2-Stichprobentest für Anteilswerte

- Wir betrachten den Anteilswert (Prozentsatz) für ein interessierendes Ereignis in zwei verschiedenen Grundgesamtheiten (π_1, π_2)
- Ziel: Auf der Basis von Stichprobenerhebungen zu entscheiden, ob die beiden Grundgesamtheiten die gleichen Anteilswerte aufweisen oder sich unterscheiden
- Frage: Wie ist die Differenz der Stichprobenanteile verteilt?

Theoretische Grundlage

- Wir betrachten 2 unabhängige Stichproben vom Umfang n_1, n_2 mit beobachteten Anteilswerten p_1,p_2

- Frage: Wie ist die Differenz der Stichprobenanteile verteilt?

$$E(p_1 - p_2) = \pi_1 - \pi_2$$

$$Var(p_1 - p_2) = Var(p_1) + Var(p_2) = \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}$$

$$(p_1 - p_2) \sim N(\pi_1 - \pi_2, \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}})$$

Beispiel

- Angenommen wir vergleichen zwei Städte mit jeweils 10.000 Einwohnern.
- In der Stadt-1 leben 3.000 Angehörige einer Religionsgemeinschaft. (π_1=30%)
- In der Stadt-2 leben nur 2.500 Angehörige einer Religionsgemeinschaft. (π_2=25%)
- Versucht man diesen Unterschied auf der Basis einer Stichprobe von jeweils 100 Personen in den beiden Städten zu erheben, so ergeben sich folgende Ergebnisse

Beispiel:

Grundgesamtheit 1

$N_1 = 10.000$
$X_1 = 3.000$
$\pi_1 = 30\%$

Grundgesamtheit 2

$N_2 = 10.000$
$X_2 = 2.500$
$\pi_2 = 25\%$

Stichprobe 1

$n_1 = 100$
$E(p_1) = 0,30$
$Var(p_1) = 0,0021$

Stichprobe 2

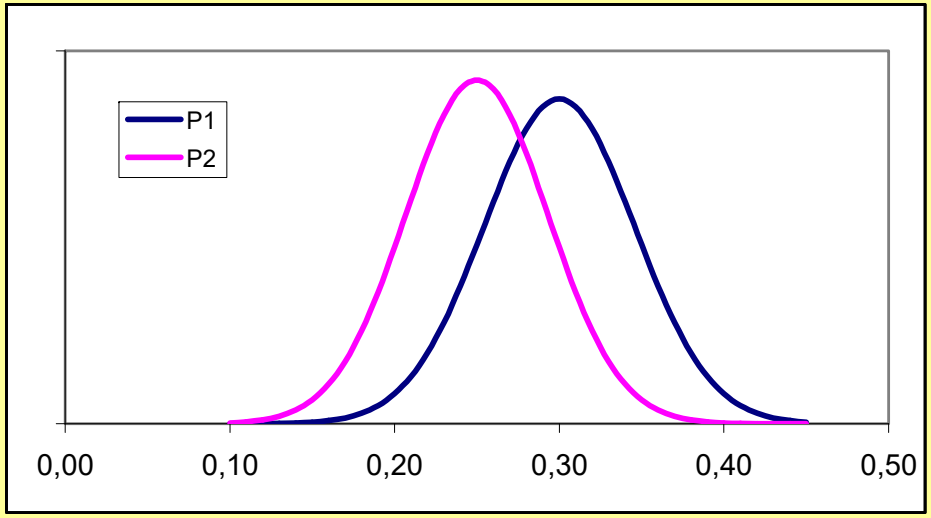
$n_2 = 100$
$E(p_2) = 0,25$
$Var(p_2) = 0,001875$

