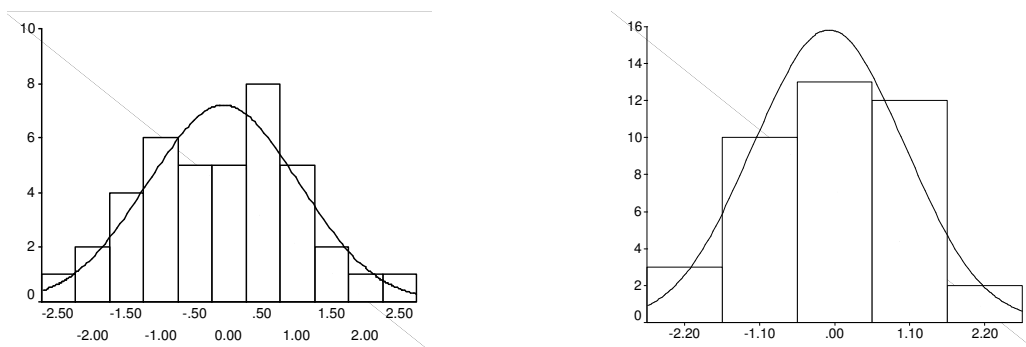
große Teilintervalle K_1, \dots, K_k der Länge $L = (b - a)/k$ zerlegt. Über der Klasse K_j wird ein Rechteck der Höhe h_j errichtet.



Dabei sind folgende Werte für h_j üblich

$$h_j = \begin{cases} n_j & \text{Anzahl der } x_i \text{ in der Klasse } K_j \text{ (absolute Häufigkeiten),} \\ n_j/n & \text{relative Häufigkeiten,} \\ \frac{n_j}{nL} & \text{Dichteschätzung.} \end{cases}$$

Mit der dritten Wahl von h_j wird die durch das Histogramm definierte Treppenfunktion $h(x)$ eine Dichtefunktion, d.h. $\int_a^b h(x) dx = 1$. Das erlaubt den Vergleich mit über das Histogramm gelegten theoretischen Dichten.



Histogramm mit Normalverteilungskurve von 40 simulierten Werten aus der $N(0, 1)$ -Verteilung. Links mit 11 Klassen (automatisch bestimmt), rechts mit 5 Klassen (manuell bestimmt). Diese Daten werden im Kapitel 3 analysiert.

Wahl des Intervalls $[a, b]$ und der Klassenzahl k : Viele Programmpakete bieten eine automatische Bestimmung von $[a, b]$ und der Klassenzahl k an. Diese erfolgt nach mehr oder weniger komplizierten Regeln. Eine einfache Regel lautet

$$[a, b] = [\min(x_i), \max(x_i)], \quad k = \begin{cases} \lfloor 2\sqrt{n} \rfloor & \text{für } n \leq 100 \\ \lfloor 10\log_{10}(n) \rfloor & \text{für } n > 100 . \end{cases}$$

Bei kleinen Stichprobenumfängen wird die Klassenzahl oft zu groß gewählt. Meist besteht die Möglichkeit sie manuell zu verändern.

Graphische Darstellung mittels Boxplot

Empirische Quantile: Sei x_1, \dots, x_n eine Stichprobe und $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ die zugehörige geordnete Stichprobe. Für $0 < p < 1$ heißt

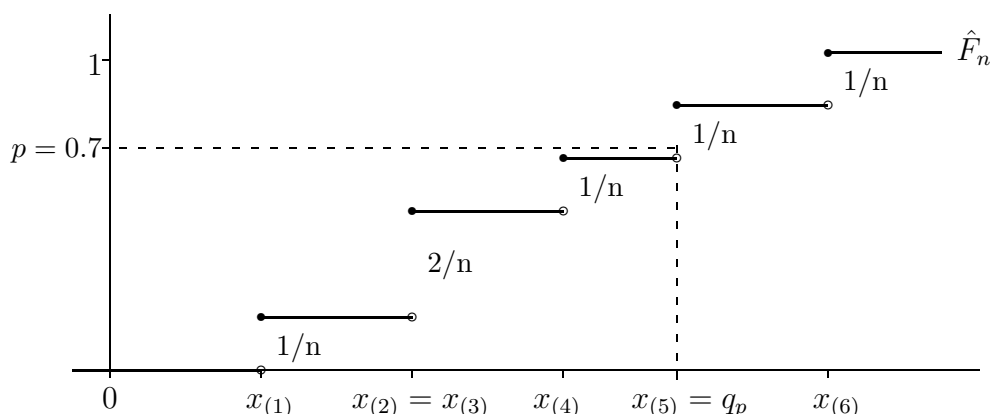
$$q_p = \begin{cases} \frac{1}{2}(x_{(np)} + x_{(np+1)}) & \text{falls } np \text{ ganzzahlig,} \\ x_{(\lfloor np \rfloor + 1)} & \text{falls } np \text{ nicht ganzzahlig,} \end{cases}$$

das p -te **empirische Quantil** (oder p -tes empirisches Perzentil). $q_{0.5}$ heißt **Median**, die Quantile $q_{0.25}, q_{0.5}, q_{0.75}$ nennt man **Quartile**, $iqr = q_{0.75} - q_{0.25}$ heißt **Interquartilsrange**. **Bedeutung von q_p :** ca. $p \cdot 100\%$ der Meßwerte sind kleiner als q_p .

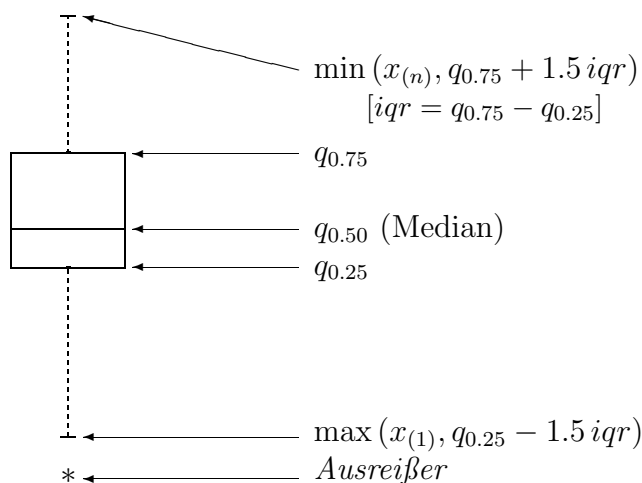
Die empirischen Quantile hängen eng mit der **empirischen Verteilungsfunktion** zusammen. Diese ist für eine Stichprobe x_1, \dots, x_n durch

$$\hat{F}_n(x) = \frac{1}{n} \times \{\text{Anzahl der } x_j \text{ kleiner gleich } x\} \quad (x \in \mathbb{R})$$

definiert. \hat{F}_n ist eine Treppenfunktion, die bei jedem Stichprobenwert $x_{(j)}$ der geordneten Stichprobe $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ um den Wert $1/n$ springt und zwischen den Stichprobenwerten konstant bleibt.

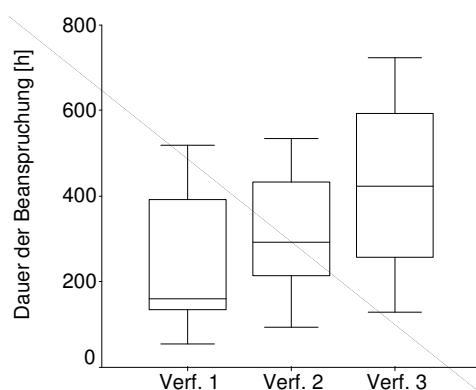


Der **Boxplot** ist eine graphische Darstellung einer Stichprobe die auf den empirischen Quartilen beruht.



Die Vorteile des Boxplots sind unter anderem:

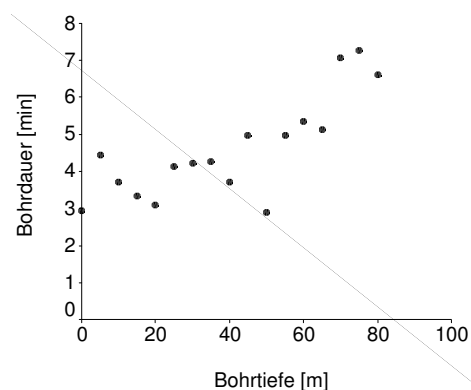
- Schneller Überblick über Charakteristika wie Median, Schiefe der Verteilung.
- Ideal zum Vergleich mehrerer Stichproben.
- Keine subjektive Wahl von Parametern.



Zur Erzeugung eines bedruckten Stoffes stehen drei Verfahren zur Verfügung. Die Boxplots beschreiben für jedes Verfahren die Verteilung der Variable “Dauer der Beanspruchung bis eine sichtbare Abnutzung entsteht”. Dieses Beispiel wird in Kapitel 4 analysiert.

Scatterplot (Streudiagramm)

Dient zur graphischen Beschreibung des Zusammenhangs zweier Merkmale. Als Daten liegen **Paare** von Meßwerten $(x_1, y_1), \dots, (x_n, y_n)$ vor. Diese werden als Punkte in der (x, y) -Ebene dargestellt.



Es soll die Abhängigkeit einer Methode zur Gesteinsbohrung von der Bohrtiefe untersucht werden. Der Scatterplot gibt die Bohrdauer für den Vortrieb der Bohrung um 1 Meter in verschiedenen Bohrtiefen an. Dieses Beispiel wird in Kapitel 5 analysiert.

Der Quantil-Quantil-Plot (QQ-Plot)

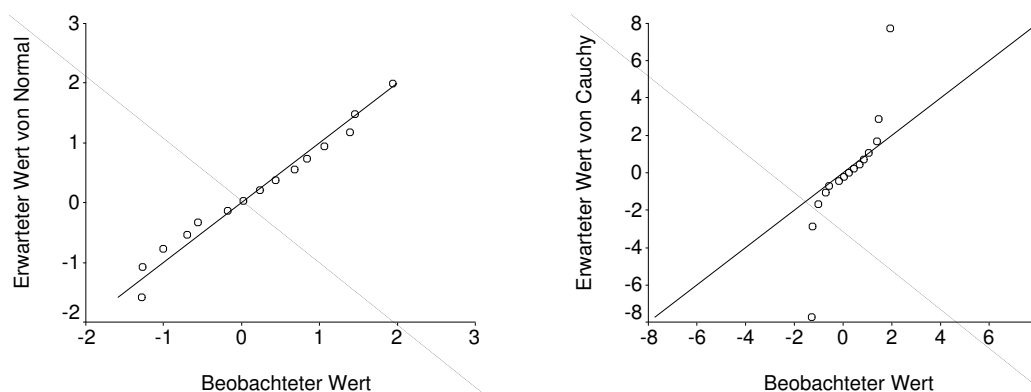
Der QQ-Plot ist ein Instrument der Explorativen Daten Analyse. Mit ihm können Verteilungsannahmen graphisch überprüft werden. Sei x_1, \dots, x_n eine Stichprobe und $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ die geordnete Stichprobe. Weiters sei F eine vorgegebene (invertierbare) Verteilungsfunktion.

