

Vorlesung Statistik – WiSe 2016/17

Teil 4 Beispiele statistischer Tests I 24.11.2016

Prof. Dr. Peter Martus
UKT Tübingen

Studiengänge
Molekulare Medizin FS 3(BSc)
Biochemie FS 3(BSc)
Medizininformatik FS 3 (BSc)

Vorlesung 4

- Durchführung einer randomisierten Studie
- Der Chi-Quadrat Test
- Der t-Test für eine Stichprobe
- Der t-Test für abhängige Stichproben
- Der t-Test für unabhängige Stichproben

Teil 1

Durchführung einer randomisierten Studie

Ablauf eines realistischen Randomized Clinical Trials (RCT)

Schritt 1

Ein Arzt entscheidet anhand von Anamnese und Diagnose, ob ein bestimmter Patient für die Studie geeignet ist.

Schritt 2

Der Patient wird über die Studie informiert und unterschreibt, dass er an der Studie teilnehmen will.

Schritt 3 (Variante 1)

Vor der Studie wurden neue- und Vergleichsmedikamente identisch verpackt und durchnummeriert. Die nächste Charge wird für den neuen Patienten verwendet, weder Arzt noch Patient wissen, welches Medikament verabreicht wird (doppelblinde Studie).

Schritt 3 (Variante 2)

Die Studienleitung fordert per Fax (oder mit Hilfe von TANKs via internet) die Information an, welche Therapie durchgeführt werden soll (offene Studie, z.B. Operation vs. Bestrahlung).

Ablauf eines realistischen Randomized Clinical Trials (RCT)

Schritt 4

Der Patient wird therapiert, das Ergebnis in eine Datenbank eingetragen.

Schritt 5

Die Datenbank wird final überprüft, Fehler korrigiert. Dies sollte auch bei offenen Studie (z.B. durch Weglassen bestimmter Variablen) verblindet erfolgen.

Schritt 6

Die Daten werden final „entblindet“ und statistisch ausgewertet. Die Auswertungsmethodik muss bereits vor der Studie so detailliert wie möglich festgelegt werden.

Der Durchführung der Studie geht eine lange Planungsphase voraus. In dieser Phase werden u.a. die Zielpopulation und die Stichprobengröße, der Outcome und die Struktur der Datenbank festgelegt.

Teil 2

Der Chi-Quadrat Test

Analyse des RCT - Beispiel

Darstellung der Heilungsraten in einer Kreuztabelle

Es werden je 200 Patienten mit Standard und neuer Therapie behandelt. Unter Standard werden 110 Patienten geheilt, unter neuer Therapie 140. Wir erzeugen die folgende Kreuztabelle:

Heilung Therapie	nein	ja	Gesamt
Standard (Kontrolle)	90 45%	110 55%	200 100%
Neu (Experimentell)	60 30%	140 70%	200 100%
Gesamt	150 37.5%	250 62.5%	400 100%

In dieser Tabelle sind die Heilungsraten 55% und 70% sogenannte „Zeilenprozent“.

Analyse des RCT - Beispiels

Vergleich der Heilungsraten mit dem Chi-Quadrat Test

Wir müssen klären, ob der Unterschied 55% vs. 70% im Bereich der Zufallsschwankung liegt oder nicht. Dies geht prinzipiell so ähnlich wie bei einer Stichprobe, allerdings lassen sich die relevanten Wahrscheinlichkeiten nicht mehr mit einfachen Balkendiagrammen darstellen.

Wir erhalten aber am Ende genau wie bei einer Stichprobe einen p-Wert (Überschreitungswahrscheinlichkeit) und schließen einen reinen Zufallseffekt aus, wenn dieser p-Wert maximal 5% beträgt.

In unserem Beispiel erhalten wir eine p-Wert $< 10^{-6}$. Der Unterschied ist also signifikant, wir haben bewiesen, dass die neue Therapie besser ist.

Analyse des RCT - Beispiel

Berechnung der erwarteten Häufigkeiten

Wenn wir davon ausgehen, dass die Unterschiede zwischen den Studienarmen zufällig sind, würden wir die (für beide Arme identische) Heilungsrate auf $250/400 = 62.5\%$ schätzen.

Heilung Therapie	nein	ja	Gesamt
Standard (Kontrolle)	90 $37.5\% \cdot 200 = 75$	110 $62.5\% \cdot 200 = 125$	200 100%
Neu (Experimentell)	60 $37.5\% \cdot 200 = 75$	140 $62.5\% \cdot 200 = 125$	200 100%
Gesamt	150 37.5%	250 62.5%	400 100%

Auf Basis der 62.5% berechnen wir für beide Studienarme die zu erwartenden Häufigkeiten (unterer Eintrag). Der Vergleich beobachtet vs. erwartet ist die Basis des Chi-Quadrat Tests

Berechnung des Chi-Quadrat Tests

Im Gegensatz zum einführenden Beispiel können wir Wahrscheinlichkeiten nicht direkt berechnen, sondern benötigen einen Zwischenschritt, die Berechnung der sog. „Prüfgröße“. Für unser Beispiel lautet die Prüfgröße:

$$\frac{(90 - 75)^2}{75} + \frac{(110 - 125)^2}{125} + \frac{(60 - 75)^2}{75} + \frac{(140 - 125)^2}{125}$$

3 + 1.8 + 3 + 1.8

Die Prüfgröße hat einen speziellen Namen: „Chi-Quadrat“, abgekürzt χ^2 . Wenn der χ^2 -Wert ≥ 3.84 ist, dann ist das Ergebnis signifikant. Die Zahl 3.84 wird auch als kritischer Wert bezeichnet. Das Ergebnis hier ist 9.6, d.h. wir haben statistisch bewiesen, dass die neue Therapie besser ist als die alte.

