INDUKTIVE STATISTIK

SCHÄTZVERFAHREN 3
Punktschätzungen 3
Begriff der Punktschätzung 3
Kriterien zur Güte einer Schätzung 3
Spezielle Schätzfunktionen 3
Intervallschätzungen 4
Bedeutung eines Konfidenzintervalls 4
Konfidenzintervalle für einen Erwartungswert 5
Konfidenzintervalle für eine Wahrscheinlichkeit 6
Bemerkungen 7
Bedeutung des Stichprobenumfangs 7
PRINZIP EINES STATISTISCHEN TESTS 8
Durchführung eines Tests 8
Funktion eines statistischen Tests 8
Formulieren der Hypothesen 8
Fehlerarten 8
Stichprobenumfang 10
Testentscheidung und Konsequenzen 10
Basis der Testentscheidung 10
p-Wert und Konfidenzintervall 11
Manipulation des Testergebnisses 11
Multiples Testen 12
Klassifikation der Testmethoden 12
Statistische Tests 13
LAGETESTS 14
t-Tests 14
t-Test für eine Stichprobe 14
t-Test für zwei verbundene Stichproben 14
t-Test für zwei unverbundene Stichproben 15
Welch-Test 16
Voraussetzungen der t-Lagetests 16
Andere Anwendungen des t-Tests 17
Rangsummentests 17
Wilcoxon-Test für eine Stichprobe 17
Wilcoxon-Test für zwei verbundene Stichproben 18
U-Test von Mann und Whiney 19
Vergleich zwischen Rangsummentests und t-Tests 20
Vorzeichentests 20
Vorzeichentest für eine Stichprobe 20
Vorzeichentest für zwei verbundene Stichproben 21
Vergleich mit anderen Lagetests 21
Ausblick auf komplexere Methoden 22
Mehrstichprobentests 22
Multiple Methoden 22
TESTS ZUM VERGLEICH VON HÄUFIGKEITEN 23
Binomialtest für eine Stichprobe 23
Chi²-Tests 24
Chi²-Vierfelder-Test 24
Median-Test 25
Chi²-Test für k*l Felder 26

Assoziationsmaße für qualitative Merkmale 27
McNemar-Test.....28
Chi²-Anpassungstest 29
Logranktest.....30
Exakter Test nach Fisher 30
Logistische Regression 31

SCHÄTZVERFAHREN

Punktschätzungen

Begriff der Punktschätzung

- Punktschätzung - Verfahren, bei dem ein unbekannter Parameter durch einen einzigen Wert geschätzt wird
  - Funktionalparameter einer Grundgesamtheit werden durch die entsprechenden Kenngrößen einer zufälligen Stichprobe geschätzt
  - z.B. Mittelwert als Schätzwert für den Erwartungswert
  - z.B. Wahrscheinlichkeit geschätzt durch die relative Häufigkeit
- Schätzfunktion – Vorschrift, nach der aus den Daten einer Stichprobe n ein angenäherter Wert für den unbekannten Parameter berechnet wird
  - Schätzfunktion für den Erwartungswert

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- Schätzwerte – Werte, die die Schätzfunktion in Abhängigkeit von der jeweiligen Stichprobe annimmt

Kriterien zur Güte einer Schätzung

- Erwartungswert kann durch den Mittelwert optimal geschätzt werden
  - symmetrische Verteilung: Erwartungswert könnte durch empirischen Median geschätzt werden
- empirische Varianz, empirische Kovarianz wird durch n-1 dividiert

Güte einer Schätzung wird durch die 4 Kriterien beurteilt

Erwartungstreue

- Durchschnitt (Erwartungswert) aller theoretisch denkbaren Schätzwerte aus den Stichproben n sollen mit dem unbekannten Parameter übereinstimmen
  - keinen zu hohen oder zu niedrigen Wert
- erwartungstreue Schätzung heißt unverzerrt

Konsistenz

- je größer der Stichprobenumfang, desto genauer sollte die Schätzung sein
- Schätzer ist konsistent, wenn dessen Varianz für große n gegen 0 geht

Effizienz

- Varianz des Schätzers sollte möglichst gering sein
  - je geringer, desto präziser ist die Schätzung
- hohe Effizienz – kleine Stichprobe liefert brauchbaren Schätzwert

Exhaustivität

- Schätzer ist exhaustiv (erschöpfend), wenn alle Informationen, die in den Daten einer Stichprobe enthalten sind, berücksichtigt werden

Spezielle Schätzfunktionen

Erwartungswert

- Mittelwert $\bar{x}$ einer Stichprobe, der den Erwartungswert μ der Grundgesamtheit schätzt wird betrachtet
  - Schätzung ist erwartungstreu und konsistent
    (Konsistenz ergibt sich aus dem Gesetz der großen Zahlen)

$$E(\bar{X}) = \mu \qquad und \qquad Var(\bar{X}) = \frac{\sigma^2}{n} \xrightarrow[n\to\infty]{} 0$$

Median

- Verteilung stetig und symmetrisch → $\tilde{X}$ ist erwartungstreuer Schätzer für $\tilde{\mu}$
  - hier: $\mu = \tilde{\mu}$ → bei der Normalverteilung ist der Median $\tilde{X}$ ein erwartungstreuer Schätzer für den Erwartungswert μ
- Varianz des Medians
  - $\tilde{X}$ ist konsistente Schätzung
  - Varianz von $\tilde{X}$ist größer als die Varianz von $\bar{X}$ → Mittelwert ist effizientere Schätzer für μ
- empirische Median ist nicht erschöpfend, weil nicht alle Stichprobenwerte in dessen Berechnung einfließen

$$Var(\tilde{X}) = \frac{\pi}{2} * \frac{\sigma^2}{n} \xrightarrow[n\to\infty]{} 0$$

Varianz

- daraus berechnete Standardabweichung ist bei quantitativen Merkmalen das am häufigsten benutzte Streuungsmaß

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

- Schätzung ist
  - erwartungstreu
  - konsistent

$$E(S^2) = \sigma^2 \qquad Var(S^2) = \frac{2\sigma^4}{n-1} \xrightarrow[n\to\infty]{} 0$$

- Schätzung der Standardabweichung σ durch S
  - konsistent
  - nicht erwartungstreu

Wahrscheinlichkeit

- p wird über die relative Häufigkeit geschätzt → deren Erwartungswert ist p $X_i \sim B(1, p)$
- Schätzung ist
  - erwartungstreu
  - konsistent (Gesetz der großen Zahlen)

$$E\left(\sum_{i=1}^{n}\frac{X_i}{n}\right) = \frac{1}{n}E\left(\sum_{i=1}^{n}X_i\right) = \frac{np}{n} = p$$

Parameter der bivariaten Statistik

- Schätzung der Kovarianz
  - erwartungstreu
  - konsistent
- Schätzung der Parameter der Regressionsgeraden
  - erwartungstreu
  - konsistent
- Schätzfunktion für den Pearson'schen Korrelationskoeffizienten
  - nicht erwartungstreu
  - konsistent

Intervallschätzungen

Bedeutung eines Konfidenzintervalls

- Schätzwert enthält keine Information darüber, wie sehr er vom „wahren" Parameter der Grundgesamtheit abweicht
  - prinzipiell keine exakten Angaben möglich, da der gesuchte Parameter unbekannt ist
  - Vermutung lässt zu, dass sich der Parameter in der näheren Umgebung des Schätzwerts befindet
- Genauigkeit der Schätzung kann durch Konstruktion eines Konfidenzintervalls aus den Daten der Stichprobe gewonnen werden
  - Intervall soll erzeugt werden, das den genauen Parameter überdeckt

- jedoch möglich, aus den Daten der Stichprobe ein Konfidenzintervall zu erzeugen, das den genauen Parameter nicht enthält
- Irrtumswahrscheinlichkeit α wird vor der Bestimmung des Konfidenzintervalls festgelegt
  - üblicherweise α=5%
  - in besonderen Fällen α=1%, α=0,1%
- 2 Möglichkeiten bei der Konstruktion eines Konfidenzintervalls
  - Konfidenzwahrscheinlichkeit 1–α – Erhalt eines Intervalls, das den unbekannten Parameter enthält
    Irrtumswahrscheinlichkeit α=5% → Konfidenzwahrscheinlichkeit 1–α=95%
  - Wahrscheinlichkeit α – Erhalt eines Intervalls, das den unbekannten Parameter <u>nicht</u> enthält

Konfidenzintervalle für einen Erwartungswert

- Schätzung über den Mittelwert $\bar{x}$
- Konfidenzintervall auf dem Niveau 1–α=95%
  - Wahrscheinlichkeit, dass ein Erwartungswert, der kleiner als die linke oder größer als die rechte Intervallsgrenze ist, zu $\bar{x}$ geführt hat, beträgt jeweils 2,5% → α=5%
  - mit Wahrscheinlichkeit von 95% erhält man ein Konfidenzintervall, das den unbekannten Erwartungswert μ überdeckt

$$\alpha = 5\% \quad \left[\bar{x} - \frac{1{,}96 * \sigma}{\sqrt{n}}; \bar{x} + \frac{1{,}96 * \sigma}{\sqrt{n}}\right] \qquad \alpha = 1\% \quad \left[\bar{x} - \frac{2{,}58 * \sigma}{\sqrt{n}}; \bar{x} + \frac{2{,}58 * \sigma}{\sqrt{n}}\right]$$

- zweiseitiges Konfidenzintervall auf dem (1- α)-Niveau definiert durch die Intervallmitte und die Grenzen

$$\left[\bar{x} - z_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}; \bar{x} + z_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}\right]$$

- Index $1 - \frac{\alpha}{2}$ bezeichnet das jeweilige Quantil der Standardnormalverteilung
  - $\alpha = 5\% \quad z_{1-\frac{\alpha}{2}} = z_{0,975} = 1{,}96$
- Standardabweichung σ der Grundgesamtheit meist nicht bekannt
  - man könnte σ durch die empirische Standardabweichung s ersetzten → bei kleinen Stichproben würde dies zu einer weiteren Ungenauigkeit der Schätzung führen
- Lösung: t-Verteilung
  - Zufallsvariable X normalverteilt, dann lassen sich die Quantile der Standardnormalverteilung durch die t-Werte ersetzen
  - f=n-1 bezeichnet die Anzahl der Freiheitsgrade der jeweiligen t-Verteilung
    (für jedes f gibt es eine spezielle t-Verteilung)
  - $1 - \frac{\alpha}{2}$ gibt das Quantil an
  - $t_{n-1;1-\frac{\alpha}{2}}$ … Werte in Tabellen

$$\left[\bar{x} - \frac{t_{n-1;1-\frac{\alpha}{2}} * s}{\sqrt{n}}; \bar{x} + \frac{t_{n-1;1-\frac{\alpha}{2}} * s}{\sqrt{n}}\right]$$

