Hypothesentests

Wissenschaftliche Methodik I

Björn Voß

Institut für Biomedizinische Genetik - RNA Biologie & Bioinformatik

Skalenniveaus von Merkmalen

1. **Nominalskala** Merkmalsausprägungen sind Namen oder Kategorien (z.B. Haarfarbe, Religion)
2. **Ordinalskala** Ausprägungen können geordnet werden (z.B. Schulnoten, Qualitätsnorm für Äpfel)
3. **Intervallskala** Abstände zwischen Ausprägungen sind aussagekräftig (z.B. Temperatur in Celsius, Jahreszahlen)
4. **Verhältnisskala** Quotienten zwischen Ausprägungen können interpretiert werden (z.B. Temperatur in K, Alter in Jahren)

Diskrete/ Qualitative Merkmale

Stetige/ Quantitative Merkmale

Diskret vs. Stetig

Diskrete und stetige Merkmale werden mit unterschiedlichen statistischen Methoden untersucht!

	Diskret	Stetig
Kennzahlen	Häufigkeiten	Erwartungswert, Varianz
Plot	Kuchen-, Stabdiagramm	Histogramm
Test	χ^2-Anpassungstest	t-Test

Interaktionen:

	Diskret	Stetig
Diskret	Kontingenztafel, Mosaikplot, χ^2-Koeffizient und χ^2-Test	F, Boxplot, ANOVA
Stetig	F, Boxplot, ANOVA	Korrelation, Streudiagramm, Regression

Hypothesen

Bei einer Untersuchung will man in der Regel eine inhaltliche, nach wissenschaftlichen Kriterien aufgestellte Hypothese auf ihre Gültigkeit in der untersuchten Population hin untersuchen.

Aus diesen **inhaltlichen** Hypothesen werden **statistische** Hypothesen über die Verteilung (und deren Parameter) eines interessierenden Merkmals generiert. Dies ermöglicht nun eine mathematische Überprüfung und Entscheidung über die inhaltliche Hypothese zu treffen.

Inhaltliche Hypothesen

Arten von inhaltlichen Hypothesen

- Unterschieds- und Zusammenhangshypothesen:
  - doppelte Düngermenge $\Rightarrow$ veränderter Ertrag?
  - veränderter Ertrag $\Rightarrow$ wegen unterschiedlicher Sonneneinstrahlung?
- Ungerichtete und gerichtete Hypothesen
  Pflanzenschutzmittel verändert Ertrag vs. erhöht Ertrag
- Unspezifische und spezifische Hypothesen
  Düngung erhöht den Ertrag vs. Düngung erhöht den Ertrag um 10%
- Äquivalenzhypothesen
  Ertrag mit Dünger A unterscheidet sich um höchtens 10% vom Ertrag mit Dünger B

Hypothesentests

Es ist oft von Interesse Vermutungen über einen Parameter oder eine Verteilung in der Grundgesamtheit zu überprüfen.

Die Vermutung wird in Bezug auf die Grundgesamtheit aufgestellt – überprüft wird sie jedoch unter Verwendung einer Stichprobe. Ob von der Stichprobe auf die Grundgesamtheit geschlossen werden kann, muss man mit Methoden der Wahrscheinlichkeitstheorie berechnen.

Tests, die keine genaueren Annahmen über die Verteilung der Zufallsvariablen $X_1, \ldots, X_n$ machen, heißen **nicht-parametrisch**. Werden Annahmen über den Verteilungstyp gemacht, so heißen die Tests **parametrisch**.

Alternativ- und Nullhypothese

Um inhaltliche Hypothesen überprüfen zu können, muss man sie in statistische Hypothesen übersetzen. Zu diesem Zweck wird ein statistisches Hypothesenpaar formuliert das aus der **Alternativ**- und der **Nullhypothese** besteht.

Die **Alternativhypothese** wird dabei so formuliert, dass sie die inhaltliche Hypothese in Form von Annahmen über die Merkmalsverteilung und Verteilungsparameter widerspiegelt.

Beispiel:
Aus der inhaltlichen Hypothese, dass die doppelte Düngermenge den Ertrag beeinflusst folgt die Alternativhypothese:

$$H_1 : \mu_D \neq \mu_0$$

wobei μ_D : mittlerer Ertrag bei doppelter Düngermenge und μ_0 : mittlerer Ertrag bei normaler Düngung.

Statistische Nullhypothese

Aus der statistischen Alternativhypothese (H_1) ergibt sich direkt die Nullhypothese (H_0). Sie behauptet, dass die zur Alternativhypothese komplementäre Aussage richtig ist.

Beispiel:
Aus der inhaltlichen Hypothese, dass die doppelte Düngermenge den Ertrag beeinflusst folgt die Alternativhypothese:

$$H_1 : \mu_D \neq \mu_0$$

wobei μ_D : mittlerer Ertrag bei doppelter Düngermenge und μ_0 : mittlerer Ertrag bei normaler Düngung.

Die sich ergebende Nullhypothese lautet daher:

$$H_0 : \mu_D = \mu_0$$

Da i.A. mehr Informationen oder zumindest begründete Annahmen über die Nullhypothese vorliegen (z.B. bekannter Mittelwert, Verteilungstyp) besteht das grundsätzliche Vorgehen beim statistischen Testen darin, die Gültigkeit der Nullhypothese zu untersuchen.

Hierfür werden Teststatistiken verwendet, deren Verteilung bei Gültigkeit der Nullhypothese bekannt ist. Die Prüfung beinhaltet nun sich die Vereinbarkeit von Messdaten und Teststatistik anzuschauen. Ist diese nicht gegeben, so wird im Umkehrschluss auf die Gültigkeit der Alternativhypothese geschlossen.