Allgemeine Bezeichnungen

Für die allgemeine Formel des Chi-Quadrat Tests verwenden wir die folgenden Bezeichnungen:

Heilung Therapie	nein	ja	Gesamt
Standard (Kontrolle)	h_{11} (= 90) e_{11} (= 75)	h_{12} (= 110) e_{12} (= 125)	$h_{11} + h_{12} = 200$ $(= e_{11} + e_{12})$
Neu (Experimentell)	h_{21} (= 60) e_{21} (= 75)	h_{22} (= 140) e_{22} (= 125)	$h_{21} + h_{22} = 200$ $(= e_{21} + e_{22})$
Gesamt	$h_{11} + h_{21} = 150$ $(= e_{11} + e_{21})$	$h_{12} + h_{22} = 250$ $(= e_{12} + e_{22})$	$h_{11} + h_{12} + h_{21} + h_{22}$ $= n = 400 =$ $e_{11} + e_{12} + e_{21} + e_{22}$

Berechnung des Chi-Quadrat Tests

Bezeichnen wir die beobachteten Häufigkeiten mit h_{11} , h_{12} , h_{21} , h_{22} und die erwarteten mit e_{11} , e_{12} , e_{21} , e_{22} , dann lautet die Formel:

$$\chi^2 = \frac{(h_{11} - e_{11})^2}{e_{11}} + \frac{(h_{12} - e_{12})^2}{e_{12}} + \frac{(h_{21} - e_{21})^2}{e_{21}} + \frac{(h_{22} - e_{22})^2}{e_{22}}$$

Schließlich hier noch die Formel für die erwarteten Häufigkeiten, die man aber einfacher direkt aus der Gesamthäufigkeit (hier: 62.5%) berechnen kann.

$$\begin{aligned} e_{11} &= \frac{(h_{11} + h_{12})(h_{11} + h_{21})}{n}, & e_{12} &= \frac{(h_{11} + h_{12})(h_{12} + h_{22})}{n}, \\ e_{21} &= \frac{(h_{21} + h_{22})(h_{11} + h_{21})}{n}, & e_{22} &= \frac{(h_{21} + h_{22})(h_{12} + h_{22})}{n} \end{aligned}$$

Eine vereinfachte Formel zur Berechnung des Chi-Quadrat Tests

In der vereinfachten Formel muss man die erwarteten Häufigkeiten nicht berechnen:

$$\chi^2 = \frac{(h_{11} + h_{12} + h_{21} + h_{22})(h_{11}h_{22} - h_{21}h_{12})^2}{(h_{11} + h_{12})(h_{21} + h_{212})(h_{11} + h_{21})(h_{12} + h_{22})}$$

Im Beispiel

$$\chi^2 = \frac{400 * (90 * 140 - 110 * 60)^2}{(200)(200)(150)(250)} = 9.6$$

Die allgemeine Formel des Chi-Quadrat Tests

Die vereinfachte Formel lässt sich nicht auf größere Tabellen verallgemeinern, die ursprüngliche dagegen schon. Für k Zeilen und l Spalten erhält man:

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

Ein Beispiel mit vier Zeilen und drei Spalten wäre der Vergleich von 4 Tumor-Therapien mit drei möglichen Ergebnissen:

vollständige Tumorrückbildung,
teilweise Tumorrückbildung,
keine Tumorrückbildung.

Freiheitsgrade beim Chi-Quadrat Test

Der kritische Wert für Tabellen mit 2 Zeilen und 2 Spalten war 3.84. Für 4 Zeilen und 3 Spalten wäre er 12.59.

Welcher Vergleichswert verwendet wird, hängt von den Freiheitsgraden der Häufigkeitstabelle ab.

Diese werden mit der Formel

$$(\text{Anzahl Zeilen} - 1) * (\text{Anzahl Spalten} - 1)$$

berechnet. Für 6 Freiheitsgrade ist der kritische Wert also 12.59.

Zusammenfassung

- Abgesehen vom einfachsten Beispiel des Tests der vorherigen Vorlesung können Wahrscheinlichkeiten für Studienergebnisse i.a. nicht direkt berechnet werden.
- Wir müssen eine **Prüfgröße** berechnen und diese mit einem **kritischen Wert** vergleichen. Diese beiden Begriffe gehören zur Basisterminologie eines jeden statistischen Tests wie Nullhypothese, Alternative, Fehler 1.Art etc.!
- Der Chi-Quadrat Test wird für den Vergleich von Häufigkeiten in zwei oder mehreren Gruppen benötigt.
- Grundidee ist der Vergleich von beobachteten mit erwarteten Häufigkeiten.
- Die Prüfgröße lässt sich nach einer allgemeinen Formel (siehe Folie 11) berechnen. Die Zahl der Freiheitsgrade und damit der kritische Werte hängt jedoch von der Zahl der Zeilen und Spalten der zugehörigen Tabelle ab.

Statistische Terminologie

Forschungshypothese (einseitig):

Das neue Medikament ist besser als das alte.

Nullhypothese H_0 :

Die beiden Medikamente sind gleich gut.

Statistische Alternative H_1 : (zweiseitig):

Die beiden Medikamente sind nicht gleich gut.

