

Grundlagen der Mathematischen Statistik

Anpassungs- und Unabhängigkeitstests

Uwe Menzel, 2018
uwe.menzel@matstat.org
www.matstat.org

Goodness-of-fit Test (Anpassungstest)

- testet ob ein statistisches Modell zutrifft, d.h. ob das Modell mit beobachteten Daten (Stichprobe) konform ist oder nicht
 - folgen die Beobachtungen einer vermuteten Verteilung?

Modell: fairer Würfel, d.h. $p_i = 1/6$ $i = 1 \dots 6$

Stichprobe: $n = 120$ Würfe



Augenzahl	1	2	3	4	5	6	
E_i : erwartete Häufigkeiten	20	20	20	20	20	20	$E_i = p_i \cdot n$
O_i : beobachtete Häufigkeiten	21	16	28	24	19	12	$\sum O_i = 120$

eine Restriktion

Angesichts dieser Beobachtungen:

- können wir weiterhin glauben, dass das Modell (fairer Würfel) stimmt?
- ... oder müssen wir diese (Null-) Hypothese verwerfen?

www.matstat.org

Goodness-of-fit Test

Teststatistik: $\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$

O_i : observierte Werte (Stichprobe)
 E_i : erwartete Werte (Modell)
 k : Anzahl der Gruppen ($k = 6$ für Würfel)

- Wenn die Nullhypothese (H_0) zutrifft und die Observationen ungefähr den erwarteten Werten entsprechen, sollte χ^2 klein sein, da dann $O_i \approx E_i$
- Wenn die Observationen dagegen stark von den Modellwerten abweichen, wird χ^2 groß und wir neigen dazu, die Nullhypothese zu verwerfen.

Verteilung der Teststatistik unter H_0 :

- Wenn H_0 zutrifft und die erwarteten Häufigkeiten in keiner Kategorie zu klein sind ($E_i \geq 5$), ist die Teststatistik χ^2 -verteilt mit f Freiheitsgraden
- f hängt von der Anzahl der Restriktionen und der Anzahl der Parameter ab, die aus der Stichprobe geschätzt werden müssen, um das Modell aufstellen zu können (siehe weiter hinten)

unter H_0 :

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(f)$$

Sowohl Teststatistik als auch Verteilung werden üblicherweise mit χ^2 bezeichnet!

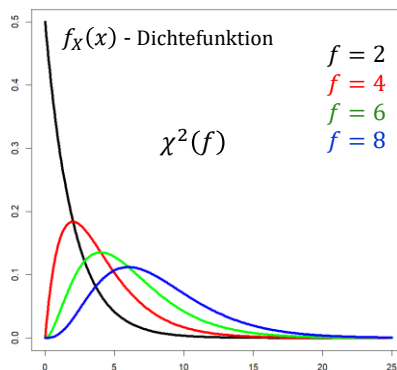
www.matstat.org

Goodness-of-fit Test (Freiheitsgrade)

Die Dichtefunktion $\chi^2(f)$ ist von der Anzahl der Freiheitsgrade (f) abhängig.

Anzahl der Freiheitsgrade
für Anpassungstest:

$$f = k - r - p$$



k = Anzahl der Kategorien ($k = 6$ für Würfel)

r = Anzahl der Restriktionen
($r = 1$ für Würfel da $\sum O_i = 120$)

p = Anzahl der Parameter die aus der Stichprobe geschätzt werden müssen, um das Modell aufstellen zu können ($p = 0$ für Würfel, siehe Anhang)

oft sind $r = 1$ und $p = 0$

⇒ $f = k - 1$ z. B. Würfel

www.matstat.org

Goodness-of-fit Test (Anpassungstest)

Teststatistik χ^2 für zwei Stichproben (Würfelbeispiel):

	1	2	3	4	5	6
E_i	20	20	20	20	20	20
O_i	21	16	28	24	19	12



$$\begin{aligned}\chi^2 &= \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{1^2}{20} + \frac{4^2}{20} + \frac{8^2}{20} + \frac{4^2}{20} + \frac{1^2}{20} + \frac{8^2}{20} \\ &= \underline{\underline{8.1}}\end{aligned}$$

Unterschied zwischen Modell und
Observation klein $\rightarrow \chi^2$ klein.

	1	2	3	4	5	6
E_i	20	20	20	20	20	20
O_i	0	3	2	1	1	113



$$\begin{aligned}\chi^2 &= \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{20^2}{20} + \frac{17^2}{20} + \frac{18^2}{20} + \frac{19^2}{20} + \frac{19^2}{20} + \frac{93^2}{20} \\ &= \underline{\underline{519.2}}\end{aligned}$$

Unterschied zwischen Modell und
Observation groß $\rightarrow \chi^2$ groß.

www.matstat.org

Goodness-of-fit Test

Die Teststatistik χ^2 ist $\chi^2(f)$ -verteilt wenn H_0 wahr ist. Unter H_0 sollte eine Observation von χ^2 also nicht zu weit vom Zentrum der Dichtefunktion für $\chi^2(f)$ entfernt liegen. Wenn die Observation dahingegen zu weit in der rechtsseitigen Flanke der Dichtefunktion liegt, verwerfen wir die Nullhypothese, denn die Observation ist unwahrscheinlich unter H_0 . Die **kritische Region** ist also:

$$\Omega_{krit} = \{ \chi^2 > \chi^2_{\alpha}(f) \} \quad \chi^2_{\alpha}(f): \text{Quantil für } \chi^2\text{-Verteilung mit } f \text{ Freiheitsgraden, für Signifikanzniveau } \alpha$$

- H_0 : Modell und Observation sind konform
- H_a : Modell und Observation sind **nicht** konform

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \quad \text{Teststatistik}$$

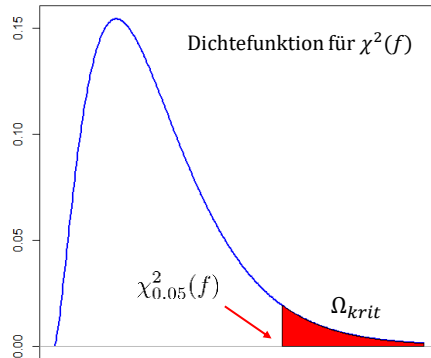
- χ^2 ist immer positiv
- H_0 wird verworfen wenn χ^2 zu große Werte annimmt.

www.matstat.org

Goodness-of-fit Test

$$\Omega_{krit} = \{ \chi^2 > \chi^2_{\alpha}(f) \}$$

$\chi^2_{\alpha}(f)$: Quantil für χ^2 -Verteilung mit f Freiheitsgraden, für Signifikanzniveau α



Teststatistik:

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

O_i : observierte Werte (Stichprobe)

E_i : erwartete Werte (Modell)

$f = k - 1$

www.matstat.org

Goodness-of-fit Test (Anpassungstest)

Beispiel 1:

- Ist das Risiko für einen Herzinfarkt über die Wochentage gleichverteilt – oder ist ein Montag (oder irgendein anderer Tag) vielleicht gefährlicher?
- Mit anderen Worten: ist die Abweichung des Montags-Wertes nur normales "Rauschen" oder ist sie statistisch signifikant?

Sonntag	Montag	Dienstag	Mittwoch	Donnerstag	Freitag	Samstag
24	36	27	26	32	26	29

$n = 200$; Anzahl der Patienten

$H_0 : p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = \frac{1}{7}$ Modell: Gleichverteilung

$E_i = n \cdot p_i = \frac{200}{7} \approx 28.57$ erwartete Anzahl Fälle pro Tag, nach dem Modell

Wenn Herzinfarkte gleichverteilt über die Wochentage sind und man 200 Fälle registriert, dann sollte man ungefähr 28 bis 29 Fälle pro Tag beobachten ...

www.matstat.org

Goodness-of-fit Test

Sonntag	Montag	Dienstag	Mittwoch	Donnerstag	Freitag	Samstag
24	36	27	26	32	26	29

$$E_i = n \cdot p_i = \frac{200}{7} \approx 28.57 \quad \text{erwartete Anzahl Fälle pro Tag, nach dem Modell}$$

$$\begin{aligned} \chi^2 &= \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \quad \text{Teststatistik} \\ &= \frac{(24 - 28.57)^2}{28.57} + \frac{(36 - 28.57)^2}{28.57} + \frac{(27 - 28.57)^2}{28.57} + \frac{(26 - 28.57)^2}{28.57} \\ &\quad + \frac{(32 - 28.57)^2}{28.57} + \frac{(26 - 28.57)^2}{28.57} + \frac{(29 - 28.57)^2}{28.57} = \frac{103.71}{28.57} = \underline{\underline{3.63}} \end{aligned}$$

$$\text{Kritische Region: } \Omega_{krit} = \{ \chi^2 > \chi_{\alpha}^2(f) \} = \{ \chi^2 > \chi_{0.05}^2(6) \} = \{ \chi^2 > 12.59 \}$$

$$\text{Freiheitsgrade: } f = 7 - 1 = 6 \quad (\text{eine Restriktion: } n = 200)$$