Wichtig:
Ergibt die Prüfung, dass die Messdaten und die Teststatistik miteinander vereinbar sind, man sagt „die Nullhypothese kann nicht verworfen werden“, dann ist das **kein** Beweis für die Gültigkeit der Nullhypothese!

Fehlentscheidungen

Bei einem statistischen Testproblem H_0 gegen H_1 und einem geeigneten statistischen Test spricht man von einem

Fehler 1. Art , wenn H_0 verworfen wird, obwohl H_0 wahr ist
Fehler 2. Art , wenn H_0 beibehalten wird, obwohl H_1 wahr ist

Es gibt folgende Ausgänge bei einem statistischen Test:

	Entscheidung für H_0	Entscheidung für H_1
H_0 wahr	richtig	falsch Fehler 1. Art (α-Fehler)
H_1 wahr	falsch Fehler 2. Art (β-Fehler)	richtig

Ein statistischer Test heißt **Test zum Signifikanzniveau** α (wobei $0 < \alpha < 1$) oder Signifikanztest, falls:

$$P(H_1 \text{ annehmen } | H_0 \text{ wahr}) \leq \alpha$$

d.h.

$$P(\text{Fehler 1. Art}) \leq \alpha$$

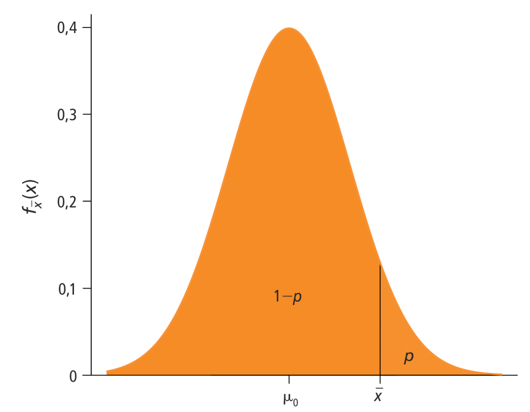
Typische Werte für das Signifikanzniveau α sind $0.05, 0.01, 0.001$.

Im Falle einer Ablehnung der Nullhypothese sagt man, dass das Ergebnis statistisch signifikant zum Niveau α sei.

Die Wahrscheinlichkeit für einen Fehler 2. Art kann man meist nicht kontrollieren. Diese Ungleichbehandlung der Fehler 1. und 2. Art ist der Grund dafür, dass die zu sichernde Behauptung als Alternativhypothese formuliert wird – und nicht umgekehrt.

Der p-Wert

Der **p-Wert** oder die **Überschreitungswahrscheinlichkeit** ist definiert als die Wahrscheinlichkeit, unter H_0 den beobachteten Prüfgrößenwert oder einen in Richtung der Alternative extremeren Wert zu beobachten.



Dichtefunktion

Der Inhalt der Fläche rechts von $\bar{x}$ entspricht dem p-Wert und gibt die Wahrscheinlichkeit für $X \geq \bar{x}$ unter H_0 an.

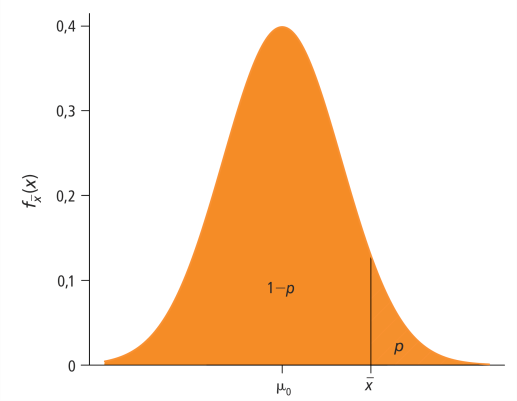
Einseitige und zweiseitige Fragestellungen

Bei der Beurteilung eines statistischen Testergebnisses muss man zwischen gerichteten und ungerichteten Hypothesen unterscheiden.

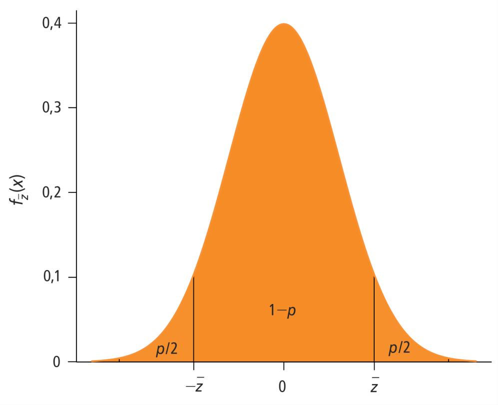
Ungerichtete inhaltliche Hypothesen resultieren in zweiseitigen Alternativhypothesen (z.B. $H_1 : \mu_D \neq \mu_0$), da negative als auch positive Abweichungen gleichermaßen zu berücksichtigen sind. Bei gerichteten inhaltlichen Hypothesen ergeben sich einseitige Alternativhypothesen (z.B. $H_1 : \mu_D > \mu_0$).

p-Wert bei ein- bzw. zweiseitigen Tests

einseitig



zweiseitig



Statistische Signifikanz

Mit dem Signifikanzniveau α wird festgelegt, mit welcher Wahrscheinlichkeit man bei dem durchgeführten Test die Nullhypothese **fälschlicherweise** ablehnen möchte (Fehler 1. Art). Den Vergleich von p-Wert und Signifikanzniveau α kann man nun folgendermaßen bewerten:

- $p > \alpha$: H_0 kann nicht verworfen werden $\Rightarrow$ Ablehnung von H_1
- $p \leq \alpha$: H_0 kann verworfen werden $\Rightarrow$ Annahme von H_1

Der Wert der Teststatistik für den gilt:

$$p_c = \alpha \text{ (einseitig) bzw. } p_{|c|} = \frac{\alpha}{2} \text{ (zweiseitig)}$$

wird als **kritischer Wert c** bezeichnet und er ergibt sich aus den Quantilen der der Nullhypothese zugrundeliegenden Verteilung, also z.B. dem $z_{0.95}$-Quantil der Standardnormalverteilung.