Hypothese: x_1, \dots, x_n ist Stichprobe einer Verteilung mit Verteilungsfunktion F , d.h. x_1, \dots, x_n sind Realisationen von unabhängigen Zufallsvariablen X_1, \dots, X_n deren Verteilungsfunktion F ist.

Der QQ-Plot besteht aus den n Punkten $(x_{(i)}, u_i)$, $1 \leq i \leq n$, wobei

$$\begin{aligned} x_{(i)} & \quad \frac{i-1/2}{n} \text{ – tes empirisches Quantil,} \\ u_i = F^{-1}\left(\left(i - \frac{1}{2}\right)/n\right) & \quad \frac{i-1/2}{n} \text{ – tes theoretisches Quantil der } F\text{-Verteilung} \end{aligned}$$

Ist die Hypothese **richtig**, dann liegen die Punkte $(x_{(i)}, u_i)$ in der Nähe der ersten Mediane (man kann zeigen, daß $E(X_{(i)}) \approx u_i$ falls die Hypothese richtig ist). Ist sie **falsch**, dann werden die Punkte von der ersten Mediane stärker abweichen.



QQ-Plot von 15 simulierten Daten aus der $N(0, 1)$ -Verteilung. Im linken Bild wurde die zu testende theoretische Verteilung gleich der $N(0, 1)$ -Verteilung gewählt und im rechten Bild gleich der Cauchy-Verteilung.

Kapitel 2

Tests unter Normalverteilungsannahmen

2.1 Die Idee des statistischen Tests

Gegeben sei eine Stichprobe x_1, \dots, x_n , d.h. eine Realisation von n unabhängigen identisch verteilten Zufallsvariablen X_1, \dots, X_n . Von der Verteilungsfunktion F der X_i sei nur bekannt, daß sie aus einer vorgegeben Menge Ω von Verteilungsfunktionen stammt (z.B. aus der Menge aller Normalverteilungen mit unbekanntem μ und σ^2). Ein statistischer Test ist ein Verfahren um **Hypothesen** (Vermutungen) über die tatsächlich zugrundeliegende Verteilungsfunktion F zu untersuchen. Der Tester stellt dazu eine **Nullhypothese** H_0 ($H_0 \subseteq \Omega$) auf (z.B. H_0 : F ist eine Normalverteilung mit Mittel $\mu = 0$ und Varianz $\sigma^2 = 1$). Auf Grund der Stichprobe x_1, \dots, x_n stellt der Test entweder die Behauptung auf, daß die **Nullhypothese** zutrifft (d.h. $F \in H_0$), oder er stellt die Behauptung auf, daß die **Gegenhypothese** $F \in H_1 = \Omega \setminus H_0$ zutrifft (d.h. $F \notin H_0$). Ein solcher Test ist durch die Angabe eines **Ablehnungsbereiches** (oder **kritischen Bereiches**) $K \subseteq \mathbb{R}^n$ eindeutig beschrieben, wobei folgende Entscheidungsregel vereinbart wird

$$\begin{aligned} (x_1, \dots, x_n) \in K & \implies H_0 \text{ verwerfen (d.h. der Test behauptet } F \notin H_0) \\ (x_1, \dots, x_n) \notin K & \implies H_0 \text{ kann nicht verworfen werden} \\ & \text{(d.h. der Test behauptet } F \in H_0) \end{aligned}$$

Dabei können zwei verschiedene Fehlentscheidungen getroffen werden:

Fehler 1. Art: H_0 ist richtig, aber der Test verwirft H_0 .

Fehler 2. Art: H_0 ist falsch, aber der Test verwirft H_0 nicht.

Natürlich möchte man die Wahrscheinlichkeit des Auftretens beider Fehler möglichst klein halten. Diese Forderung läßt sich aber nicht erfüllen. Zum Beispiel legt bei den nachfolgend beschriebenen klassischen Tests die Wahrscheinlichkeit des Fehlers 1. Art den ganzen Test eindeutig fest. Damit ist auch die Wahrscheinlichkeit des Fehlers 2. Art festgelegt. Man gibt sich deshalb eine Zahl α mit $0 < \alpha < 1$ (das **Signifikanzniveau**, übliche Werte sind $\alpha = 0.01$ oder $\alpha = 0.05$) vor und legt den kritischen Bereich so fest, daß die Wahrscheinlichkeit des Fehlers 1. Art $\leq \alpha$ ist. Über den Fehler 2. Art hat man keine

Kontrolle! Daraus ergibt sich folgende für die Interpretation von Testergebnissen wichtige Konsequenz:

Bei einem statistischen Test ist die Entscheidung “verwerfe H_0 ” die einzige Entscheidung über deren Fehlerwahrscheinlichkeit eine präzise Aussage gemacht wird; sie ist kleiner gleich dem vorgegebenen Signifikanzniveau α .

Hat ein Test die Nullhypothese H_0 verworfen, dann kann man sich natürlich nicht sicher sein, daß die Gegenhypothese H_1 tatsächlich richtig ist. Aber man weiß, daß man diese Entscheidung mit einem Verfahren getroffen hat, das nur in $\alpha \cdot 100\%$ der Fälle eine falsche Entscheidung fällt (wenn es H_0 verwirft). Hat der Test die Nullhypothese nicht verworfen, ist keinerlei quantitative Aussage über Irrtumswahrscheinlichkeiten möglich. Daher vermeidet man Sprechweisen wie “die Nullhypothese H_0 wird angenommen” und sagt besser “ H_0 kann nicht verworfen werden”.

Der Einstichproben-Gauß-Test

Beispiel: Eine Maschine stellt zylinderförmige Aluminiumbolzen her, deren Durchmesser (in mm) (approximativ) durch eine $N(\mu, \sigma_0^2)$ -verteilte Zufallsvariable beschrieben wird. Die Maschine hat eine fixe bekannte Genauigkeit $\sigma_0 = 0.5$ mm, der mittlere Durchmesser μ kann eingestellt werden. Die Maschine wird auf $\mu = 20$ mm eingestellt und dann 20 Bolzen nachgemessen.

19.46 20.23 20.18 19.14 20.58 19.94 19.37 19.64 20.27 20.22
18.64 20.55 19.48 19.93 19.66 19.00 20.14 19.46 19.25 18.71

Ist die Maschine korrekt eingestellt?

In dem Beispiel wird angenommen, daß die obigen Zahlenwerte x_1, \dots, x_n , $n = 20$, eine Realisation von unabhängigen normalverteilten Zufallsvariablen X_1, \dots, X_n mit unbekanntem μ und **bekanntem** Varianz $\sigma_0^2 = 0.25$ sind. Die zu überprüfenden Hypothesen lauten (mit $\mu_0 = 20$)

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

Zur Festlegung des kritischen Bereiches K betrachte man folgende **Teststatistik**

$$Z = \frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}} = \sqrt{n} \frac{\bar{X} - \mu_0}{\sigma_0}.$$

Z hat folgende Eigenschaften:

- Ist H_0 **richtig**, dann ist zu erwarten, daß $|Z|$ im Mittel kleine Werte annimmt (beachte $E(Z) = 0$).
- Ist H_0 **falsch**, dann wird $|Z|$ im Mittel größere Werte annehmen (beachte $E(Z) = \sqrt{n}(\mu_1 - \mu_0)/\sigma_0 \neq 0$, wenn $E(X_i) = \mu_1 \neq \mu_0$ ist).

- Z enthält keine unbekanntenen Größen, kann also aus der Stichprobe berechnet werden.
- Wenn H_0 richtig ist, ist Z eine $N(0, 1)$ -verteilte Zufallsvariable.

Diese Eigenschaften legen es nahe, den Test wie folgt festzulegen:

$$\begin{aligned} |z| \geq c &\implies H_0 \text{ verwerfen,} \\ |z| < c &\implies H_0 \text{ kann nicht verworfen werden.} \end{aligned}$$

Dabei ist $c > 0$ ein noch festzulegender Parameter und z die Realisation der Zufallsvariable Z . Da z eine Funktion von x_1, \dots, x_n ist, bedeutet diese Festsetzung, daß der kritische Bereich K des Tests die spezielle Form $K = \{(x_1, \dots, x_n) \in \mathbb{R}^n : |z(x_1, \dots, x_n)| \geq c\}$ hat. Dabei wird c so festgelegt, daß die Wahrscheinlichkeit des Fehlers 1. Art gleich einem vorgegebenen Signifikanzniveau α ist

$$\begin{aligned} \alpha &= P(\text{Fehler 1. Art}) = P(H_0 \text{ verwerfen, wenn } H_0 \text{ richtig ist}) \\ &= P(|Z| \geq c, \text{ wenn } Z \text{ } N(0, 1)\text{-verteilt ist}) \\ &= 2(1 - \Phi(c)). \end{aligned}$$

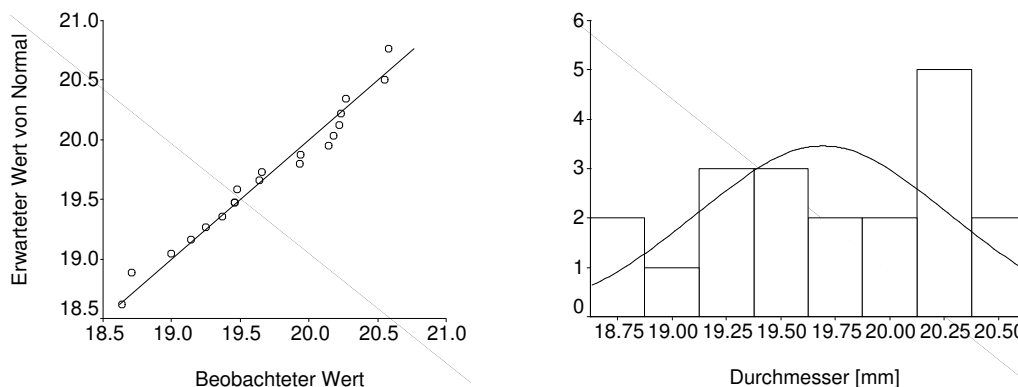
Daraus folgt

$$c = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) = u_{1-\frac{\alpha}{2}},$$

wobei $u_{1-\frac{\alpha}{2}}$ das $(1 - \frac{\alpha}{2})$ -Quantil der $N(0, 1)$ -Verteilung bezeichnet.

| | | | | |
|------------|-------|-------|-------|-------|
| α | 0.95 | 0.975 | 0.99 | 0.995 |
| u_α | 1.645 | 1.960 | 2.326 | 2.576 |

Zum Beispiel: Der QQ-Plot (mit der Hypothese einer Normalverteilung) zeigt, daß die Modellierung des Bolzendurchmessers durch eine normalverteilte Zufallsvariable gerechtfertigt ist. Die Interpretation des Histogramms ist nicht so einfach. Bei einem Stichprobenumfang von $n = 20$ sind aber die hier auftretenden Abweichungen von der Normalverteilungsdichte zu erwarten. Man vergleiche dazu auch Kapitel 3.