  - $\frac{s}{\sqrt{n}}$ … Schätzung für den Standardfehler des Mittelwerts $\frac{\sigma}{\sqrt{n}}$
  - einseitige Konfidenzintervalle (an einer Seite offen)

$$\left(-\infty; \bar{x} + \frac{t_{n-1;1-\frac{\alpha}{2}} * s}{\sqrt{n}}\right] \qquad \left[\bar{x} - \frac{t_{n-1;1-\frac{\alpha}{2}} * s}{\sqrt{n}}; +\infty\right)$$

  - Voraussetzung: Grundgesamtheit ist unendlich groß
  - Stichprobe des Umfangs n wird aus einer endlichen Grundgesamtheit des Umfangs N gezogen → Standardfehler muss korrigiert werden
  - Endlichkeitskorrektur ergibt sich aus der Varianz der hypergeometrischen Verteilung
  - Grenzen des Konfidenzintervalls bei einer endlichen Grundgesamtheit
  - bei großen Grundgesamtheiten mit N/n ≥ 100 → Endlichkeitskorrektur nimmt Wert nahe 1 an und ist deshalb vernachlässigbar

$$\bar{x} \pm t_{n-1;1-\frac{\alpha}{2}} * s * \sqrt{\frac{N-n}{n * (N-1)}}$$

<u>Beispiel</u>
Körpergröße n=23
$\bar{x} \pm s = (181{,}22 \pm 7{,}12)cm$
Konfidenzintervalle:
$[178{,}14cm; 184{,}29cm]\ (\alpha = 0{,}05)\ mit\ \ t_{22;0{,}975} = 2{,}074$
$[177{,}14cm; 185{,}40cm]\ (\alpha = 0{,}01)\ mit\ \ t_{22;0{,}995} = 2{,}819$

$$\left[181{,}22 - \frac{2{,}074 * 7{,}12}{\sqrt{23}}; 181{,}22 + \frac{2{,}074 * 7{,}12}{\sqrt{23}}\right] = [178{,}141; 184{,}299]$$

$$\left[181{,}22 - \frac{2{,}819 * 7{,}12}{\sqrt{23}}; 181{,}22 + \frac{2{,}819 * 7{,}12}{\sqrt{23}}\right] = [177{,}034; 185{,}405]$$

- Konfidenzintervall ist schmaler, wenn höherer Stichprobenumfang und geringere Streuung der Daten
- 95%-Intervalle sind schmaler als 99%-Intervalle
  - t-Quantile für α =0,05 sind kleiner als für α =0,01

Konfidenzintervalle für eine Wahrscheinlichkeit

- als Punktschätzer für eine Wahrscheinlichkeit p dient die relative Häufigkeit
  - X … Häufigkeit des Ereignisses A bei n Zufallsexperimenten

$$\hat{p} = \frac{X}{n}$$

- Grenzen des Konfidenzintervalls
  - Voraussetzung: $n\hat{p} > 5$ und $n(1 - \hat{p}) > 5$
  - → Stichprobenumfang darf nicht zu klein und die relativen Häufigkeiten sollten nicht zu extrem sein

$$\hat{p} \pm \left(\frac{1}{2n} + z_{1-\frac{\alpha}{2}} * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

  - $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ … Standardfehler der Schätzung
  - bei endlicher Grundgesamtheit: Multiplikation von $\sqrt{\frac{N-n}{(N-1)}}$ mit dem Standardfehler
  - $\frac{1}{2n}$ … Stetigkeitskorrektur

<u>Beispiel</u>

Anteil weiblicher Studenten: $\hat{p} = \frac{X}{n} = \frac{48}{71} = 0{,}676$
Mehr als die Hälfte der Studenten weiblich, oder höherer Anteil zufällig bedingt?
95%-Konfidenzintervall

$$\frac{48}{71} \pm \left(\frac{1}{142} + 1{,}96 * \sqrt{\frac{\frac{48}{71} * \frac{23}{71}}{71}}\right) \quad \rightarrow \quad [0{,}560; 0{,}792]$$

→ Annahme, dass tatsächlich mehr als die Hälfte Frauen sind

Bemerkungen

Bedeutung des Stichprobenumfangs

- Präzision einer Schätzung wird durch die Breite des Konfidenzintervalls ausgedrückt
  - je schmaler das Intervall, desto genauer ist die Schätzung
  - sehr breites Konfidenzintervall → für praktische Zwecke unbrauchbar

Breite des zweiseitigen Konfidenzintervalls für den Erwartungswert

- t-Wert ≈ 2 für $\alpha = 5\%$ und n ≥ 10
- Standardabweichung σ meist nicht bekannt
- bei gleicher Standardabweichung σ ist der vierfache Stichprobenumfang erforderlich, um die Breite des Intervalls zu halbieren

$$BK = \frac{2 * t_{n-1;1-\frac{\alpha}{2}} * s}{\sqrt{n}}$$

Breite des Konfidenzintervalls für die Wahrscheinlichkeit p

$$BK = 2 * z_{1-\frac{\alpha}{2}} * \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{1}{n}}$$

<u>Beispiel</u>
n= 71 Studenten
60 Studenten haben Rhesusfaktor „positiv" (85%)
Konfidenzintervall mit $\alpha = 5\%$

$$\frac{60}{71} \pm \left(\frac{1}{142} + 1{,}96 * \sqrt{\frac{\frac{60}{71} * \frac{11}{71}}{71}} \right) = [0{,}75; 0{,}94]$$

n=284 (4-fache Stichprobenumfang)
240 Personen haben Rhesusfaktor „positiv" (selber Schätzwert $\hat{p} = \frac{240}{284} \approx 85\%$)
→ Konfidenzintervall ist schmaler und Schätzung präziser

$$\frac{240}{284} \pm \left(\frac{1}{568} + 1{,}96 * \sqrt{\frac{\frac{240}{284} * \frac{44}{284}}{284}} \right) = [0{,}80; 0{,}89]$$

3 Faktoren für die Präzision der Schätzung

- Irrtumswahrscheinlichkeit α
  - für $\alpha = 5\%$ ergibt sich ein schmaleres Intervall als für $\alpha = 1\%$
- Standardabweichung s
  - je homogener die Grundgesamtheit, desto kleiner sind die Standardabweichung und die Breite des Intervalls
- Stichprobenumfang n
  - Schätzung umso präziser, je höher der Stichprobenumfang

Prinzip eines statistischen Tests

Durchführung eines Tests

Funktion eines statistischen Tests

- Hypothese – Vermutung, die präzise formuliert wird
- Überprüfung einer Hypothese
  - Erarbeitung eines theoretischen Hintergrund, zum Untermauern der Hypothese mit sachlichen Argumenten
  - statistisch absichern → Erhebung relevanter Daten und Analyse mit geeigneten Testmethoden
- Vermutung
  - z.B. Babys, deren Mütter während der Schwangerschaft einem bestimmten Risiko ausgesetzt waren, wiegen im Durchschnitt weniger (als 3500g)
- kein Unterschied
  - mittleres Geburtsgewicht der Risiko-Babys ≈ 3.500g
- Unterschied
  - Babies wiegen durchschnittlich weniger als 3.500g (Unterschied nicht zufällig bedingt)
- → komplementäre Aussagen, ergänzen sich und schließen sich gegenseitig aus

Statistischer Test
… hilft Grenze festzulegen, ab wann Abweichung vom Sollwert nicht zufällig ist

- 2 komplementäre Hypothesen aufstellen
- Auswahl eines für die Fragestellung passenden Tests
- Testgröße wird aus den Daten einer Stichprobe berechnet
  - objektive und nachvollziehbare Entscheidung zugunsten von einer der beiden Hypothesen kann getroffen werden

Formulieren der Hypothesen

- Hypothesen vor Durchführung des Tests inhaltlich so präzise wie möglich formulieren
- **Alternativhypothese** H_1 – Hypothese, die eine innovative Aussage beinhaltet und Althergebrachtes in Frage stellt
- **Nullhypothese** H_0 – „Es gibt keinen Unterschied"
  - kann komplementär zur Alternativhypothese formuliert werden
  - Nullhypothese ist in der Testtheorie die Basis
- zweiseitige/ungerichtete Hypothese
  - über die Richtung eines Unterschieds wird nicht ausgesagt
  - z.B. $H_1\colon\ \mu \neq 3.500$
- einseitige/gerichtete Hypothese
  - z.B. $H_1\colon\ \mu < 3.500$

<u>Beispiel „Risiko-Babys"</u>
$H_0\colon\ \mu = 3.500 \quad H_1\colon\ \mu < 3.500$
→ Alternativhypothese vereint in sich alle Hypothesen außer die Nullhypothese
→ $\mu > 3.500$ wird nicht in Betracht gezogen

Fehlerarten

- Testentscheidung hängt von der Prüfgröße ab
  - Prüfgröße wird aus den Stichprobenwerten ermittelt
- Testverfahren könnte zu einer Fehlentscheidung führen

α Fehler (Fehler 1. Art)
- Nullhypothese ist richtig
- Alternativhypothese wird fälschlicherweise genommen
- kontrollierbarer Fehler - kann nur bei Gültigkeit der Nullhypothese auftreten und diese ist eindeutig formuliert

- Wahrscheinlichkeitsverteilung der Prüfgröße kann angegeben werden

- Zufallsvariable unter H_0 ist t-verteilt

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

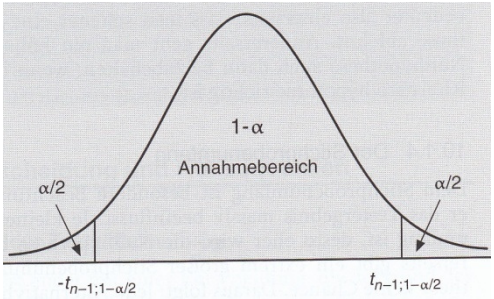
- Prüfgröße t
  - Werte zwischen –∞ und +∞