Prinzipien des Testens

1. Schritt: Quantifizierung der Fragestellung

2. Schritt: Formulierung der Modellannahmen

3. Schritt: Festlegung der Null- und Alternativhypothese

4. Schritt: Wahl des Signifikanzniveaus

5. Schritt: Wahl einer Prüfgröße (Teststatistik), die in der Lage ist, zwischen H_0 und H_1 zu differenzieren. Bestimmung der Verteilung der Prüfgröße unter der Nullhypothese. Konstruktion des Ablehnungsbereichs.

6. Schritt: Berechnung des Wertes der Prüfgröße für die konkrete Stichprobe

7. Schritt: Testentscheidung

Prinzipien des Testens

1. Schritt: Quantifizierung der Fragestellung

2. Schritt: Formulierung der Modellannahmen

3. Schritt: Festlegung der Null- und Alternativhypothese

4. Schritt: Wahl des Signifikanzniveaus

5. Schritt: Wahl einer Prüfgröße (Teststatistik), die in der Lage ist, zwischen H_0 und H_1 zu differenzieren. Bestimmung der Verteilung der Prüfgröße unter der Nullhypothese. Konstruktion des Ablehnungsbereichs.

6. Schritt: Berechnung des Wertes der Prüfgröße für die konkrete Stichprobe

7. Schritt: Testentscheidung

Parametrische Tests
Sie setzen voraus, dass die Verteilung des untersuchten Merkmals in der Population bekannt ist, also die Art der Verteilung (normalverteilt, exponential, . . .) und die zugehörigen Parameter (also z.B. μ, σ^2, λ, . . .). Bekanntester Vertreter ist der *t*-Test nach Student.

Nicht-parametrische Tests
Kommen zum Einsatz wenn nicht genügend Informationen über die „echte" Verteilung der Merkmale vorliegen. Weiterhin können sie auch auf ordinal- oder nominalskalierte Daten angewandt werden. Häufig werden diese Tests zur Überprüfung des Verteilungstyps angewendet, z.B. der Kolmogorow-Smirnow-Test.

Da jedoch parametrische Tests trotz Verletzung ihrer Annahmen häufig eine bessere *Power* bieten als nicht-parametrische, kommen letztere eher selten zum Einsatz.

Einige parametrische Tests und ihre Voraussetzungen

Test	Test bezgl.	Voraussetzungen
eine Stichprobe		
Einstichproben Gauß-Test	Mittelwert	Normal oder ZGS*; σ^2 bekannt
Einstichproben t-Test	Mittelwert	Normal oder ZGS; σ^2 **nicht** bekannt
zwei unabhängige Stichproben		
Zweistichproben Gauß-Test	Mittelwert	Normal oder ZGS; σ^2 bekannt & gleich
Zweistichproben t-Test	Mittelwert	Normal oder ZGS; σ^2 unbekannt, aber gleich
Welch-Test	Mittelwert	Normal oder ZGS; σ^2 unbekannt & ungleich
mehrere unabhängige Stichproben		
ANOVA	Mittelwerte	Normalvert. und gleiche Varianzen

*Zentraler Grenzwertsatz; kurz: Summen beliebig verteilter Zufallsvariablen sind normalverteilt.

Überprüfung von Vorraussetzungen

Sind die Vorraussetzungen für einen statistischen Test nicht gegeben, so lässt sich zwar immer noch ein p-Wert berechnen, dessen Aussagekraft geht jedoch gegen null. Da während der Berechnung des Wertes der Teststatistik die Vorraussetzungen nicht überprüft werden, muss dies explizit vorab getan werden.

Beispiele häufiger Vorraussetzungen

- Unabhängigkeit (ist allermeist durch zufällige Auswahl gewährleistet)
- Normalverteilung
- Varianzhomogenität

Test auf Normalverteilung

- Grafisch mit einem Normal-Quantil-Plot
- Mit dem Kolmogorow-Smirnow-Test

Sind meine Daten normalverteilt?
Um das (graphisch) zu überprüfen kann man die Häufigkeitsverteilung der Beobachtungen einer Variable X direkt mit einer Normalverteilung vergleichen.
Einfacher ist es die Quantile der Häufigkeitsverteilung mit den Quantilen der Standardnormalverteilung zu vergleichen, also einen sog. **Normal-Quantil-Plot** anzufertigen.

$x_{(1)}, \dots, x_{(n)}$	geordnete Stichprobe
$z_{\frac{1}{n}}, \dots, z_{\frac{n}{n}}$	Quantile der Standardnormalverteilung
	oder besser
$z_{\frac{1-0.5}{n}}, \dots, z_{\frac{n-0.5}{n}}$	Quantile der Standardnormalverteilung

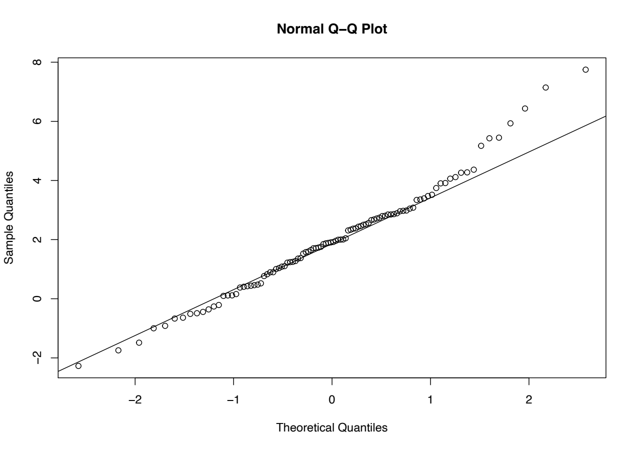
Normal-Quantil-Plots

Der Normal-Quantil-Plot besteht aus den Punkten

$$(z_{(1)}, x_{(1)}), \ldots, (z_{(n)}, x_{(n)})$$

im z-x-Koordinatensystem.
Ist die empirische Verteilung der Beobachtungen annähernd normalverteilt, so liegen die Punkte $(z_{(i)}, x_{(i)})$ des NQ-Plots nahe an oder auf der Winkelhalbierenden $z \propto x$.