Es ist die Hypothese $H_0 : \mu = 20$ ($\mu_0 = 20$) mit Hilfe des Gauß-Testes zu untersuchen. Dazu berechnet man

$$n = 20, \quad \bar{x} = 19.693, \quad z = -2.746.$$

Wählt man $\alpha = 0.01$, wird wegen $|z| > u_{1-\alpha/2} = u_{0.995} = 2.576$ die Nullhypothese $\mu_0 = 20$ zum Signifikanzniveau 0.01 verworfen. Wir behaupten daher, daß die Maschine nicht korrekt eingestellt ist. Wir wissen zwar nicht, ob diese Behauptung richtig ist, aber wir stellen sie doch mit einer gewissen Überzeugung auf, da wir mit einem Verfahren zu diese Behauptung gekommen sind, das in 99% der Fälle die richtige Entscheidung fällt.

Der P-Wert oder Signifikanz-Wert

Der P-Wert liefert eine alternative Darstellung der Entscheidungsregel eines Tests, die in vielen Programmpaketen verwendet wird. Der **P-Wert** oder **Signifikanz-Wert** ist definiert als

$p(x_1, \dots, x_n)$ = kleinstes Signifikanzniveau, bei dem H_0 gerade noch verworfen wird.

Der Test nimmt dann folgende Form an:

$$\begin{aligned} p(x_1, \dots, x_n) \leq \alpha &\implies H_0 \text{ verwerfen,} \\ p(x_1, \dots, x_n) > \alpha &\implies H_0 \text{ kann nicht verworfen werden.} \end{aligned}$$

Bezeichnet z die Realisation von Z im Gauß-Test, dann ist $p = p(x_1, \dots, x_n)$ durch $|z| = u_{1-\frac{p}{2}}$ festgelegt, d.h. $\Phi(|z|) = 1 - \frac{p}{2}$ und daher

$$p = 2(1 - \Phi(|z|)).$$

Es gilt

$$|z| \geq u_{1-\frac{\alpha}{2}} \text{ ist äquivalent zu } p \leq \alpha.$$

Im obigen Beispiel erhält man $p = 2(1 - \Phi(2.746)) = 0.006$. Man wird also selbst bei einem Signifikanzniveau von 0.006 noch behaupten, daß die Maschine falsch eingestellt ist.

2.2 Einstichproben-Tests

In den folgenden Kapiteln werden klassische Testverfahren beschrieben. Dabei werden folgende Bezeichnungen verwendet

| Verteilung | Verteilungsfunktion | α -Quantil |
|------------------------|---------------------|---------------------|
| $N(0, 1)$ -Verteilung | $\Phi(x)$ | u_α |
| t_n -Verteilung | $F_{t_n}(x)$ | $t_{n;\alpha}$ |
| χ_n^2 -Verteilung | $F_{\chi_n^2}(x)$ | $\chi_{n;\alpha}^2$ |
| $F_{m,n}$ -Verteilung | $F_{m,n}(x)$ | $F_{m,n;\alpha}$ |

Bemerkung: Ist etwa X eine t_n -verteilte Zufallsvariable, dann gilt $P(X \leq x) = F_{t_n}(x)$, $P(X \leq t_{n;\alpha}) = \alpha$ und $t_{n;\alpha} = F_{t_n}^{-1}(\alpha)$.

Beim Einstichproben-Problem liegt eine Stichprobe x_1, \dots, x_n reeller Meßwerte vor. Wir setzen voraus, daß x_1, \dots, x_n Realisierungen von unabhängigen und identisch normalverteilten Zufallsvariablen X_1, \dots, X_n sind. Es soll der Mittelwert oder die Varianz dieser Verteilung getestet werden.

Gauß-Test (Mittelwerttest bei bekannter Varianz)

1. Voraussetzungen:

X_1, \dots, X_n unabhängig, $N(\mu, \sigma_0^2)$ -verteilt mit unbekanntem μ und bekanntem σ_0^2 .

2. Teststatistik:

$$Z = \frac{\bar{X} - \mu_0}{\sigma_0 / \sqrt{n}} \quad N(0, 1) - \text{verteilt, wenn } \mu = \mu_0$$

$$p = 2(1 - \Phi(|z|)) \quad \text{P-Wert für die zweiseitige Hypothese } H_0 : \mu = \mu_0$$

3. Hypothesen und Entscheidungsregeln:

| | | | |
|--------------------------------------------------------|----------------------------------------------|-------------------------------------------------------------------|---------------------------------------------------------------|
| H_0 | $\mu = \mu_0$ | $\mu \leq \mu_0$ | $\mu \geq \mu_0$ |
| H_1 | $\mu \neq \mu_0$ | $\mu > \mu_0$ | $\mu < \mu_0$ |
| H_0 verwerfen, wenn oder H_0 verwerfen, wenn | $ z \geq u_{1-\alpha/2}$ $p \leq \alpha$ | $z \geq u_{1-\alpha}$ $z \geq 0$ und $\frac{p}{2} \leq \alpha$ | $z \leq u_\alpha$ $z \leq 0$ und $\frac{p}{2} \leq \alpha$ |

Beispiel: siehe oben

t-Test (Mittelwerttest bei unbekannter Varianz)

1. Voraussetzungen:

X_1, \dots, X_n unabhängig, $N(\mu, \sigma^2)$ -verteilt mit unbekanntem μ und σ^2 .

2. Teststatistik:

$$T = \frac{\bar{X} - \mu_0}{S / \sqrt{n}} \quad t_{n-1} - \text{verteilt, wenn } \mu = \mu_0$$

$$p = 2(1 - F_{t_{n-1}}(|t|)) \quad \text{P-Wert für die zweiseitige Hypothese } H_0 : \mu = \mu_0$$

Dabei bezeichnet $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ die empirische Varianz.

3. Hypothesen und Entscheidungsregeln:

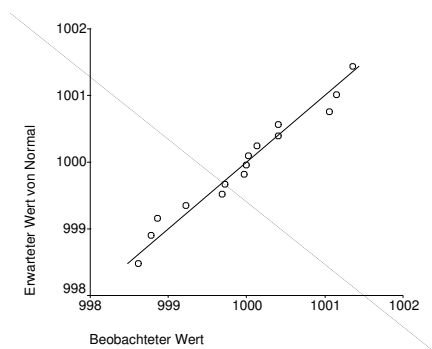
| | | | |
|--------------------------------------------------------|--------------------------------------------------|-----------------------------------------------------------------------|---------------------------------------------------------------------|
| H_0 | $\mu = \mu_0$ | $\mu \leq \mu_0$ | $\mu \geq \mu_0$ |
| H_1 | $\mu \neq \mu_0$ | $\mu > \mu_0$ | $\mu < \mu_0$ |
| H_0 verwerfen, wenn oder H_0 verwerfen, wenn | $ t \geq t_{n-1;1-\alpha/2}$ $p \leq \alpha$ | $t \geq t_{n-1;1-\alpha}$ $t \geq 0$ und $\frac{p}{2} \leq \alpha$ | $t \leq t_{n-1;\alpha}$ $t \leq 0$ und $\frac{p}{2} \leq \alpha$ |

Beispiel: Um die Genauigkeit eines neuen Meßgerätes zur Messung von Entfernungen im Gelände zu kontrollieren, wurde eine Strecke von genau 1000 Meter 15 mal vermessen.

1000.0 1001.1 998.6 999.7 999.2 999.7 998.9 1001.1 1001.4 998.8
 1000.0 1000.4 1000.0 1000.4 1000.1

Man prüfe die Hypothese, daß das Messgerät im Mittel die richtige Entfernung angibt.

Bei Meßfehlern ist fast immer die Annahme der Normalverteilung gerechtfertigt. Mit Hilfe des QQ-Plots überzeugt man sich, daß im konkreten Fall die Daten nicht gegen die Normalverteilungsannahme sprechen.



Unter der Annahme einer Normalverteilung ist die zweiseitige Hypothese $H_0 : \mu = 1000$ zu testen. Man erhält

$$n = 15 \quad \bar{x} = 999.960 \quad s = 0.850 \quad t = -0.182.$$

Wegen

$$p = 2(1 - F_{t_{14}}(0.182)) = 0.858 > \alpha$$

kann die Hypothese nicht verworfen werden. Die Daten geben keinen Hinweis darauf, daß das Meßgerät im Mittel falsche Werte liefern würde. Die alternative Entscheidungsregel führt wegen $t_{14;0.9725} = 2.145$ beim Signifikanzniveau $\alpha = 0.05$ zur selben Entscheidung.

χ^2 -Streuungstest

1. Voraussetzungen:

X_1, \dots, X_n unabhängig, $N(\mu, \sigma^2)$ -verteilt mit unbekanntem μ und σ^2 .

2. Teststatistik:

$$Y = \frac{(n-1)S^2}{\sigma_0^2}$$

χ_{n-1}^2 - verteilt, wenn $\sigma = \sigma_0$

$$p = \min(2F_{\chi_{n-1}^2}(y), 2(1 - F_{\chi_{n-1}^2}(y)))$$

P-Wert für die zweiseitige Hypothese

$H_0 : \sigma = \sigma_0$

Dabei bezeichnet $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ die empirische Varianz.

3. Hypothesen und Entscheidungsregeln:

| | | | |
|-----------------------|----------------------------------------------------------------------------|-----------------------------------------------------|-----------------------------------------------------|
| H_0 | $\sigma = \sigma_0$ | $\sigma \leq \sigma_0$ | $\sigma \geq \sigma_0$ |
| H_1 | $\sigma \neq \sigma_0$ | $\sigma > \sigma_0$ | $\sigma < \sigma_0$ |
| H_0 verwerfen, wenn | $y \leq \chi_{n-1;\alpha/2}^2$ oder $y \geq \chi_{n-1;1-\alpha/2}^2$ | $y \geq \chi_{n-1;1-\alpha}^2$ | $y \leq \chi_{n-1;\alpha}^2$ |
| H_0 verwerfen, wenn | $p \leq \alpha$ | $s^2 \geq \sigma_0^2$ und $\frac{p}{2} \leq \alpha$ | $s^2 \leq \sigma_0^2$ und $\frac{p}{2} \leq \alpha$ |

Beispiel: Im Beispiel zum t-Test auf Seite 12 wurde ein Meßgerät zur Entfernungsmessung getestet. Der t-Test ergab, daß man davon ausgehen kann, daß das Gerät im Mittel den richtigen Meßwert liefert. Es soll nun untersucht werden, ob das neue Meßgerät eine höhere Genauigkeit aufweist als ein herkömmliches Gerät, das auf diese Entfernung eine Streuung von $\sigma_0 = 1.3$ Meter aufweist.