Fehler 1. Art:

Es wird ein signifikanter Unterschied in den Heilungsraten beider Medikament gefunden, obwohl sie gleich gut sind. Dieser Fehler wurde auf 0.05 festgelegt.

Fehler 2. Art/Power:

Es wird kein Unterschied bewiesen, obwohl das neue Medikament besser ist. Dieser Fehler kann nur mit Vorwissen berechnet werden (siehe VL Fallzahlschätzung)

Statistische Terminologie

Prüfgröße:

χ^2 wie auf Folie 12

Wert der Prüfgröße:

$$\chi^2 = 9.6$$

Freiheitsgrade

$$df = (2-1)*(2-1) = 1$$

Kritischer Wert:

beträgt laut Tabelle/Computer 3.84

Testentscheidung:

Das Ergebnis war signifikant. Es wurde bewiesen, dass die neue Therapie der herkömmlichen überlegen ist.

Die Testentscheidung wird also wieder einseitig formuliert, auch wenn die Testprozedur zweiseitig durchgeführt wurde.

Exakte mathematische Darstellung

Statistisches Modell:

Die Heilungswahrscheinlichkeiten für einen Patienten werden mit p_{standard} und p_{neu} bezeichnet.

In beiden Gruppen ist die Zahl der Heilungen binomialverteilt, also $P(k) = b(n, p, k)$ für $k = 1, \dots, n$ (hier n jeweils 200) mit $p = p_{\text{standard}}$ bzw. $p = p_{\text{neu}}$.

Nullhypothese

$$p_{\text{standard}} = p_{\text{neu}}$$

Alternative

$$p_{\text{standard}} \neq p_{\text{neu}}$$

Wir werden in dieser Vorlesung die „Prosaform“ der Hypothesen verwenden und das jeweilige statistische Modell ebenfalls nur in Prosa beschreiben. In der VL über Wahrscheinlichkeitsrechnung werden wir die exakten mathematischen Formulierungen angeben.

Teil 3

Der t-Test für eine Stichprobe

Beispiel

Im Rahmen der Studie zur Nierenfunktionsbestimmung musste der Kit zur Messung des Cystatin C Werts geändert werden.

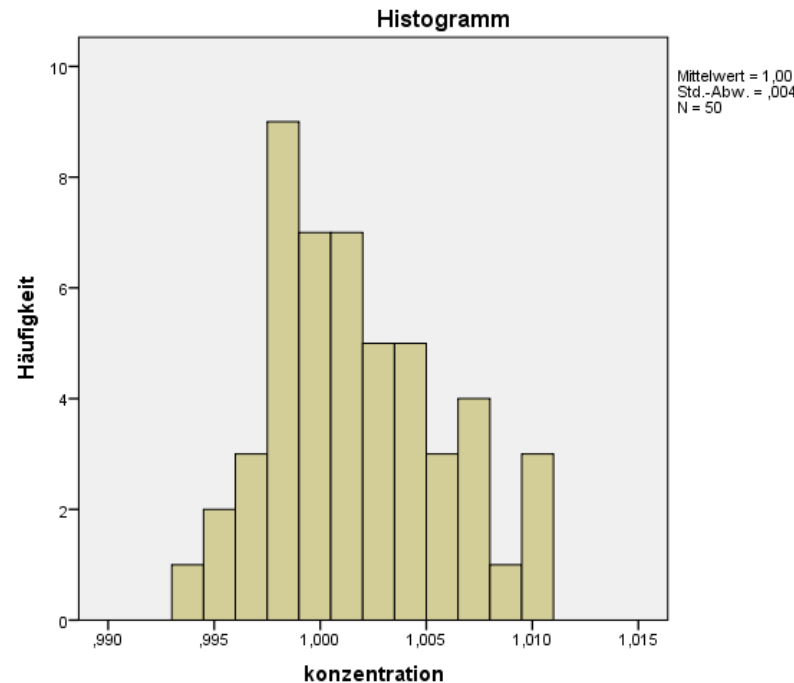
Es besteht aber der Verdacht, dass der neue Kit systematisch höhere Werte liefert als der alte. Um dies dem Hersteller nachzuweisen, wurden 50 Normproben mit der Konzentration 1.0 mg/dl mit dem Kit gemessen. Es wird angenommen, dass die Messwerte normalverteilt sind.

Ergebnisse:

1,0040	0,9988	0,9993	1,0008	1,0069	1,0024	0,9977	1,0104
1,0014	1,0023	0,9998	1,0085	1,0048	1,0037	1,0076	0,9980
1,0011	1,0013	1,0027	1,0050	0,9997	0,9983	0,9983	1,0037
0,9975	0,9998	1,0050	0,9973	0,9955	1,0102	1,0070	1,0010
0,9961	1,0033	0,9958	1,0019	0,9990	0,9979	0,9985	1,0066
1,0056	1,0002	1,0028	0,9944	1,0015	0,9966	0,9991	1,0105
1,0049	0,9992						

Beispiel (Fortsetzung)

Die Analyse der Messreihe ergab einen Mittelwert von 1.0017 bei einer Streuung von 0.00407.



Voraussetzung für die Anwendung des t-Tests ist die Normalverteilung der Daten. Das Histogramm hat zwar nicht „perfekte“ Glockenform, aber wir nehmen Normalverteilung an.