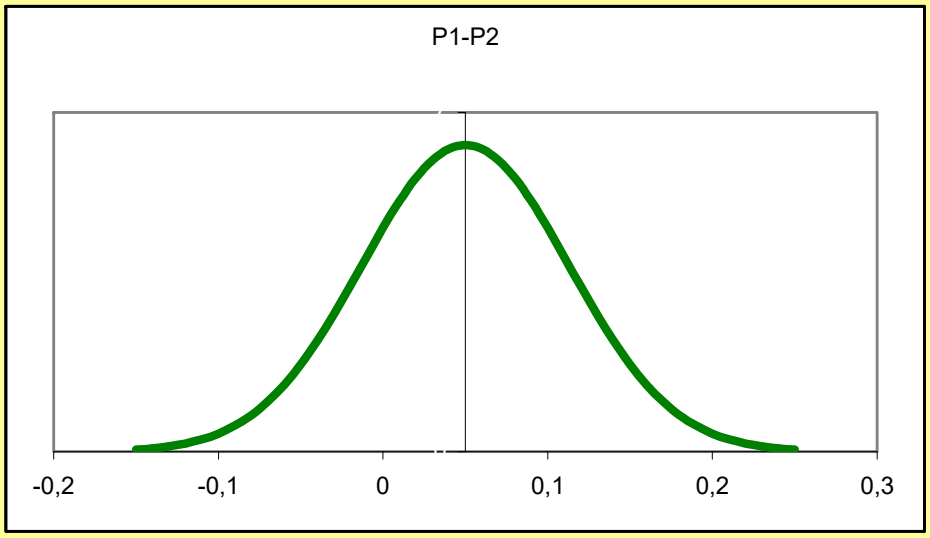
Differenz

$E(p_1-p_2) = 0,05$
$Var(p_1-p_2) = 0,003975$

$Prob(p_1 < p_2) = 0,21387382$

$Prob(p_1 - p_2 > 0,15) = 0,05635796$





Interpretation

- Wir können davon ausgehen, dass die Differenz der beiden Stichprobenanteilswerte im Erwartungswert 5% betragen wird.
- Weiters können wir auch die statistisch zu erwartende Schwankung durch eine Normalverteilung quantifizieren

2-Stichprobentest für Anteilswerte

- $H_0: \pi_1 = \pi_2 = \pi$
- $H_1: \pi_1 \neq \pi_2$ bei zweiseitigen Fragestellungen bzw.
- $H_1: \pi_1 > \pi_2$ oder $H_1: \pi_1 < \pi_2$ bei einseitigen Fragestellungen
- Unter Gültigkeit von H_0

$$\hat{\pi} = p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

Die Teststatistik standardisiert die Differenz der Anteilswerte

$$Z = \frac{(p_1 - p_2)}{\sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}}} = \frac{(p_1 - p_2)}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = N(0,1)$$

Beispiel:

- Zwei Gruppen A und B, die aus jeweils 100 Personen bestehen, leiden an einer Krankheit
- Gruppe A bekommt ein Heilmittel; Gruppe B (Kontrollgruppe) bekommt ein Placebo
- In A werden 75 geheilt; In B werden 65 geheilt
- Kann bei einer Irrtumswahrscheinlichkeit von 0,01 die Wirksamkeit des Medikaments nachgewiesen werden ?
- $H_0: \pi_A \leq \pi_B$ $H_1: \pi_A > \pi_B$ (einseitiger Test)
- Kritischer Wert: $z_c = z_{0,99} = 2,33$

Beispiel

- p_1=75/100=0,75 p_2=65/100=0,65
- p=(75+65)/(100+100)=0,70

$$Z = \frac{(0{,}75 - 0{,}65)}{\sqrt{0{,}7(1-0{,}7)\left(\frac{1}{100}+\frac{1}{100}\right)}} = \frac{0{,}1}{\sqrt{0{,}0042}} = \frac{0{,}1}{0{,}0648} = 1{,}54$$

- p-value (empirisches Signifikanzniveau) 0,0614
- Nicht signifikant; H_0 kann nicht abgelehnt werden
- Nachweis der Wirksamkeit konnte nicht erbracht werden; d.h. aber nicht, dass bewiesen wurde, dass das Medikament unwirksam ist.