Tabelle

www.matstat.org

Goodness-of-fit Test

$$\text{Kritische Region: } \Omega_{krit} = \{ \chi^2 > \chi_{\alpha}^2(f) \} = \{ \chi^2 > \chi_{0.05}^2(6) \} = \{ \chi^2 > 12.59 \}$$

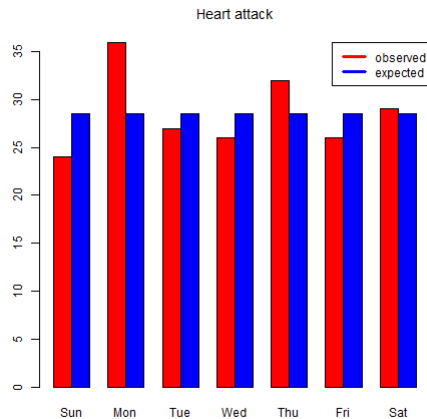
$$\text{Quantile für } \chi^2\text{-Verteilung, } \chi_{\alpha}^2(f) \quad \alpha = 0.05$$

$f \backslash \alpha$	0.25	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
1	1.3233	2.7055	3.8415	5.0239	6.6349	7.8794	10.828	12.116
2	2.7726	4.6052	5.9915	7.3778	9.2103	10.597	13.816	15.202
3	4.1083	6.2514	7.8147	9.3484	11.345	12.838	16.266	17.730
4	5.3853	7.7794	9.4877	11.143	13.277	14.860	18.467	19.997
5	6.6257	9.2364	11.070	12.833	15.086	16.750	20.515	22.105
6	7.8408	10.645	12.592	14.449	16.812	18.548	22.458	24.103
7	9.0371	12.017	14.067	16.013	18.475	20.278	24.322	26.018
8	10.219	13.362	15.507	17.535	20.090	21.955	26.124	27.868

Fazit: $\chi^2 = 3.63 < 12.59 \rightarrow$ Die Observation befindet sich **nicht** in der kritischen Region. Die Nullhypothese wird **nicht** verworfen. Wir können **nicht** nachweisen, dass die Anzahl der Herzinfarkte an irgendeinem Wochentag signifikant höher ist als an anderen Tagen.

www.matstat.org

Goodness-of-fit Test (Anpassungstest)



Barplot:
Vergleich "observed"
und "expected"

Montags unterscheidet sich der beobachtete Wert ziemlich viel von dem nach dem Modell erwarteten. Ein χ^2 -Test hat jedoch gezeigt, dass die "Montagsabweichung" nicht statistisch signifikant ist (auf Signifikanzniveau 0.05).

www.matstat.org

Goodness-of-fit Test (Anpassungstest)



Beispiel 2:

- **Hypothese/Modell:** Vögel bevorzugen keine besondere Baumsorte um Samen zu suchen, sie suchen in allen Baumsorten mit gleicher Wahrscheinlichkeit – desto mehr Bäume einer gewissen Sorte vorkommen, desto mehr Vögel befinden sich bei dieser Sorte.
- **Stichprobe:** $n = 156$ Vögel wurden in einem Wald beobachtet

Mannan, R.W., and E.C. Meslow. 1984. Bird populations and vegetation characteristics in managed and old-growth forests, northeastern Oregon. *J. Wildl. Manage.* **48**: 1219-1238.

	Edeltanne	Kiefer	Küstentanne	Lärche	Summe
Volumen der Krone	54%	40%	5%	1%	100%
erwartet	$156 \cdot 0.54 = 84.24$	$156 \cdot 0.4 = 62.4$	$156 \cdot 0.05 = 7.8$	$156 \cdot 0.01 = 1.56$	156
beobachtet	70	79	3	4	156

1.56 zu klein! → Küstentanne und Lärche werden zusammengefasst

www.matstat.org

Goodness-of-fit Test (Anpassungstest)



$n = 156$	Edeltanne	Kiefer	Küstentanne & Lärche
expected	84,24	62,4	9,36
observed	70	79	7

zusammen
ist $E_i \geq 5$

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{(70 - 84.24)^2}{84.24} + \frac{(79 - 62.4)^2}{62.4} + \frac{(7 - 9.36)^2}{9.36} = \underline{7.418} \quad \text{Teststatistik}$$

$$\Omega_{krit} = \{ \chi^2 > \chi^2_{\alpha}(f) \} = \{ \chi^2 > \chi^2_{0.05}(2) \} = \{ \chi^2 > 5.99 \} \quad \text{Kritische Region}$$

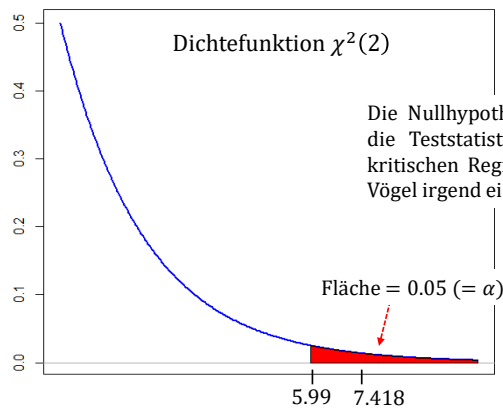
www.matstat.org

Goodness-of-fit Test (Anpassungstest)



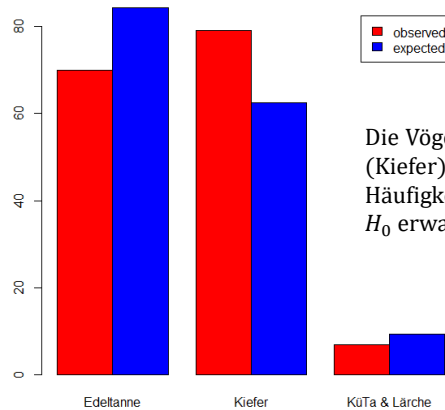
$$\Omega_{krit} = \{ \chi^2 > \chi^2_{\alpha}(f) \} = \{ \chi^2 > \chi^2_{0.05}(2) \} = \{ \chi^2 > 5.99 \} \quad \text{Kritische Region}$$

$$\chi^2 = 7.418 \quad \text{Observation}$$



www.matstat.org

Goodness-of-fit Test (Anpassungstest)



Die Vögel bevorzugten Kategorie 2 (Kiefer) – hier ist die beobachtete Häufigkeit größer als die unter H_0 erwartete.

www.matstat.org

Unabhängigkeitstest

- testet, ob kategoriale Zufallsvariablen unabhängig sind
 - sind zwei Variablen zur Klassifikation einer Population unabhängig?

Beispiel:

- zwei Variablen zur Klassifikation (einer zufällig gewählten Person): Geschlecht und Rauchen (2 nominale Variablen)
- Hängt Rauchen vom Geschlecht ab?
- Annahme (**Modell**): Rauchen und Geschlecht sind unabhängig
- **Stichprobe**: $n = 250$ zufällig gewählte Personen; klassifiziere nach den beiden oben genannten Variablen:

	weibl.	männl.
R	20	30
R^*	80	120

jede Person wird in genau eine der Zellen der Tabelle eingeordnet

R = Raucher/in

R^* = Nichtraucher/in

www.matstat.org

Unabhängigkeitstest

Kontingenztafel (Kreuztafel): Reihen- und Zeilensummen einfügen!

	weibl.	männl.	Σ	
R	20	30	50	← 50 Raucher/innen
R*	80	120	200	← 200 Nichtraucher/innen
Σ	100	150	250	← $n = 250$

Nullhypothese: Rauchen hängt nicht vom Geschlecht ab, beide Variablen sind unabhängig.

Wenn die Nullhypothese stimmt, sollte der Anteil der Raucher bei Frauen und Männern ungefähr gleich sein (und damit auch ungefähr gleich dem Gesamtanteil der Raucher für beide Geschlechter).