Daten normalverteilt?




```
## Normal−Quantil−Plot in R
qqnorm(x)
qqline(x)
```

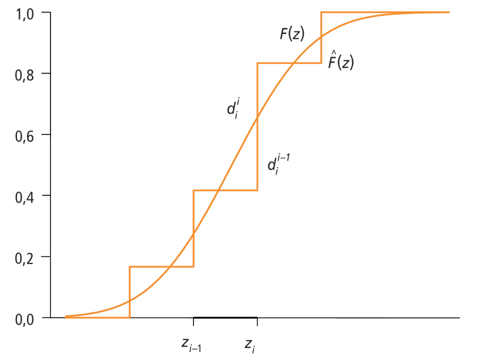
Kolmogorow-Smirnow-Test auf Normalverteilung

Gegeben seien die aufsteigend geordeneten Werte $x_1, \ldots, x_n$ mit dem arithmetischen Mittel $\bar{x}$ und der Standardabweichung s einer ZV X. Die Nullhypothese lautet:

$$H_0 : X \sim N(\mu, \sigma^2), \qquad \hat{\mu} = \bar{x}, \qquad \hat{\sigma} = s$$

Für den Test wird die standardisierte ZV Z mit $z_i = \frac{x_i - \bar{x}}{s}, i = (1, \ldots, n)$ berechnet und deren relativen Summenhäufigkeiten $H^{rel}(z_i) = \frac{H(z_i)}{n}$ bestimmt, wobei $H(z_i)$ die Anzahl der Werte mit $z \leq z_i$ bezeichnet.

$H^{rel}(z_i)$ ist eine Treppenfunktion vergleichbar mit der Verteilungsfunktion der Standardnormalverteilung $\phi(z)$.



Der interessierende Wert ist die größte Distanz $d_{max} = \max(d_i^j, d_i^{j-1})$ zwischen $H^{rel}(z_i)$ und $\phi(z)$.

Werte der Teststatistik D_n ergeben sich aus $d_n = d_{max} \cdot \sqrt{n}$ und folgen einer Kolmogorov-Smirnof-Verteilung mit Lilliefors-Korrektur.

```
> install.packages("nortest")
> library(nortest)
> lillie.test(rnorm(30,2,3))
        Lilliefors
            (Kolmogorov−Smirnov)
            normality test
data:   rnorm(30, 2, 3)
D = 0.090286, p−value = 0.7699
```

$p > 0.05 \Rightarrow$ NV kann sein

Daten nicht normalverteilt – Was nun?

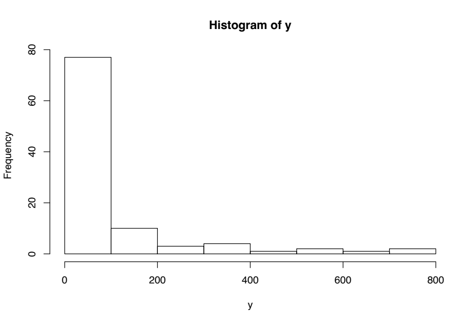
Transformation
Man kann Daten durch eine geeignete Transformation so umwandeln, dass die transformierten Daten normalverteilt sind. Auf diese transformierten Daten darf man dann alle statistischen Methoden für normalverteilte Daten, also z.B. t-Test, anwenden.

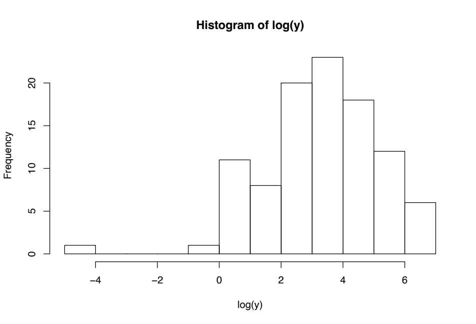
Welche Transformationen?
Es dürfen nur Transformationen durchgeführt werden, die die Reihenfolge der Daten nicht verändert. Mögliche Funktionen:

- Logarithmierung
- Wurzel
- Kehrwert
- Addition einer Konstanten (z.B. wichtig als Vorbereitung für Logarithmierung)

Transformation

Ob und welche Transformation geeignet ist, kann nur empirisch (anschauen der Daten, ausprobieren) ermittelt werden.






```
> lillie.test(y)
    Lilliefors
        (Kolmogorov−Smirnov)
        normality test
data:   y
D = 0.28015, p−value < 2.2e−16
```

```
> lillie.test(log(y))
    Lilliefors
        (Kolmogorov−Smirnov)
        normality test
data:   log(y)
D = 0.0602, p−value = 0.4994
```

Der zentrale Grenzwertsatz
$X_1, \ldots, X_n$ seien unabhängig identisch verteilte Zufallsvariablen mit

$$E(X_i) = \mu \quad \text{und} \quad \textbf{Var}(X_i) = \sigma^2$$

Dann konvergiert die Verteilungsfunktion $F_n(z) = P(Z_n \leq z)$ der standardisierten Summe

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma}$$

für $n \to \infty$ an jeder Stelle $z \in \mathbb{R}$ gegen die Verteilungsfunktion $\phi(z)$ der Standardnormalverteilung:

$$F_n(z) \to \phi(z) \quad (n \to \infty)$$

Zentraler Grenzwertsatz

In Worten:
Unabhängig wie die Verteilung von $X_1, \ldots, X_n$ ist, solange sie die gleiche Verteilung haben, ist ihre Summe für große n ($n \to \infty$) normalverteilt.

