Schließende Statistik

Am Beispiel des
Chi-Quadrat-Tests

Sind Männer Angsthasen ?

<u>Frage:</u> Gibt es einen Zusammenhang zwischen dem Geschlecht und dem Wunsch nach Lokalanästhesie vor einer Zahnbehandlung? D.h., unterscheidet sich der Anteil der Männer, die Lokalanästhesie wünschen vom Anteil der Frauen?

Betrachten zwei Behauptungen:

<u>Nullhypothese,</u> H_0:
Es besteht **<u>kein</u>** Zusammenhang zwischen Geschlecht und dem Wunsch nach Lokalanästhesie. (Die Anteile sind gleich)

<u>Alternativhypothese,</u> H_A:
Es besteht ein Zusammenhang zwischen Geschlecht und dem Wunsch nach Lokalanästhesie. (Die Anteile unterscheiden sich)

Sind Männer Angsthasen ?

Daten der letzten beiden Tage aus Ihrer Praxis:

	Lokalanästhesie gewünscht	nicht gewünscht	Summe
Männer	11	9	20
Frauen	7	13	20
Summe	18	22	40

Anteil Lokalanästhesien bei Männern: 11/20 = 55%
Anteil Lokalanästhesien bei Frauen: 7/20 = 35%
Anteil Lokalanästhesien gesamt: 18/40 = 45%

Sind Männer Angsthasen ?

Angenommen, der Wunsch nach Lokalanästhesie ist tatsächlich unabhängig vom Geschlecht (d.h. bei Männern und Frauen gleich ausgeprägt), welche „Zellenhäufigkeiten“ würden wir erwarten?

	Lokalanästhesie gewünscht	nicht gewünscht	Summe
Männer	?	?	20
Frauen	?	?	20
Summe	18	22	40

Anteil Lokalanästhesien gesamt: 18/40 = 45%

Sind Männer Angsthasen ?

Angenommen, der Wunsch nach Lokalanästhesie ist tatsächlich unabhängig vom Geschlecht (d.h. bei Männern und Frauen gleich ausgeprägt), welche „Zellenhäufigkeiten“ würden wir erwarten?

	Lokalanästhesie gewünscht	nicht gewünscht	Summe
Männer	9	11	20
Frauen	9	11	20
Summe	18	22	40

<u>zB:</u> Zelle links oben: 20*18/40 = 9.

Chi-Quadrat-Test

<u>Gesucht:</u> Maßzahl, die den Abstand zwischen der “beobachteten” und der unter H_0 “erwarteten” Tabelle erfaßt.

beobachtet

	LA ja	LA nein	Summe
Männer	11	9	20
Frauen	7	13	20
Summe	18	22	40

unter H_0 erwartet

	LA ja	LA nein	Summe
Männer	9	11	20
Frauen	9	11	20
Summe	18	22	40

- Berechne $\frac{(\text{beobachtet - erwartet})^2}{\text{erwartet}}$ für alle 4 Zellen
- Aufsummieren der erhaltenen Werte zu <u>einer</u> Zahl:

$$\chi^2 = \frac{(11-9)^2}{9} + \frac{(9-11)^2}{11} + \frac{(7-9)^2}{9} + \frac{(13-11)^2}{11} = 1.616$$

	LA ja	LA nein	Summe
Männer	11	9	20
Frauen	7	13	20
Summe	18	22	40

Frage: Wie wahrscheinlich ist es, dieses spezielle Resultat $(\chi^2 = 1.616)$ oder ein "noch extremeres" zu beobachten, unter der Annahme, dass kein Zusammenhang besteht (H_0)?

„Noch extremer" ?:

	LA ja	LA nein	Summe
Männer	$\uparrow$	$\downarrow$	20
Frauen			20
Summe	18	22	40

	LA ja	LA nein	Summe
Männer	11	9	20
Frauen	7	13	20
Summe	18	22	40

Frage: Wie wahrscheinlich ist es, dieses spezielle Resultat $(\chi^2 = 1.616)$ oder ein "noch extremeres" zu beobachten, unter der Annahme, dass kein Zusammenhang besteht (H_0)?