Für obige Tabelle:

- Der Gesamtanteil der Raucher beträgt 20% (50/250, letzte Spalte).
 - Der Anteil von Rauchern bei Frauen beträgt 20%.
 - Der Anteil von Rauchern bei Männern beträgt 20%.
- Die Variablen "Geschlecht" und "Rauchen" sind unabhängig.

www.matstat.org

Unabhängigkeitstest

Unter H_0 (Unabhängigkeit):

	w.	m.	Σ
R	20	30	50
R*	80	120	200
Σ	100	150	250

allgemein:

	w.	m.	Σ
R	a_{11}	a_{12}	Z_1
R*	a_{21}	a_{22}	Z_2
Σ	S_1	S_2	n

Unabhängigkeit erfordert (für den Anteil Raucher):

$$\frac{20}{100} \cong \frac{50}{250} \quad \text{Frauen: Anteil Raucher} \cong \text{Gesamtanteil Raucher}$$

$$\frac{30}{150} \cong \frac{50}{250} \quad \text{Männer: Anteil Raucher} \cong \text{Gesamtanteil Raucher}$$



Kriterien für Unabhängigkeit:

$$\frac{a_{11}}{S_1} \cong \frac{Z_1}{n} \rightarrow a_{11} \cong \frac{Z_1 \cdot S_1}{n}$$

$$\frac{a_{12}}{S_2} \cong \frac{Z_1}{n} \rightarrow a_{12} \cong \frac{Z_1 \cdot S_2}{n}$$

www.matstat.org

Unabhängigkeitstest

Unter H_0 (Unabhängigkeit):

	w.	m.	Σ
R	a_{11}	a_{12}	Z_1
R*	a_{21}	a_{22}	Z_2
Σ	S_1	S_2	n

$$a_{11} \cong \frac{Z_1 \cdot S_1}{n} \quad a_{12} \cong \frac{Z_1 \cdot S_2}{n} \quad a_{21} \cong \frac{Z_2 \cdot S_1}{n} \quad a_{22} \cong \frac{Z_2 \cdot S_2}{n}$$

Formeln gelten für den Anteil Raucher/-innen.

analoge Formeln gelten auch für den Anteil Nichtraucher/-innen.

Bei **Unabhängigkeit**:
eine Zelle der Tabelle $\cong$ Zeilensumme $\cdot$ Spaltensumme $/ n$

www.matstat.org

Unabhängigkeitstest

Auch wenn die Nullhypothese stimmt, können wir nicht erwarten, dass die Anteile **genau** gleich sind, da die Stichprobe vom Zufall abhängt:

Realistische Tabelle:

	w.	m.	Σ
R	24	26	50
R*	76	124	200
Σ	100	150	250

Realistische Stichprobe: Eine solche Kreuztabelle kann sich ergeben, selbst wenn beide Variablen unabhängig sind ...

Ideale Tabelle:

	w.	m.	Σ
R	20	30	50
R*	80	120	200
Σ	100	150	250

... denn die Kreuztabelle weicht ja nicht viel von der **idealen** (unter H_0 erwarteten) **Tabelle** ab.

www.matstat.org

Unabhängigkeitstest

Erst wenn die Stichprobenwerte stark von den unter H_0 erwarteten Werten abweichen sind wir geneigt, die Nullhypothese zu verwerfen.

Stichprobe:

	w.	m.	Σ
R	4	46	50
R*	96	104	200
Σ	100	150	250

Ergibt sich eine solche Kreuztabelle können wir vermuten, dass die Variablen **Geschlecht und Rauchen nicht unabhängig** sind.

Ideale Tabelle:

	w.	m.	Σ
R	20	30	50
R*	80	120	200
Σ	100	150	250

... denn die Stichprobe weicht stark von der unter H_0 erwarteten Tabelle ab

www.matstat.org

Unabhängigkeitstest

- **Nullhypothese:** die Variablen sind unabhängig
- **Alternative Hypothese:** die Variablen sind abhängig

Die Nullhypothese wird verworfen, wenn die aus der Stichprobe abgeleitete Kreuztabelle zu stark von der unter H_0 erwarteten (idealen) Kreuztabelle abweicht.

Teststatistik: $\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ O_{ij} : observierte Werte in Zelle i, j
 E_{ij} : erwartete Werte in Zelle i, j

Wenn die beobachteten Tabellenwerte stark von den unter H_0 erwarteten Tabellenwerten abweichen, wird χ^2 groß. **Wir verwerfen die Nullhypothese also für große Werte von χ^2 .**

Verteilung der Teststatistik unter H_0 :

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2(f)$$

χ^2 -Verteilung mit f Freiheitsgraden

www.matstat.org

Unabhängigkeitstest

Verteilung der Teststatistik unter H_0 : $\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2(f)$ χ^2 -Verteilung mit f Freiheitsgraden

Voraussetzungen für χ^2 -Verteilung: $E_i \geq 5$ in jeder Zelle (n groß)

Freiheitsgrade: $f = (r - 1) \cdot (c - 1)$ $r = \text{Anzahl der Zeilen der Kreuztabelle}$
 $c = \text{Anzahl der Spalten der Kreuztabelle}$
 (jeweils ohne Marginalsummen!)

Die Teststatistik χ^2 ist $\chi^2(f)$ -verteilt wenn H_0 wahr ist. Unter H_0 sollte eine Observation von χ^2 also nicht zu weit vom Zentrum der Dichtefunktion für $\chi^2(f)$ entfernt liegen. Wenn die Observation dahingegen zu weit in der rechtsseitigen Flanke der Dichtefunktion liegt, verwerfen wir die Nullhypothese. Die **kritische Region** ist also:

$$\Omega_{krit} = \{ \chi^2 > \chi^2_{\alpha}(f) \} \quad \chi^2 \text{ ist immer positiv; } H_0 \text{ wird verworfen wenn } \chi^2 \text{ groß wird.}$$

www.matstat.org

Unabhängigkeitstest

beobachtet:

	w.	m.	Σ
R	24	26	50
R*	76	124	200
Σ	100	150	250

erwartet (unter H_0):

	w.	mm.	Σ
R	20	30	50
R*	80	120	200
Σ	100	150	250

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \quad \text{Teststatistik}$$

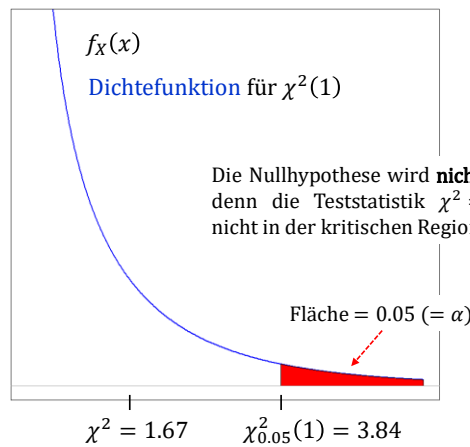
$$= \frac{(24 - 20)^2}{20} + \frac{(26 - 30)^2}{30} + \frac{(76 - 80)^2}{80} + \frac{(124 - 120)^2}{120} = \frac{103.71}{28.57} = \underline{\underline{1.67}}$$

$$\Omega_{krit} = \{ \chi^2 > \chi^2_{\alpha} \} = \{ \chi^2 > \chi^2_{0.05}(1) \} = \{ \chi^2 > \underline{\underline{3.84}} \}$$

H_0 wird **nicht** verworfen. Es kann nicht gezeigt werden (auf Signifikanzniveau $\alpha = 0.05$), dass Rauchen und Geschlecht assoziiert sind.

www.matstat.org

Unabhängigkeitstest



H_0 wird **nicht** verworfen. Es konnte nicht gezeigt werden (auf Signifikanzniveau $\alpha = 0.05$), dass Rauchen und Geschlecht assoziiert sind.

www.matstat.org

Unabhängigkeitstest

Beispiel: Bevorzugen zwei Moränenarten verschiedene Lebensräume ?

Young, R.F., and H.E. Winn. 2003. Activity patterns, diet, and shelter site use for two species of moray eels, *Gymnothorax moringa* and *Gymnothorax vicinus*, in Belize. Copeia 2003: 44-55.



Stichprobe: Es wurde registriert, wie oft sich zwei Moränenarten in bestimmten Habitaten aufhielten:

	<i>G. moringa</i>	<i>G. vicinus</i>	Σ
Gras	127	116	243
Sand	99	67	166
Grenze	264	161	425
Σ	490	344	834

Zwei nominale Variablen:

- Moränenart
- Habitat

www.matstat.org

Unabhängigkeitstest

Beispiel: Bevorzugen zwei Moränenarten verschiedene Lebensräume ?