Beispiel (Fortsetzung)

Der t-Test ist durch zwei Überlegungen motiviert:

- Es soll beurteilt werden, ob die Abweichung von 0.0017 gegenüber dem Normwert von 1.0 reiner Zufall gewesen sein kann.
- Die Information über die zufällige Variabilität der Messergebnisse können wir nur aus der Stichprobe ziehen.
- Wir setzen daher die mittlere Abweichung ins Verhältnis zur Streuung, berechnen also $0.0017 / 0.00407 \approx 0.42$. Der Abstand zwischen Messwerten und vorgegebenem Wert beträgt also 0.42 Streuungen.
- Die Ergebnisse sind natürlich umso überzeugender je größer die Stichprobe ist.
- In der Vorlesung Abschnitt „Wahrscheinlichkeitsrechnung“ werden wir sehen, dass es gute Gründe gibt, die Fallzahl nicht proportional sondern nur mit dem Faktor $\sqrt{n}$ in den Test einfließen zu lassen.

Berechnungen

Die Prüfgröße für den t-Test wird mit folgender Formel berechnet:

$$t = \frac{1.0017 - 1.0}{0.402} \sqrt{50} = 2.912$$

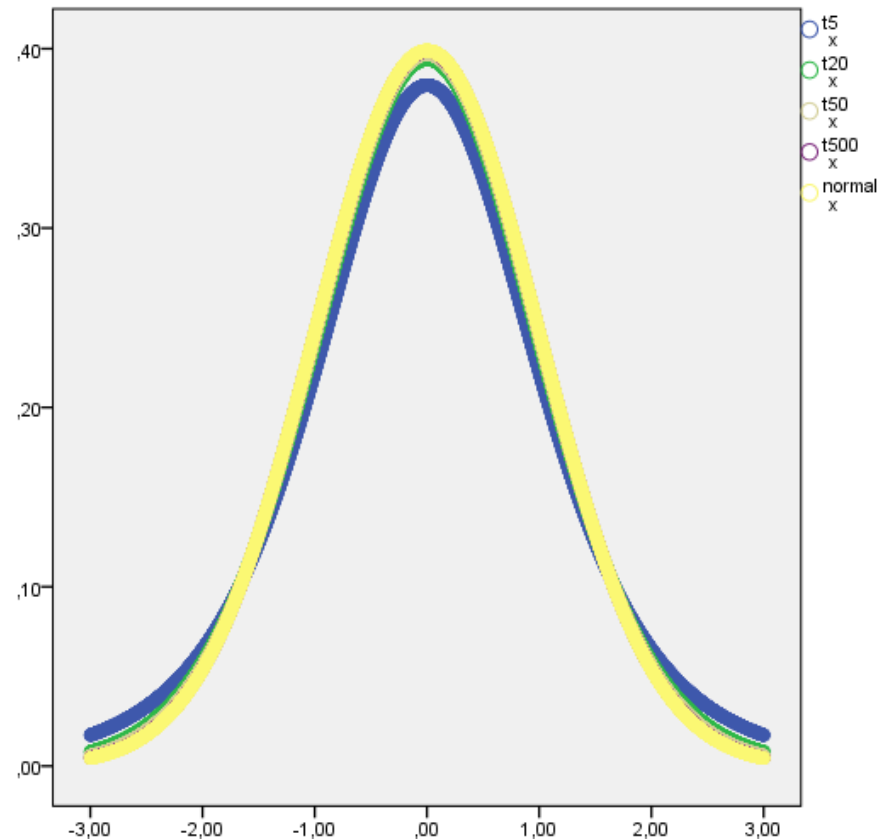
Bzw. allgemein

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

Dabei ist μ der (theoretische, angeblich korrekte) Mittelwert, $\bar{x}$ und s sind Mittelwert und Streuung aus der Stichprobe, n bezeichnet den Stichprobenumfang.

Kritischer Wert, Signifikanz und p-Wert

Die Prüfgröße t ist t-verteilt, d.h. „in etwa“ normalverteilt. (Würde man die Streuung kennen und nicht aus der Stichprobe entnehmen, wäre sie normalverteilt, solange nur die Daten normalverteilt sind).



Kritischer Wert, Signifikanz und p-Wert

Bei normalverteilter Prüfgröße wäre der kritische Wert 1.96 (Warum?).

Die t-Verteilung und damit kritische Werte, Signifikanz und p-Werte hängen von der Fallzahl ab. Je größer die Fallzahl, desto genauer die Schätzung der Streuung und desto näher ist die t-Verteilung an der Normalverteilung.

Auch beim t-Test kann man Freiheitsgrade definieren. Beim t-Test für eine Stichprobe vom Umfang n sind sie $n-1$.

Der kritische Wert beträgt für 49 Freiheitsgrade 2.01, das Ergebnis war signifikant, weil die Prüfgröße (2.912) in der Stichprobe größer als der kritische Wert war.

Der p-Wert war 0.005.

Statistische Terminologie

Forschungshypothese (einseitig):

Der neue Kit produziert höhere Werte als der alte.

Nullhypothese H_0 :

Der neue Kit produziert im Mittel Messwerte von 1.0.

Statistische Alternative H_1 : (zweiseitig):

Der neue Kit produziert im Mittel Messwerte ungleich 1.0.

Fehler 1.Art:

Es wird eine signifikante Abweichung von 1.0 gefunden, obwohl der neue Kit korrekt misst. Dieser Fehler wurde auf 0.05 festgelegt.