„Noch extremer" ?:

	LA ja	LA nein	Summe
Männer	12	8	20
Frauen	6	14	20
Summe	18	22	40

	LA ja	LA nein	Summe
Männer	9	11	20
Frauen	9	11	20
Summe	18	22	40

Männer	10	10	20
Frauen	8	12	20
Summe	18	22	40

Männer	11	9	20
Frauen	7	13	20
Summe	18	22	40

Männer	12	8	20
Frauen	6	14	20
Summe	18	22	40

Männer	13	7	20
Frauen	5	15	20
Summe	18	22	40

Männer	14	6	20
Frauen	4	16	20
Summe	18	22	40

Männer	15	5	20
Frauen	3	17	20
Summe	18	22	40

Männer	16	4	20
Frauen	2	18	20
Summe	18	22	40

Männer	17	3	20
Frauen	1	19	20
Summe	18	22	40

Männer	18	2	20
Frauen	0	20	20
Summe	18	22	40

0

	LA ja	LA nein	Summe
Männer	9	11	20
Frauen	9	11	20
Summe	18	22	40

0,404

Männer	10	10	20
Frauen	8	12	20
Summe	18	22	40

1,616

Männer	11	9	20
Frauen	7	13	20
Summe	18	22	40

3,636

Männer	12	8	20
Frauen	6	14	20
Summe	18	22	40

6,465

Männer	13	7	20
Frauen	5	15	20
Summe	18	22	40

10,10

Männer	14	6	20
Frauen	4	16	20
Summe	18	22	40

14,55

Männer	15	5	20
Frauen	3	17	20
Summe	18	22	40

19,80

Männer	16	4	20
Frauen	2	18	20
Summe	18	22	40

25,86

Männer	17	3	20
Frauen	1	19	20
Summe	18	22	40

32,73

Männer	18	2	20
Frauen	0	20	20
Summe	18	22	40

Chi-Quadrat-Test

Frage: Wie wahrscheinlich ist es, dieses spezielle Resultat $(\chi^2 = 1.616)$ oder ein "noch extremeres" zu beobachten, wenn kein Zusammenhang besteht (H_0)?

$\rightarrow$ **p-Wert !**

Antwort: (mittels Computer)

Gesuchte Wahrscheinlichkeit für $\chi^2 = 1.616$ $\rightarrow$ p = 0,2036

Durchführung des Tests mittels Computer

1) Formulieren von H_0 und H_A
2) Signifikanzniveau festlegen, z.B.: $\alpha = 0.05$
3) Passenden Test auswählen und durchführen
4) p-Wert berechnen
5) p-Wert mit Signifikanzniveau vergleichen
6) Testentscheidung:

   wenn $p < \alpha$ $\rightarrow$ H_0 verwerfen

   wenn $p > \alpha$ $\rightarrow$ H_0 nicht verwefen

Frequency	1	2	Total
1	11	9	20
2	7	13	20
Total	18	22	40

Statistics for Table of geschl by Lokanae

Statistic	DF	Value	Prob
Chi-Square	1	1.6162	0.2036

Zweites Beispiel
zum Chi-Quadrat Test

Fluorid und Zahneruptionsverhalten

Frage: Hat der Fluoridgehalt des Grundwassers einen Einfluß auf das Zahneruptionsverhalten bei Mädchen?

Nullhypothese, H_0:

Es besteht **kein** Zusammenhang zwischen dem Fluoridgehalt des Grundwassers und dem Anteil der 12-jährigen Mädchen mit eruptiertem „45er".

Alternativhypothese, H_A:

Es besteht ein Zusammenhang.