	G. moringa	G. vicinus	Σ
Gras	127	116	243
Sand	99	67	166
Grenze	264	161	425
Σ	490	344	834

H_0 : Die Arten bevorzugen gleiche Habitate.

H_a : Die Arten bevorzugen unterschiedliche Habitate.

Unter H_0 erwarten wir, dass die prozentualen Aufenthaltsanteile für Gras, Sand und Grenze für beide Arten ungefähr gleich sind, also z. B. 30% Gras, 20% Sand und 50% Grenze für **beide** Arten.

Berechnung der unter H_0 erwarteten (idealen) Kreuztabelle:

$$E_{ij} = \frac{Z_i \cdot S_j}{n}$$

E_{ij} : erwartete Werte in Zelle i, j
 Z_i : Zeilensumme für Zeile i
 S_j : Spaltensumme für Spalte j

Unabhängigkeitstest

beobachtet

O_{ij}	G. moringa	G. vicinus	Σ
Gras	127	116	243
Sand	99	67	166
Grenze	264	161	425
Σ	490	344	834

erwartet

E_{ij}	G. moringa	G. vicinus	Σ
Gras	142.8	100.2	243
Sand	97.5	68.5	166
Grenze	249.7	175.3	425
Σ	490	344	834

$$\frac{243 \cdot 344}{834}$$

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Teststatistik

Es werden immer entsprechende Zellen beider Tabellen "verglichen".

$$= \frac{(127 - 142.8)^2}{142.8} + \frac{(116 - 100.2)^2}{100.2} + \frac{(99 - 97.5)^2}{97.5} + \frac{(67 - 68.5)^2}{68.5} + \frac{(264 - 249.7)^2}{249.7} + \frac{(161 - 175.3)^2}{175.3} = \underline{\underline{6.26}}$$

$$f = (r - 1) \cdot (c - 1) = 2 \cdot 1 = 2$$

Freiheitsgrade

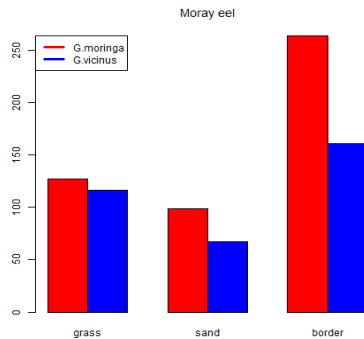
$$\Omega_{krit} = \{\chi^2 > \chi^2_{\alpha}\} = \{\chi^2 > \chi^2_{0.05}(2)\} = \{\chi^2 > \underline{\underline{5.99}}\}$$

www.matstat.org

Unabhängigkeitstest

Teststatistik: $\chi^2 = 6.26$ Kritische Region: $\Omega_{krit} = \{ \chi^2 > 5.99 \}$

Die Teststatistik liegt in der kritischen Region. H_0 wird verworfen. Die Variablen Art und Habitat sind **abhängig**, verschiedene Arten bevorzugen verschiedene Lebensräume.



Die Proportionen für die 3 Habitats sind so unterschiedlich dass die Nullhypothese verworfen werden kann.



?chisq.test

www.matstat.org

Fishers Exakter Test

- testet ob kategoriale Zufallsvariablen unabhängig sind (wie χ^2 -Test oben)
- Unterschied zu χ^2 -Test:
 - keine Normalapproximation (deshalb "exakter" Test)
 - daher auch anwendbar für $E_i \leq 5$ (bzw. gerade dann!)
- H_0 : keine Assoziation zwischen den Variablen
- H_a : die Variablen sind abhängig



Im Folgenden soll der **p-Wert** für den Fisher-Test berechnet werden.

Zur Erinnerung (siehe Vorlesung **F11**):

Der **p-Wert** ist die Wahrscheinlichkeit für mindestens so extreme Testresultate wie das aktuell beobachtete, berechnet unter der Annahme dass die Nullhypothese wahr ist.

Im Gegensatz zu allen anderen bisher in der Vorlesung behandelten Fällen ist das Testresultat beim Fisher-Test kein numerischer Wert einer Teststatistik, sondern die aus der Stichprobe hervorgehende Kreuztabelle selbst (die "observierte Tabelle").

www.matstat.org

Fisher-Test

Um den p -Wert zu erhalten,

- muss also die Wahrscheinlichkeit des Zustandekommens der beobachteten Tabelle unter H_0 berechnet werden,
- außerdem müssen die Wahrscheinlichkeiten (unter H_0) noch extremerer ("weiter von H_0 entfernten") Tabellen hinzugerechnet werden.

Beispiel: Erfolg für 2 Therapieformen A und B (5 Personen in Gruppe A, 7 Personen in Gruppe B, 8 werden gesund, 4 werden nicht gesund)

	gesund	krank	Σ
A	2	3	5
B	6	1	7
Σ	8	4	12

"observierte Tabelle"
= Testresultat

- 40% gesund werdend für Therapie A
- 85% gesund werdend für Therapie B
- 67% gesund werdend insgesamt

Ist der beobachtete Unterschied der Anteile gesund gewordener statistisch signifikant?

www.matstat.org

Fisher-Test

	gesund	krank	Σ
A	2	3	5
B	6	1	7
Σ	8	4	12

Die Wahrscheinlichkeit des Zustandekommens dieser Tabelle "unter H_0 " muss berechnet werden.

Die Wahrscheinlichkeit der beobachteten Tabelle unter H_0 kann berechnet werden, indem die **Marginalsummen** (Σ) als **gegeben** vorausgesetzt werden. Wir setzen also voraus, dass von den 12 Patienten

- 5 Patienten mit Therapie A behandelt wurden
- 7 Patienten mit Therapie B behandelt wurden
- 8 Patienten gesund wurden
- 4 Patienten nicht gesund wurden.

Wir gehen also von vornherein davon aus, dass 8 Patienten gesund werden. Dies entspricht gewissermaßen der Nullhypothese: die Therapie hat keinen Einfluss auf den zukünftigen Gesundheitszustand.

www.matstat.org

Fisher-Test

Die Wahrscheinlichkeit der beobachteten Tabelle unter H_0 ist gleich der Wahrscheinlichkeit, dass von den 8 (sowieso) gesund werdenden genau 2 "zufällig" zur Therapie A ausgewählt werden. Die Wahrscheinlichkeit, dass eine Zelle der Tabelle bei gegebener Marginalverteilung einen bestimmten Wert annimmt, ist **hypergeometrisch** verteilt (Vorlesung F4):

	gesund	krank	Σ
A	$a_{11} = 2$	$a_{12} = 3$	5
B	$a_{21} = 6$	$a_{22} = 1$	7
Σ	8	4	12

Tabellenfelder:

$a_{11} = \{A, \text{gesund}\}$

$a_{12} = \{A, \text{krank}\}$

$a_{21} = \{B, \text{gesund}\}$

$a_{22} = \{B, \text{krank}\}$

$$P(a_{11} = 2) = \frac{\binom{8}{2} \cdot \binom{4}{3}}{\binom{12}{5}}$$

- **Nenner:** Wir wählen 5 von insgesamt 12 Patienten für Therapie A.
- **Zähler, Faktor 1:** Unter diesen 5 sind 2 von den insgesamt 8 gesund werdenden
- **Zähler, Faktor 2:** Unter diesen 5 sind 3 von den insgesamt 4 krank bleibenden

www.matstat.org

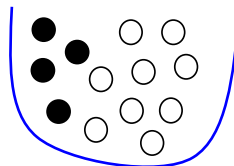
Fisher-Test

Analogie zum Urnenmodell (Vorlesung F4):

Man kann sich vorstellen, dass die Namen aller Patienten auf Kugeln geschrieben werden, die in eine Urne gelegt werden. Es sollen nun 5 Patienten für die Therapie A ausgewählt werden, indem 5 Kugeln ohne Zurücklegen gezogen werden. Unter den zu ziehenden Kugeln tragen 8 Namen von Patienten, die sowieso gesund werden, und 4 tragen Namen von Patienten, die nicht gesund werden. Wir suchen die Wahrscheinlichkeit, dass unter den 5 gezogenen Kugeln genau 2 Namen von gesund werdenden Patienten sind.

$$P(a_{11} = 2) = \frac{\binom{8}{2} \cdot \binom{4}{3}}{\binom{12}{5}}$$

↑
P("ziehe zwei weiße")