Bei den meisten Hypothesentests wird der Mittelwert untersucht, also die Summe $\frac{1}{n}\sum_{i=1}^{n} X_i$. Der zentrale Grenzwertsatz garantiert, dass die Verteilung des Mittelwertes gegen eine Normalverteilung konvergiert. Ist n also groß genug, so gilt die Normalverteilungsannahme immer mehr.

Faustregel: Ist $n > 30$ so gilt der Test approximativ, d.h. der Fehler 1. Art ist näherungsweise gleich α für $\mu = \mu_0$.

Sowohl der t-Test als auch die Varianzanalyse (ANOVA) fordern, dass die Varianz der Stichproben gleich ist, d.h. **Varianzhomogenität** besteht.

Ein robuster (gegenüber Abweichungen von der Normalverteilung) und daher weit verbreiteter Test auf Varianzhomogenität ist der **Levene-Test**.

Die Grundidee des Tests ist:
Die Varianz beruht auf Unterschieden der Messwerte zum arithmetischen Mittel. Unterschiede in den Varianzen lassen sich also durch Unterschiede in den mittleren betragsmäßigen Abweichungen der Messwerte zu den Mittelwerten belegen. Gibt es einen signifikanten Unterschied bei den mittleren betragsmäßigen Abweichungen zum Stichprobenmittelwert zwischen den Stichproben?

Der eigentliche statistische Test erfolgt auf den Differenzen in Form eines t-Tests oder einer Varianzanalyse. Diese fordern allerdings wiederum eine Normalverteilung, die im Falle der Differenzen nur approximativ gegeben ist.

Als Alternative kann der Levene-Test auch mittels eines nicht-parametrischen Tests (z.B. dem U-Test) durchgeführt werden.

Wie bei anderen statistischen Tests auch, kann der Levene-Test einseitig als auch zweiseitig durchgeführt werden, wobei letzteres im Falle der Varianzhomogenität sinnvoller ist, da sowohl geringere als auch größere Varianz einen Unterschied bedeutet.

Test auf Varianzhomogenität – Beispiel

```
> library(car)
> x <- rnorm(100,mean=3,sd=2)
> y <- rnorm(100,mean=3,sd=4)
> group <- as.factor(c(rep(1, length(x)), rep(2, length(y))))
> leveneTest(c(x,y),group)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value    Pr(>F)
group   1  37.783 4.276e-09 ***
      198
> y <- rnorm(100,mean=3,sd=2.2)
> leveneTest(c(x,y),group)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value  Pr(>F)
group   1  4.5741 0.03368 *
      198
> y <- rnorm(100,mean=3,sd=2.1)
> leveneTest(c(x,y),group)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   1  0.4456 0.5052
      198
> y <- rnorm(100,mean=3,sd=2.0)
> leveneTest(c(x,y),group)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   1  1.4258 0.2339
      198
```

Welcher Test für welche Daten?

Entscheidungsbaum nach Bortz†

1. Ist die Variable nicht stetig skaliert?
   → Falls ja, dann nicht-parametrisch testen. STOP.
2. Eine grafische Überprüfung der Voraussetzungen durchführen. Sind die Testvoraussetzungen deutlich verletzt?
   → Falls ja, dann prüfen, ob man mit einer Variablentransformation die Verletzung beheben kann. Macht eine entsprechende Transformation keinen Sinn, dann nicht-parametrisch testen. STOP.
3. Sind Testverzerrungen aufgrund der Stichprobencharakteristika zu erwarten?
   → Falls ja, dann nicht-parametrisch testen. STOP.
   → Sonst parametrisch testen.
4. Wird die Alternativhypothese H_1 angenommen?
   → Falls ja, dann die Alternativhypothese H_1 annehmen. STOP.
5. Überprüfung der Voraussetzungen des Tests mittels entsprechender Tests. Ist mindestens eine Voraussetzungen nicht erfüllt?
   → Falls ja, dann die Nullhypothese H_0 beibehalten. STOP.
6. Zusätzlich nicht-parametrisch testen. Wird das Ergebnis des parametrischen Test bestätigt?
   → Falls ja, dann die Nullhypothese H_0 beibehalten. STOP.
7. Es wird die Alternativhypothese H_1 angenommen. STOP.

† Jürgen Bortz, Christof Schuster: Statistik für Human- und Sozialwissenschaftler. 7. Auflage. Springer, Berlin 2010, ISBN 978-3-642-12769-4.

Einige statistische Tests

Spezielle Tests

„Statistische Tests gibt es wie Sand am Meer“. Wir beschränken uns auf einige Testverfahren zu ausgewählten Standardproblemen.

1. **Einstichprobentest**
   Untersuchung einer Verteilung eines eindimensionalen Merkmals:
   H_0: Die zu erwartende Miete in einem Wohnviertel beträgt 8 Euro$/m^2$.
   H_0: Der Würfel ist fair.
2. **Zweistichprobentest**
   Vergleich von Parametern aus zwei Populationen
   H_0: Das Mietniveau in den Wohnvierteln A und B ist gleich.
   H_0: Der Blutdruck ist vor und nach einer Therapie gleich.
3. **Mehrere Stichproben**
   Varianzanalyse (ANOVA) $\rightarrow$ nächste 2 Wochen (Prof. Heyer)!
4. **Zusammenhanganalyse**
   H_0: Geschlecht und Parteipräferenz sind unabhängig
   H_0: Mietpreis $\propto$ Wohnfläche $\rightarrow$ letzte Woche (Prof. Heyer)!