Realisierung mit R

R Console

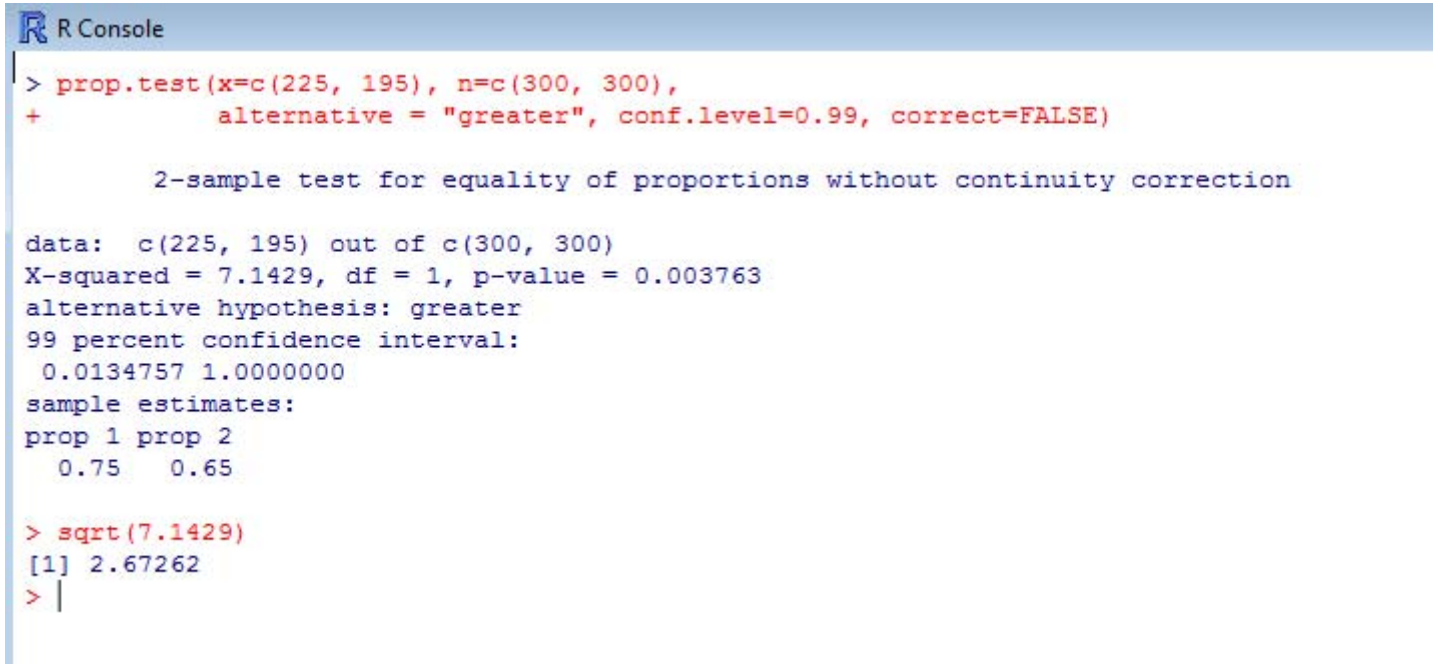
```
> # Two Sample test
> prop.test(x=c(75, 65), n=c(100, 100),
+           alternative = "greater", conf.level=0.99, correct=FALSE)

        2-sample test for equality of proportions without continuity correction

data:  c(75, 65) out of c(100, 100)
X-squared = 2.381, df = 1, p-value = 0.06141
alternative hypothesis: greater
99 percent confidence interval:
 -0.04986448  1.00000000
sample estimates:
prop 1 prop 2
  0.75   0.65

> sqrt(2.381)
[1] 1.543049
> |
```

Modifiziertes Beispiel

- Wir betrachten nun $n_1=n_2=300$ Personen in jeder Gruppe
- In A werden nun 225, in B195 Personen geheilt
- $p_1=0,75$ $p_2=0,65$ $p=0,7$

$$Z = \frac{(0,75-0,65)}{\sqrt{0,7(1-0,7)\left(\frac{1}{300}+\frac{1}{300}\right)}} = \frac{0,1}{\sqrt{0,0014}} = \frac{0,1}{0,0374} = 2,67$$

- Jetzt kann die H_0 bei einem $\alpha=0,01$ verworfen werden


R Console
> prop.test(x=c(225, 195), n=c(300, 300),
+ alternative = "greater", conf.level=0.99, correct=FALSE)
2-sample test for equality of proportions without continuity correction
data: c(225, 195) out of c(300, 300)
X-squared = 7.1429, df = 1, p-value = 0.003763
alternative hypothesis: greater
99 percent confidence interval:
0.0134757 1.0000000
sample estimates:
prop 1 prop 2
0.75 0.65
> sqrt(7.1429)
[1] 2.67262
>