Will man als Besitzer des alten Meßgerätes das neue Meßgerät kaufen, dann wird man die einseitige Nullhypothese $H_0 : \sigma \geq \sigma_0$ testen. Im Fall der Ablehnung von H_0 hat man dann die statistisch gesicherte Aussage (mit Fehlerwahrscheinlichkeit!), daß das neue Gerät eine höhere Genauigkeit aufweist als das alte Gerät. In unserem Beispiel erhält man

$$n = 15 \quad s = 0.850 \quad y = 5.985 \quad \frac{p}{2} = F_{\chi_{14}^2}(5.985) = 0.0331.$$

Der P-Wert zeigt, daß die Hypothese $\sigma \geq 1.3$ zum Signifikanzniveau $\alpha = 0.05$ verworfen wird. Der Käufer kann sich also ziemlich sicher sein, ein besseres Gerät zu erwerben als er schon besitzt. Zur gleichen Entscheidung gelangt man mit dem Quantil $\chi_{14;0.05}^2 = 6.571$.

In einer anderen Position ist der Erzeuger des neuen Gerätes, der mit dem Test die Überlegenheit seines Gerätes dokumentieren möchte. Er wird die Nullhypothese $H_0 : \sigma \leq \sigma_0$ testen. Im Fall der Ablehnung weiß er bei einem Signifikanzniveau von $\alpha = 0.05$, daß die für ihn ungünstige Fehlentscheidung “Der Test behauptet, daß das Gerät nicht besser ist, obwohl es in Wahrheit eine höhere Genauigkeit aufweist” nur in 5% der Fälle gefällt wird. Wegen $s^2 < 1.3^2$ kann diese Nullhypothese im vorliegenden Fall nicht verworfen werden.

Die beiden gegensätzlichen Risiken, die hier auftreten nennt man das **Konsumentenrisiko** und das **Produzentenrisiko**. Die Wahl der zu testenden Nullhypothese richtet sich danach, welches der beiden Risiken als schwerwiegender empfunden wird. Die schwerwiegendere Fehlentscheidung sollte immer der Fehler 1. Art sein, denn nur dieser hat eine kontrollierte Fehlerwahrscheinlichkeit α . Diese Problematik tritt bei allen Tests mit einseitiger Hypothese auf.

2.3 Zweistichproben-Tests

Es liegen zwei Stichproben x_1, \dots, x_m und y_1, \dots, y_n aus zwei **unabhängigen** Normalverteilungen vor. Die Mittelwerte oder Varianzen der beiden Verteilungen sollen verglichen werden.

Gauß-Test (Mittelwertvergleich bei bekannten Varianzen)

1. Voraussetzungen:

$X_1, \dots, X_m, Y_1, \dots, Y_n$ sind unabhängig,

X_1, \dots, X_m sind $N(\mu_X, \sigma_X^2)$ -verteilt mit unbekanntem μ_X und bekanntem σ_X^2 ,

Y_1, \dots, Y_n sind $N(\mu_Y, \sigma_Y^2)$ -verteilt mit unbekanntem μ_Y und bekanntem σ_Y^2 .

2. Teststatistik:

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} \quad N(0, 1) - \text{verteilt, wenn } \mu_X = \mu_Y$$

$$p = 2(1 - \Phi(|z|)) \quad \text{P-Wert für die zweiseitige Hypothese } H_0 : \mu_X = \mu_Y$$

3. Hypothesen und Entscheidungsregeln:

| H_0 | $\mu_X = \mu_Y$ | $\mu_X \leq \mu_Y$ | $\mu_X \geq \mu_Y$ |
|-----------------------|---------------------------|------------------------------------------|------------------------------------------|
| H_1 | $\mu_X \neq \mu_Y$ | $\mu_X > \mu_Y$ | $\mu_X < \mu_Y$ |
| H_0 verwerfen, wenn | $ z \geq u_{1-\alpha/2}$ | $z \geq u_{1-\alpha}$ | $z \leq u_\alpha$ |
| oder | | | |
| H_0 verwerfen, wenn | $p \leq \alpha$ | $z \geq 0$ und $\frac{p}{2} \leq \alpha$ | $z \leq 0$ und $\frac{p}{2} \leq \alpha$ |

t-Test (Mittelwertvergleich bei unbekanntem Varianzen)

1. Voraussetzungen:

$X_1, \dots, X_m, Y_1, \dots, Y_n$ sind unabhängig,

X_1, \dots, X_m sind $N(\mu_X, \sigma_X^2)$ -verteilt mit unbekanntem μ_X und σ_X^2 ,

Y_1, \dots, Y_n sind $N(\mu_Y, \sigma_Y^2)$ -verteilt mit unbekanntem μ_Y und σ_Y^2 .

2a. Teststatistik, wenn $\sigma_X^2 = \sigma_Y^2$ (exakter Test):

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{S_P^2 \left(\frac{1}{m} + \frac{1}{n}\right)}} \quad t_\nu - \text{verteilt mit } \nu = m + n - 2, \text{ wenn } \mu_X = \mu_Y$$

$$p = 2(1 - F_{t_\nu}(|t|)) \quad \text{P-Wert für die Hypothese } H_0 : \mu_X = \mu_Y$$

Dabei bezeichnet $S_P^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}$ die **gepoolte** empirische Varianz und F_{t_ν} die Verteilungsfunktion der t_ν -Verteilung.

2b. Teststatistik, wenn $\sigma_X^2 \neq \sigma_Y^2$ (approximativer Test):

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_X^2}{m} + \frac{s_Y^2}{n}}} \quad \text{approx. } t_\nu - \text{verteilt, wenn } \mu_X = \mu_Y$$

$$p = 2(1 - F_{t_\nu}(|t|)) \quad \text{P-Wert für die zweiseitige Hypothese } H_0 : \mu_1 = \mu_2$$

Dabei bezeichnen S_X^2 und S_Y^2 die empirischen Varianzen der beiden Stichproben und

$$\nu = \frac{\left(\frac{S_X^2}{m} + \frac{S_Y^2}{n}\right)^2}{\frac{1}{m-1} \left(\frac{S_X^2}{m}\right)^2 + \frac{1}{n-1} \left(\frac{S_Y^2}{n}\right)^2}.$$

3. Hypothesen und Entscheidungsregeln:

| | | | |
|-------------------------------|-------------------------------|------------------------------------------|------------------------------------------|
| H_0 | $\mu_X = \mu_Y$ | $\mu_X \leq \mu_Y$ | $\mu_X \geq \mu_Y$ |
| H_1 | $\mu_X \neq \mu_Y$ | $\mu_X > \mu_Y$ | $\mu_X < \mu_Y$ |
| H_0 verwerfen, wenn oder | $ t \geq t_{\nu;1-\alpha/2}$ | $t \geq t_{\nu;1-\alpha}$ | $t \leq t_{\nu;\alpha}$ |
| H_0 verwerfen, wenn | $p \leq \alpha$ | $t \geq 0$ und $\frac{p}{2} \leq \alpha$ | $t \leq 0$ und $\frac{p}{2} \leq \alpha$ |

Beispiel: Bei zwei Computerlaufwerken sind folgende Zugriffszeiten (in Millisekunden) gemessen worden:

| disk 1 (X) | | | | disk 2 (Y) | | | |
|------------|----|----|-----|------------|----|----|----|
| 59 | 92 | 54 | 102 | 71 | 38 | 47 | 53 |
| 73 | 60 | 73 | 75 | 63 | 48 | 41 | 68 |
| 74 | 84 | 47 | 33 | 40 | 60 | 44 | 39 |
| 61 | | | | 34 | 75 | 86 | |

Unterstützen die Daten die Hypothese, daß die mittleren Zugriffszeiten der beiden Laufwerke unterschiedlich sind?

Es ist die Hypothese $H_0 : \mu_X = \mu_Y$ zu testen. Aus den Daten berechnet man

$$\begin{aligned} n_1 &= 13 & \bar{x} &= 68.231 & s_X &= 18.660 \\ n_2 &= 15 & \bar{y} &= 53.800 & s_Y &= 15.808. \end{aligned}$$

Um den exakten Test durchführen zu können, hat man $\sigma_X = \sigma_Y$ anzunehmen. Diese Hypothese kann mit dem F -Test getestet werden (man vergleiche dazu das Beispiel zum F -Test auf Seite 16). Bei einem Signifikanzniveau von $\alpha = 0.05$ wird die Hypothese der Gleichheit der Varianzen nicht verworfen. Zur Durchführung des (exakten) t -Tests berechnet man

$$s_P^2 = \frac{12s_X^2 + 14s_Y^2}{26} = 295.163 \quad s_P = 17.180 \quad t = 2.217 \quad \nu = 26.$$

Daraus ergibt sich ein P-Wert von $p = 2(1 - F_{t_{26}}(2.217)) = 0.0356$. Die Hypothese der Gleichheit der Erwartungswerte wird also zum Niveau $\alpha = 0.05$ verworfen, d.h. es besteht ein statistisch signifikanter Unterschied zwischen den mittleren Laufzeiten der beiden Laufwerke. Zur gleichen Entscheidung gelangt man, wenn man $|t| = 2.217$ mit dem Quantil $t_{\nu,1-\alpha/2} = t_{26;0.975} = 2.056$ vergleicht.

Alternativ kann der (asymptotische) t -Test durchgeführt werden. Man erhält

$$\sqrt{\frac{s_X^2}{m} + \frac{s_Y^2}{n}} = 6.591 \quad t = 2.189 \quad \nu = 23.706.$$

Daraus erhält man den P-Wert $p = 2(1 - F_{t_{23.706}}(2.189)) = 0.0387$, sodaß auch dieser Test H_0 zum Signifikanzniveau $\alpha = 0.05$ verwirft. Zur selben Entscheidung gelangt man mit dem Quantil $t_{23.706;0.975} = 2.065$ (hat man nur Tabellen der t -Verteilung mit ganzen Freiheitsgraden ν zur Verfügung, muß ν gerundet werden: $t_{24;0.975} = 2.064$).

F-Test (Vergleich der Varianzen)

1. Voraussetzungen:

$X_1, \dots, X_m, Y_1, \dots, Y_n$ sind unabhängig,
 X_1, \dots, X_m sind $N(\mu_X, \sigma_X^2)$ -verteilt mit unbekanntem μ_X und σ_X^2 ,
 Y_1, \dots, Y_n sind $N(\mu_Y, \sigma_Y^2)$ -verteilt mit unbekanntem μ_Y und σ_Y^2 .

2. Teststatistik:

$$F = \frac{S_X^2}{S_Y^2} \qquad F_{m-1, n-1} \text{ - verteilt, wenn } \sigma_X = \sigma_Y$$

$$p = \min(2F_{m-1, n-1}(f), 2(1 - F_{m-1, n-1}(f))) \qquad \text{P-Wert für die zweiseitige Hypothese}$$

$$H_0 : \sigma_X = \sigma_Y$$

Dabei bezeichnen S_X^2 und S_Y^2 die empirischen Varianzen der beiden Stichproben, f die Realisation von F und $F_{m-1, n-1}$ die Verteilungsfunktion der $F_{m-1, n-1}$ -Verteilung¹.