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Unter der Nullhypothese betragen die entsprechenden Wahrscheinlichkeiten ($\alpha = 5\%$)

Bereich der Prüfgröße t		**Entscheidung für**	**Wahrscheinlichkeit**
$t < t_{n-1;\frac{\alpha}{2}} < 0$	kritischer Bereich	H_1	$\frac{\alpha}{2} = 2{,}5\%$
$t_{n-1;1-\frac{\alpha}{2}} \leq t \leq t_{n-1;\,1-\frac{\alpha}{2}}$	Annahmebereich	H_0	$1 - \alpha = 95\%$
$t > t_{n-1;1-\frac{\alpha}{2}} > 0$	kritischer Bereich	H_1	$\frac{\alpha}{2} = 2{,}5\%$

- kritischer Bereich: Prüfgröße $|t| > t_{n-1;1-\frac{\alpha}{2}}$
  - wegen der Symmetrie der t-Verteilung $t_{n-1;1-\frac{\alpha}{2}} = -t_{n-1;1-\frac{\alpha}{2}}$
- Entscheidungsregel
  - falls $|t| \leq t_{n-1;1-\frac{\alpha}{2}}$ → Nullhypothese wird beibehalten
  - falls $|t| > t_{n-1;1-\frac{\alpha}{2}}$ → Alternativhypothese wird genommen
- $\pm t_{n-1;1-\frac{\alpha}{2}}$ → kritische Werte
  - trennen den Annahme- vom kritischen Bereich



Annahme- und Ablehnungsbereich beim t-Test
(zweiseitige Fragestellung)

- H_1: $\mu > \mu_0$
  - Ablehnung der Nullhypothese, falls $t > t_{n-1;1-\frac{\alpha}{2}}$
- H_1: $\mu < \mu_0$
  - Annahme der Alternativhypothese, falls Prüfgröße t negativ und $t < \pm t_{n-1;1-\frac{\alpha}{2}}$
- Ablehnungsbereich mit Fläche α bei einseitigen Fragestellungen nur auf einer Seite der Dichtefunktion der t-Verteilung
- kritische Werte abhängig von
  - der Anzahl der Freiheitsgrade $f = n - 1$
  - Irrtumswahrscheinlichkeit α
  - einseitiger oder zweiseitiger Test
- falls Nullhypothese richtig, wird mit 95%iger Wahrscheinlichkeit die richtige Entscheidung getroffen
  - $\alpha\ \text{Fehler} = 5\%$ → Signifikanzniveau, α -Niveau
- aus der Größe des α Fehlers ergibt sich
  - Annahmebereich für die Nullhypothese
  - kritischer Bereich (Ablehnungsbereich für die Nullhypothese)

β Fehler (Fehler 2. Art)
- Alternativhypothese ist richtig
- Nullhypothese wird fälschlicherweise genommen
- je mehr sich der unbekannte Erwartungswert μ und der Sollwert μ_0 unterscheiden, desto kleiner ist der β-Fehler
- β-Fehler im Gegensatz zum α–Fehler nicht abschätzbar
- kann bei kleinem Stichprobenumfang sehr groß sein
- β-Fehler kann durch Wahl des α –Fehlers beeinflusst werden
  - je größer α, umso größer ist der kritische Bereich → umso kleiner ist β
- kleiner α-Wert
  - Ablehnung einer richtigen Nullhypothese seltener
  - höheres Risiko, Nullhypothese fälschlicherweise beizubehalten

Stichprobenumfang
- beeinflusst Testergebnis massiv
- je kleiner Stichprobenumfang, desto eher wird die Nullhypothese beibehalten
- bei extrem großem Stichprobenumfang, wird der Nullhypothese keine Chance gegeben
- sequenzielle Testverfahren
  - Stichprobenumfang wird vor dem Testen nicht als fixe Größe festgelegt, sondern als Zufallsvariable aufgefasst
  - β und α werden vorab festgelegt
  - Test wird zunächst mit minimalen Stichprobenumfang durchgeführt, dieser wird dann um 1 erhöht
  - Prozedur wird solange durchgeführt, bis Testentscheidung möglich

Optimaler Stichprobenumfang hängt ab von
- α-Fehler (üblich $\alpha = 0{,}05$)
- β-Fehler (üblich β=0,20)
- Art der Daten und deren Skalenniveau
- Streuung der Daten
- speziellen Test
- Größe des nachzuweisenden Effekts

Testentscheidung und Konsequenzen

Basis der Testentscheidung
- Test endet mit Testentscheidung
  - Prüfgröße liegt im kritischen Bereich → Alternativhypothese
    $\alpha = 10\%$ - schwach signifikant
    $\alpha = 5\%$ - signifikant
    $\alpha = 1\%$ - hoch-signifikant
    $\alpha = 0{,}1\%$ - höchst signifikant
  - Prüfgröße liegt im Annahmebereich → Nullhypothese

- Wahrscheinlichkeit eines Tests, eine richtige Alternativhypothese zu treffen
  - quantifiziert Güte, Teststärke, Trennschärfe oder Macht
    $1 - \beta$

Testentscheidung bei einem statistischen Test

Testentscheidung	**Wahrscheinlichkeit**	
	H_0 gilt	H_2 gilt
für H_0	richtige Entscheidung $1 - \alpha$	Fehler 2. Art β
für H_1	Fehler 1. Art α	richtige Entscheidung $1 - \beta$
Summe	1	1

p-Wert und Konfidenzintervall

- Prüfgröße, die aus dem t-Test für eine Stichprobe resultiert, ist betragsmäßig größer als das Quantil $t_{n-1;0,975}$ → Unterschied ist auf dem 1%-Niveau abgesichert

p-Wert

- quantifiziert Wahrscheinlichkeit, dass das gefundene Testergebnis zustande kommt, wenn in Wirklichkeit die Nullhypothese richtig ist
  (Wahrscheinlichkeit, dass das Testergebnis reiner Zufallsbefund ist)
- p kleiner als das zuvor festgelegte Signifikanzniveau α → Alternativhypothese
- besagt lediglich, ob statisch signifikanter Unterschied existiert
  - sinnvoll, Konfidenzintervall zum p-Wert berechnen
- alle Werte im Konfidenzintervall sind für die Größe des Unterschieds in Betracht zu ziehen
  - je schmaler das Intervall, desto präziser ist die Schätzung, desto einfacher ist die Interpretation des Testergebnisses
- keine Aussage, ob kein praktisch relevanter Unterschied existiert oder dieser nur nicht nachgewiesen werden kann, wenn
  - kleiner Stichprobenumfang
  - nicht-signifikantes Ergebnis
  - breites Konfidenzintervall

<u>Beispiel</u>
20 Babys von Risiko-Patienteninnen
Geburtsgewicht $\bar{x} \pm s = (3280 \pm 490)$
Durchschnittswert 3.500g

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{3280 - 3500}{\frac{490}{\sqrt{20}}} = -2{,}0079$$

- p=0,0509
  kritischer Wert: $t_{19;0,975} = 2{,}093$ (2seitiger Test)
  $\alpha = 5\%$-Niveau → Nullhypothese
- kritischer Punkt: : $t_{19;0,95} = 1{,}729$ (1seitiger Test)
  p=0,0255 (wird halbiert)
  einseitiges Konfidenzintervall

$$\left(-\infty;\ \bar{x} + \frac{t_{n-1;1-\alpha} * s}{\sqrt{n}}\right] = (-\infty; 3469]$$

- 20 Babies wiegen durchschnittlich zwar weniger, aber Unterschied ist nicht gravierend

Anpassungstest

- Nachweis, dass empirische Verteilung mit theoretischer Verteilung vereinbar ist
  - Nullhypothese wird erst bei p ≥ 0,10 angenommen

Äquivalenztest

- z.B. bei Bioverfügbarkeitsstudien, um therapeutische Gleichwertigkeit zweier Behandlungen zu prüfen

Manipulation des Testergebnisses

- signifikantes Ergebnis lässt sich leichter publizieren als ein nicht-signifikantes

Manipulation des Ergebnisses durch

- Auswahl des Tests
  - alle in Frage kommenden Testverfahren ausprobieren und dieses wählen, welches den besten p-Wert hat
- Ein- oder zweiseitiges Testen
  - einseitig testen nachdem man zweiseitig getestet hat, dann halbiert sich der p-Wert des zweiseitigen Tests und man erhält statt einem nicht-signifikanten Ergebnis ein signifikantes

Multiples Testen

- bei k Tests beträgt der α–Fehler etwa $k\alpha$
- Bonferroni-Korrektur
  - einzelnes Testergebnis ist erst dann signifikant, wenn $p < k\alpha$
  - Nachteil: β-Wert erhöht sich enorm

Klassifikation der Testmethoden

Anzahl der Stichproben

- 1-Stichprobentest
  - empirische Kenngröße wird mit vorgegebenem Sollwert verglichen (z.B. Mittelwert)
- 2-Stichprobentest
- Mehrstichprobentest
  - Stichproben werden bezüglich eines bestimmten Parameters (z.B. Mittelwert) miteinander verglichen

Art der Stichproben

- verbundene (abhängige)
  - Stichproben haben denselben Umfang
  - → paarig – jeder Wert der einen Stichprobe bildet mit Wert der anderen Stichprobe inhaltlich Paar
- unverbundene (unabhängige)
  - bezüglich Beobachtungseinheiten unabhängig voneinander
  - Umfänge können unterschiedlich sein

Funktion des Tests

- Lagetests
  - Vergleich von Lagemaßen
- Wahrscheinlichkeitstests
  - Vergleich einer relevanten Häufigkeit mit einer vorgegebenen Wahrscheinlichkeit
- Homogenitätstests
  - Vergleich mehrerer Stichproben bezüglich einer Häufigkeitsverteilung
- Dispersionstests
  - Prüfung von Streumaßen
- Unabhängigkeitstests
  - Prüfen der Unabhängigkeit zweier Merkmale
- Anpassungstests
  - Vergleich einer empirischen Verteilung mit einer theoretischen (z.B. Normalverteilung, Poissonverteilung)
- Tests zum Vergleich von Überlebenszeitkurven

Prüfgrößen

- Unterscheidung von t-Tests, Rangsummentests, Vorzeichentests, Chi^2-Tests, Binomialtests