Rechenschema mit Excel

alpha=	0,05							
z-Wert=	1,9600	beidseitig						
z-Wert=	1,6449	einseitig						
Stichprobe 1	n_1=	400	X_1=	196		emax=	0,04898927	
	p_1=	**0,49**				UG=	44,1%	
	Var(p_1)=	0,0006				OG=	53,9%	
	Std.Abw.(p_1)=	0,0250						
Stichprobe 2	n_2 =	400	X_2=	155		emax=	0,04774267	
	p_2=	**0,39**				UG=	33,98%	
	Var(p_2)=	0,0006				OG=	43,52%	
	Std.Abw.(p_2)=	0,0244						
Differenz								
	p_1-p_2=	**0,10**				Teststatistik=	2,9211	
	p=	0,44				p-value=	0,0017	einseitig
	Var(p)=	0,0012				p-value=	0,0035	beidseitig
	Std.Abw.(p)=	0,0351						

χ^2-Unabhängigkeitstest

Der χ^2-Unabhängigkeitstest erlaubt es, zu testen, ob zwei nominalskalierte Merkmale voneinander unabhängig sind oder nicht.

Dabei werden die Abweichungen der beobachteten Häufigkeiten in einer Kreuztabelle von den unter der Unabhängigkeitshypothese erwarteten Häufigkeiten evaluiert.

Unter der Unabhängigkeitshypothese ergeben sich die erwarteten relativen Häufigkeiten in einer Zelle i,j durch Multiplikation der zugehörigen relativen Randhäufigkeiten bzw. sind die bedingten Verteilungen konstant und gleich der Randverteilung.

Notation

Wir betrachten eine I×m Kreuztabelle

Zeilenindex i (1,..., I) Spaltenindex j (1,..., m)

n_{ij}	$n_{1.}$ $n_{i.}$ $n_{I.}$
$n_{.1}$ $n_{.j}$ $n_{.m}$	$n_{..}$

Bedingung für Unabhängigkeit

$$P(A \cap B) = P(A) \cdot P(B)$$

Wahrscheinlichkeit einer Zelle ist das Produkt der Zeile- und der Spaltenwahrscheinlichkeit

Bei Unabhängigkeit gilt daher: $\frac{n_{ij}}{n_{..}} = \frac{n_{i.}}{n_{..}} \cdot \frac{n_{.j}}{n_{..}}$

$$\rightarrow \qquad n_{ij} = \frac{n_{i.} \cdot n_{.j}}{n_{..}}$$

Bei Unabhängigkeit ergibt sich die absolute Häufigkeit einer Zelle als das Produkt der Zeilen- und der Spaltenhäufigkeit dividiert durch die Gesamtanzahl

Spezialfall k x 2 Tabelle

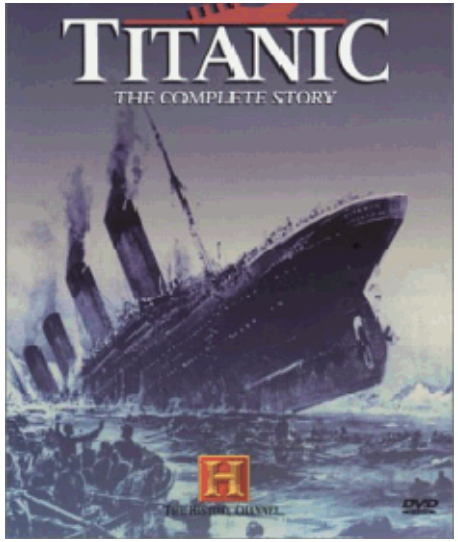
- Im Falle einer k x 2 Tabelle kann der Test auf Unabhängigkeit als direkte Erweiterung des 2-Stichprobentests für Anteilswerte interpretiert werden.
- Wir betrachten 2 Merkmale A (k…Kategorien), B (2… Kategorien)
- Die Nullhypothese, dass die beiden Merkmale A und B unabänig sind, impliziert, dass in den k von A definierten Gruppen die Anteilswerte p(B=1) bzw. p(B=2) gleich sind.
- Falls k=2 ist ➔ 2x2 Tabelle ("Vierfeldertafel") entspricht der Test auf Unabhängigkeit exakt dem 2-Stichprobentest für Anteilswerte.