$w = 8$ gesund werdende

$s = 4$ krank bleibende

$N = 12$ Gesamtanzahl der Kugeln

$n = 5$ Anzahl gezogener Kugeln

www.matstat.org

Fisher-Test

Es scheint willkürlich, dass ausgerechnet die Wahrscheinlichkeit für $a_{11} = 2$ berechnet wurde. Tatsächlich spielt es jedoch keine Rolle, für welche der 4 Zellen die Wahrscheinlichkeit berechnet wird:

	gesund	krank	Σ
A	$a_{11} = 2$	$a_{12} = 3$	5
B	$a_{21} = 6$	$a_{22} = 1$	7
Σ	8	4	12

Tabellenfelder:

$a_{11} = \{A, \text{gesund}\}$

$a_{12} = \{A, \text{krank}\}$

$a_{21} = \{B, \text{gesund}\}$

$a_{22} = \{B, \text{krank}\}$

$$P(a_{12} = 3) = \frac{\binom{4}{3} \cdot \binom{8}{2}}{\binom{12}{5}} \quad P(a_{11} = 2) = \frac{\binom{8}{2} \cdot \binom{4}{3}}{\binom{12}{5}} \quad \text{wählen 5 von 12 für Therapie A}$$

$$P(a_{21} = 6) = \frac{\binom{8}{6} \cdot \binom{4}{1}}{\binom{12}{7}} \quad P(a_{22} = 1) = \frac{\binom{4}{1} \cdot \binom{8}{6}}{\binom{12}{7}} \quad \text{wählen 7 von 12 für Therapie B}$$

Alle diese Wahrscheinlichkeiten sind gleich, da $\binom{n}{k} = \binom{n}{n-k}$

www.matstat.org

Fisher-Test

- Wenn die Wahrscheinlichkeit für nur eine Zelle der obigen Tabelle berechnet wurde, hat man die Wahrscheinlichkeit der gesamten Tabelle, denn bei festgelegten Marginalsummen ergeben sich dann alle anderen Zellen zwangsläufig.
- Das ist gleichbedeutend mit der Tatsache, dass die obige Tabelle nur einen **Freiheitsgrad** hat.
- Im Allgemeinen beträgt die **Anzahl der Freiheitsgrade** für Tabellen mit c Spalten und r Zeilen $f = (c - 1) \cdot (r - 1)$

	gesund	krank	Σ
A	$a_{11} = 2$	$a_{12} = ??$	5
B	$a_{21} = ??$	$a_{22} = ??$	7
Σ	8	4	12

Tabellenfelder:

$a_{11} = \{A, \text{gesund}\}$

$a_{12} = \{A, \text{krank}\}$

$a_{21} = \{B, \text{gesund}\}$

$a_{22} = \{B, \text{krank}\}$

Mit $a_{11} = 2$ folgt automatisch (da Marginalsummen feststehen):

$$\rightarrow a_{21} = 8 - 2 = 6 \quad \rightarrow a_{12} = 5 - 2 = 3 \quad \rightarrow a_{22} = 4 - a_{12} = 1$$

www.matstat.org

Fisher-Test

	gesund	krank	Σ
A	$a_{11} = 2$	$a_{12} = 3$	$Z_1 = 5$
B	$a_{21} = 6$	$a_{22} = 1$	$Z_2 = 7$
Σ	$S_1 = 8$	$S_2 = 4$	$N = 12$

Tabellenfelder:

$a_{11} = \{A, \text{gesund}\}$
 $a_{12} = \{A, \text{krank}\}$
 $a_{21} = \{B, \text{gesund}\}$
 $a_{22} = \{B, \text{krank}\}$

$$P(a_{11} = 2) = \frac{\binom{8}{2} \cdot \binom{4}{3}}{\binom{12}{5}} = P(\text{"Tabelle"})$$

$f = 1$ (ein Freiheitsgrad,
schon eine einzige Zelle
bestimmt alle anderen)

Allgemein:

gilt auch für höher-
dimensionale Tabellen !

$$P(\text{"Tabelle"}) = \frac{\binom{S_1}{a_{11}} \cdot \binom{S_2}{a_{12}}}{\binom{N}{Z_1}} = \frac{S_1! \cdot S_2! \cdot Z_1! \cdot Z_2!}{N! \cdot a_{11}! \cdot a_{12}! \cdot a_{21}! \cdot a_{22}!}$$

www.matstat.org

Fisher-Test

	gesund	krank	Σ
A	$a_{11} = 2$	$a_{12} = 3$	$Z_1 = 5$
B	$a_{21} = 6$	$a_{22} = 1$	$Z_2 = 7$
Σ	$S_1 = 8$	$S_2 = 4$	$N = 12$

Tabellenfelder:

$a_{11} = \{A, \text{gesund}\}$
 $a_{12} = \{A, \text{krank}\}$
 $a_{21} = \{B, \text{gesund}\}$
 $a_{22} = \{B, \text{krank}\}$

$$P(\text{"Tabelle"}) = \frac{\binom{S_1}{a_{11}} \cdot \binom{S_2}{a_{12}}}{\binom{N}{Z_1}} = \frac{S_1! \cdot S_2! \cdot Z_1! \cdot Z_2!}{N! \cdot a_{11}! \cdot a_{12}! \cdot a_{21}! \cdot a_{22}!} \quad \text{hypergeometrische Verteilung}$$

$$= \frac{8! \cdot 4! \cdot 5! \cdot 7!}{12! \cdot 2! \cdot 3! \cdot 6! \cdot 1!} = \frac{14}{99} = \underline{\underline{0.141414}}$$

Die Wahrscheinlichkeit der beobachteten Tabelle unter Annahme der Nullhypothese beträgt $P_1 = 0.141414$.

www.matstat.org

Fisher-Test

	gesund	krank	Σ
A	$a_{11} = 2$	$a_{12} = 3$	$Z_1 = 5$
B	$a_{21} = 6$	$a_{22} = 1$	$Z_2 = 7$
Σ	$S_1 = 8$	$S_2 = 4$	$N = 12$

Tabellenfelder:

$a_{11} = \{A, \text{gesund}\}$

$a_{12} = \{A, \text{krank}\}$

$a_{21} = \{B, \text{gesund}\}$

$a_{22} = \{B, \text{krank}\}$

```
## Hypergeometric Distribution

?dhyper

dhyper(x = 2, m = 8, n = 4, k = 5)
# 0.1414141

# explicit calculation:
p = choose(8,2) * choose(4,3) / choose(12,5)
p # 0.1414141
```



www.matstat.org

Fisher-Test



Minitab

Hypergeometric Distribution

☒ Probability
☐ Cumulative probability
☐ Inverse cumulative probability

Population size (N): 12
 Event count in population (M): 8
 Sample size (n): 5

☐ Input column:
 Optional storage:

☒ Input constant: 2
 Optional storage:

Select OK Cancel

Calc / Probability Distributions

Hypergeometric with $N = 12$, $M = 8$, and $n = 5$
 $x \quad P(X = x)$
 2 0,141414

www.matstat.org

Fisher-Test

Um den p -Wert zu erhalten,

- muss die Wahrscheinlichkeit des Zustandekommens der observierten Tabelle unter H_0 berechnet werden ... *ok, done!*
- außerdem müssen die Wahrscheinlichkeiten (unter H_0) noch extremerer ("weiter von H_0 entfernten") Tabellen hinzugerechnet werden.

Um auch Punkt 2 zu erfüllen, kann man nun zwei Wege beschreiten:

- 1) Überlegen, welche Tabellen noch extremer unter der Nullhypothese sind und deren Wahrscheinlichkeiten addieren (siehe Anhang).
- 2) Einfach die Wahrscheinlichkeiten **aller** Tabellen unter H_0 berechnen und dann alle Wahrscheinlichkeiten addieren, die kleiner oder gleich der Wahrscheinlichkeit der beobachteten Tabelle sind.