Statistische Tests

Funktion des Tests	Bezeichnung	Testgegenstand
Lagetest für eine Stichprobe	t-Test	X normalverteilt
	Wilcoxon-Test	X symmetrisch verteilt
	Vorzeichentest	Variable X
Lagetest für zwei verbundene Stichproben	t-Test	Differenz X-Y normalverteilt
	Wilcoxon-Test	Differenz X-Y symmetrisch verteilt
	Vorzeichentest	Differenz X-Y
Lagetest für zwei unverbundene Stichproben	t-Test	X und Y normalverteilt mit gleicher Varianz
	Welch-Test	X und Y normalverteilt
	U-Test	X und Y gleiche Verteilungsform
	Median-Test	X und Y ordinal skaliert
Dispersionstest	F-Test	2 Varianzen
Unabhängigkeitstest	t-Test	Korrelationskoeffizient
Wahrscheinlichkeitstest	Binomialtest	Alternativmerkmal
Homogenitätstest für zwei unverbundene Stichproben, Unabhängigkeitstest	Vierfeldertest	2 Alternativmerkmale
	Chi²-Test	2 qualitative Merkmale
	Fisher's exakter Test	2 qualitative Merkmale
Homogenitätstest für zwei verbundene Stichproben	McNemar-Test	Alternativmerkmal
Anpassungstest	Chi²-Anpassungstest	empirische Verteilung
Vergleich von Überlebenszeiten	Logrank-Test	Überlebenszeitkurven

LAGETESTS

t-Tests

- = parametrische Tests– bei bekannter Verteilung der Zufallsvariablen werden nur noch bestimmte Parameter überprüft
- Voraussetzung: normalverteilte Grundgesamtheiten

t-Test für eine Stichprobe

- Vergleich des Mittelwerts $\bar{x}$ einer Stichprobe mit einem vorgegebenen Sollwert μ_0
- Voraussetzung
  - Stichprobenwerte x_i sind Realisationen einer normalverteilten Zufallsvariablen $X \sim N(\mu,\sigma^2)$
- Prüfgröße t

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

t-Test für zwei verbundene Stichproben

- Lagetest zur Überprüfung der Gleichheit von zwei Erwartungswerten z
- Voraussetzung
  - zwei verbundene Stichproben des Umfangs n mit Wertepaaren (x_i, y_i), die aus Grundgesamtheiten mit den Erwartungswerten μ_1 und μ_2 stammen
  - Differenzen $d_i = x_i - y_i$ sind Realisationen einer normalverteilen Zufallsvariablen D mit dem Erwartungswert δ
- Hypothesen bei zweiseitiger Fragestellung

$$H_0\colon\ \delta = 0 \qquad H_1\colon \delta \neq 0$$

- Hypothesen bei einseitiger Fragestellung

$$H_1\colon \delta > 0 \quad oder \quad H_1\colon \delta < 0$$

- unter der Nullhypothese erwartet man für die Differenzen d_i den Mittelwert $\bar{d} = 0$
- Prüfgröße

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$$

- s_d … empirische Standardabweichung der Differenzen d_i
- Ablehnung der Nullhypothese, falls …
  - bei 2seitiger Fragestellung: $|t| > t_{n-1;1-\frac{\alpha}{2}}$
  - bei 1seitiger Fragestellung:
    für $H_1\colon \delta > 0$ $\quad t > t_{n-1;1-\frac{\alpha}{2}}$
    für $H_1\colon \delta < 0$ $\quad t < -t_{n-1;1-\frac{\alpha}{2}}$
- Konfidenzintervall
  - zur Abschätzung des „wahren" Unterschieds der Größe

$$\left[\bar{d} - \frac{t_{n-1;1-\frac{\alpha}{2}} * s_d}{\sqrt{n}}; \bar{d} + \frac{t_{n-1;1-\frac{\alpha}{2}} * s_d}{\sqrt{n}}\right]$$

<u>Beispiel</u>

- Körpergewichte von 10 Personen vor und nach der Diät
- Mittelwert vor der Diät 93,9kg
- Mittelwert nach der Diät 91,2kg
- mittlere Differenz (2,68 ± 3,32)kg
- Prüfgröße t

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} = \frac{2{,}68}{\frac{3{,}32}{\sqrt{10}}} = \frac{2{,}68}{1{,}04987} = 2{,}55$$

- $t_{9;0,975}$=2,262 (siehe Tabelle B)

- p-Wert=0,0312
- Unterschied ist signifikant auf dem Niveau $\alpha = 0{,}05$
- Konfidenzintervall für die Differenz

$$\left[\bar{d} - \frac{t_{n-1;1-\frac{\alpha}{2}} * s_d}{\sqrt{n}}; \bar{d} + \frac{t_{n-1;1-\frac{\alpha}{2}} * s_d}{\sqrt{n}}\right] = \left[2{,}68 - \frac{2{,}262 * 3{,}32}{\sqrt{10}}; 2{,}68 + \frac{2{,}262 * 3{,}32}{\sqrt{10}}\right] = [0{,}3052; 5{,}0548]$$

- Testergebnis ist zwar signifikant, höherer Stichprobenumfang würde aber zu kleinerem Konfidenzintervall und genaueren Schätzung führen

t-Test für zwei unverbundene Stichproben

- zwei unverbundene Stichproben der Umfänge n_1 und n_2
- Daten beider Stichproben entstammen normalverteilten Grundgesamtheiten mit derselben Varianz
  - $X \sim N(\mu, \sigma^2)$
  - $Y \sim N(\mu, \sigma^2))$
- beide Verteilungen sollten dieselbe Form haben
- höchstens Unterscheidung bezüglich Erwartungswerte
- Nullhypothese

$$H_0\colon\ \mu_1 = \mu_2$$

- Prüfgröße ($n_1 \neq n_2$)

$$t = \frac{\bar{x} - \bar{y}}{s * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- zwei unabhängige Mittelwerte fließen in Berechnung mit ein, deshalb
  - Anzahl der Freiheitsgrade $f = n_1 + n_2 - 2$
- Ablehnung der Nullhypothese, falls
  - bei zweiseitiger Fragestellung:
    $|t| > t_{f;1-\frac{\alpha}{2}}$
  - bei einseitiger Fragestellung:
    kritischer Wert $+t_{f;1-\alpha}$ bzw. $t_{f;\alpha} = -t_{f;1-\alpha}$
- s^2 … mittlere Varianz
  - Berechnung aus der gewichteten Mittelung der beiden empirischen Varianzen s_1^2und s_2^2 ($n_1 \neq n_2$)

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

- Grenzen des zweiseitigen Konfidenzintervalls ($n_1 \neq n_2$)

$$\bar{x} - \bar{y} \pm t_{n_1+n_2-2;1-\frac{\alpha}{2}} * s * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- bei <u>gleichen</u> Stichprobenumfängen $n=n_1=n_2$

$$t = \frac{\bar{x} - \bar{y}}{s * \sqrt{\frac{2}{n}}}$$

$$s^2 = \frac{s_1^2 + s_2^2}{2}$$

<u>Beispiel</u>

- Körpergrößen männlicher und weiblicher Studenten
- Mittelwerte $\bar{x}_m = 181{,}22cm$ und $\bar{x}_w = 169{,}06cm$
- Standardabweichungen $s_m = 7{,}12cm$ und $s_w = 6{,}60cm$
- Stichprobenumfänge $n_1=23$ und $n_2=48$
- Unterschied zufällig oder signifikant?
- Anzahl der Freiheitsgrade $f = 23 + 48 - 2 = 69$

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{22 * 7{,}12^2 + 47 * 6{,}60^2}{69} cm^2 = 45{,}835cm^2$$

- Prüfgröße

$$t = \frac{\bar{x} - \bar{y}}{s * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{181{,}22 - 169{,}06}{\sqrt{\frac{45{,}835}{23} + \frac{45{,}835}{48}}} = \frac{12{,}16}{1{,}717} = 7{,}083$$

- kritischer Wert $t_{69;0,975}$=1,995
- kritischer Wert ist wesentlich kleiner als Prüfgröße
- p-Wert p < 0,0001
- Ergebnis ist hoch signifikant
- Konfidenzintervall für mittlere Differenz [8,73cm;15,58cm]

Welch-Test

- Alternative zum t-Test für zwei unverbundene Stichproben
- Gleichheit der Varianzen (=Homoskedazität) der beiden Grundgesamtheiten <u>nicht</u> vorausgesetzt
- empirische Stichprobenvarianzen s_1^2und s_2^2 sind Schätzwerte für die Varianzen der Grundgesamtheiten
- Prüfgröße

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- Anzahl der Freiheitsgrade

$$f = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

- Verwendung des Welch-Tests sinnvoll, wenn
  - Varianzen der Grundgesamtheit unbekannt
  - Gleichheit nicht angenommen werden kann
- Bedingungen des t-Tests erfüllt → Welch-Test hat geringere Power
- Welch-Test nimmt ungleiche Varianzen und damit verschiedene Verteilungsformen an

Voraussetzungen der t-Lagetests

- Grundvoraussetzung: Normalverteilung der Zufallsvariablen (bezogen auf Grundgesamtheit)
- Überprüfung anhand der Stichprobe, ob Normalverteilung ausgeschlossen
  - Histogramm
    Daten symmetrisch oder eher schief verteilt
  - Mittelwert und Median
    weichen beide Parameter stark voneinander ab → schiefe Verteilung
  - Schiefe und Kurtosis
    Werte um 0 → Daten normalverteilt
  - Anpassungstest
    Prüfgröße fällt in Annahmebereich → Normalverteilung ist nicht abgesichert, lediglich nicht ausgeschlossen
- t-Test ist robust gegenüber Abweichungen von der Normalverteilung
- t-Test für eine Stichprobe
  - n < 10 → sind Daten nicht normalverteilt, sollte anderer Lagetest herangezogen werden
  - n ≥ 10 → Daten müssen lediglich annähernd symmetrisch verteilt sein
  - n ≥ 25 → Stichprobenmittelwerte müssen nach dem zentralen Grenzwertsatz normalverteilt sein
- t-Test für zwei verbundene Stichproben
  - n ≥ 10 → Differenzen d_i müssen annähernd symmetrisch verteilt sein
    (Bedingung erfüllt, wenn Variablen X und Y ungefähr gleiche Verteilungsform)
- t-Test für zwei unverbundene Stichproben
  - beide Stichprobenumfänge müssen mindestens 10, bei nicht symmetrischen Verteilungen mindestens 20 und ähnlich groß sein
  - Zufallsvariablen X und Y sollen ungefähr denselben Verteilungstyp haben