Beispiel – Mietspiegel
Die Quadratmetermiete für Wohnungen in Stadt A unter 50 m^2, die nach 1983 gebaut wurden, soll untersucht werden. Eine Teilstichprobe von $n = 11$ Wohnungen ergab:

i	1	2	3	4	5	6	7	8	9	10	11
x_i	13.22	6.81	10.22	14.03	8.04	10.16	9.43	13.07	13.63	5.05	11.63

In der Stadt B liegt der Durchschnittswert bei 8 Euro$/m^2$. Es soll überprüft werden, ob der Quadratmeterpreis in Stadt A signifikant größer ist als der Vergleichswert aus Stadt B.

Die Quadratmetermieten werden als normalverteilt angesehen. Der Erwartungswert μ ist der interessierende Parameter, σ sei **nicht bekannt**.

Einstichproben-t-Test

Seien $X_1, \ldots, X_n$ unabhängig $N(\mu, \sigma^2)$-verteilte Zufallsvariablen. Wir betrachten folgende Testprobleme über den Parameter μ:

- $H_0 : \mu = \mu_0$ gegen $H_1 : \mu \neq \mu_0$
- $H_0 : \mu \geq \mu_0$ gegen $H_1 : \mu < \mu_0$
- $H_0 : \mu \leq \mu_0$ gegen $H_1 : \mu > \mu_0$

Bei vorgegebenem Signifikanzniveau α und mit der Teststatistik

$$T = \frac{\bar{X}_n - \mu_0}{\sqrt{S_n^2/n}} = \sqrt{n}\frac{\bar{X}_n - \mu_0}{S_n} \quad \text{(Beachte: } T \sim t_{n-1}, \text{ falls } \mu = \mu_0\text{)}$$

wird die Nullhypothese abgelehnt, falls

- $|T| > t_{n-1,1-\alpha/2}$
- $T < -t_{n-1,1-\alpha}$
- $T > t_{n-1,1-\alpha}$

Wir führen den t-Test zum Signifikanzniveau $\alpha = 0.05$ durch.
Hypothese:

$$H_0 : \mu \leq 8 \leq \mu_0 \quad \text{gegen} \quad H_1 : \mu > 8$$

Teststatistik:

$$T = \frac{\bar{X}_n - \mu_0}{\sqrt{S^2/n}}$$

Der kritische Wert zum Niveau $\alpha = 0.05$ ist

$$t_{n-1,1-\alpha} = t_{10,0.95} = 1.8125$$

Berechnung des kritischen Wertes in R:

```
> qt(0.95, df=10)
[1] 1.812461
```

Berechnung des Wertes der Teststatistik:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{11}(13.22 + 6.81 + \ldots + 11.63) = 10.4809,$$

$$\sum_{i=1}^{n} x_i^2 = (13.22^2 + \ldots + 11.63^2) = 1296.587,$$

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2\right) = \frac{1}{10}(1296.5871 - 11 \cdot 10.4809^2) = 8.8245$$

$$t = \sqrt{n}\frac{\bar{x} - \mu_0}{\sqrt{s^2}} = \sqrt{11}\frac{10.4809 - 8}{\sqrt{8.8245}} = 2.77$$

Testentscheidung: Da

$$t = 2.77 > 1.8125$$

ist, wird die Nullhypothese abgelehnt. Die Preise in Stadt A sind signifikant höher als 8.

Einstichproben-t-Test mit R

```
> x <- c(13.22,6.81,10.22,14.03,8.04,10.16,9.43,13.07,13.63,5.05,11.63)
> t.test(x,mu=8,alternative="greater")

        One Sample t-test

data:  x
t = 2.7699, df = 10, p-value = 0.009895
alternative hypothesis: true mean is greater than 8
95 percent confidence interval:
 8.857557      Inf
sample estimates:
mean of x
 10.48091
```

Beispiel: Autopreise
US-Behörden werfen japanischen Autoherstellern vor, ihre Autos in Japan teurer zu verkaufen als in den USA, also die US-Verkäufe zu subventionieren. Ein Ökonom sammelt Verkaufspreise (in Tausend US-\$) von 50 Wagen in den USA und 30 in Japan.

Wir bezeichnen mit $x_1, \ldots, x_{50}$ die Preise in den USA und mit $y_1, \ldots, y_{30}$ die in Japan. Es haben sich folgende Werte ergeben:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = 16.596, \quad s_X = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} = 1.981$$

$$\bar{y} = \frac{1}{m} \sum_{i=1}^{m} y_i = 17.250, \quad s_Y = \sqrt{\frac{1}{m-1} \sum_{i=1}^{m} (y_i - \bar{y})^2} = 1.865$$

Es ist in der Tat $\bar{x} < \bar{y}$. Kann dies Zufall sein, oder müssen wir niedrigere Verkaufspreise in den USA annehmen?

Statistisches Modell:

X Verkaufspreis in den USA, Y Verkaufspreis in Japan.
Zu vergleichen sind

- $E(X) = \mu_X$: Durschnittspreis in den USA
- $E(Y) = \mu_Y$: Durschnittspreis in Japan

Also betrachten wir deren Differenz:

$$\Delta := \mu_X - \mu_Y$$

Seien $n = 50$ und $m = 30$. Wir betrachten

$$X_1, \ldots, X_n \sim N(\mu_X, \sigma_X^2) \quad \text{und} \quad Y_1, \ldots, Y_m \sim N(\mu_Y, \sigma_Y^2)$$

und nehmen an, dass $X_1, \ldots, X_n$ und $Y_1, \ldots, Y_m$ stochastisch unabhängig sind.

Da die $X_1, \ldots, X_n$ bzw. $Y_1, \ldots, Y_m$ jeweils eigene Parameter haben, spricht man hier von einem **Zweistichprobenproblem**.