Im Folgenden soll Weg 2) gezeigt werden → berechne alle möglichen Tabellen:

	gesund	krank	Σ
A	a_{11}	a_{12}	5
B	a_{21}	a_{22}	7
Σ	8	4	12

Dazu kann die Zelle a_{12} als Pivot-Element gewählt werden, denn a_{12} kann nur 5 verschiedene Werte annehmen: 0, 1, 2, 3, 4 und 5. (Die anderen Zellen ergeben sich wiederum automatisch aus a_{12})

www.matstat.org

Fisher-Test

a_{11}	a_{12}	Z_1
a_{21}	a_{22}	Z_2
S_1	S_2	N

$$P(\text{"Tabelle"}) = \frac{S_1! \cdot S_2! \cdot Z_1! \cdot Z_2!}{N! \cdot a_{11}! \cdot a_{12}! \cdot a_{21}! \cdot a_{22}!}$$

5	$a_{12} = 0$	5
3	4	7
8	4	12

$$P = 0.070707$$

4	$a_{12} = 1$	5
4	3	7
8	4	12

$$P = 0.353535$$

3	$a_{12} = 2$	5
5	2	7
8	4	12

$$P = 0.424242$$

2	$a_{12} = 3$	5
6	1	7
8	4	12

$$P = 0.141414$$

(Beobachtung)

1	$a_{12} = 4$	5
7	0	7
8	4	12

$$P = 0.010101$$

Der p -Wert wird damit:

$$p = 0.1414 + 0.0707 + 0.0101$$

$$\underline{\underline{p \approx 0.222}}$$

www.matstat.org

Berechnung der Wahrscheinlichkeiten in R:

```
C:\Users\Uwe\Desktop\TALKS_POSTERS\Magdeburg_Talk\R_scripts_for_Magdeburg_Talk.R - R Editor

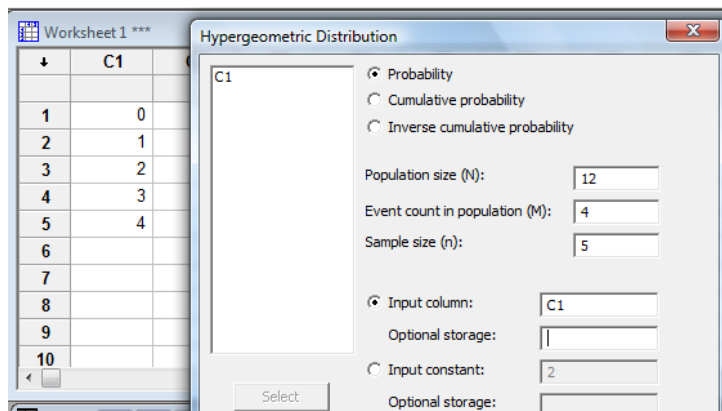
## Hypergeometrische Verteilung

white_urn = 8      # the number of white balls in the urn
black_urn = 4      # the number of black balls in the urn
total_drawn = 5    # the number of balls drawn from the urn

for (a12 in 0:4) {
  white_drawn = total_drawn - a12
  p = dhyper(x = white_drawn, m = white_urn, n = black_urn, k = total_drawn)
  p = round(p, 6)
  cat(paste("a12 =", a12, "  p =", p, "\n"))
}

# a12 = 0    p = 0.070707
# a12 = 1    p = 0.353535
# a12 = 2    p = 0.424242
# a12 = 3    p = 0.141414
# a12 = 4    p = 0.010101
```

www.matstat.org



Hypergeometric with $N = 12$, $M = 4$, and $n = 5$

x	$P(X = x)$
0	0,070707
1	0,353535
2	0,424242
3	0,141414
4	0,010101

www.matstat.org



Durchführung des Tests in R:

```
## Fischers Exakter Test

A = matrix(c(2,3,6,1), nrow = 2, byrow = TRUE)
A
#      [,1] [,2]
# [1,]    2    3
# [2,]    6    1

fisher.test(A)

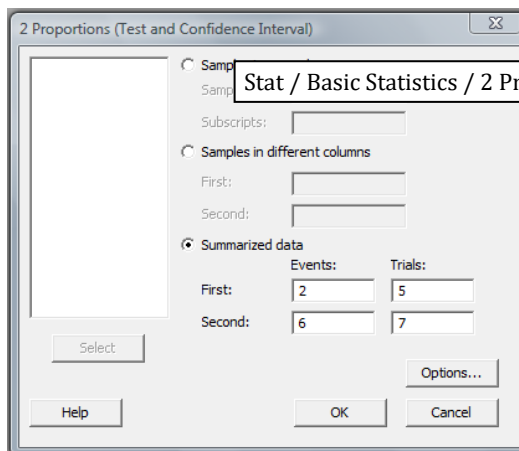
#      Fisher's Exact Test for Count Data
#
# data:  A
# p-value = 0.2222
# alternative hypothesis: true odds ratio is not equal to 1
# 95 percent confidence interval:
# | 0.001830735 2.834757137
```

2	3	5
6	1	7
8	4	12

(Beobachtung)

www.matstat.org

Durchführung des Tests mit Minitab:



Minitab



Normalapproximation
ungeeignet

Test for difference = 0 (vs not = 0): $Z = -1,79$ P-Value $\in 0,074$
 Fisher's exact test: P-Value = 0,222
 * NOTE * The normal approximation may be inaccurate for small samples.

www.matstat.org

Fisher-Test

	gesund	krank	Σ
A	$a_{11} = 2$	$a_{12} = 3$	5
B	$a_{21} = 6$	$a_{22} = 1$	7
Σ	8	4	12

$p. value = 0.2222$

- 40% Gesunde in Gruppe A
- 85% Gesunde in Gruppe B

Nullhypothese (H_0): $p_1 = p_2$ (gleiche Proportionen, gleiche Erfolgsrate)

Der p -Wert ist groß ($p > 0.05$). Die Nullhypothese kann daher **nicht** zurückgewiesen werden. Es konnte **nicht** gezeigt werden, dass der Unterschied in den beobachteten Proportionen statistisch signifikant ist, d.h. wir müssen weiter davon ausgehen, dass beide Therapieformen gleich erfolgreich sind (Therapieform und Therapieerfolg sind **unabhängig**).

Dies ist in Anbetracht der beobachteten Proportionen erstaunlich. Die Ursache liegt in der schlechten Versuchsplanung – **die Stichprobe ist zu klein**.

www.matstat.org

p -Wert für eine große Stichprobe bei gleichen Proportionen

```
## Fischers Exakter Test - grosse Stichprobe

A = matrix(c(200,300,600,100), nrow = 2, byrow = TRUE)
A
#      [,1] [,2]
# [1,] 200 300
# [2,] 600 100

fisher.test(A)

#      Fisher's Exact Test for Count Data

# data:  A
# p-value < 2.2e-16
# alternative hypothesis: true odds ratio is not equal to 1
# 95 percent confidence interval:
# | 0.08335038 0.14790301
```

www.matstat.org

Vergleich des p -Wertes für eine kleine und eine große Stichprobe bei gleichen Proportionen

	gesund	krank	Σ
A	2	3	5
B	6	1	7
Σ	8	4	12



- 40% Gesunde in Gruppe A
- 85% Gesunde in Gruppe B

$p.value = 0.222$

	gesund	krank	Σ
A	200	300	500
B	600	100	700
Σ	800	400	1200



$p.value < 10^{-16}$

Aufgrund der kleinen Stichprobe reicht die Trennschärfe ("Power") des Testes nicht aus, um die Nullhypothese zu verwerfen. Dies hätte man eigentlich voraussehen können. Für dieses Beispiel hätte überhaupt nur die extremste Konfiguration einen p -Wert unter 0.05 ergeben können – für $a_{12} = 4$ (siehe nächste Seite).

www.matstat.org

$$a_{12} = 4$$

Nur diese Beobachtung hätte Signifikanz (mit $\alpha = 0.05$) ergeben können:

- 20% Gesunde in Gruppe A
- 100% Gesunde in Gruppe B

	gesund	krank	Σ
A	1	4	5
B	7	0	7
Σ	8	4	12

```

C:\Users\Uwe\Desktop\TALKS_POSTERS\Magdeburg_Talk\R_scripts_for_Magdeburg_Talk.R - R Editor

## Fischers Exakter Test - der "extremste" Fall

A = matrix(c(1,4,7,0), nrow = 2, byrow = TRUE)
A
#      [,1] [,2]
# [1,]    1    4
# [2,]    7    0

fisher.test(A)
#
#      Fisher's Exact Test for Count Data
#
# data:  A
# p-value = 0.0101

```

www.matstat.org

Anhang

Anpassungs- und Unabhängigkeitstests

Uwe Menzel, 2018
uwe.menzel@matstat.org
www.matstat.org

Goodness-of-fit Test (Freiheitsgrade)

Die Dichtefunktion $\chi^2(f)$ ist von der Anzahl der Freiheitsgrade (f) abhängig.