3. Hypothesen und Entscheidungsregeln:

| | | | |
|-----------------------|------------------------------------------------------------------------------|--------------------------------------------------|--------------------------------------------------|
| H_0 | $\sigma_X = \sigma_Y$ | $\sigma_X \leq \sigma_Y$ | $\sigma_X \geq \sigma_Y$ |
| H_1 | $\sigma_X \neq \sigma_Y$ | $\sigma_X > \sigma_Y$ | $\sigma_X < \sigma_Y$ |
| H_0 verwerfen, wenn | $f \leq F_{m-1, n-1; \alpha/2}$ oder $f \geq F_{m-1, n-1; 1-\alpha/2}$ | $f \geq F_{m-1, n-1; 1-\alpha}$ | $f \leq F_{m-1, n-1; \alpha}$ |
| H_0 verwerfen, wenn | $p \leq \alpha$ | $S_X^2 \geq S_Y^2$ und $\frac{p}{2} \leq \alpha$ | $S_X^2 \leq S_Y^2$ und $\frac{p}{2} \leq \alpha$ |

Beispiel: Im Beispiel zum t-Test auf Seite 15 soll die Hypothese $H_0 : \sigma_X = \sigma_Y$ getestet werden. Es gilt $m = 13, n = 15$ und

$$s_X^2 = 348.196 \qquad s_Y^2 = 249.893 \qquad f = 1.393.$$

Daraus folgt $p = 2(1 - F_{12, 14}(1.393)) = 0.548$. Bei jedem Signifikanzniveau $\alpha < 0.548$ wird die Hypothese der Gleichheit der Varianzen nicht verworfen. Die alternative Entscheidungsregel führt mit $\alpha = 0.05$ wegen

$$F_{12, 14; 0.025} = 0.312 < f < 3.050 = F_{12, 14; 0.975}$$

zur selben Entscheidung.

¹Für ganze Zahlen $m, n \geq 1$ ist die $F_{m, n}$ -Verteilung wie folgt definiert: Ist X eine χ_m^2 -verteilte Zufallsvariable und Y eine von X unabhängige χ_n^2 -verteilte Zufallsvariable, dann ist der Quotient $\frac{X/m}{Y/n}$ per Definition $F_{m, n}$ -verteilt. Die Parameter m und n werden auch Freiheitsgrade der F -Verteilung genannt.

Kapitel 3

Der Chi-Quadrat Anpassungstest

Anpassungstests sind statistische Tests, die Verteilungsannahmen untersuchen. Eine mögliche Nullhypothese ist etwa :

“Den Daten liegt eine Normalverteilung zugrunde”.

Zur graphischen Überprüfung von Verteilungsannahmen haben wir schon den QQ-Plot kennengelernt. Der wesentliche Unterschied zum Chi-Quadrat Anpassungstest ist die Tatsache, daß diese (im Fall der Ablehnung der Nullhypothese) eine quantitative Aussage erlauben (Signifikanz). Für eine Ablehnung der Nullhypothese sind allerdings oft recht große Stichproben nötig.

Der Chi-Quadrat Anpassungstest ist seiner Konstruktion nach ein Test zur Untersuchung der Verteilung eines **diskreten** Merkmals. Wir werden anschließend sehen, wie er auch zur Analyse stetiger Merkmale herangezogen werden kann.

Beispiel 1: Mit einem Würfel wurde 100 mal gewürfelt. Dabei sind die Augenzahlen mit folgenden absoluten Häufigkeiten aufgetreten

| | | | | | | |
|------------------|----|----|----|----|----|----|
| Augenzahl j | 1 | 2 | 3 | 4 | 5 | 6 |
| Häufigkeit n_j | 14 | 13 | 16 | 18 | 26 | 13 |

Auf Seite 2 sind diese Daten durch ein Balkendiagramm und ein Kreisdiagramm graphisch dargestellt. Spricht die offensichtliche Unregelmäßigkeit der Daten für die Annahme, daß der Würfel verfälscht ist, oder sind diese Schwankungen auch bei einem fairen Würfel zu erwarten?

1. Voraussetzungen:

Es ist eine Stichprobe x_1, \dots, x_n aus einer diskreten Verteilung gegeben, d.h. Realisationen von unabhängigen diskreten Zufallsvariablen X_1, \dots, X_n , die nur die Werte ξ_1, \dots, ξ_k annehmen.

2. Nullhypothese H_0 :

Die Werte ξ_j treten mit Wahrscheinlichkeit p_j auf ($1 \leq j \leq k$). Dabei sind p_1, \dots, p_k fest vorgegeben mit $p_j \geq 0$ und $p_1 + \dots + p_k = 1$.

3. Teststatistik und P-Wert:

$$Y = \sum_{j=1}^k \frac{(N_j - np_j)^2}{np_j} \quad \text{approximativ } \chi_{k-1}^2\text{-verteilt, wenn } H_0 \text{ richtig}$$

$$p = 1 - F_{\chi_{k-1}^2}(y) \quad \text{approximativer P-Wert für } H_0$$

Dabei bezeichnet $N_j \leq n$ die Anzahl der X_i mit Wert ξ_j (die absolute Häufigkeit von ξ_j).

4. Entscheidungsregel:

$$H_0 \text{ verwerfen} \quad \iff \quad y \geq F_{\chi_{k-1;1-\alpha}^2} \quad \iff \quad p \leq \alpha.$$

Hier bezeichnet y die Realisation von Y .

Bemerkungen:

- Ist H_0 richtig, dann wird Y im Mittel (wegen $E(N_j) = np_j$) kleine Werte annehmen. Deshalb verwirft der Test H_0 , wenn Y große Werte annimmt.
- Die Verteilung von Y ist nur approximativ χ_{k-1}^2 . Damit die Approximation hinreichend gut ist, sollte $n \geq 30$ und $np_j \geq 5$ für alle j gelten (das ist nur eine Faustregel, die nicht strikt erfüllt sein muß).
- Um für ein **stetiges** Merkmal die Hypothese

H_0 : Die Verteilungsfunktion ist gleich einem vorgegebenen F

testen zu können, geht man wie folgt vor. Man unterteilt die reelle Zahlengerade in k disjunkte Intervalle $(-\infty, a_1] \cup (a_1, a_2] \cup \dots \cup (a_{k-1}, \infty)$ und testet mit dem Chi-Quadrat-Anpassungstest die durch diese Aufteilung definierte diskrete Verteilung

$$p_j = F(a_j) - F(a_{j-1}) \quad (1 \leq j \leq k, \quad a_0 = -\infty, \quad a_k = \infty).$$

N_j ist dann die Anzahl der beobachteten Werte die in $(a_{j-1}, a_j]$ liegen. Die Aufteilung der Zahlengerade in k Intervalle ist dabei im Prinzip beliebig. Wegen der Approximationsgüte sollte aber in jedem Intervall $np_j \geq 5$ gelten (muß nicht strikt erfüllt sein). Andererseits sollte die Aufteilung in Intervalle nicht zu grob sein (dies führt zu einem Informationsverlust).

- Hängt F noch von Parametern ab, die erst aus den Daten geschätzt werden müssen, dann verändert sich die Verteilung der Teststatistik. Y ist dann approximativ χ_{k-l-1}^2 -verteilt, wobei l die Anzahl der geschätzten Parameter bezeichnet ($l < k$). Ist etwa die Nullhypothese

H_0 : F ist eine Normalverteilung mit unbekanntem μ und σ^2

zu testen, dann sind μ und σ^2 durch \bar{x} und s^2 zu schätzen. Y ist approximativ χ_{k-3}^2 -verteilt. Daraus folgt: H_0 wird verworfen, wenn $y \geq F_{\chi_{k-3;1-\alpha}^2}$.

Zum Beispiel 1: Die Zufallsvariable X_i = “Ausgang des i -ten Würfelwurfes” nimmt nur die Werte $\xi_1 = 1, \dots, \xi_6 = 6$ an. Ist der Würfel fair, dann gilt $p_j = P(X_i = j) = \frac{1}{6}$ für $j = 1, \dots, 6$. Es ist also die Nullhypothese

$$H_0 : p_1 = \dots = p_6 = \frac{1}{6}$$

zu testen. Die folgende Tabelle stellt die notwendigen Größen zur Berechnung der Teststatistik zusammen

| j | n_j | p_j | np_j | $n_j - np_j$ |
|-----|-------|-------|--------|--------------|
| 1 | 14 | 1/6 | 16.667 | -2.667 |
| 2 | 13 | 1/6 | 16.667 | -3.667 |
| 3 | 16 | 1/6 | 16.667 | -0.667 |
| 4 | 18 | 1/6 | 16.667 | 1.333 |
| 5 | 26 | 1/6 | 16.667 | 9.333 |
| 6 | 13 | 1/6 | 16.667 | -3.667 |

Daraus erhält man

$$y = 7.4 \quad p = 1 - F_{\chi^2_5}(7.4) = 0.193 > 0.05$$

Die Nullhypothese kann also zum Signifikanzniveau $\alpha = 0.05$ nicht verworfen werden. Das bedeutet, daß die hier auftretenden Schwankungen der absoluten Häufigkeiten auch bei einem fairen Würfel zu erwarten sind. Tatsächlich sind die Daten durch Simulation einer exakten diskreten Gleichverteilung auf den Zahlen $1, \dots, 6$ entstanden.

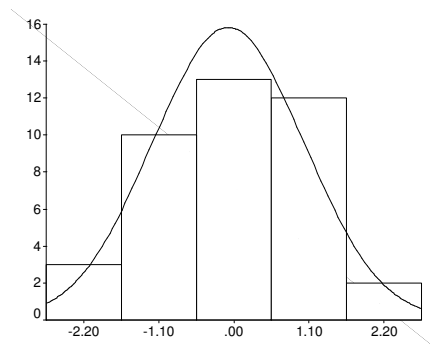
Beispiel 2: Es wurden 40 Werte aus einer $N(0, 1)$ -Verteilung simuliert.

-2.31 -1.59 0.41 0.49 0.57 -0.77 -0.22 -1.17 -1.27 -1.77
 -0.77 -0.51 0.14 0.73 1.03 -0.46 -1.16 0.76 0.95 -1.46
 0.68 -0.10 -0.35 0.34 1.42 -1.40 1.40 1.18 2.05 0.56
 -0.51 0.97 2.29 -1.87 0.62 -1.20 -0.80 -0.04 0.13 -0.30

Diese Daten sind auf Seite 1 mit Hilfe zweier Histogramme dargestellt. Wir wollen die Klasseneinteilung des zweiten Histogramms benutzen, um die Hypothese

$$H_0 : \text{Die Daten stammen aus einer Normalverteilung}$$

zu testen.