Fehler 2.Art/Power:

In der Studie wird keine Abweichung gefunden, obwohl der neue Kit falsch misst. Dieser Fehler könnte nur mit Vorwissen berechnet werden (siehe VL Fallzahlschätzung)

Statistische Terminologie

Prüfgröße:

t wie auf Folie 24

Wert der Prüfgröße:

$t = 2.912$

Freiheitsgrade

$df = 50 - 1 = 49$

Kritischer Wert:

beträgt laut Tabelle/Computer 2.01.

Testentscheidung:

Das Ergebnis war signifikant. Es wurde bewiesen, dass der neue Kit systematisch höhere Werte produziert als der alte.

Die Testentscheidung wird also wieder einseitig formuliert, auch wenn die Testprozedur zweiseitig durchgeführt wurde.

Teil 4

Der t-Test für abhängige Stichproben

Beispiel

Im vergangenen Beispiel vermuteten wir, dass ein neuer Kit zur Bestimmung des Serum-Cystatin s systematisch höhere Werte als der alte liefert. Wir haben hierfür mit dem neuen Kit normierte Proben gemessen und die Ergebnisse gegen den Normwert verglichen.

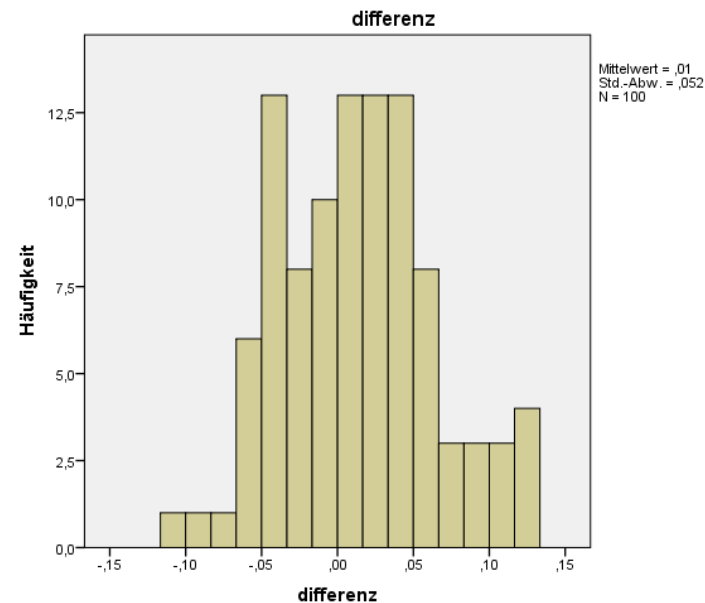
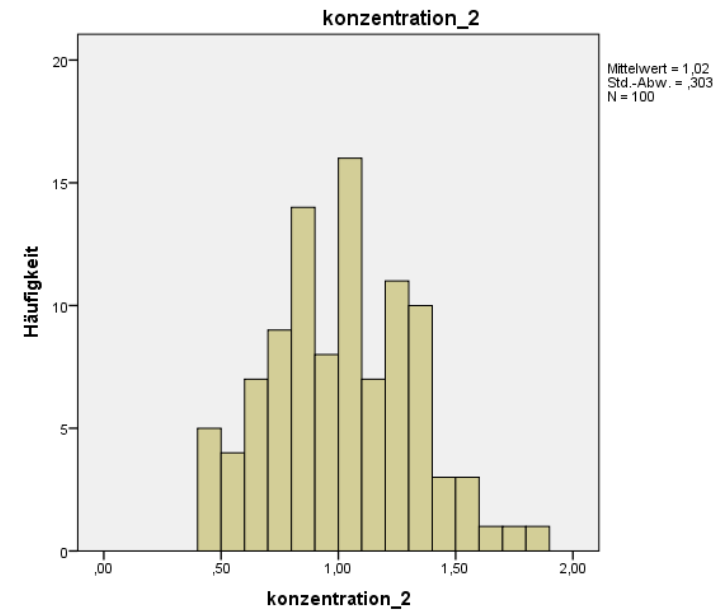
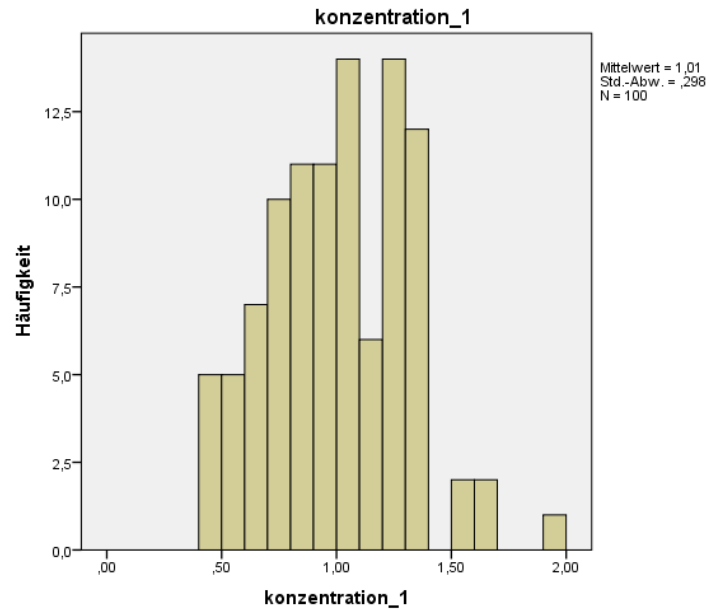
Ein anderer Ansatz wäre gewesen, anstelle der wiederholten Messung von Normproben „reale“ Proben zweifach zu bestimmen, mit dem neuen und dem herkömmlichen Kit.