Gesamter Datensatz

CLASS * SURVIVED * SEX Kreuztabelle

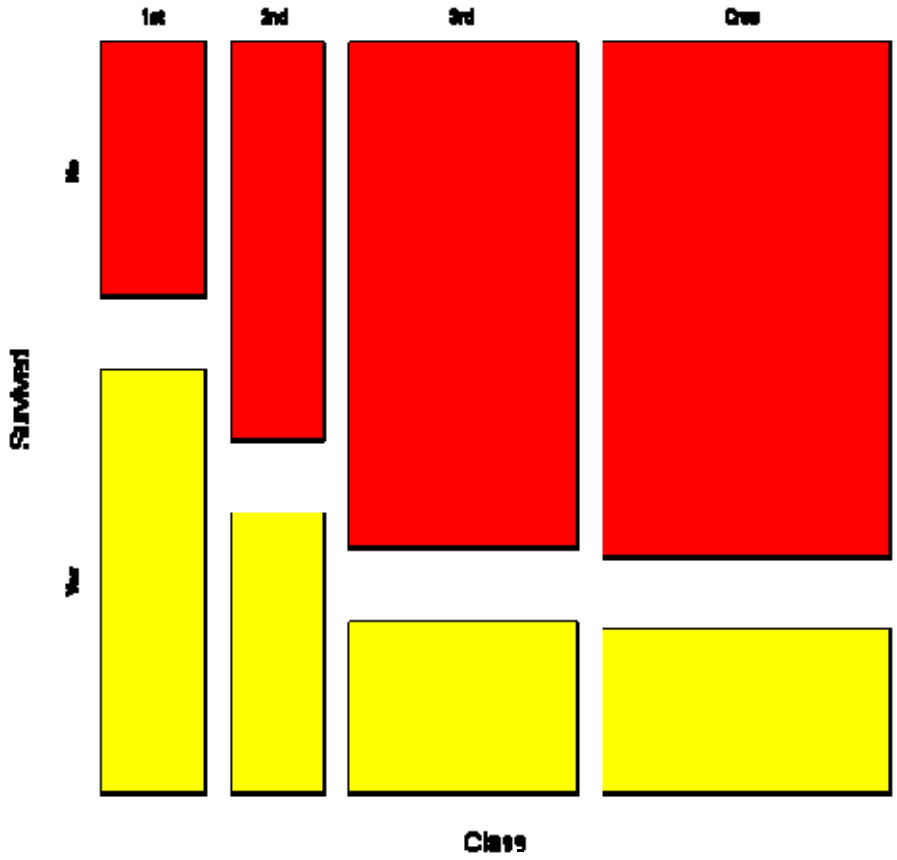
Anzahl

SEX			SURVIVED		Gesamt
			No	Yes	
Male	CLASS	1st Class	118	62	180
		2nd Class	154	25	179
		3rd Class	422	88	510
		Crew	670	192	862
	Gesamt		1364	367	1731
Female	CLASS	1st Class	4	141	145
		2nd Class	13	93	106
		3rd Class	106	90	196
		Crew	3	20	23
	Gesamt		126	344	470



Visualisierung mittels Mosaic-Plot

Zusammenhang: Überleben x Passagierklasse



Beispiel

Es soll untersucht werden, ob ein signifikanter Zusammenhang zwischen der beiden Merkmale „Unterkunftsklasse“ am Schiff und „Überleben des Passagiers“ besteht.

Merkmal überlebt:

	Anzahl	rel. Häufigkeit
NEIN	1.490	0,68
JA	711	0,32
	2.201	

Merkmal Unterkunftsklasse:

	Anzahl	rel. Häufigkeit
1st Class	325	14,8%
2nd Class	285	12,9%
3rd Class	706	32,1%
Crew	885	40,2%
	2.201	

Zusammenhang?

Ausgangsdaten & Fragestellung

Beobachtete Häufigkeiten

		überlebt		Zeilen-summe
		NEIN	JA	
Klasse	1st Class	122	203	325
	2nd Class	167	118	285
	3rd Class	528	178	706
	Crew	673	212	885
Spaltensumme		1.490	711	2.201

Zeilenprozent beobachtet

		überlebt		Zeilen-summe
		NEIN	JA	
Klasse	1st Class	37,5%	62,5%	100,0%
	2nd Class	58,6%	41,4%	100,0%
	3rd Class	74,8%	25,2%	100,0%
	Crew	76,0%	24,0%	100,0%
Spaltensumme		67,7%	32,3%	100,0%

Ist der Anteil der Überlebenden in den 4 Personengruppen gleich?