Die Parameter μ und σ^2 sind durch

$$\bar{x} = -0.00820 \quad s^2 = 1.225 \quad (s = 1.107)$$

zu schätzen. Die Wahrscheinlichkeit, daß eine $N(\mu, \sigma^2)$ -verteilte Zufallsvariable einen Wert in der Klasse $(a_{j-1}, a_j]$ annimmt ist

$$p_j = \Phi\left(\frac{a_j - \mu}{\sigma}\right) - \Phi\left(\frac{a_{j-1} - \mu}{\sigma}\right).$$

Die folgende Tabelle stellt die notwendigen Daten zur Durchführung des Chi-Quadrat-Anpassungstests zusammen.

| j | Klasse | n_j | p_j | np_j | $n_j - np_j$ |
|-----|--------------------|-------|-------|--------|--------------|
| 1 | $(-\infty, -1.65]$ | 3 | 0.069 | 2.76 | 0.24 |
| 2 | $(-1.65, -0.55]$ | 10 | 0.243 | 9.72 | 0.28 |
| 3 | $(-0.55, 0.55]$ | 13 | 0.381 | 15.24 | -2.24 |
| 4 | $(0.55, 1.65]$ | 12 | 0.240 | 9.60 | 2.4 |
| 5 | $(1.65, \infty)$ | 2 | 0.067 | 2.68 | -0.68 |

Daraus berechnet man den Wert der Teststatistik $y = 1.131$ und den P-Wert

$$p = 1 - F_{\chi^2_2}(1.131) = 0.432.$$

Die Nullhypothese kann also wie erwartet nicht verworfen werden.

Kapitel 4

Einfache Varianzanalyse

Die einfache Varianzanalyse (**ANOVA** - Analysis of Variance) ist eine Verallgemeinerung des t -Tests zum Vergleich der Mittelwerte zweier Meßreihen. Sie dient zum Vergleich der Mittelwerte von k Meßreihen. Mit ihr wird geprüft, ob die Annahme gleicher Erwartungswerte in allen k Meßreihen gerechtfertigt ist. Grundvoraussetzung dabei ist, daß die Meßreihen von unabhängigen normalverteilten Zufallsvariablen stammen, die alle die selbe (unbekannte) Varianz haben.

Beispiel: Zur Erzeugung eines bedruckten Stoffes stehen drei alternative Druckverfahren zur Verfügung. Die Qualität der Verfahren soll verglichen werden. Dazu werden mit den drei Verfahren jeweils 10 Stoffstücke bedruckt. Die bedruckten Stoffstücke werden anschließend in einer langsam rotierenden (mit Stoff ausgekleideten) Trommel so lange beansprucht, bis eine sichtbare Abreibung festgestellt werden kann. Die folgende Tabelle enthält die Dauer der Beanspruchung (in Stunden), bis eine sichtbare Abnutzung aufgetreten ist.

| | | | | | | | | | | |
|-------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Verfahren 1 | 148 | 76 | 393 | 520 | 236 | 134 | 55 | 166 | 415 | 153 |
| Verfahren 2 | 513 | 264 | 433 | 94 | 535 | 327 | 214 | 135 | 280 | 304 |
| Verfahren 3 | 335 | 643 | 216 | 536 | 128 | 723 | 258 | 380 | 594 | 465 |

Gibt es zwischen den drei Verfahren einen signifikanten Qualitätsunterschied?

Im allgemeinen liegen k Stichproben mit eventuell unterschiedlichen Stichprobenumfängen n_1, \dots, n_k vor :

$$x_{11}, \dots, x_{1n_1} \quad x_{21}, \dots, x_{2n_2} \quad \dots \quad x_{k1}, \dots, x_{kn_k} .$$

Es wird angenommen, daß die x_{ij} Realisationen von unabhängigen normalverteilten Zufallsvariablen X_{ij} ($1 \leq i \leq k$, $1 \leq j \leq n_i$) sind, die alle die selbe unbekanntes Varianz σ^2 besitzen (**Varianzhomogenität**). Es bezeichne μ_i den Erwartungswert der i -ten Meßreihe. Zu untersuchen ist die Nullhypothese

$$H_0 : \quad \mu_1 = \mu_2 = \dots = \mu_k .$$

Die Varianzanalyse beruht auf einer Zerlegung der in den Daten steckenden Gesamtvariabilität ($SS = \mathbf{S}$ um of \mathbf{S} quares) in einen Variabilitätsanteil, der durch unterschiedliche

Gruppenmittel erklärt wird ($SST = \mathbf{S}$ um of \mathbf{S} quares of \mathbf{T} reatment), und einen Variabilitätsanteil der durch Variabilität innerhalb der Gruppen erklärt wird ($SSE = \mathbf{S}$ um of \mathbf{S} quares of \mathbf{E} rrors, Restvariabilität). Es gilt

$$SS = SST + SSE,$$

wobei

$$\begin{aligned} SS &= \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2 && \text{Gesamtvariabilität} \\ SST &= \sum_{i=1}^k n_i (\bar{X}_i - \bar{X})^2 && \text{Variabilität zwischen den Gruppen} \\ SSE &= \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 && \text{Variabilität innerhalb der Gruppen.} \end{aligned}$$

Dabei bezeichnet \bar{X}_i das Mittel in der i -ten Gruppe und \bar{X} das Gesamtmittel, d.h. mit $n = n_1 + \dots + n_k$

$$\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}, \quad \bar{X} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} X_{ij}.$$

Zur Festlegung des kritischen Bereiches betrachte man folgende **Teststatistik**

$$F = \frac{\frac{1}{k-1} SST}{\frac{1}{n-k} SSE} = \frac{MST}{MSE}.$$

Dabei bezeichnet

$$MST = \frac{1}{k-1} SST \quad \text{und} \quad MSE = \frac{1}{n-k} SSE$$

die **mittleren Quadratsummen** von SST und SSE . F hat folgende Eigenschaften

- Es gilt

$$E(MSE) = \sigma^2 \quad \text{und} \quad E(MST) = \sigma^2 + \frac{1}{k-1} \sum_{i=1}^k n_i (\mu_i - \bar{\mu})^2,$$

wobei $\bar{\mu} = \frac{1}{n} \sum_{i=1}^k n_i \mu_i$ ein gewichtetes Mittel der μ_i bezeichnet. Daraus folgt: Ist H_0 **richtig**, dann gilt $\bar{\mu} = \mu_i$ für alle i und daher $E(MST) = \sigma^2$. In diesem Fall wird man erwarten, daß MST und MSE im Mittel nicht sehr verschieden sind (d.h. F ist im Mittel nicht groß). Ist H_0 hingegen **falsch**, dann ist $E(MST) > \sigma^2$ und man wird erwarten, daß MST im Mittel größer als MSE ist (d.h. F ist im Mittel groß).

- F ist $F_{k-1, n-k}$ -verteilt, wenn H_0 richtig ist.

Diese Eigenschaften legen es nahe, den Test wie folgt festzulegen:

$$\begin{aligned} f \geq c &\implies H_0 \text{ verwerfen,} \\ f < c &\implies H_0 \text{ kann nicht verworfen werden.} \end{aligned}$$

Dabei bezeichnet f die sich aus der Stichprobe ergebende Realisation der Zufallsvariable F . Die Konstante c setzt man so fest, daß das Signifikanzniveau des Tests gleich einem vorgegebenen α ist, d.h.

$$\alpha = P(\text{Fehler 1. Art}) = P(F \geq c, \text{ wenn } H_0 \text{ richtig}) = 1 - F_{k-1, n-k}(c).$$

Daraus folgt, daß c gleich dem $(1 - \alpha)$ -Quantil $c = F_{k-1, n-k; 1-\alpha}$ der F -Verteilung mit $k - 1$ und $n - k$ Freiheitsgraden ist. Alternativ verwirft der Test die Nullhypothese genau dann, wenn der P-Wert $p = 1 - F_{k-1, n-k}(f)$ kleiner gleich α ist.

Einfache Varianzanalyse (Zusammenfassung)

1. Voraussetzungen:

- Es liegen k Stichproben vom Umfang n_1, \dots, n_k vor :

$$x_{11}, \dots, x_{1n_1} \quad x_{21}, \dots, x_{2n_2} \quad \dots \quad x_{k1}, \dots, x_{kn_k}$$

- Die x_{ij} sind Realisationen von unabhängigen normalverteilten Zufallsvariablen X_{ij} ($1 \leq i \leq k$, $1 \leq j \leq n_i$), die alle die selbe unbekannte Varianz σ^2 besitzen.
- Die Zufallsvariablen der i -ten Meßreihe X_{i1}, \dots, X_{in_i} sind $N(\mu_i, \sigma^2)$ -verteilt mit unbekanntem μ_i (Erwartungswert der i -ten Meßreihe).

2. Nullhypothese:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$$

3. Teststatistik:

$$F = \frac{\frac{1}{k-1} SST}{\frac{1}{n-k} SSE} = \frac{MST}{MSE} \quad F_{k-1, n-k}\text{-verteilt, wenn } H_0 \text{ richtig.}$$

$$p = 1 - F_{k-1, n-k}(f) \quad \text{P-Wert für } H_0.$$

Dabei bezeichnet f die Realisation von F und $F_{k-1, n-k}$ die Verteilungsfunktion der F -Verteilung mit $k - 1$ und $n - k$ Freiheitsgraden.

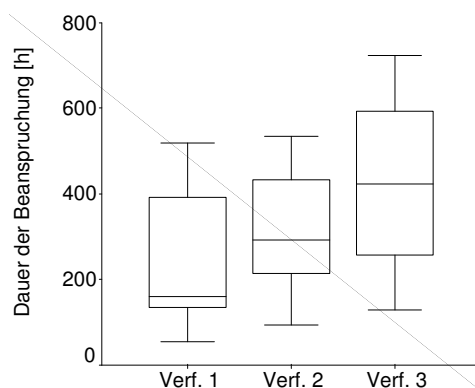
4. Entscheidungsregel:

$$H_0 \text{ verwerfen} \iff f \geq F_{k-1, n-k; 1-\alpha} \iff p \leq \alpha.$$

Die für die Varianzanalyse notwendigen Größen werden üblicherweise in einer sogenannten ANOVA-Tabelle zusammengefaßt (df steht für "degrees of freedom" = Freiheitsgrade):

| | SS | df | mittlere SS | F | P-Wert |
|-----------------------|-------|---------|-------------------------|-----------------------|--------|
| zwischen den Gruppen | SST | $k - 1$ | $MST = \frac{SST}{k-1}$ | $F = \frac{MST}{MSE}$ | p |
| innerhalb der Gruppen | SSE | $n - k$ | $MSE = \frac{SSE}{n-k}$ | | |
| gesamt | SS | $n - 1$ | | | |

Zum Beispiel: Die folgende Graphik, bestehend aus drei Boxplots, ermöglicht einen visuellen Vergleich der drei Verteilungen. Insgesonders sieht man, daß die Voraussetzung der Varianzhomogenität der Daten nicht unplausibel ist.