Auf diese Art wurden 100 Proben doppelt getestet. Die Ergebnisse finden Sie in einer Textdatei in ILIAS.

Wir erhielten folgende Einzelergebnisse (Mittelwert $\pm$ Streuung):

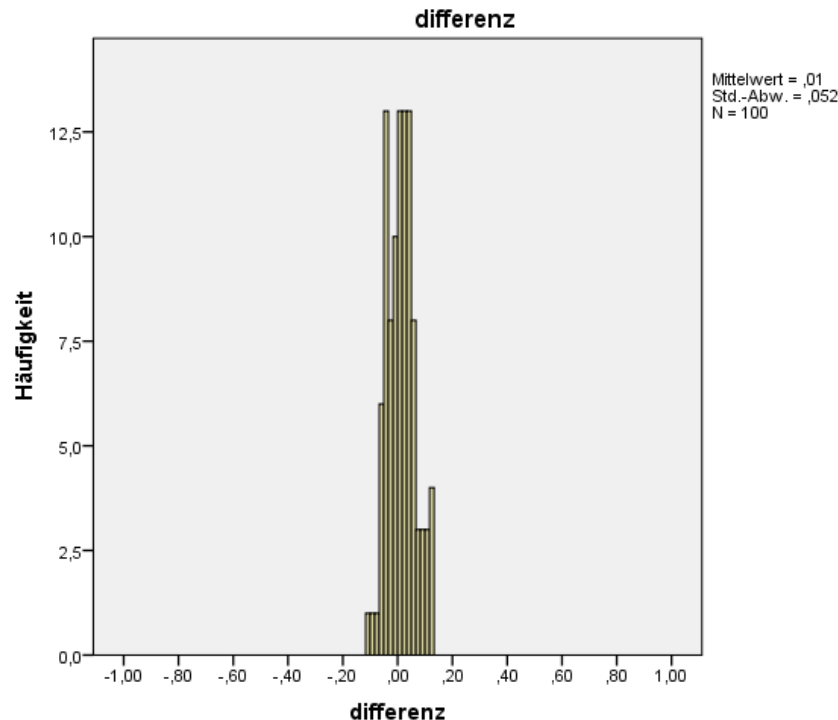
Messungen alter Kit:	1.0063 ± 0.298
Messungen neuer Kit:	1.0190 ± 0.303
Differenzen der Messwerte:	0.0127 ± 0.0517

Fortsetzung Beispiel



Fortsetzung Beispiel

Wählt man für die Differenzengraphik den gleichen Maßstab wie für die Einzelwerte, erkennt man, dass die Streuung der Differenzen deutlich niedriger liegt, als diejenige der Einzelwerte.



Für die Anwendung eines T-Tests genügt es, dass die Differenzen normalverteilt sind, die beiden Messreihen für sich genommen müssen nicht normalverteilt sein.

Der t-Test für abhängige Stichproben

Messungen alter Kit: 1.0063 ± 0.298

Messungen neuer Kit: 1.0190 ± 0.303

Differenzen der Messwerte: 0.0127 ± 0.0517

Wir benötigen wieder eine t-Test Formel der Art „Mittelwertsdifferenz/Streuung*Fallzahlfaktor“.

$$t = \frac{\Delta}{S} f(n)$$

Die Mittelwertsdifferenz Δ ist unproblematisch, wir berechnen sie entweder als Differenz der Mittelwerte $1.0190 - 1.0063$ oder direkt als 0.0127 , dem Mittelwert der Differenzenstichprobe.

Aber welche Fallzahl (200 Messwerte oder 100 Differenzen?) und welche der Streuungen sollten wir verwenden?

Der t-Test für abhängige Stichproben

Lösung:

Wir dürfen die Streuung der Differenzen verwenden, aber entsprechend auch nur die Fallzahl der Differenzen, also $n = 100$.

$$t = \frac{0.0127}{0.0517} \sqrt{100} \approx 2.467$$

Der zugehörige p-Wert (t-Verteilung mit 99 Freiheitsgraden) ist 0.015.

Wir haben also auch mit dieser Versuchsreihe statistisch bewiesen, dass der neue Kit systematisch höhere Werte produziert als der alte.

Statistische Terminologie

Forschungshypothese (einseitig):

Der neue Kit produziert höhere Werte als der alte.

Nullhypothese:

Der neue Kit produziert im Mittel gleich hohe Werte wie der alte.

Statistische Alternative (zweiseitig):

Der neue Kit weist eine systematische Abweichung zum alten Kit auf.

Fehler 1.Art:

Es wird eine Abweichung gefunden, obwohl der neue Kit korrekt misst. Dieser Fehler wurde auf 0.05 festgelegt.

Fehler 2.Art/Power:

In der Studie wird keine Abweichung gefunden, obwohl der neue Kit falsch misst. Dieser Fehler könnte nur mit Vorwissen berechnet werden (siehe VL Fallzahlschätzung)

Statistische Terminologie

Prüfgröße:

t wie auf Folie 21

Wert der Prüfgröße:

$t = 2.912$

Freiheitsgrade:

$df = 100 - 1 = 99$

Kritischer Wert:

beträgt laut Tabelle/Computer 1.984

Testentscheidung:

Das Ergebnis war signifikant. Es wurde bewiesen, dass der neue Kit systematisch höhere Werte produziert als der alte.