Anzahl der Freiheitsgrade
für Anpassungstest:

$$f = k - r - p$$

- k = Anzahl der Kategorien (gesund, krank = 2 Kategorien)
- r = Anzahl der Restriktionen (z. B.: Summe aller Beobachtungen steht fest, siehe Beispiel Würfel)
- p = Anzahl der Parameter, die aus der Stichprobe geschätzt werden müssen, um das Modell aufstellen zu können (siehe folgende Seiten)

Im folgenden sollen 2 Fälle behandelt werden:

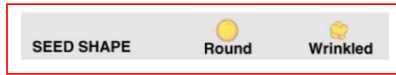
- 1) es muss **kein** Parameter aus der Stichprobe geschätzt werden ($p = 0$)
- 2) es muss **ein** Parameter aus der Stichprobe geschätzt werden ($p = 1$)

www.matstat.org

Anpassungstest

1) Es muss **kein** Parameter aus der Stichprobe geschätzt werden ($p = 0$)

Mendel's Experiment:



Zwei Allele: A, a **Allel A ist dominant**

Mendel kreuzte nur heterozygote (Aa) Bohnen (beide Allel-Frequenzen = 0.5):

	A	a
A	AA	Aa
a	Aa	aa

Die Genotypen AA und Aa werden rund,
nur der Genotyp aa wird faltig.

- Wenn das Mendelsche Gesetz gilt, dann wird ein Verhältnis von 3:1 für runde:faltige erwartet
- **Observation:** 423 runde Bohnen, 133 faltige Bohnen in der Folgegeneration

www.matstat.org

Anpassungstest

1) Es muss **kein** Parameter aus der Stichprobe geschätzt werden ($p = 0$)

	rund	faltig	Σ
O_i	423	133	556
E_i	$556 \cdot 3/4 = 417$	$556 \cdot 1/4 = 139$	556

$n = 556$ Stichprobe

$r = k - 1 = 1$

1 Freiheitsgrad

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} = \frac{(423 - 417)^2}{417} + \frac{(133 - 139)^2}{139} = \underline{\underline{0.345}} \quad \text{Test-statistik}$$

$$\Omega_{krit} = \{ \chi^2 > \chi^2_{\alpha}(f) \} = \{ \chi^2 > \chi^2_{0.05}(1) \} = \{ \chi^2 > 3.84 \} \quad \text{Kritische Region}$$

Die Teststatistik liegt **nicht** in der kritischen Region. Die Nullhypothese wird **nicht** verworfen. Es konnte **kein signifikanter Unterschied** zwischen der Beobachtung und den erwarteten Ergebnissen nachgewiesen werden. Die Beobachtung ergab keinen Widerspruch zum Modell (zum Mendelschen Gesetz).

www.matstat.org

Anpassungstest

2) Es muss **ein** Parameter aus der Stichprobe geschätzt werden ($p = 1$)

- Oft kennt man jedoch die Allel-Frequenzen (p, q) nicht von vornherein wie in Mendels Experiment (dort war $p = q = 0.5$ bekannt, weil nur heterozygote Bohnen (mit Genotyp Aa) gekreuzt wurden)
- Hardy-Weinberg-Modell**: die Allel-Frequenzen bleiben über die Generationen konstant

Seien die Allel-Frequenzen in der 1. Generation $f(A) = p$ und $f(a) = q$, **beide unbekannt**. Für die Folgegeneration gilt:

$$f(AA) = p^2$$

$$f(Aa) = 2 \cdot p \cdot q$$

$$f(aa) = q^2$$

	A	a
A	$AA (p^2)$	$Aa (pq)$
a	$Aa (pq)$	$aa (q^2)$

www.matstat.org

Anpassungstest

Um die unter dem Modell erwarteten Häufigkeiten AA , Aa und aa berechnen zu können, müssen zunächst die Allel-Frequenzen p und q **aus der Stichprobe berechnet** werden:

Beispiel: http://en.wikipedia.org/wiki/Hardy-Weinberg_principle, etwas geändert.



Callimorpha dominula
[Linnaeus](#), 1758

Stichprobe:

Genotyp	White-spotted (AA)	Intermediate (Aa)	Little spotting (aa)	$\Sigma (=n)$
beobachtet	1329	268	15	1612

$$p = \frac{2 \cdot \text{Anzahl}(AA) + \text{Anzahl}(aA)}{2 \cdot n} = \frac{2 \cdot 1329 + 268}{2 \cdot 1612} = 0.907 \quad \text{Frequenz für Allel } A$$

$$q = \frac{2 \cdot \text{Anzahl}(aa) + \text{Anzahl}(aA)}{2 \cdot n} = 1 - p = 0.093 \quad \text{Frequenz für Allel } a$$

(bei n Individuen haben wir insgesamt $2 \cdot n$ Allele)

Anpassungstest

Berechnung der unter dem HW-Modell **erwarteten** Häufigkeiten ("E = expected"):

$$\begin{aligned} E(AA) &= n \cdot p^2 = 1612 \cdot 0.907^2 = 1326.11 \\ E(Aa) &= n \cdot 2 \cdot pq = 271.95 \\ E(aa) &= n \cdot q^2 = 13.94 \end{aligned} \quad \left. \vphantom{\begin{aligned} E(AA) \\ E(Aa) \\ E(aa) \end{aligned}} \right\} \text{diese Werte müssen den beobachteten Werten (Tabelle oben) gegenübergestellt werden}$$

Genotyp:	White-spotted (AA)	Intermediate (Aa)	Little spotting (aa)	$\Sigma (= n)$
O_i	1329	268	15	1612
E_i	1326.11	271.95	13.94	1612

Teststatistik:

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} = \frac{(1329 - 1326.11)^2}{1326.11} + \frac{(268 - 271.95)^2}{271.95} + \frac{(15 - 13.94)^2}{13.94} = \underline{0.144}$$

www.matstat.org

Anpassungstest

Anzahl der Freiheitsgrade für Anpassungstest:

$$f = k - r - p$$

- Anzahl der Kategorien: $k = 3$ (Genotypen AA, Aa, aa)
- Anzahl der Restriktionen: $r = 1$ (Gesamtanzahl $n = 1612$)
- Anzahl der Parameter, die aus der Stichprobe geschätzt werden müssen, um das Modell aufstellen zu können: $p = 1$ (wenn eine der beiden Allel-Frequenzen berechnet wurde, ergibt sich die andere automatisch: $q = p - 1$)

$$\Rightarrow f = 3 - 1 - 1 = 1 \quad \text{allgemein für HW-Modell: } f = \text{Anzahl Genotypen minus Anzahl der Allele}$$

$$\Omega_{krit} = \{\chi^2 > \chi^2_{\alpha}\} = \{\chi^2 > \chi^2_{0.05}(1)\} = \{\chi^2 > 3.84\} \quad \text{Kritische Region}$$

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} = 0.144 \quad \text{Teststatistik}$$

Die Teststatistik liegt **nicht** in der kritischen Region. Die Beobachtung ergab keinen Widerspruch zum Hardy-Weinberg-Modell.

Fisher-Test

Beispiel S. 31: Therapieerfolg für 2 Therapieformen (A und B)

Ermittlung der extremen Ereignisse mit χ^2 :

- Oben wurde vorgeschlagen, solche Resultate als extremer (als die Beobachtung) anzusehen, die unter H_0 eine kleinere Wahrscheinlichkeit als die Beobachtung haben.
- Eine **alternative Methode** betrachtet diejenigen Tabellen als extremer, die einen größeren "Abstand" zur unter H_0 erwarteten Tabelle als die beobachtete Tabelle haben. Der "Abstand" aller möglichen Tabellen zu der unter H_0 erwarteten Tabelle lässt sich mit Hilfe von χ^2 quantifizieren:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad \chi^2 = \text{Abstand einer Tabelle zur unter } H_0 \text{ erwarteten Tabelle}$$

O_{ij} : aktueller Wert in der Zelle i, j einer gegebenen Tabelle

$E_{i,j}$: unter der Nullhypothese erwarteter Wert in der Zelle i, j

Wichtig: χ^2 dient hier **nicht** als Teststatistik, sondern nur zur "Messung" des Abstandes aller Tabellen zur H_0 -Tabelle.

www.matstat.org

Fisher-Test

Um die $E_{i,j}$ zu berechnen, müssen wir zunächst die Tabelle finden, die wir unter H_0 erwarten würden:

Nullhypothese (H_0): $p_1 = p_2$ (gleicher Anteil Gesunder für A und B)

	gesund	krank	Σ
A	E_{11}	...	5
B	E_{21}	...	7
Σ	8	4	12

Wenn der Anteil der Gesunden in beiden Gruppen gleich groß ist, muss gelten:

$$\frac{E_{11}}{5} = \frac{8}{12} \quad \text{und} \quad \frac{E_{21}}{7} = \frac{8}{12} \quad H_0: \text{Der Anteil der Gesunden ist in beiden Gruppen } 8/12 = 2/3$$