Es ist die Hypothese $H_0 : \mu_1 = \mu_2 = \mu_3$ zu testen. Die Mittel und Standardabweichungen in den einzelnen Gruppen sind

| | n_i | \bar{x}_i | s_i |
|-------------|-------|-------------|---------|
| Verfahren 1 | 10 | 229.6 | 158.196 |
| Verfahren 2 | 10 | 309.9 | 147.874 |
| Verfahren 3 | 10 | 427.8 | 196.818 |
| gesamt | 30 | 322.4 | 182.836 |

Die ANOVA-Tabelle lautet

| | SS | df | mittlere SS | F | P-Wert |
|-----------------------|------------|------|---------------|-------|--------|
| zwischen den Gruppen | 198772.467 | 2 | 99386.233 | 3.482 | 0.045 |
| innerhalb der Gruppen | 770670.900 | 27 | 28543.367 | | |
| gesamt | 969443.367 | 29 | | | |

Da der P-Wert kleiner als 0.05 ist, wird die Nullhypothese zum Signifikanzniveau $\alpha = 0.05$ verworfen, d.h. die drei Verfahren liefern signifikant unterschiedliche Druckqualität. Aus der Tabelle der Mittelwerte entnimmt man, daß das Verfahren 3 die besten Ergebnisse liefert.

Kapitel 5

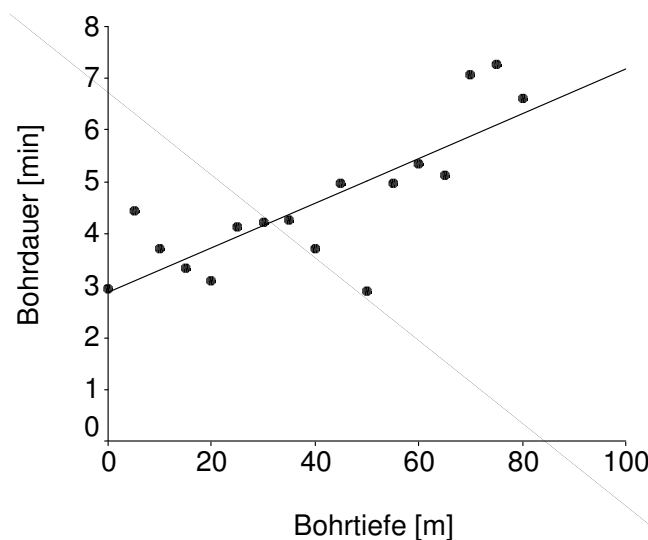
Lineare Regression

Beispiel: Es soll die Abhängigkeit einer bestimmten Methode zur Gesteinsbohrung von der Bohrtiefe untersucht werden. Dazu wird während einer Bohrung gemessen, wie lange man in verschiedenen Bohrtiefen x_i für den Vortrieb der Bohrung um 1 Meter benötigt. y_i bezeichne die in der Bohrtiefe x_i gemessene Bohrdauer in Minuten.

| | | | | | | | | | |
|-------|------|------|------|------|------|------|------|------|------|
| x_i | 0 | 5 | 10 | 15 | 20 | 25 | 30 | 35 | 40 |
| y_i | 2.94 | 4.45 | 3.71 | 3.34 | 3.10 | 4.13 | 4.23 | 4.27 | 3.71 |

| | | | | | | | | |
|-------|------|------|------|------|------|------|------|------|
| x_i | 45 | 50 | 55 | 60 | 65 | 70 | 75 | 80 |
| y_i | 4.97 | 2.90 | 4.97 | 5.35 | 5.12 | 7.07 | 7.27 | 6.61 |

Besteht eine approximativ lineare Beziehung zwischen der Bohrdauer und der Bohrtiefe?



Scatterplot mit Regressionsgerade

Das Modell der linearen Regression

Es liegen n Paare von Meßwerten (x_i, y_i) , $1 \leq i \leq n$, vor. Dabei sind die x_i **vorgegebene** nicht zufällige Werte, die y_i hingegen gemessene Variable deren Werte vom Zufall (im

obigen Beispiel etwa von der Inhomogenität des Gesteins) und von x_i abhängen. Man bezeichnet daher x als die **unabhängige Variable** oder **erklärende Variable** und y als die **abhängige Variable** oder **Responsevariable**.

Die y_i werden als Realisationen von Zufallsvariablen der Form

$$Y_i = a + bx_i + E_i$$

modelliert. Dabei sind a und b Parameter. Die E_i sind Zufallsvariable, die die Abweichung von der linearen Beziehung beschreiben. Es wird angenommen, daß sie folgende Eigenschaften besitzen:

- Die Zufallsvariablen E_i sind **unabhängig**.
- Die Zufallsvariablen E_i sind $N(0, \sigma^2)$ -verteilt mit einer unbekanntem aber für alle i gleichen Varianz σ^2 (**Varianzhomogenität**).

Das bedeutet, daß die Y_i unabhängige normalverteilte Zufallsvariable sind, die alle die gleiche Varianz σ^2 besitzen, und deren Erwartungswerte $\mu_i = E(Y_i)$ die lineare Beziehung

$$\mu_i = a + bx_i$$

erfüllen. Das gesamte Modell hat drei Parameter a , b und σ^2 . Diese müssen aus den Daten geschätzt werden.

Schätzung der Parameter

Die Schätzung der Parameter erfolgt nach der **Methode der kleinsten Quadrate**. Man wählt als Schätzer von a , b diejenigen Zahlen \hat{a} , \hat{b} , welche die Quadratsumme

$$\sum_{i=1}^n (y_i - a - bx_i)^2$$

minimieren. Die Lösung dieses Minimierungsproblems lautet

$$\hat{b} = \frac{s_{xy}}{s_x^2}, \quad \hat{a} = \bar{y} - \hat{b}\bar{x},$$

wobei \bar{x} und \bar{y} die arithmetischen Mittel der x_i und y_i bezeichnet und

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}),$$

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Durch \hat{a} und \hat{b} wird die **Regressionsgerade (Ausgleichsgerade)**

$$y = \hat{a} + \hat{b}x$$

festgelegt. Sie ist diejenige Gerade, die minimalen mittleren quadratischen Abstand von den Punkten (x_i, y_i) hat. Hat man \hat{a} und \hat{b} berechnet, wird die Varianz σ^2 durch

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{a} - \hat{b}x_i)^2 = \frac{1}{n-2} \sum_{i=1}^n r_i^2$$

geschätzt. Man bezeichnet die Größe $r_i = y_i - \hat{a} - \hat{b}x_i$ als das **Residuum** an der Stelle x_i .

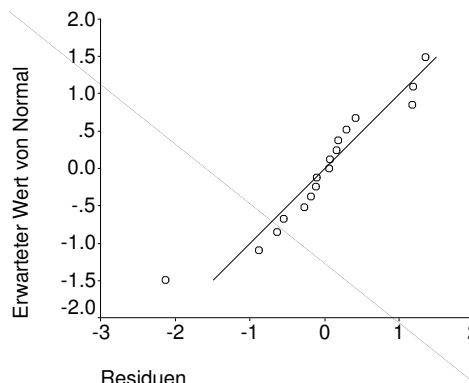
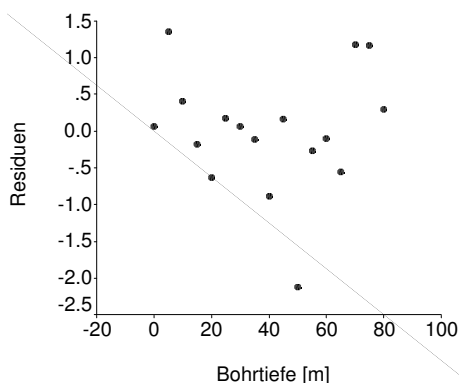
Zum Beispiel: Aus den Daten erhält man

$$n = 17 \quad \bar{x} = 40.000 \quad \bar{y} = 4.596 \quad s_x = 25.249 \quad s_y = 1.368 \quad s_{xy} = 27.413$$

und daraus die Schätzer

$$\hat{b} = 0.0430 \quad \hat{a} = 2.876 \quad \hat{\sigma}^2 = 0.738.$$

Den Schätzer $\hat{\sigma}^2$ entnimmt man bei vielen Programmpaketen der ANOVA Tabelle (siehe weiter unten). Der folgende Scatterplot der Residuen zeigt, daß die vorausgesetzte Varianzhomogenität annähernd erfüllt ist. Mit einem QQ-Plot der Residuen (gegen eine Normalverteilung) kann graphisch getestet werden, ob die Daten der Normalverteilungsannahme widersprechen.



Eigenschaften der Schätzer

- Alle drei Schätzer \hat{a} , \hat{b} , $\hat{\sigma}^2$ sind erwartungstreu.
- Die Teststatistiken

$$T_a = \frac{\hat{a} - a_0}{s.e.(\hat{a})} \quad \text{mit} \quad s.e.(\hat{a}) = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}}$$

$$T_b = \frac{\hat{b} - b_0}{s.e.(\hat{b})} \quad \text{mit} \quad s.e.(\hat{b}) = \frac{\hat{\sigma}}{\sqrt{(n-1)s_x^2}}$$

sind unter der Hypothese $a = a_0$ (bzw. $b = b_0$) beide t_{n-2} verteilt (*s.e.* steht für "standard error").

- Die Teststatistik

$$Y = \frac{(n-2)\hat{\sigma}^2}{\sigma_0^2}$$

ist χ_{n-2}^2 -verteilt, wenn $\sigma = \sigma_0$.

Konfidenzintervalle und Hypothesentests

Die Verteilungseigenschaften von T_a , T_b und Y ermöglichen die Konstruktion von Konfidenzintervallen mit vorgegebener Überdeckungswahrscheinlichkeit $1 - \alpha$.

Konfidenzintervall für a : $[\hat{a} - s.e.(\hat{a})t_{n-2;1-\alpha/2}, \hat{a} + s.e.(\hat{a})t_{n-2;1-\alpha/2}]$

Konfidenzintervall für b : $[\hat{b} - s.e.(\hat{b})t_{n-2;1-\alpha/2}, \hat{b} + s.e.(\hat{b})t_{n-2;1-\alpha/2}]$

Konfidenzintervall für σ^2 : $\left[\frac{(n-2)\hat{\sigma}^2}{\chi_{n-2;1-\alpha/2}^2}, \frac{(n-2)\hat{\sigma}^2}{\chi_{n-2;\alpha/2}^2} \right]$

Analog zum t -Test und χ^2 -Streuungstest können Hypothesen über die Parameter mit folgenden Entscheidungsregeln getestet werden:

| | | | |
|-----------------------|---------------------------------|---------------------------------|----------------------------------------------------------------------------|
| H_0 | $a = a_0$ | $b = b_0$ | $\sigma = \sigma_0$ |
| H_1 | $a \neq a_0$ | $b \neq b_0$ | $\sigma \neq \sigma_0$ |
| H_0 verwerfen, wenn | $ t_a \geq t_{n-2;1-\alpha/2}$ | $ t_b \geq t_{n-2;1-\alpha/2}$ | $y \leq \chi_{n-2;\alpha/2}^2$ oder $y \geq \chi_{n-2;1-\alpha/2}^2$ |

Zum Beispiel: Man erhält

$$\hat{a} = 2.877 \quad s.e.(\hat{a}) = 0.399 \quad \hat{b} = 0.0430 \quad s.e.(\hat{b}) = 0.00850 \quad \hat{\sigma}^2 = 0.738.$$

Die 95%-Konfidenzintervalle ($\alpha = 0.05$) für a , b , σ^2 ermitteln sich wegen $t_{15;0.975} = 2.131$ und $\chi_{15;0.025}^2 = 6.262$, $\chi_{15;0.975}^2 = 27.488$ zu

| Parameter | 95%-Konfidenzintervall |
|------------|------------------------|
| a | [2.027, 3.727] |
| b | [0.025, 0.061] |
| σ^2 | [0.403, 1.768] |

Die relativ großen Konfidenzintervalle resultieren aus dem geringen Stichprobenumfang von nur $n = 17$ Punkten. Die Hypothesen $H_0 : a = 0$ (d.h. $a_0 = 0$) und $H_0 : b = 0$ (d.h. $b_0 = 0$) werden wegen

$$t_a = 7.210 \quad t_b = 5.059$$

und $t_{15;0.995} = 2.947$ beide zum Signifikanzniveau $\alpha = 0.01$ verworfen.