Die Testentscheidung wird also wieder einseitig formuliert, auch wenn die Testprozedur zweiseitig durchgeführt wurde.

Vergleich der beiden t-Tests für eine und für abhängige Stichproben

Man müsste im vorausgegangenen Beispiel mit den zwei Messwerten gar nicht die Einzelwerte kennen, es würde vollkommen ausreichen, die Differenzstichprobe zu analysieren.

Generell ist ein t-Test für 2 abhängige Stichproben äquivalent zu einem t-Test für 1 Stichprobe, man muss nur die Differenzenstichprobe bilden.

Der Vergleichswert für die Differenzenstichprobe ist immer Null, für eine Stichprobe aber beliebig.

Teil 4

Der t-Test für unabhängige Stichproben

Beispiel

In der Studie zur Nierenfunktion wurde vermutet, dass die Kreatininwerte bei Frauen systematisch niedriger liegen als bei Männern.

Zu diesem Zweck wurden Serumkreatininwerte bei 325 Männern und 243 Frauen miteinander verglichen.

Wir erhielten folgende Einzelergebnisse (Mittelwert $\pm$ Streuung):

Serumkreatinin Männer: 1.0940 ± 0.3397

Serumkreatinin Frauen: 0.8372 ± 0.2105

Allerdings waren Schiefe und Kurtosis jeweils außerhalb des Intervalls $[-1,1]$:

Schiefe / Kurtosis Männer: $1.764 / 3.893$

Schiefe / Kurtosis Frauen: $1.174 / 1.439$

Weil die Voraussetzung für den t-Test Normalverteilung der Werte in beiden Stichproben ist, entschied man sich dafür, die logarithmierten Werte „auszuprobieren“.

Fortsetzung Beispiel

Für die logarithmierten Werte erhielten wir

log. Serumkreatinin Männer: 0.0501 ± 0.2718

log. Serumkreatinin Frauen: -0.2060 ± 0.2338

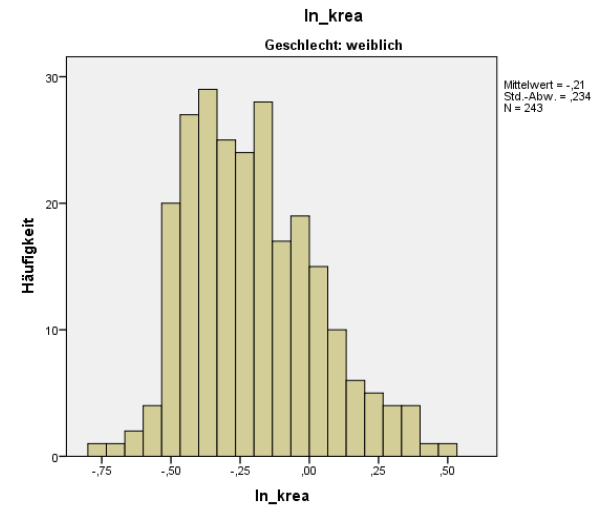
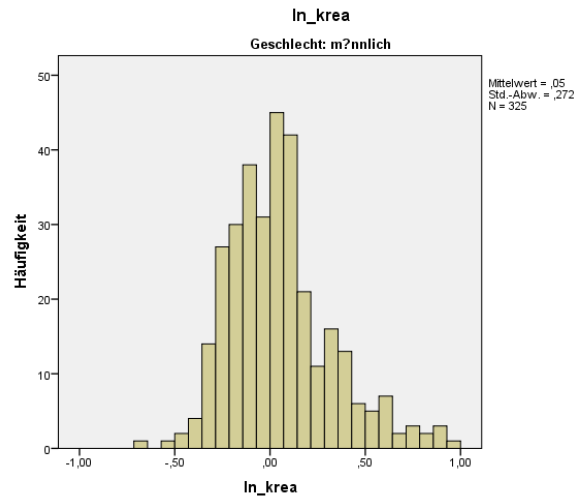
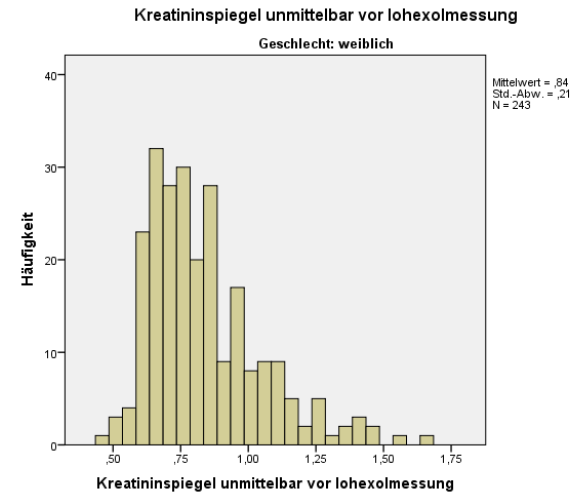
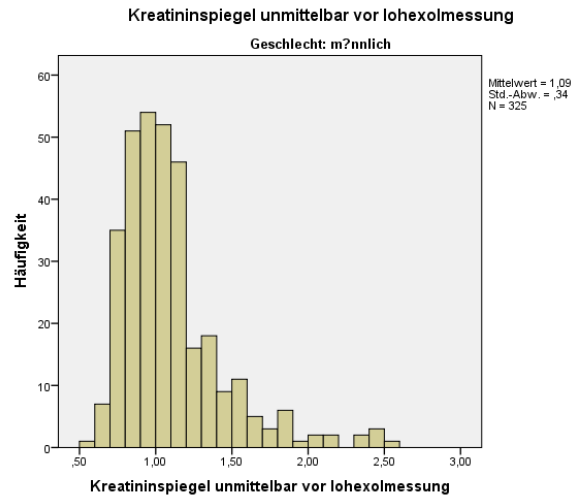
Hier waren Schiefe und Kurtosis jeweils im Intervall $[-1,1]$:

Schiefe / Kurtosis Männer: $0.812 / 0.881$

Schiefe / Kurtosis Frauen: $0.553 / 0.005$

Aus diesem Grund entschied man sich dafür, die logarithmierten Werte zu verwenden, vergleiche auch die Histogramme auf der nächsten Folie.