Schätzen von Δ mittels:

$$\hat{\Delta} = \bar{X} - \bar{Y}.$$

Für diesen Schätzer gilt:

$$E(\hat{\Delta}) = E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_X - \mu_Y$$

$$\mathbf{Var}(\hat{\Delta}) = \mathbf{Var}(\bar{X} - \bar{Y}) = \mathbf{Var}(\bar{X}) + \mathbf{Var}(\bar{Y}) = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}$$

Der Schätzer ist als Linearkombination von unabhängigen normalverteilten Zufallsvariablen wieder normalverteilt:

$$\hat{\Delta} = \bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right)$$

$$\Rightarrow Z = \frac{\hat{\Delta} - E(\hat{\Delta})}{\sqrt{\mathbf{Var}(\hat{\Delta})}} = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim N(0, 1)$$

Problem: σ_X^2 und σ_Y^2 sind im Beispiel und im Allgemeinen unbekannt.

Es gibt drei Lösungsansätze:

- Unbekannte, aber gleiche Varianzen
  $\Rightarrow$ Zweistichproben-t-Test
- Approximatives Vorgehen bei großen Stichproben und beliebigen Varianzen
  $\Rightarrow$ Approx. Zweistichproben-Gauß-Test
- Korrektur der Anzahl der Freiheitsgrade bei unbekannten, beliebigen Varianzen und beliebigen Stichprobengrößen
  $\Rightarrow$ t-Test mit Welch-Korrektur

Problem: σ_X^2 und σ_Y^2 sind im Beispiel und im Allgemeinen unbekannt.

Es gibt drei Lösungsansätze:

- Unbekannte, aber gleiche Varianzen
  $\Rightarrow$ Zweistichproben-t-Test
- Approximatives Vorgehen bei großen Stichproben und beliebigen Varianzen
  $\Rightarrow$ Approx. Zweistichproben-Gauß-Test
- Korrektur der Anzahl der Freiheitsgrade bei unbekannten, beliebigen Varianzen und beliebigen Stichprobengrößen
  $\Rightarrow$ t-Test mit Welch-Korrektur

1. Lösungsansatz: Unbekannte, aber gleiche Varianzen

Annahme: $\sigma_X^2 = \sigma_Y^2 = \sigma^2$

Dann ist:

$$\mathbf{Var}(\hat{\Delta}) = \mathbf{Var}(\bar{X}) + \mathbf{Var}(\bar{Y}) = \frac{\sigma^2}{n} + \frac{\sigma^2}{m} = \left(\frac{1}{n} + \frac{1}{m}\right)\sigma^2$$

Und falls $\mu_X - \mu_Y = \delta_0$, kann gezeigt werden, dass

$$T = \frac{\bar{X} - \bar{Y} - \delta_0}{\sqrt{(\frac{1}{n} + \frac{1}{m})S_p^2}}$$

Hier wird die gemeinsame Varianz σ^2 geschätzt durch die **„gepoolte“ Schätzung der Varianz**

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

Diese Schätzung funktioniert, wenn in zwei Stichproben die gleiche Varianz herrscht, aber womöglich verschiedene Mittelwerte. Daher gehen zwei Freiheitsgrade verloren!

Zweistichproben-t-Test – unbekannte, aber gleiche Varianzen

Sei $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ und seien $X_1, \ldots, X_n \sim N(\mu_X, \sigma^2)$ und $Y_1, \ldots, Y_m \sim N(\mu_Y, \sigma^2)$. Außerdem seien $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ unabhängig. Wir betrachten folgende Testprobleme:

1. $H_0 : \mu_X - \mu_Y = \delta_0$ gegen $H_1 : \mu_X - \mu_Y \neq \delta_0$
2. $H_0 : \mu_X - \mu_Y \geq \delta_0$ gegen $H_1 : \mu_X - \mu_Y < \delta_0$
3. $H_0 : \mu_X - \mu_Y \leq \delta_0$ gegen $H_1 : \mu_X - \mu_Y > \delta_0$

Bei vorgegebenem Signifikanzniveau α und basierend auf der Teststatistik

$$T = \frac{\bar{X} - \bar{Y} - \delta_0}{\sqrt{(\frac{1}{n} + \frac{1}{m})S_p^2}}, \text{ wobei } S_p^2 = \frac{1}{n+m-2}\left((n-1)S_X^2 + (m-1)S_Y^2\right)$$

fällt die Entscheidung für H_1 im Testproblem,

1. falls $|t| > t_{n+m-2,1-\alpha/2}$
2. falls $t < -t_{n+m-2,1-\alpha}$
3. falls $t > t_{n+m-2,1-\alpha}$

Beispiel: Autopreise

Wir gehen davon aus, die Daten sind näherungsweise normalverteilt mit gleichen Varianzen – die beiden Schätzer sind etwa gleich groß.

Hypothesen: $H_0 : \mu_X - \mu_Y \geq \delta_0 = 0$ gegen $H_1 : \mu_X - \mu_Y < \delta_0 = 0$
Bei einem Niveau $\alpha = 0.05$ ergibt sich der kritische Wert

$$-t_{78,0.95} = -1.66$$

Teststatistik

$$s_p^2 = \frac{1}{n+m-2}\left((n-1)s_X^2 + (m-1)s_Y^2\right) = \frac{49 \cdot 1.981^2 + 29 \cdot 1.865^2}{49+29} = 3.7585$$

$$t = \frac{\bar{x} - \bar{y} - \delta_0}{\sqrt{(\frac{1}{n} + \frac{1}{m})s_p^2}} = \frac{16.596 - 17.250}{\sqrt{3.7585 \cdot (\frac{1}{50} + \frac{1}{30})}} = -1.4607$$

Da $t = -1.46 > -1.66 = -t_{78,0.95}$ wird H_0 nicht verworfen. Ein signifikanter Preisunterschied ist nicht nachweisbar.