Signifikanzniveau:

5%-Signifikanzniveau, d.h. $\alpha = 0{,}05$

Fluorid und Zahneruptionsverhalten

Die Daten einer Studie mit 91 Patienten

Anzahl der Mädchen	Fluoridgehalt <0.02 ppm (F0)	>1 ppm (F1)	Summe
mit eruptiertem „45er"	41	17	58
mit noch nicht eruptiertem „45er"	20	13	33
Summe	61	30	91

Anteil der Mädchen mit eruptiertem „45er" in F1-Gemeinden: 41/61 =67 %
Anteil der Mädchen mit eruptiertem „45er" in F1-Gemeinden: 17/30 =57 %
Anteil der Mädchen mit eruptiertem „45er" gesamt: 58/91 = 64%

Fluorid und Zahneruptionsverhalten

Angenommen, das Eruptionsverhalten ist tatsächlich unabhängig vom Fluoridgehalt, welche „Zellhäufigkeiten" würden wir erwarten?

Anzahl der Mädchen	Fluoridgehalt <0.02 ppm (F0)	>1 ppm (F1)	Summe
mit eruptiertem „45er"	?	?	58
mit noch nicht eruptiertem „45er"	?	?	33
Summe	61	30	91

Anteil der Mädchen mit eruptiertem „45er" gesamt: 58/91 = 64%

Fluorid und Zahneruptionsverhalten

Angenommen, Eruptionsverhalten ist tatsächlich unabhängig vom Fluoridgehalt, welche „Zellhäufigkeiten" würden wir erwarten?

Anzahl der Mädchen	Fluoridgehalt <0.02 ppm (F0)	>1 ppm (F1)	Summe
mit eruptiertem „45er"	38,88	19,12	58
mit noch nicht eruptiertem „45er"	22,12	10,88	33
Summe	61	30	91

zB: Zelle links oben: 61 •58 /91=38,88

Chi-Quadrat-Test

Gesucht: Maßzahl, die den Abstand zwischen der "beobachteten" und der unter H_0 "erwarteten" Tabelle erfaßt.

Beobachtete Häufigkeiten

	F0	F1	Summe
45er	41	17	58
kein 45er	20	13	33
Summe	61	30	91

Erwartete Häufigkeiten

	F0	F1	Summe
45er	38,88	19,12	58
kein 45er	22,12	10,88	33
Summe	61	30	91

- Berechne $\frac{(\text{beobachtet - erwartet})^2}{\text{erwartet}}$ für alle 4 Zellen

- Aufsummieren der erhaltenen Werte zu <u>einer</u> Zahl:

$$\chi^2 = \frac{(41-38{,}88)^2}{38{,}88} + \frac{(17-19{,}12)^2}{19{,}12} + \frac{(20-22{,}12)^2}{22{,}12} + \frac{(13-10{,}88)^2}{10{,}88} = 0{,}97$$

Fluorid und Zahneruptionsverhalten

Frage: Wie wahrscheinlich ist es, dieses spezielle Resultat ($\chi^2 = 0{,}97$) oder ein "noch extremeres" zu beobachten, wenn kein Zusammenhang besteht (H_0)?

Antwort:

Entweder wieder mittels Computer $\chi^2 = 0{,}97$ → p = 0,32

Da p = 0,32 > α = 0,05 , können wir H_0 auf dem 5%-Signifikanzniveau nicht verwerfen!

→ Wir können keinen Zusammenhang zwischen Fluoridgehalt und Zahneruptionsverhalten nachweisen !

Fluorid und Zahneruptionsverhalten

Oder, mittels Tabelle:

	Signifikanzniveau			
	20%	10%	5%	1%
für 2x2-Tafel:	**1.642**	**2,71**	**3,84**	**6,63**

Da $\chi^2 = 0{,}97 < 3{,}84$, können wir H_0 auf dem 5%-Signifikanzniveau nicht verwerfen!

→ Schließen: Wir können keinen Zusammenhang zwischen Fluoridgehalt und Zahneruptionsverhalten nachweisen !

Verschiedene Fragestellungen - verschiedene Tests

Auswahl des Testverfahrens abhängig von:

- Design /Fragestellung:
  - Gruppenvergleich
  - vor-nach Vergleich
  - Prüfung von Zusammenhängen, Abhängigkeiten
  - ...
- Skalenniveau der Zielvariable(n)
- Anzahl zu vergleichender Gruppen
- Verteilungsannahmen
- ...