Fortsetzung Beispiel



Der t-Test für unabhängige Stichproben

Die Formel für den t-Test für unabhängige Stichproben hat die gleiche Struktur wie bei den bereits besprochenen t-Tests.

$$t = \frac{\text{Differenz}}{SD} f(\text{Fallzahl})$$

$$t = \frac{\Delta}{S} f(n)$$

Allerdings sind die Terme für Streuung und Fallzahl komplizierter als die bisherigen.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{pooled}} \sqrt{\frac{n_1 \bullet n_2}{n_1 + n_2}}$$

Die Formel für die gepoolte Streuung steht auf der nächsten Folie.

Die Formel für die gepoolte Streuung

$$s_{pooled} = \sqrt{\frac{1}{n_1 + n_2 - 2} \left[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 \right]}$$

Man kann sich die Formel so erklären, dass die Varianzen der Strichprobe gewichtet nach Stichprobenumfang gemittelt werden. Für die Streuung wird dann einfach die Wurzel gezogen.

Bei gleichen Stichprobengrößen ist die Formel für die gepoolte Varianz der Mittelwert der Einzelvarianzen:

$$s_{pooled}^2 = \frac{1}{n + n - 2} \left[(n - 1)s_1^2 + (n - 1)s_2^2 \right] = \frac{1}{2} \cdot \left[\frac{n - 1}{n - 1} s_1^2 + \frac{n - 1}{n - 1} s_2^2 \right] = \frac{1}{2} \cdot [s_1^2 + s_2^2]$$

Die Formel ist nur gültig, wenn die Unterschiede der Varianzen in beiden Gruppen zufällig sind. Für systematisch verschiedene Varianzen gibt es eine Modifikation des t-Tests (Welsh-Test). Statistikprogramme berechnen automatisch beide Varianten.

Ergebnisse, Freiheitsgrade und Hypothesenprüfung

log. Serumkreatinin Männer: 0.0501 ± 0.2718 , $n_1 = 325$

log. Serumkreatinin Frauen: -0.2060 ± 0.2338 , $n_2 = 243$

$$s_{pooled} = 0.2562$$

$$t = \frac{0.0501 - (-0.2060)}{s_{pooled}} \sqrt{\frac{325 \cdot 243}{325 + 243}} = 11.787$$

Das Ergebnis ist signifikant zum 5% Niveau mit einem p-Wert von $< 10^{-28}$.

Statistische Terminologie

Forschungshypothese (einseitig):

Frauen haben niedrigere Kreatininwerte als Männer.

Nullhypothese:

Frauen und Männer haben gleiche Kreatininwerte.

Statistische Alternative (zweiseitig):

Frauen und Männer haben unterschiedliche Kreatininwerte.

Fehler 1.Art:

Es wird ein Unterschied zwischen Männern und Frauen gefunden, obwohl sie gleiche Kreatininwerte haben. Dieser Fehler wurde auf 0.05 festgelegt.

Fehler 2.Art/Power:

Es wird kein Unterschied zwischen Männern und Frauen gefunden, obwohl in Wahrheit einer existiert. Dieser Fehler könnte nur mit Vorwissen berechnet werden (siehe VL Fallzahlschätzung)

Statistische Terminologie

Prüfgröße:

† wie auf Folie 42/43

Wert der Prüfgröße:

† = 11.787

Freiheitsgrade:

$df = 325 + 243 - 2 = 566$

Kritischer Wert:

beträgt laut Tabelle/Computer 1.964

Testentscheidung:

Das Ergebnis war signifikant. Es wurde bewiesen, dass Männer höhere Kreatininwerte als Frauen haben.

Die Testentscheidung wird also wieder einseitig formuliert, auch wenn die Testprozedur zweiseitig durchgeführt wurde.

Zusammenfassung

Allen drei behandelten t-Tests liegt die Idee zugrunde, für normalverteilte Daten eine beobachtete Mittelwertsdifferenz mit Hilfe der beobachteten Streuung zu normieren und den Einfluss der Fallzahl durch eine Wurzelfunktion zu berücksichtigen.

Während die Berechnung der jeweiligen Mittelwertsdifferenz unmittelbar einleuchtend ist, unterscheiden sich sowohl die Formeln für die Streuungen (Streuung der Messwerte, der Differenzen oder gepoolt aus den Messwerten der Gruppen) als auch die Fallzahlformeln (Wurzel n oder Wurzel aus $(n_1 * n_2) / (n_1 + n_2)$).

Die Normalverteilungsannahme gilt entweder für die Messwerte (eine Stichprobe, unabhängige Stichproben) oder für die Differenzen (abhängige Stichproben).

Die t-Tests für eine und für abhängige Stichproben unterscheiden sich nur marginal.