Zweistichproben-t-Test (gleiche Varianz) mit R

Mit den Zahlen des obigen Beispiels:

```
> t.test(x, y, var.equal=T, alternative='less')

    Two Sample t-test

data: x and y
t = -1.4607, df = 78, p-value = 0.07405
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
    -Inf -0.2211939
sample estimates:
mean of x mean of y
  16.27413  17.24061
```

χ^2-Anpassungstest für kategoriale Merkmale

Beispiel: Parteipräferenzen
In einem Land mit drei Parteien wurde zuletzt vor vier Jahren gewählt. Eine Woche vor der aktuell anstehenden Wahl ergab eine Stichprobenbefragung vom Umfang $n = 500$ eine neue Verteilung:

	i	1	2	3
Letzte Wahl	π_i	0.40	0.35	0.25
Aktuelle Umfrage	N_i	210	190	100
rel. Häufigkeit	f_i	0.42	0.38	0.20

Hat sich die Wahlpräferenz gegenüber der letzten Wahl (signifikant) verändert?

Ein **Anpassungstest** untersucht, ob Daten zu einer bestimmten Verteilung passen. Hier ist es also ein Vergleich zweier diskreter Verteilungen, nämlich der Stimmenverteilung bei der letzten Wahl mit der Verteilung in den Daten, die aus der Stichprobenbefragung stammen.

Statistisches Modell

Sei X eine Zufallsvariable mit Träger $\{1, 2, 3\}$ und wahrer aber unbekannter Verteilung:

$$P(X = i) = p_i, \ i = 1, 2, 3$$

Die Hypothesen für den Test sind dann:

$$H_0 : p_1 = \pi_1 \text{ und } p_2 = \pi_2 \text{ und } p_3 = \pi_3 \quad \text{gegen} \quad H_1 : H_0 \text{ ist falsch}$$

Wir betrachten $N = 500$ unabhängige Kopien $X_1, \ldots, X_n$ und die Anzahl der Wähler der Stichprobe, die sich für die Partei i entschieden haben

$$N_i = \#\{X_j = i | j = 1, \ldots, n\} \Rightarrow N_i \sim \text{Bin}(500, p_i)$$

Die relativen Häufigkeiten $\hat{p}_i = N_i/n$ sind geeignete Schätzer für p_i.

Die Nullhypothese ist hierbei, dass sich die Verteilung der Stimmen im Vergleich zur vorherigen Wahl nicht verändert hat. Dann sollte unter der Nullhypothese die folgende Summe klein sein:

$$\chi^2 = \sum_{i=1}^{3} \frac{(N_i - n\pi_i)^2}{n\pi_i}$$

Es gilt: "Große" Werte von χ^2 treten auf bei großen Abweichungen zwischen den Wahrscheinlichkeiten π_i und den relativen Häufigkeiten N_i/n. Bei großer Übereinstimmung sind die Werte von χ^2 hingegen "klein".

Die zugehörige χ^2-Verteilung hat als Parameter die Anzahl der Freiheitsgrade. Die N_i sind nicht unabhängig, da $N_3 = N - N_1 - N_2$, und somit ist die Anzahl der Freiheitsgrade $k - 1 = 2$.

χ^2-Anpassungstest

Seien $X_1, \ldots, X_n$ unabhängig und verteilt wie X, wobei X diskret mit Träger $\mathcal{T} = \{1, \ldots, k\}$. Wir betrachten die Hypothesen

$$\begin{aligned} & H_0 : P(X = i) = \pi_i, i = 1, \ldots, k \\ \text{gegen} \quad & H_1 : P(X = i) \neq \pi_i, \text{ für mindestens ein } i \in \mathcal{T} \end{aligned}$$

Betrachte die Teststatistik $\chi^2 = \sum_{i=1}^{k} \frac{(N_i - n\pi_i)^2}{n\pi_i}$.

Unter H_0 gilt approximativ $\chi^2 \sim \chi^2_{k-1}$, falls $n\pi_i \geq 5$ für alle i ist[‡].

Bei vorgegebenem α fällt die Entscheidung für H_1, falls $\chi^2 > \chi^2_{k-1,1-\alpha}$, wobei $\chi^2_{k-1,1-\alpha}$ das $(1-\alpha)$-Quantil der χ^2_{k-1}-Verteilung ist.

[‡]Es reicht: $n\pi_i \geq 1$ für alle i und $n\pi_i \geq 5$ für mind. 80% der i.

Beispiel: Parteipräferenz (Fortsetzung)

Hypothese:

$$
\begin{aligned}
& H_0 : P(X = i) = \pi_i, i = 1, \ldots, 3 \\
\text{gegen} \quad & H_1 : P(X = i) \neq \pi_i, \text{ für mindestens ein } i \in \mathcal{T}
\end{aligned}
$$

Voraussetzung: $n \cdot \pi_i \geq 5$? Ja, denn $200, 175, 125 \geq 5$

Teststatistik: $\chi^2 = \sum_{i=1}^{3} \frac{(N_i - n\pi_i)^2}{n\pi_i} \sim \chi^2_2$

Bei einem Signifikanzniveau von $\alpha = 0.05$ ergibt sich der kritische Wert:

$$c = \chi^2_{2,0.95} = 5.99$$

Wert der Teststatistik

$$\chi^2 = \frac{(210-200)^2}{200} + \frac{(190-175)^2}{175} + \frac{(100-125)^2}{125} = 6.79$$

Da $\chi^2 > 5.99$ wird H_0 verworfen, d.h. das Wahlverhalten hat sich signifikant geändert.

χ^2-Anpassungstest in R

```
> x <- c(210,190,100)
> y <- c(200,175,125)
> chisq.test(x,p=y/sum(y))
In chisq.test(x, p) : Chi-Quadrat-Approximation kann inkorrekt sein

    Chi-squared test for given probabilities

data:  x
X-squared = 6.7857, df = 2, p-value = 0.03361
```

Der p-Wert ist $0.03361 < 0.05$ und damit wird H_0 verworfen.