Chi-Quadrat-Test

Anwendbar auf nominal und ordinalskalierte Merkmale, und auf metrisch skalierte nach einer Klasseneinteilung.

Voraussetzungen:

- wir interessieren uns für die Unabhängigkeit zweier Merkmale;
- unabhängige Beobachtungseinheiten;
- die erwarteten Häufigkeiten sollten in jeder Zelle mindestens 5 sein (ansonsten Fishers exakter Test)

Allgemeines zum Testprinzip: Fehler erster und zweiter Art

		tatsächlich gilt H_0	tatsächlich gilt H_A
Testentscheidung	H_0	$1-\alpha$	β
	H_A	α	$1-\beta$

Fehler erster Art (α) :
Wahrscheinlichkeit die Nullhypothese fälschlicherweise zu verwerfen.

Fehler zweiter Art (β) :
Wahrscheinlichkeit die Nullhypothese fälschlicherweise nicht zu verwerfen.

Allgemeines zum Testprinzip: Fehler erster und zweiter Art

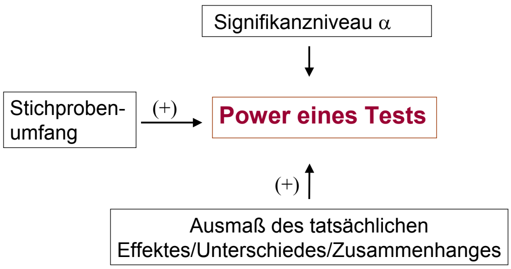
		tatsächlich gilt H_0	tatsächlich gilt H_A
Testentscheidung	H_0	$1-\alpha$	β
	H_A	α	$1-\beta$

Signifikanzniveau α wird "von vorne herein festgelegt"

zB.: $\alpha = 0{,}05$ →

Güte bzw. **Power** eines Tests $(1-\beta)$:
Wahrscheinlichkeit für die korrekte Verwerfung der Nullhypothese.

Allgemeines zum Testprinzip:
Power eines statistischen Tests



Erhöhung der Power durch Erhöhung der Fallzahl

	LA ja	LA nein	Summe
Männer	11	9	20
Frauen	7	13	20
Summe	18	22	40

Erhöhung der Power durch Erhöhung der Fallzahl

Annahme: Verfügen über Daten von letzten 20 Tagen:

	LA ja	LA nein	Summe
Männer	110	90	200
Frauen	70	130	200
Summe	180	220	400

Chi-Quadrat-Test

Annahme: Verfügen über Daten von letzten 20 Tagen:

	LA ja	LA nein	Summe
Männer	110	90	200
Frauen	70	130	200
Summe	180	220	400

	LA ja	LA nein	Summe
Männer	90	110	200
Frauen	90	110	200
Summe	180	220	400

$$\chi^2 = \frac{(110-90)^2}{90} + \frac{(90-110)^2}{110} + \frac{(70-90)^2}{90} + \frac{(130-110)^2}{110} = 16.16$$

p-Wert < 0,0001 → Verwerfen H_0 → Geschlechtsabhängigkeit!

t-Test für *verbundene* Stichproben

Thyroxin-Index-Änderung durch TSH-Injektion

In einer klinischen Studie wurde der freie Thyroxin-Index (FTI) von 12 Patienten **vor und nach** der Injektion des Hypophysenhormons TSH bestimmt.

Das Ziel der Studie war es festzustellen, ob TSH eine Veränderung des Thyroxin-Index bewirkt.

Nullhypothese H_0: Im Mittel keine Änderung des FTI nach TSH-Injektion.