- Vortests, zur Überprüfung der Voraussetzungen
  Gleichheit der Varianzen
  Normalverteilung
  → F-Test – Prüfung der Gleichheit zweier Varianzen (gibt lediglich an, ob diese ausgeschlossen werden kann)

Absicherung des Unterschieds mit einem t-Test
- hoher Stichprobenumfang
- großer Unterschied zwischen den Mittelwerten
- geringe Streuung der Daten

Möglichkeiten, wenn die Voraussetzungen nicht erfüllt sind
- Transformation in normalverteilte Daten
- Test mit schwächeren Voraussetzungen wählen

Andere Anwendungen des t-Tests

- Testen, ob sich empirischer Korrelationskoeffizient nach Pearson signifikant von 0 unterscheidet
- Überprüfung des Steigungskoeffizienten der Regressionsgeraden
- Prüfgröße

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

- Freiheitsgrade $f = n - 2$
- Ablehnung der Nullhypothese, falls
  - $t > t_{n-2;1-\alpha} \quad bzw. \quad t < t_{n-2;1-\alpha}$
- Angabe eines Konfidenzintervall für ein empirisch ermitteltes r sinnvoll
- Annahme der Alternativhypothese
  - je größer der Betrag des empirischen Korrelationskoeffizienten r
  - je größer der Stichprobenumfang n
  - → desto größer der Betrag der Prüfgröße t

Rangsummentests

- alternativ zu den t-Lagetests
- verteilungsfreie (nicht-parametrische) Tests
- weniger strenge Voraussetzungen
  - setzen keine bestimmte Verteilungsform voraus
- Prüfgrößen werden aus den Rangzahlen berechnet
  - auch für ordinal-skalierte Merkmale anwendbar

Wilcoxon-Test für eine Stichprobe

- Überprüfung, ob und in welchem Maß die Werte einer Stichprobe von einem vorgegebenen Sollwert $\tilde{\mu}$ abweichen
- Nullhypothese

$$H_0: \tilde{\mu} = \tilde{\mu}_0$$

- $\tilde{\mu}$ … Median der Grundgesamtheit
- Testdurchführung
  - Berechnung für jeden Stichprobenwert Differenz zum Sollwert
  - Eliminierung der Stichprobenwerte, die mit Sollwert übereinstimmen
  - Sortierung der Differenzen d_i nach Größe ihres Betrags in aufsteigender Reihenfolge
    Versehen mit Rangzahlen → kleinste Differenz erhält Rangzahl 1, größte Rangzahl n
  - zwei oder mehr identische Differenzbeträge → mittlere Rangzahl (=verbundene Ränge)
  - Aufaddieren der Rangzahlen der negativen Differenzen und die Rangzahlen positiver Differenzen → Rangsummen R^- bzw. R^+
  - Prüfgröße R ist kleinere der beiden Rangsummen
  - aus Tabelle C: kritische Werte in Abhängigkeit vom Stichprobenumfang n und der Irrtumswahrscheinlichkeit α

  - Ablehnung der Nullhypothese, falls Prüfgröße ≤ kritischer Wert

- n > 25 – Prüfgröße approximativ normalverteilt mit dem Erwartungswert $\frac{n(n+1)}{4}$ und der Varianz $\frac{n(n+1)(2n+1)}{24}$
- Transformation der Prüfgröße in den Wert der Standardnormalverteilung → Abschätzung, ob Ergebnis signifikant
- bei zweiseitigen Fragestellungen: kritischer Wert = 1,96 (für $\alpha = 0{,}05$)

Prüfgröße R
- Wertebereich ist zwischen 0 und $\frac{n(n+1)}{4}$
- Extremfall - R=0
  - beide Rangsummen unterscheiden sich maximal
  - alle Stichprobenwerte sind kleiner/größer als der Sollwert

Voraussetzung
- symmetrische Verteilung (Normalverteilung nicht notwendig)

Wilcoxon-Test für zwei verbundene Stichproben
- Pendant zum t-Test für zwei verbundene Stichproben mit jeweils dem Umfang n
- Vergleich beider Mediane
- Nullhypothese

$$H_0: \tilde{\mu}_1 = \tilde{\mu}_2$$

- Testverfahren
  - Bildung der Differenzen $d_i = x_i - y_i$ für jedes Merkmalpaar
    Differenzen müssen symmetrisch verteilt sein
  - Eliminierung von $d_i = 0$
  - Sortierung der Differenzen d_i nach Größe ihres Betrags in aufsteigender Reihenfolge
    Versehen mit Rangzahlen → kleinste Differenz erhält Rangzahl 1, größte Rangzahl n
  - Aufaddieren der Rangzahlen der negativen Differenzen und die Rangzahlen positiver Differenzen → Rangsummen R^- bzw. R^+
  - Prüfgröße R ist kleinere der beiden Rangsummen
  - aus Tabelle C: kritische Werte
    für n > 25 ist Prüfgröße normalverteilt mit dem Erwartungswert $\frac{n(n+1)}{4}$ und Varianz $\frac{n(n+1)(2n+1)}{24}$

Prüfgröße R
- Wertebereich von R zwischen 0 und $\frac{n(n+1)}{4}$
- R = $\frac{n(n+1)}{4}$
  - Ränge teilen sich vollkommen gleichmäßig zwischen den beiden Stichproben auf
- R = 0
  - jeder Wert der einen Stichprobe ist kleiner als jeder beliebige Wert der anderen Stichprobe

Voraussetzungen
- Zufallsvariablen X und Y weisen annähernd die gleiche Verteilungsform auf → Differenzen d_i symmetrisch verteilt

Beispiel
- Diät bei 10 Personen
- Wilcoxon-Test für zwei verbundene Stichproben → Überprüfung, ob Änderung des durchschnittlichen Gewichts erfolgte
- Gewicht vor der Diät x_i
- Gewicht nach der Diät y_i

i	x_i	y_i	$d_i=x_i-y_i$	Rangzahlen für $d_i>0$	Rangzahlen für $d_i<0$
1	92,7	85,8	6,9	9	
2	86,2	83,4	2,8	5,5	
3	102,1	98,3	3,8	7	
4	85,9	83,6	2,3	3	
5	96,3	91,1	5,2	8	
6	90,2	92,7	-2,5		4
7	87,5	88,6	-1,1		2
8	98,0	98,7	-0,7		1
9	89,9	87,1	2,8	5,5	
10	110,2	102,9	7,3	10	
				$R^+=48$	**$R^-=7$**

$$R^+ + R^- = \sum i = \frac{n(n+1)}{2} = 55$$

- Differenzbeträge von i=2 und i=9 stimmen überein → Ermittlung mittlerer Rangzahl
- Prüfgröße R=7
- kritischer Punkt für $\alpha = 5\%$ und n=10 (zweiseitige Fragestellung) = 8
- R < 8 → Alternativhypothese wird angenommen

U-Test von Mann und Whiney
- Alternative zum t-Test für zwei unverbundene Stichproben
- Vergleich zweier Mediane miteinander
- Nullhypothese

$$H_0: \tilde{\mu}_1 = \tilde{\mu}_2$$

- n_1 und n_2 müssen nicht ident sein
- Voraussetzungen
  - Zufallsvariable X und Y müssen in etwa gleiche Verteilungsform haben
- Durchführung
  - Sortierung aller Werte aus beiden Stichproben in aufsteigender Reihenfolge Versehen mit Rangzahlen
  - Addieren der Rangzahlen beider Stichproben → Rangsummen R_1 und R_2
  - Berechnung von U_1 und U_2

$$U_1 = n_1 * n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U_2 = n_1 * n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

  - $U_1 + U_2 = n_1 * n_2$
  - Testgröße $U = min(U_1, U_2)$
  - U ≤ kritischer Wert → Ablehnung der Nullhypothese

Prüfgröße U
- $n_1, n_2 \geq 10$ → Prüfgröße normalverteilt mit dem Erwartungswert $\frac{n_1(n_1+n_2+1)}{2}$ und Varianz $\frac{n_1 n_2}{6}$
  - n_1 ... kleinerer Umfang
- Wertebereich zwischen 0 und $\frac{n_1 n_2}{2}$
- je näher U bei 0 liegt
  - umso mehr unterscheiden sich die beiden Stichproben → Annahme der Alternativhypothese
- verbundene Ränge sind unproblematisch, wenn sie innerhalb einer Stichprobe auftreten

- Anzahl der verbundenen Ränge sollten höchstens 20% betragen

Beispiel
- Nachweis, dass männliche Studenten im Durchschnitt höheres Körpergewicht haben als weibliche
- 10 Studenten, 12 Studentinnen
- Merkmal „Körpergewicht" → keine Normalverteilung, also U-Test

Gewicht x_i	**Rang**	**Gewicht y_i**	**Rang**
61	8	48	1
69	11	52	2
70	12,5	55	3
72	14	57	4
75	16	58	5
79	18	60	6,5
82	19	60	6,5
84	20	63	9
85	21	65	10
90	22	70	12,5
		74	15
		77	17
	R_1=161,5		**R_2=91,5**

- U_1=13,5 → U
- U_2=106,5
- kritischer Wert (zweiseitiger Test, $\alpha = 5\%$) = 29 (Tabelle D)
- U < 29 → Unterschied abgesichert
- p-Wert = 0,0024
- bei t-Test p=0,0007

Vergleich zwischen Rangsummentests und t-Tests

- Rangsummentests haben schwächere Voraussetzungen als t-Test
- Rangsummentests haben breiteres Anwendungsspektrum
  - Verwendung auch für metrisch-diskrete und ordinal-skalierte Merkmale
- Rangsummentest wertet nur Reihenfolge der Daten aus
  - t-Test nutzt in den Daten enthaltene Information vollständig aus
- mittels t-Verteilung lassen sich p-Werte, Konfidenzintervalle ermitteln
  - Berechnung der Konfidenzintervalle bei Rangsummentests nicht möglich
- Verwendung eines Rangsummentests, obwohl Voraussetzungen für t-Test erfüllt → Verminderung der Power 1-β
  - Test, der zur Beibehaltung der Nullhypothese tendiert → konservativ
  - Test, der mehr Ergebnisse signifikant werden lässt als dem festgelegten α–Niveau entspricht → progressiv