↑
Anteil Gesunder
in Gruppe A
↑
Anteil Gesunder
in Gruppe B

www.matstat.org

Fisher-Test

Um die $E_{i,j}$ zu berechnen, müssen wir zunächst die Tabelle finden, die wir unter H_0 erwarten würden:

Nullhypothese (H_0): $p_1 = p_2$ (gleicher Anteil Kranker für A und B)

	gesund	krank	Σ
A	...	E_{12}	5
B	...	E_{22}	7
Σ	8	4	12

Analog muss in beiden Gruppen der Anteil der Kranken bei $1/3$ liegen:

$$\frac{E_{12}}{5} = \frac{4}{12} \quad \text{und} \quad \frac{E_{22}}{7} = \frac{4}{12} \quad H_0: \text{Der Anteil der Kranken ist in beiden Gruppen } 4/12 = 1/3$$

↑
Anteil Kranker
in Gruppe A
↑
Anteil Kranker
in Gruppe B

www.matstat.org

Fisher-Test

$$\frac{E_{11}}{5} = \frac{8}{12} \quad \frac{E_{12}}{5} = \frac{4}{12} \quad \frac{E_{21}}{7} = \frac{8}{12} \quad \frac{E_{22}}{7} = \frac{4}{12}$$

Allgemein gilt für die "erwartete" Tabelle:

$$E_{11} = \frac{Z_1 \cdot S_1}{N} \quad E_{21} = \frac{Z_2 \cdot S_1}{N} \quad E_{12} = \frac{Z_1 \cdot S_2}{N} \quad E_{22} = \frac{Z_2 \cdot S_2}{N}$$

Jede unter H_0 erwartete Zelle ergibt sich also aus **Zeilensumme mal Spaltensumme geteilt durch Gesamtzahl**:

	gesund	krank	Σ
A	$10/3$	$5/3$	5
B	$14/3$	$7/3$	7
Σ	8	4	12

$$E_{ij} = \frac{Z_i \cdot S_j}{N}$$

(Idealerweise) unter H_0 erwartete Tabelle

www.matstat.org

Fisher-Test

	gesund	krank	Σ
A	$10/3$	$5/3$	5
B	$14/3$	$7/3$	7
Σ	8	4	12

$$E_{ij} = \frac{Z_i \cdot S_j}{N}$$

(Idealerweise) unter
 H_0 erwartete Tabelle

- Mit Hilfe von χ^2 berechnen wir nun den "Abstand" aller möglichen Tabellen (einschließlich der beobachteten) zu dieser Tabelle.
- Um alle "möglichen" Tabellen zu finden, benutzen wir wiederum das Element a_{12} als Pivot.
- Das Element a_{12} kann aufgrund der vorgegebenen Marginalsummen nur die Werte 0,1,2,3 und 4 annehmen.
- Ist a_{12} festgelegt, ergeben sich die anderen Zellen zwangsläufig.

www.matstat.org

Fisher-Test

Beispiel: Abstand ideale (H_0 -)Tabelle zur beobachteten Tabelle:

beobachtete Tabelle

	gesund	krank	Σ
A	2	3	5
B	6	1	7
Σ	8	4	12

H_0 - Tabelle

	gesund	krank	Σ
A	$10/3$	$5/3$	5
B	$14/3$	$7/3$	7
Σ	8	4	12

$$\chi_0^2 = \sum_{i,k=1}^K \frac{(O_{ik} - E_{ik})^2}{E_{ik}} = \frac{(2 - 10/3)^2}{10/3} + \frac{(3 - 5/3)^2}{5/3} + \frac{(6 - 14/3)^2}{14/3} + \frac{(1 - 7/3)^2}{7/3}$$

$\chi_0^2 = \underline{2.743}$ "Abstand" der beobachteten Tabelle zur unter H_0 erwarteten Tab.

Es werden nun die χ^2 aller möglichen Tabellen berechnet und mit χ_0^2 verglichen. Ist $\chi^2 \geq \chi_0^2$ wird die zugehörige Tabelle als extremer als die Beobachtung angesehen und deren Wahrscheinlichkeit unter H_0 geht in die Berechnung des p -Wertes ein.

www.matstat.org

Berechnung des Abstandes für alle Tabellen

```

C:\Users\Uwe\Desktop\TALKS_POSTERS\Magdeburg_Talk\R_scripts_for_Magdeburg_Talk.R - R Editor

## Chi-squared for test of independence                                Uwe Menzel, 2010

S1 = 8 # Spalte 1
S2 = 4 # Spalte 2
Z1 = 5 # Zeile 1
Z2 = 7 # Zeile 2
N = 12 # total

E11 = Z1 * S1 / N
E12 = Z1 * S2 / N
E21 = Z2 * S1 / N
E22 = Z2 * S2 / N

E = matrix(c(E11, E12, E21, E22), nrow = 2, byrow = TRUE)

for (a12 in 0:4) {
  a11 = Z1 - a12
  a22 = S2 - a12
  a21 = S1 - a11
  A = matrix(c(a11, a12, a21, a22), nrow = 2, byrow = TRUE)
  chi2 = sum((E-A)^2/E)
  chi2 = round(chi2, 3)
  cat(paste("a12 =", a12, " chi2 =", chi2, "\n"))
}

```

Berechnung des χ^2 -"Abstandes" für alle Tabellen



	C1-T	C2-T	C3
	treatment	state	count
1	control	sick	3
2	therapy	sick	1
3	control	healthy	2
4	therapy	healthy	6
5			
6			
7			
8			
9			
10			

Cross Tabulation and Chi-Square

Categorical variables:

For rows: treatment

For columns: state

For layers:

Frequencies are in: count (optional)

Display

☒ Counts
 ☐ Row percents
 ☐ Column percents
 ☐ Total percents

Chi-Square...

Other Stats...

Options...

OK

Cancel

Stat / Tables / Cross Tabulation and Chisquare

www.matstat.org

χ^2 und p für alle möglichen Tabellen

Tabelle	χ^2	p									
<table><tr><td>5</td><td>0</td><td>5</td></tr><tr><td>3</td><td>4</td><td>7</td></tr><tr><td>8</td><td>4</td><td>12</td></tr></table>	5	0	5	3	4	7	8	4	12	4.286	0.070707
5	0	5									
3	4	7									
8	4	12									
<table><tr><td>4</td><td>1</td><td>5</td></tr><tr><td>4</td><td>3</td><td>7</td></tr><tr><td>8</td><td>4</td><td>12</td></tr></table>	4	1	5	4	3	7	8	4	12	0.686	0.353535
4	1	5									
4	3	7									
8	4	12									
<table><tr><td>3</td><td>2</td><td>5</td></tr><tr><td>5</td><td>2</td><td>7</td></tr><tr><td>8</td><td>4</td><td>12</td></tr></table>	3	2	5	5	2	7	8	4	12	0.171	0.424242
3	2	5									
5	2	7									
8	4	12									
<table><tr><td>2</td><td>3</td><td>5</td></tr><tr><td>6</td><td>1</td><td>7</td></tr><tr><td>8</td><td>4</td><td>12</td></tr></table>	2	3	5	6	1	7	8	4	12	2.743	0.141414
2	3	5									
6	1	7									
8	4	12									
<table><tr><td>1</td><td>4</td><td>5</td></tr><tr><td>7</td><td>0</td><td>7</td></tr><tr><td>8</td><td>4</td><td>12</td></tr></table>	1	4	5	7	0	7	8	4	12	8.4	0.010101
1	4	5									
7	0	7									
8	4	12									

Möglich sind alle Tabellen, bei denen die Zeilen- und Spaltensummen beibehalten werden.

$$\chi^2 = \sum_{i,j=1} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad E_{ij} = \frac{Z_i \cdot S_j}{N}$$

$$P(\text{"table"}) = \frac{S_1! \cdot S_2! \cdot Z_1! \cdot Z_2!}{N! \cdot a_{11}! \cdot a_{12}! \cdot a_{21}! \cdot a_{22}!}$$

Rot markiert: kleinere Wahrscheinlichkeit unter H_0 verglichen mit Beobachtung sowie größeres χ^2 als Beobachtung (hier stimmen beide Kriterien genau überein). Die entsprechenden Wahrscheinlichkeiten tragen zum p -Wert bei:

$$p = 0.0707 + 0.1414 + 0.0101 \approx \underline{\underline{0.222}}$$