Alternativhypothese H_0: TSH-Injektion verändert FTI im Mittel

Signifikanzniveau: 5%-Signifikanzniveau, d.h. α = 0,05

Patient	Vor Behandung	Nach Behandlung	Differenz $\bar{x}$
1	1,04	1,06	-0,02
2	0,88	0,91	-0,03
3	1,02	1,13	-0,11
4	1,09	1,17	-0,08
5	0,85	1,00	-0,15
6	0,97	1,26	-0,29
7	0,92	1,11	-0,19
8	1,17	1,05	0,12
9	1,07	0,98	0,09
10	0,85	0,95	-0,10
11	0,96	1,04	-0,08
12	0,84	1,05	-0,21
arithmetisches Mittel			-0,09

Thyroxin-Index-Änderung durch TSH-Injektion

Weist die Vor/Nach-Differenz von $\bar{x} = -0.09$ auf eine Verringerung des mittleren FTI durch die TSH-Injektion hin? (Ist die Verringerung des FTI statistisch signifikant?)

Die Antwort hängt vom SEM ab:

Die Differenzen haben Standardabweichung s = 0,118. Der Standardfehler des Mittelwerts ist $SEM = 0{,}118/\sqrt{12} = 0{,}034$.

Thyroxin-Index-Änderung durch TSH-Injektion

t-Teststatistik: $T = \dfrac{\bar{x}}{SEM} = -2{,}65$

Testentscheidung: *(mit Computer)*

Wenn die Nullhypothese gilt, d.h. die TSH-Injektion keine Thyroxin-Änderung bewirkt, ist die Wahrscheinlichkeit, dass der Absolutwert der Teststatistik $|T|$ einen Wert größer oder gleich 2,65 annimmt, p = 0,023.

Da $p < 0{,}05$, ist die Vor/Nach-Differenz signifikant von 0 verschieden!
D.h. wir können schließen, dass die TSH-Injektion den FTI im Durchschnitt verringert.

Thyroxin-Index-Änderung durch TSH-Injektion

Testentscheidung: *(mit Tabelle)*

	Signifikanzniveau		
	10%	5%	1%
n=12 Patienten	**1.796**	**2.201**	**3.106**

Da $|T| = 2{,}65 > 2.201$, ist die Vor/Nach-Differenz signifikant von 0 verscheiden!

D.h. wir können schließen, dass die TSH-Injektion den FTI im Durchschnitt verringert.

Verbundener t-Test

Zum Vergleich der Mittelwerte zweier metrisch skalierter Merkmale.

Voraussetzungen:

- die zu vergleichenden Werte stammen von der gleichen Beobachtungseinheit, oder von zwei gepaarten Beobachtungseinheiten (matched pairs);
- hinreichend grosse Zahl an Beobachtungen oder symmetrische Verteilung der Daten.

t-Test mit unverbundener Stichprobe

Pilocarpin versus Chibro-Timoptol

In einer Therapie-Studie ging es um den Vergleich zweier Medikamente (Pilocarpin und Chibro-Timoptol) zur Behandlung des Weitwinkelglaukoms. Wirksamkeitskriterium sind die Druckwerte nach 120 min.

Nullhypothese H_0: Beide Medikamente führen zu denselben durchschnittlichen Druckwerten

Alternativhypothese H_0: Eines der Medikamente führt zu einem geringeren durchschnittlichen Druck als das andere.

Signifikanzniveau: 5%-Signifikanzniveau, d.h. $\alpha = 0{,}05$

Die Daten

Chibro-Timopol: 2,2 2,2 2,3 2,4 2,4 2,1 2,7 2,6 2,4 2,5 2,6 2,3 2,1 2,0 2,5 2,3

$n_1 = 16$, $\bar{x}_1 = 2{,}350$, $s_1 = 0{,}200$

Pilocarpin: 2,6 2,8 3,1 3,3 2,9 2,5 2,8 2,9 3,1 3,2 3,2 3,0 2,7 3,0 3,0 2,7

$n_2 = 16$, $\bar{x}_2 = 2{,}925$, $s_2 = 0{,}229$

Differenz der arithmetischen Mittel:

$$\bar{x}_1 - \bar{x}_2 = 2{,}350 - 2{,}925 = -0{,}575$$

Pilocarpin versus Chibro-Timoptol

Weist der mittlere Therapieunterschied von -0,575 auf eine stärkere Wirkung von Chibro-Timoptol hin?