Vorzeichentests

Vorzeichentest für eine Stichprobe

- Überprüfung, ob und in welchem Maß die Werte einer Stichprobe von einem vorgegebenen Sollwert $\tilde{\mu}$ abweichen
- Nullhypothese

$$H_0: \tilde{\mu} = \tilde{\mu}_0$$

- Testverfahren
  - Beurteilung jedes Stichprobenwerts, ob er größer (positives Vorzeichen) oder kleiner (negatives Vorzeichen) als der Sollwert ist
  - Eliminierung der Werte, die mit Sollwert ident sind
  - Zählen der Anzahl der positiven und negativen Vorzeichen → kleinere Zahl ist Prüfgröße k
  - (Nullhypothese – Übereinstimmung der Anzahl der positiven und negativen Vorzeichen)
  - Testentscheidung: Vergleich der Prüfgröße mit dem kritischen Wert (Tabelle F)

- Prüfgröße ist unter der Nullhypothese binomialverteilt mit Erwartungswert $\mathrm{n} * 0{,}5$
- Ermittlung des Annahmebereichs für die Prüfgröße k bei kleinen Stichproben
  - Berechnung der einzelnen Wahrscheinlichkeiten

$$P(X = k) = \binom{n}{k} * 0{,}5^n$$

- bei größeren Stichprobenumfängen n ≥ 36 → Binomialverteilung lässt sich durch Normalverteilung mit Erwartungswert $n * 0{,}5$ und Varianz $n * 0{,}25$ approximieren
  - Schranken für den Annahmebereich

$$0{,}5 \pm (1{,}96 * \sqrt{0{,}25 * n} + 0{,}5)$$

Vorzeichentest für zwei verbundene Stichproben

- Voraussetzung
  - Vergleichbarkeit der Zufallsvariablen der beiden Stichproben in irgendeiner Weise gegeben → Differenz muss nicht exakt quantifiziert werden
- jedem Beobachtungspaar kann ein positives oder negatives Vorzeichen zugeordnet werden
- Nullhypothese
  - etwa gleich viele positive wie negative Vorzeichen sind nötig

$$P(X < Y) = P(X > Y) = 0{,}5$$

- Testverfahren
  - Zuordnung des passenden Vorzeichens zum Beobachtungspaar
  - Stichprobenwerte, die sich nicht unterscheiden, werden nicht berücksichtigt
  - Zählen der Anzahl positiver bzw. negativer Vorzeichen → kleinere Zahl ist Prüfgröße
  - Testentscheidung: Vergleich der Prüfgröße mit kritischem Wert (Tabelle F)

<u>Beispiel</u>
Körpergewicht von n=10

i	x_i	y_i	$d_i=x_i-y_i$	positive Vorzeichen	negative Vorzeichen
1	92,7	85,8	6,9	+	
2	86,2	83,4	2,8	+	
3	102,1	98,3	3,8	+	
4	85,9	83,6	2,3	+	
5	96,3	91,1	5,2	+	
6	90,2	92,7	-2,5		-
7	87,5	88,6	-1,1		-
8	98,0	98,7	-0,7		-
9	89,9	87,1	2,8	+	
10	110,2	102,9	7,3	+	
				$R^+=7$	**$R^-=3$**

k = 3
Annahmebereich, Intervall liegt zwischen 2 und 8
k liegt innerhalb des Intervalls → Beibehaltung der Nullhypothese

Vergleich mit anderen Lagetests

- Vorzeichentest beinhaltet quasi keine Voraussetzungen
- nützt allerdings nicht alle Informationen der Stichprobendaten
- Vorzeichentest → geringere Power als t-Test oder Rangsummentest
- „Schnelltest"

Ausblick auf komplexere Methoden

Mehrstichprobentests

- t-Lagetests und Rangsummentests auf eine oder zwei Stichproben beschränkt

Vergleich von mehr als zwei unverbundenen Stichproben bezüglich einer quantitativen Zielgröße:

1-faktorielle Varianzanalyse
- Erweiterung des t-Tests für mehr als zwei unverbundene Stichproben
- Voraussetzung
  - normalverteilte Grundgesamtheit mit gleichen Varianzen
- Vergleich der Varianz der k Mittelwerte $\bar{x}_i$ $(i = 1, \dots k)$ mit der Varianz der Abstände $(x_{ij} - \bar{x}_i)$
  - Anzahl der Stichproben k ≥ 2
- Quotient dieser Varianzen folgt einer F-Verteilung
  - Wert = 1 – alle Stichproben sind aus derselben Grundgesamtheit
  - je größer F, desto mehr streuen Mittelwerte → Alternativhypothese

Kruskal-Wallis-Test
- Erweiterung des U-Tests für mehr als zwei Stichproben

Vergleich von mehr als zwei verbundenen Stichproben bezüglich einer quantitativen Zielgröße:

Varianzanalyse mit Messwiederholungen
- Vergleich von mehr als zwei Zeitpunkten

Friedman-Test
- basiert auf der Analyse von Rangsummen
- Verallgemeinerung des Wilcoxon-Tests für mehr als zwei verbundene Stichproben

Multiple Methoden

- Einfluss mehrerer Einflussgrößen auf eine Zielgröße wird simultan untersucht

Zwei- oder mehrfaktorielle Varianzanalyse
- Zielgröße hängt nicht nur von einer Gruppierungsvariablen (z.B. Therapieform), sondern darüber hinaus von weiteren qualitativen Faktoren (z.B. Geschlecht, Diagnose, Schweregrad der Krankheit) ab

Varianzanalyse mit Messwiederholungen
- Involvieren mehrerer Gruppierungs- und Messwiederholungsfaktoren

Multiple Regressionsanalyse
- quantitative Zielgröße wird von mehreren quantitativen Einflussgrößen bestimmt

Allgemeines lineares Modell
- Untersuchung des Einflusses mehrerer Faktoren, die sowohl quantitativ als auch qualitativ sein können
- lineare Gleichung, mit der ein Wert für die Zielgröße in Abhängigkeit der signifikanten Einflussfaktoren geschätzt werden kann

$$y = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_k x_k$$

Tests zum Vergleich von Häufigkeiten

Binomialtest für eine Stichprobe

- Stichprobe mit n Beobachtungseinheiten
- Stichprobenwerte sind Ausprägungen eines Alternativmerkmals
- Überprüfung, ob die relative Häufigkeit der Ausprägung A mit einer vorgegebenen Wahrscheinlichkeit p_0 vereinbar ist
- Hypothesen

$$H_0: p = p_0 \qquad H_1: p \neq p_0 \ (bei\ zweiseitiger\ Fragestellung)$$

Testentscheidung

- Zählen der Beobachtungseinheiten der Stichprobe mit der Ausprägung A; Anzahl=X
- relative Häufigkeit $\hat{p} = \frac{X}{n}$ ist Schätzwert für die Wahrscheinlichkeit p der Grundgesamtheit
- Nullhypothese
  - Anzahl X ist binomialverteilt mit Erwartungswert np_0
  - Berechnung der Wahrscheinlichkeiten P(X=k)

$$P(X = k) = \binom{n}{k} * p^k * q^{n-k} \quad für (k = 0, \dots, n$$

  - Konstruktion eines Bereichs, in den X bei Gültigkeit der Nullhypothese mit einer Wahrscheinlichkeit von $1 - \alpha = 95\%$ fallen würde
- falls n hinreichend groß mit $np_0(1 - p_0) \geq 9$
  - Binomialverteilung von X kann durch Normalverteilung mit Erwartungswert np_0 und Varianz $np_0(1 - p_0)$ approximiert werden
  - $\hat{p} = \frac{X}{n}$ normalverteilt mit Erwartungswert p_0 und Varianz $\frac{p_0(1-p_0)}{n}$
  - Prüfgröße einer Standardnormalverteilung
    kritischer Punkt = 1,96 (für $\alpha = 5\%$, bei zweiseitiger Fragestellung) $z_{1-\frac{\alpha}{2}}$
    einseitige Fragestellung $\pm z_{1-\alpha}$

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

<u>Beispiel</u>

- n=71 Studenten
- k=48 (weiblich)
- Nullhypothese: p=0,5 (gleich viele Männer und Frauen studieren Medizin)
- Schätzwert $\hat{p} = \frac{48}{71} = 0{,}68$
- $np_0(1 - p_0) = 71 * 0{,}5 * 0{,}5 = 17{,}75 \geq 9$ → Binomialverteilung von X kann durch eine Normalverteilung mit dem Erwartungswert $\mu = 71 * 0{,}5 = 35{,}5$ und Varianz $17{,}75$ approximiert werden
- Prüfgröße mit p_0=0,5

$$z = \frac{\frac{48}{71} - 0{,}5}{\sqrt{\frac{0{,}25}{71}}} = 2{,}967$$

- 2,967 > 1,96 → Alternativhypothese
- p-Wert = 0,0030 → Ergebnis signifikant
- Konfidenzintervall für $\hat{p}$ → welche Größenordnung der Anteil weiblicher Studenten kann angenommen werden

Chi²-Tests

- zur Analyse von Häufigkeitsunterschieden
- Häufigkeiten lassen sich bei jeder Merkmalsart und jedem Skalenniveau ermitteln
- Tests sind vielseitig verwendbar

Chi²-Vierfelder-Test

- = Chi²-Unabhängigkeitstest - Untersuchung der Unabhängigkeit zweier Alternativmerkmale
- Stichprobe n
- Häufigkeiten, die sich aus der Betrachtung zweier Alternativmerkmale ergeben
- Ausprägungen der beiden Merkmale A und $\bar{A}$, B und $\bar{B}$
  - 4 Kombinationsmöglichkeiten mit den Häufigkeiten a, b, c und d
- Vergleich der relativen Häufigkeiten zweier unabhängiger Stichproben, ob ein bestimmtes Merkmal in den beiden Stichproben gleich verteilt ist
- <u>Chi²-Homogenitätstest</u> – 2 unverbundene Stichproben werden hinsichtlich eines Merkmals miteinander verglichen
  - z.B. Männer und Frauen (2 unverbundene Stichproben) hinsichtlich des Merkmals „Rauchgewohnheiten"

Vierfeldertafel

	$\boldsymbol{A}$	$\bar{\boldsymbol{A}}$	**Randsummen**
$\boldsymbol{B}$	a	b	$n_1 = a + b$
$\bar{\boldsymbol{B}}$	c	d	$n_2 = c + d$
Randsummen	$a + c$	$b + d$	$n = a + b + c + d$