Lineare Regression und ANOVA

Wie in der Varianzanalyse kann die in den Daten steckende Gesamtvariabilität SS in einen durch die Regression erklärten Variabilitätsanteil SSR und in einen durch die Schwankung der Daten um die Regressiongerade entstehenden Variabilitätsanteil SSE zerlegt werden. Es gilt

$$SS = SSR + SSE$$

wobei

$$\begin{aligned} SS &= \sum_{i=1}^n (Y_i - \bar{Y})^2 && \text{Gesamtvariabilität} \\ SSR &= \sum_{i=1}^n (\hat{a} + \hat{b}x_i - \bar{Y})^2 && \text{durch die Regression erklärte Variabilität} \\ SSE &= \sum_{i=1}^n (Y_i - \hat{a} - \hat{b}x_i)^2 && \text{Restvariabilität.} \end{aligned}$$

Die **Teststatistik**

$$F = \frac{SSR}{\frac{1}{n-2}SSE} = \frac{SSR}{\hat{\sigma}^2}$$

hat folgende Eigenschaften:

- Es gilt

$$E(\hat{\sigma}^2) = \sigma^2 \quad \text{und} \quad E(SSR) = \sigma^2 + b^2(n-1)s_x^2.$$

Daraus folgt: Ist die Hypothese $H_0 : b = 0$ **richtig**, dann gilt $E(SSR) = \sigma^2$ und man wird erwarten, daß SSR und $\hat{\sigma}^2$ im Mittel nicht sehr verschieden sind (d.h. F ist im Mittel nicht groß). Ist H_0 hingegen **falsch**, dann ist $E(SSR) > \sigma^2$ und man wird erwarten, daß SSR im Mittel größer als $\hat{\sigma}^2$ ist (d.h. F ist im Mittel groß).

- F ist $F_{1,n-2}$ -verteilt, wenn $b = 0$.

Dies liefert eine alternative (wie man zeigen kann sogar äquivalente) Möglichkeit die Hypothese $b = 0$ zu testen:

$$H_0 : b = 0 \text{ verwerfen} \iff f \geq F_{1,n-2;1-\alpha} \iff p \leq \alpha.$$

Dabei ist der P-Wert durch $p = 1 - F_{1,n-2}(f)$ festgelegt. Die Daten für diesen Test werden wie bei der Varianzanalyse in einer ANOVA-Tabelle zusammengefaßt. Viele Programmpakete liefern bei der Behandlung der linearen Regression automatisch folgende Tabelle.

| | SS | df | mittlere SS | F | P-Wert |
|------------|-------|---------|------------------------------------|-----------------------------|--------|
| Regression | SSR | 1 | SSR | $F = \frac{SSR}{SSE/(n-2)}$ | p |
| Rest | SSE | $n - 2$ | $\frac{SSE}{n-2} = \hat{\sigma}^2$ | | |
| gesamt | SS | $n - 1$ | | | |

Zum Beispiel: In unserem Zahlenbeispiel erhält man folgende ANOVA-Tabelle

| | SS | df | mittlere SS | F | P-Wert |
|------------|--------|------|---------------|--------|--------|
| Regression | 18.860 | 1 | 18.860 | 25.548 | 0.000 |
| Rest | 11.073 | 15 | 0.738 | | |
| gesamt | 29.933 | 16 | | | |

Ihr entnimmt man, daß die Hypothese $H_0 : b = 0$ hochsignifikant verworfen wird. Außerdem liefert sie den Schätzwert $\hat{\sigma}^2 = 0.738$.

Das Bestimmtheitsmaß

Zur Beurteilung der Güte der Anpassung wird meist das **Bestimmtheitsmaß**

$$r^2 = \frac{SSR}{SS}$$

herangezogen. Es gilt $0 \leq r^2 \leq 1$. r^2 gibt den relativen Variabilitätsanteil, der durch die lineare Regression erklärt wird, an. $r^2 = 0.7$ bedeutet, daß 70% der in den Daten steckenden Variabilität durch die Regression erklärt werden (nur 30% der Variabilität ist auf die Schwankung um die Regressionsgerade zurückzuführen). Bei der hier besprochenen (einfachen) linearen Regression kann gezeigt werden, daß r^2 gleich dem Quadrat des **Pearsonschen Korrelationskoeffizienten**

$$r = \frac{s_{xy}}{s_x s_y}$$

ist. Dieser wird in den meisten Programmpaketen automatisch angegeben (aus r kann der oft nicht angegebene Wert von s_{xy} bestimmt werden).

Zum Beispiel: Wegen $r^2 = 0.630$ sind 63% der Gesamtvariabilität durch die Regression erklärt. 37% der Gesamtvariabilität sind durch die Schwankung der Meßwerte um die Regressionsgerade begründet.

Prognoseintervall für $y(x)$

Ist das Regressionsmodell aufgestellt, kann es zur **Prognose** (Vorhersage) verwendet werden. Um eine Prognose des zu einer Stelle x gehörenden y -Wertes $y(x)$ zu machen, wird angenommen, daß dieser die Realisation einer Zufallsvariable $Y(x)$ ist, die $N(a + bx, \sigma^2)$ -verteilt und unabhängig von den Zufallsvariablen Y_1, \dots, Y_n ist. Unter diesen Voraussetzungen kann gezeigt werden, daß

$$\frac{Y(x) - \hat{a} - \hat{b}x}{s.e.(Y(x))} \quad \text{mit} \quad s.e.(Y(x)) = \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1)s_x^2}}$$

eine t_{n-2} -verteilte Zufallsvariable ist. Daraus folgt, daß das zufällige Intervall

$$[\hat{a} + \hat{b}x - s.e.(Y(x))t_{n-2;1-\alpha/2}, \hat{a} + \hat{b}x + s.e.(Y(x))t_{n-2;1-\alpha/2}]$$

mit Wahrscheinlichkeit $1 - \alpha$ die Realisation $y(x)$ von $Y(x)$ enthält. Dieses Intervall nennt man daher **Prognoseintervall** von $y(x)$ mit Überdeckungswahrscheinlichkeit $1 - \alpha$.

Achtung: Prognosen sollten nur für x -Werte durchgeführt werden, die innerhalb des Intervalls $[\min(x_i), \max(x_i)]$ liegen (also in dem Bereich in dem Messungen vorliegen). Führt man Prognosen für x -Werte durch, die (weit) außerhalb dieses Intervalls liegen, dann ist damit zu rechnen, daß der tatsächliche Fehler größer ist als angegeben. In diesem Fall hat man keine Daten zur Verfügung, die die Annahme der Korrektheit des linearen Regressionsmodells unterstützen.

Zum Beispiel: Es soll für die Bohrdauer in einer Bohrtiefe von 86 Metern ein 95%-Prognoseintervall aufgestellt werden. Wegen

$$x = 86 \quad \hat{a} + \hat{b}x = 6.574 \quad s.e.(Y(x)) = 0.967$$

und $t_{15;0.975} = 2.131$ erhält man als 95%-Prognoseintervall für $y(86)$ das Intervall

$$[6.574 - 2.060, 6.574 + 2.060] = [4.514, 8.634].$$

Prognoseintervall für $E(Y(x)) = a + bx$

Manchmal ist man auch an einer Prognose des Erwartungswertes $E(Y(x)) = a + bx$ interessiert. $\hat{Y}(x) = \hat{a} + \hat{b}x$ ist ein erwartungstreuer Schätzer von $E(Y(x)) = a + bx$. Zur Konstruktion eines Prognoseintervalls verwendet man die Tatsache, daß

$$\frac{\hat{Y}(x) - a - bx}{s.e.(\hat{Y}(x))} \quad \text{mit} \quad s.e.(\hat{Y}(x)) = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1)s_x^2}}$$

eine t_{n-2} -verteilte Zufallsvariable ist. Daraus folgt, daß das zufällige Intervall

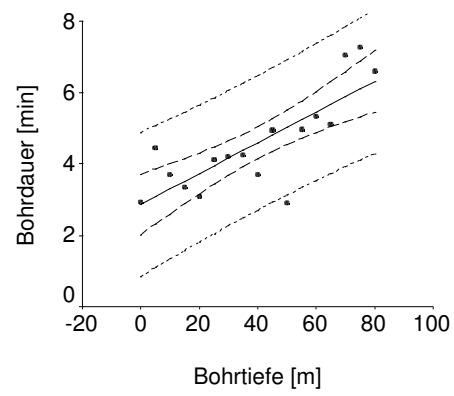
$$[\hat{a} + \hat{b}x - s.e.(\hat{Y}(x))t_{n-2;1-\alpha/2}, \hat{a} + \hat{b}x + s.e.(\hat{Y}(x))t_{n-2;1-\alpha/2}]$$

mit Wahrscheinlichkeit $1 - \alpha$ den Wert $E(Y(x)) = a + bx$ überdeckt. Dieses Intervall nennt man daher **Prognoseintervall** von $E(Y(x))$ mit Überdeckungswahrscheinlichkeit $1 - \alpha$.

Zum Beispiel: Es soll für die mittlere Bohrdauer in einer Bohrtiefe von 86 Metern ein 95%-Prognoseintervall aufgestellt werden. Wegen $s.e.(\hat{Y}(x)) = 0.444$ erhält man als 95%-Prognoseintervall für $E(Y(86))$ das Intervall

$$[6.574 - 0.946, 6.574 + 0.946] = [5.628, 7.520].$$

Man beachte, daß dieses Intervall als Prognoseintervall für den Erwartungswert von $Y(x)$ deutlich kleiner ist als das Prognoseintervall für den individuellen Wert $y(x)$.



95% Prognoseintervalle (\cdots) für den Bohrdauer $y(x)$ und 95% Prognoseintervalle ($- - -$) für die mittlere Bohrdauer $E(Y(x))$.