Die Antwort hängt wieder vom Standardfehler der Differenz der Mittelwerte (SEDM) ab:

$$SEDM = \sqrt{\frac{s_1^2 + s_2^2}{n}} = \sqrt{\frac{0{,}200^2 + 0{,}229^2}{16}} = 0{,}076$$

Pilocarpin versus Chibro-Timoptol

T-Teststatistik: $T = \dfrac{\bar{x}_1 - \bar{x}_2}{SEM} = -7{,}57$

Die Wahrscheinlichkeit 7,57 oder einen größeren absoluten Wert $|T|$ von T zu beobachten ist $p = 2 \cdot 10^{-8}$.

Testentscheidung: *(mit Computer)*

Da $p = 2 \cdot 10^{-8} < 0{,}05$, ist der Therapieunterschied signifikant!

D.h. wir können schließen, dass Chibro-Timoptol zu einem geringeren Druck führt als Pilocarpin.

Pilocarpin versus Chibro-Timoptol

Testentscheidung: *(mit Tabelle)*

	Signifikanzniveau		
	10%	5%	1%
n=12 Patienten	**1.697**	**2.042**	**2.750**

Da $|T| = 7{,}57 > 2{,}042$, ist der Therapieunterschied signifikant!

D.h. wir können schließen, dass Chibro-Timoptol zu einem geringeren Druck führt als Pilocarpin..

Unverbundener t-Test

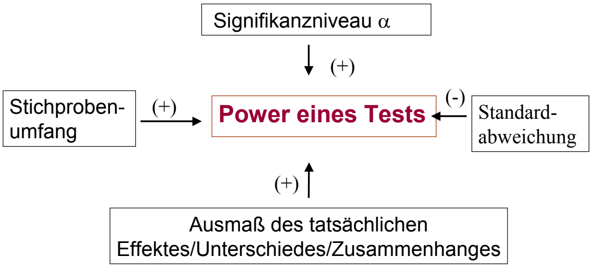
Zum Vergleich der Mittelwerte eines in zwei unabhängigen Gruppen gemessenen metrisch skalierten Merkmals.

Voraussetzungen:

- Die Beobachtungen stammen von unabhängigen Beobachtungseinheiten
- Die Daten sind normalverteilt oder die Stichprobe ist groß.

<u>Allgemeines zum Testprinzip:</u>

Power eines statistischen Tests



Fallen beim statistischen Testen

1. Es ist unzulässig, die Irrtumswahrscheinlichkeit α *nach* der Versuchauswertung so festzulegen, dass der Wert der Prüfgröße den Schwellenwert gerade übersteigt.

2. Es ist falsch, ohne weiteres den Stichprobenumfang so lange zu vergrößern, bis ein „signifikantes" Ergebnis erscheint. Dazu sind spezielle sequentielle statistische Verfahren notwendig.

3. Es ist falsch, auffällige Resultate eines umfangreichen Datenmaterials (z.B. viele verschiedene Merkmale, viele verschiedene Untergruppen) durch nachgeschobene Tests „statistisch abzusichern". Zu prüfende Hypothesen sollten vor der Auswertung (Datenerhebung) festgelegt werden.

Fallen beim statistischen Testen

4. Für zahlreiche Fragestellungen lassen sich mehrer statistische Tests anwenden. Es ist falsch, diese der Reihe nach durch zu probieren, um dann das „günstigste" Testergebnis zu verwenden.

5. Es ist falsch, eine Untersuchung so lange zu wiederholen, bis in einer der Wiederholungen das erhoffte Resultat gefunden wird (ohne die „negativen" Ergebnisse bei der Interpretation zu berücksichtigen).

6. Es ist falsch, ein Untersuchungsergebnis schon deshalb wissenschaftliche „Bedeutsamkeit" (medizinische Relevanz) zukommen zu lassen, weil das Testergebnis „statistisch signifikant" ist.