- Nullhypothese
  - Übereinstimmung aller beobachteten Häufigkeiten mit den erwarteten
  - relevante Ereignisse sind unabhängig voneinander → Multiplikationssatz gilt

$$H_0\colon\ P(A|B) = (A)$$

$$\frac{a}{a+b} = \frac{a+c}{n}$$

- Alternativhypothese
  - Abhängigkeit besteht

Testverfahren

- beobachtete Häufigkeiten a, b, c und d werden verglichen mit den Häufigkeiten, die unter der Nullhypothese zu erwarten sind
- Berechnung des Quotienten für jede Häufigkeit

$$\frac{(beobachtete\ Häufigkeit - erwartete\ Häufigkeit)^2}{erwartete\ Häufigkeit} = \frac{(B-E)^2}{E}$$

Beobachtete und erwartete Häufigkeiten beim Vierfelder-Test

	beobachtete Häufigkeit B	**unter H_0 erwartete Häufigkeit E**	$\boldsymbol{\frac{(B-E)^2}{E}}$
	a	$\frac{(a+b)(a+c)}{n}$	$\frac{(ad-bc)^2}{n*(a+b)(a+c)}$
	b	$\frac{(a+b)(b+d)}{n}$	$\frac{(ad-bc)^2}{n*(a+b)(b+d)}$
	c	$\frac{(c+d)(a+c)}{n}$	$\frac{(ad-bc)^2}{n*(c+d)(a+c)}$
	d	$\frac{(c+d)(b+d)}{n}$	$\frac{(ad-bc)^2}{n*(c+d)(b+d)}$
Summe	$\boldsymbol{n}$	$\boldsymbol{n}$	$\boldsymbol{\chi^2}$

- Prüfgröße
  - Summe dieser vier Quotienten
  - annähernd χ^2-verteilt mit einem Freiheitsgrad
  - Prüfgröße umso größer, je mehr die beobachteten von den erwarteten Häufigkeiten abweichen → Alternativhypothese

$$\chi^2 = \frac{n * (ad - bc)^2}{(a+b)(a+c)(c+d)(b+d)}$$

- Beibehaltung der Nullhypothese auf dem α-Niveau
  - Prüfgröße liegt innerhalb des Intervalls $[0, \chi^2_{1;1-\alpha}]$
  - für $\alpha = 5\%$ $\chi^2_{1;0,95} = 3{,}841$ (Tabelle E)

<u>Beispiel</u>
- n=71 Studenten
  - 2 unverbundene Stichproben → männliche und weibliche Studenten
- Betrachtung der Alternativmerkmale „Rauchen" und „Geschlecht"

	beobachtete Häufigkeiten			**erwartete Häufigkeiten**		
	Raucher	*Nichtraucher*		*Raucher*	*Nichtraucher*	
Männer	a=4	b=19	23	a=4,2	b=18,8	23
Frauen	c=9	d=39	48	c=8,8	d=39,2	48
	13	58	71	13	58	71

- 17% der Männer, 19% der Frauen rauchen
- Nullhypothese: Es besteht kein Zusammenhang zwischen Rauchen und Geschlecht

$$\chi^2 = \frac{n * (ad - bc)^2}{(a+b)(a+c)(c+d)(b+d)} = \frac{71 * (4 * 39 - 19 * 9)^2}{23 * 13 * 48 * 58} = 0{,}0192$$

- kritischer Wert = 3,841
- 0,0192 < 3,841 → kein Zusammenhang zwischen den beiden Merkmalen
- p-Wert = 0,8898
- Konfidenzintervall für die Differenz (Anteil Frauen – Anteil Männer) = [-0,18; 0,20]
  - Raucheranteil der Frauen könnte um 20% über dem der Männer liegen; er könnte ebenso gut um 18% geringer sein

Einseitiges Testen
- Vierfelder-Test → zweiseitige Fragestellung
- einseitige Fragestellung
  - z.B. „Rauchen mehr Frauen als Männer?
  - kritischer Wert: $\chi^2_{1;1-2\alpha}$

Voraussetzungen
- jede der erwarteten Häufigkeiten sollte mindestens 5 betragen
- keine der beobachteten Häufigkeiten darf 0 sein
- ansonsten: Fisher's exakter Test

Median-Test
- Anwendung des Vierfelder-Tests ist nicht auf Alternativmerkmale beschränkt
- Vergleich von zwei unabhängigen Stichproben bezüglich eines ordinal skalierten oder eines quantitativen Merkmals möglich
- Test anwendbar, auch wenn Verteilungen der Stichproben ungleich oder unbekannt
  - Voraussetzungen von t-Test oder U-Test erfüllt → diese nutzen wegen höherer Power
- Nullhypothese

$$H_0\colon \tilde{\mu}_1 = \tilde{\mu}_2$$

- Testverfahren
  - Bildung des gemeinsamen empirischen Medians $\tilde{\mu}$ aus den Daten beider Stichproben
  - Ermittlung der Häufigkeiten der Vierfeldertafel
  - Berechnung der Prüfgröße χ^2
  - Wert der Prüfgröße innerhalb von $[0, \chi^2_{1;1-\alpha}]$ → Nullhypothese beibehalten

Vierfeldertafel beim Median-Test

	$\leq \tilde{x}$	$> \tilde{x}$	
Stichprobe 1	a	b	$n_1 = a + b$
Stichprobe 2	c	d	$n_2 = c + d$
	$a + c$	$b + d$	$n = a + b + c + d$

Beispiel
- Unterscheidung der Klausurergebnisse von n_1=23 männlichen und n_2=48 weiblichen Studenten
- Median $\tilde{x} = 8$ von allen n=71 Studenten

	$\leq \tilde{x}$	$> \tilde{x}$	**Σ**
Männer	12	11	$n_1 = 23$
Frauen	24	24	$n_2 = 48$
Σ	36	35	$n = 71$

$$\chi^2 = \frac{n * (ad - bc)^2}{(a+b)(a+c)(c+d)(b+d)} = \frac{71 * (12 * 24 - 11 * 24)^2}{23 * 48 * 36 * 35} = 0{,}029$$

- $\chi^2_{1;0,95} = 3{,}841$
- 0,029 < 3,841 → Nullhypothese (kein Zusammenhang/Unterschied)
- p-Wert = 0,8639

Chi²-Test für k*l Felder
- Verallgemeinerung des Vierfelder-Unabhängigkeitstests
- beide betrachtete Merkmale haben k Ausprägungen $A_1, \dots, A_k$ bzw. l Ausprägungen $B_1, \dots, B_l$
  - Kontingenztafel mit $k * l$ Feldern
- Nullhypothese
  - kein Zusammenhang zwischen den beiden Merkmalen besteht
- Testverfahren
  - Vergleich der beobachteten mit den erwarteten Häufigkeiten
  - n_{ij} … Anzahl der Stichprobenelemente mit der Ausprägungskombination A_i und B_j
  - e_{ij} … unter H_0 erwartete Häufigkeiten
  - Prüfgröße χ^2

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

- $(k-1)(l-1)$ Freiheitsgrade
- Berechnung der erwarteten Häufigkeiten e_{ij} aus den Randsummen
- Homogenitätstest
  - Überprüfung, ob ein Merkmal mit l Ausprägungen in k Stichproben homogen verteilt ist

Voraussetzung
- Anteil der erwarteten Häufigkeiten mindestens 5

Assoziationsmaße für qualitative Merkmale

- Chi²-Unabhängigkeitstest – Nachweis der Existenz einer Assoziation zwischen zwei nominal skalierten Merkmale → keine Angabe über die Stärke des Testergebnisses
- Assoziationskoeffizienten, zur Quantifizierung der Stärke eines solchen Zusammenhangs

Phi-Koeffizient

- Beschreibung des Zusammenhangs zwischen zwei Alternativmerkmalen
- Phi-Koeffizient ist signifikant größer als 0, falls das Ergebnis des Vierfeldertests signifikant ist
- $\Phi = 0$
  - vollkommene Unabhängigkeit der Merkmale
- $\Phi = 1$
  - falls $b = c = 0$
  - Merkmal kann anderes Merkmal präzise vorhersagen

$$\Phi = \sqrt{\frac{\chi^2}{n}}$$

<u>Beispiel</u>

- klinisch-kontrollierte Studie
- 50 Patienten testen neues Medikament bzw. Standardmedikament
- Therapien erfolgreich
  - a=35 (neu)
  - c=25 (Standard)
- Therapien nicht erfolgreich
  - b=15
  - d=25
- Chi²-Test $\chi^2 = 4{,}1667$
- p=0,0412
- Stärke des Zusammenhangs $\Phi = \sqrt{\frac{4{,}1667}{100} = 0{,}204}$
- Zusammenhang ist signifikant, aber eher schwach
- Differenz der Erfolgsraten beträgt 20%
- 95%-Konfidenzintervall liegt zwischen 1,2% und 38,8%

Cramérs Index

- Verallgemeinerung von Φ für $\mathrm{k} * \mathrm{l}$-Kontingenztafeln
  - $R = min(k, l)$
  - für $R = 2$ mit Φ ident

$$CI = \sqrt{\frac{\chi^2}{n * (R - 1)}}$$

Kontingenzkoeffizient von Pearson

- ältestes und bekanntestes Assoziationsmaß
- Maximalwert von CC ist gleich $\mathrm{C_{max}} = \sqrt{\frac{(\mathrm{R}-1)}{\mathrm{R}}}$
- Nachteil: Koeffizient kann 1 nie erreichen → schwer zu interpretieren

$$CC = \sqrt{\frac{\chi^2}{n + \chi^2}}$$

McNemar-Test

- Häufigkeitstest für zwei verbundene Stichproben, die hinsichtlich eines Alternativmerkmals zu vergleichen sind
  - z.B. Patienten mit zwei verschiedenen Therapien werden nacheinander behandelt → Therapieerfolg „ja", „nein"
- Stichprobenumfang n

Vierfeldertafel beim McNemar-Test

		Stichprobe 1	
		A	$\bar{A}$
Stichprobe 2	A	a	b
	$\bar{A}$	c	d

- Nullhypothese
  - Übereinstimmung Stichproben bezüglich der Häufigkeitsverteilung $a + b = a + c$ bzw. $b = c$
- a und d … Häufigkeiten, die für die Gleichheit der Stichproben sprechen
- b und c … repräsentieren die Unterschiede
- Alternativhypothese
  - b und c weichen stark vom Durchschnittswert $\frac{b+c}{2}$ ab
- Testverfahren
  - Ermittlung der Häufigkeiten der Vierfeldertafel
  - Berechnung der Prüfgröße

$$\chi^2 = \frac{(b-c)^2}{b+c}$$

  - Prüfgröße größer als $\chi^2_{1;1-\alpha}$ → Alternativhypothese
- Stetigkeitskorrektur für $\mathrm{b} + \mathrm{c} \leq 30$
  - erforderlich, weil die Häufigkeiten b und c diskrete Werte darstellen und χ^2 eine Variable ist

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c}$$

Voraussetzung

- Anteil der erwarteten Häufigkeiten $\frac{b+c}{2}$ mindestens 5

Nullhypothese

- unterschiedliche Beurteilungen in beiden Richtungen sind gleich häufig b=c

Stichprobenumfang

- in Berechnung der Prüfgröße fließt nicht der volle Stichprobenumfang ein, lediglich die Häufigkeiten b und c
- Konfidenzintervall der Differenz benötigt gesamten Umfang n

Verallgemeinerung auf qualitative Merkmale

- McNemar-Test setzt Alternativmerkmal voraus
- bei Merkmal mit mehr als zwei Ausprägungen entsteht Matrix
  - Symmetrietest von Bowker zur Überprüfung der Symmetrie der Matrix

Verallgemeinerung auf mehrere verbundene Stichproben

- Q-Test von Cochran → wenn Patienten mehrfach nacheinander auf Alternativmerkmal hin untersucht werden

Beispiel
- n=20 Patienten, bei denen Präparat mit Placebo verglichen wird
- Beurteilung der Wirkung der Therapien

		Wirkung des Placebos	
		schwach	stark
Wirkung des Serums	schwach	a=3	b=2
	stark	c=11	d=4

- Prüfgröße

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c} = \chi^2 = \frac{(|2-11|-1)^2}{2+11} = 4{,}923 > 3{,}841$$

- p-Wert = 0,265
- $4{,}923 > 3{,}841$ → Alternativhypothese
- Serum zeigt in 15 von 20 Fällen starke Wirkung, Placebo nur 6
- Konfidenzintervall für diese Wirkungsdifferenz $[0{,}17;\ 0{,}73]$

Chi²-Anpassungstest
- Überprüfung, ob die empirische Verteilung einer Stichprobe vereinbar ist mit einer vermuteten, theoretischen Verteilung
- Vergleich der beobachteten mit den erwarteten Häufigkeiten
  - Berechnung der erwarteten Häufigkeiten, indem man für jede Ausprägung (Klasse, Gruppe) die entsprechende Wahrscheinlichkeit bestimmt und diesen Wert mit dem Stichprobenumfang multipliziert
- Anzahl der Freiheitsgrade $f = k - 1 - r$
  - k … Anzahl der gegebenen Klassen
  - Reduktion durch 1, weil generell Restriktion durch den Stichprobenumfang gegeben ist
  - Anzahl der Freiheitsgrade sind beschränkt durch die Anzahl r der Parameter, die zur Berechnung der erwarteten Häufigkeiten erforderlich sind
- Anzahl der Freiheitsgrade ist abhängig von der Verteilung (durch Nullhypothese definiert)
  - Gleichverteilung: $f = k - 1$
  - Poissonverteilung: $f = k - 2$
  - Normalverteilung: $f = k - 3$
- zur Geringhaltung des β–Fehlers, wird α–Fehler vergrößert → $\alpha = 0{,}10$
  - Annahme der Alternativhypothese erst für p > 0,10

Beispiel
- 80 Versuche: zwischen 0 und 12 Erythrozyten pro Zählkammer
- beobachtete Häufigkeiten: k mit $0 \leq k \leq 12$
- Prüfung, ob die Werte mit der Annahme einer Poissonverteilung vereinbar sind
- Berechnung des Mittelwerts aus den gegebenen Häufigkeiten, als Schätzer für den Erwartungswert $\lambda = 5{,}9125$
- Berechnung der theoretischen Wahrscheinlichkeiten für alle k

$$P(X = k) = \frac{\lambda^k}{k!} * e^{-\lambda}$$

- erwartete Häufigkeiten $\mathrm{P(X = k)} * 80$
  - erwartete Häufigkeiten müssen mindestens 5 betragen → Zusammenfassung der drei ersten und letzten

$$P(X = 3) = \frac{5{,}9125^3}{3!} * e^{-5{,}9125} = 0{,}093196 * 80 = 7{,}45566$$

- Nullhypothese
  - Anzahl der Erythrozyten in einer Zählkammer folgt einer Poisson-Verteilung

k	beobachtete Häufigkeit B	erwartete Häufigkeit E	$\frac{(B-E)^2}{E}$
0-2	5	5,28	0,01475
3	7	7,46	0,02785
4	11	11,02	0,00004
5	12	13,03	0,08167
6	16	12,84	0,77682
7	10	10,85	0,06607
8	7	8,02	0,12884
9	5	5,27	0,01346
10-11	7	5,61	0,34337
Summe	80	79,37	$\chi^2 = 1{,}45287$

- Anzahl der Freiheitsgrade $9 - 2 = 7$
- $\chi^2_{7;0,90} = 12{,}017$ (Tabelle E)
- $1{,}45287 < 12{,}017$ → Beibehaltung der Nullhypothese
- $p - Wert = 0{,}99318$

Logranktest

- basierend auf Chi²-Verteilung, zum Vergleich von Überlebenszeiten
- Vergleich von Überlebensfunktionen $S_1(t)$ und $S_2(t)$ zweier (oder mehrerer) unverbundener Stichproben
  - z.B. Überprüfung, ob sich eine Therapie oder ein prognostischer Faktor auf die Überlebenszeit oder allgemein auf die Zeit bis zum Eintreten eines bestimmten Endereignisses auswirkt
- Berücksichtigung von zensierten Daten möglich
- Vergleich zweier Überlebenszeitkurven
  - Ermittlung der Anzahl der aufgetretenen Endereignisse b_1 und b_2 in den Stichproben
  - Berechnung der Anzahl der Endereignisse e_1 und e_2, die man erwarten würde, wenn die beiden Kurven identisch wären
  - Teststatistik

- Häufigkeiten b_1 und $\mathrm{b}_2$werden durch einfaches Zählen ermittelt
- Erwartungshäufigkeiten
  - Betrachtung der beiden Stichproben
  - Zeitpunkte t_i $(i = 1, \dots, k)$, zu denen in einer der beiden Stichproben ein Endereignis stattfindet
  - Anzahl der dazugehörenden Ereignisse d_i
  - n_{1i}, n_{2i} … Beobachtungseinheiten der 1. bzw. 2. Stichprobe, die zum Zeitpunkt t_i noch leben

$$e_1 = \sum_{i=1}^{k} d_i * \frac{n_{1i}}{n_{1i} + n_{2i}} \qquad e_2 = \sum_{i=1}^{k} d_i * \frac{n_{2i}}{n_{1i} + n_{2i}}$$

- Prüfgröße

$$\chi^2 = \frac{(b_1 - e_1)^2}{e_1} + \frac{(b_2 - e_2)^2}{e_2}$$

Exakter Test nach Fisher

- Alternative zu den Vierfelder-Tests oder Chi²-Tests für $k * l$ Felder
  - wenn Voraussetzungen für diese nicht erfüllt (z.B. erwartete Häufigkeiten zu klein)
- „exakt" – p-Wert als Prüfgröße kann direkt berechnet werden
- Testverfahren (Prinzip im Vierfelderfall)
  - Anordnung der Vierfeldertafel, dass die kleinste Häufigkeit oben links steht (entspricht dann der Häufigkeit a)
  - Berechnung der Wahrscheinlichkeit P

$$P = \frac{(a+b)!\ \ (c+d)!\ \ (a+c)!\ \ (b+d)!}{n! * a! * b! * c! * d!}$$

- Wahrscheinlichkeit $P > \alpha$ → Beendigung des Tests, Beibehaltung der Nullhypothese
- ansonsten: unter Beibehaltung der Randsummen schrittweise Reduktion der Häufigkeit a um 1, bis a=0
- Berechnung der Einzelwahrscheinlichkeit P
- Aufaddierung der berechneten Einzelwahrscheinlichkeiten – wie groß die Wahrscheinlichkeit ist, die Ausgangssituation oder eine noch extremere Situation zu erhalten
  bei einseitiger Fragestellung entspricht dies dem p-Wert
- bei zweiseitiger Fragestellung wird p-Wert der einseitigen Fragestellung verdoppelt
- p-Wert < α → Ablehnung der Nullhypothese

<u>Beispiel</u>

- Vergleich zweier Gruppen von Patienten bezüglich neuer Therapie
- Überprüfung, ob Misserfolgsquoten der beiden Therapie unterschiedlich
- $\alpha = 5\%$

	Misserfolg	**Erfolg**	
Therapie 1	a=0	b=8	8
Therapie 2	c=5	d=3	8
	5	11	16

$$P(a=0) = \frac{(a+b)!\ \ (c+d)!\ \ (a+c)!\ \ (b+d)!}{n! * a! * b! * c! * d!} = \frac{8!\ \ 8!\ \ 5!\ \ 11!}{16! * 0! * 8! * 5! * 3!} = 0{,}0128$$

- $0{,}0128 < \alpha$
- a=5, c=0 → P(a=5) = 0,0128
- 0,0128+0,0128=0,256 < 0,05 → Ablehnung der Nullhypothese für $\alpha = 5\%$

Logistische Regression

- multiples Verfahren, mit dem es möglich ist, die Wahrscheinlichkeit für das Auftreten eines bestimmten Endereignisses basierend auf mehreren Einflussgrößen zu modellieren
- Einflussgrößen - qualitativ und quantitativ
- Zielgröße – binär
  (auch möglich ordinal oder nominal skaliert mit mehreren Ausprägungen)
- Wahrscheinlichkeit für das Auftreten eines Ereignisses A in Abhängigkeit von mehreren signifikanten Einflussgrößen für den Einzelfall
  - Wahrscheinlichkeit zwischen 0 und 1
  - Regressionskoeffizient a_i

$$P(A) = \frac{\exp\,(a_0 + a_1x_1 + \cdots + a_kx_k)}{1 + \exp\,(a_0 + a_1x_1 + \cdots + a_kx_k)}$$