Verallgemeinerung der Fragestellung des 2-Stichprobentests für den Vergleich von mehr als 2 Klassen

Bedingte Wahrscheinlichkeiten nicht konstant ➔ keine strikte Unabhängigkeit

Erwartete Häufigkeit

Berechnung der erwarteten Häufigkeiten bei Unabhängigkeit der Merkmale

$$Erwartete\ Häufigkeit = \frac{Zeilensumme \times Spaltensumme}{Gesamtsumme}$$

		überlebt		Zeilen-summe
		NEIN	JA	
Klasse	1st Class	???		325
	2nd Class			285
	3rd Class			706
	Crew			885
Spaltensumme		1.490	711	2.201

		überlebt		Zeilen-summe
		NEIN	JA	
Klasse	1st Class	220,0		325
	2nd Class			285
	3rd Class			706
	Crew			885
Spaltensumme		1.490	711	2.201

$$\text{Erwartete Häufigkeit}_{\text{NEIN, 1st Class}} = \frac{1.490 \times 325}{2.201} = 220{,}0$$

Erwartete Häufigkeit

Berechnung der erwarteten Häufigkeiten bei Unabhängigkeit der Merkmale

$$Erwartete\ Häufigkeit = \frac{Zeilensumme \times Spaltensumme}{Gesamtsumme}$$

		überlebt		Zeilen-summe
		NEIN	JA	
Klasse	1st Class	220,0		325
	2nd Class			285
	3rd Class		????	706
	Crew			885
Spaltensumme		1.490	711	2.201

		überlebt		Zeilen-summe
		NEIN	JA	
Klasse	1st Class	220,0		325
	2nd Class			285
	3rd Class		228,1	706
	Crew			885
Spaltensumme		1.490	711	2.201

$$\text{Erwartete Häufigkeit}_{\text{JA, 3rd Class}} = \frac{711 \times 706}{2.201} = 228{,}1$$

Erwartete Häufigkeit

Erwartete Häufigkeiten Unterkunftsklasse versus „Person hat überlebt“ bei Gültigkeit der Unabhängigkeitshypothese

		überlebt		Zeilen-summe
		NEIN	JA	
Klasse	1st Class	220,0	105,0	325
	2nd Class	192,9	92,1	285
	3rd Class	477,9	228,1	706
	Crew	599,1	285,9	885
Spaltensumme		1.490	711	2.201

Vergleich Beobachtete Häufigkeiten - Erwartete Häufigkeiten

Beobachtete Häufigkeiten ⟷ **Erwartete Häufigkeiten unter Ho**

		überlebt		Zeilen-summe
		NEIN	JA	
Klasse	1st Class	122	203	325
	2nd Class	167	118	285
	3rd Class	528	178	706
	Crew	673	212	885
Spaltensumme		1.490	711	2.201

		überlebt		Zeilen-summe
		NEIN	JA	
Klasse	1st Class	220	105	325
	2nd Class	193	92	285
	3rd Class	478	228	706
	Crew	599	286	885
Spaltensumme		1.490	711	2.201

↓ **Zeilenprozent beobachtet**

		überlebt		Zeilen-summe
		NEIN	JA	
Klasse	1st Class	37,5%	62,5%	100,0%
	2nd Class	58,6%	41,4%	100,0%
	3rd Class	74,8%	25,2%	100,0%
	Crew	76,0%	24,0%	100,0%
Spaltensumme		67,7%	32,3%	100,0%

↑ **Zeilenprozent erwartet**

		überlebt		Zeilen-summe
		NEIN	JA	
Klasse	1st Class	67,7%	32,3%	100,0%
	2nd Class	67,7%	32,3%	100,0%
	3rd Class	67,7%	32,3%	100,0%
	Crew	67,7%	32,3%	100,0%
Spaltensumme		67,7%	32,3%	100,0%

Beobachtete minus erwartete Häufigkeit

		überlebt		Zeilen-summe
		NEIN	JA	
Klasse	1st Class	-98,0	98,0	0
	2nd Class	-25,9	25,9	0
	3rd Class	50,1	-50,1	0
	Crew	73,9	-73,9	0
Spaltensumme		0	0	0

χ^2-Statistik

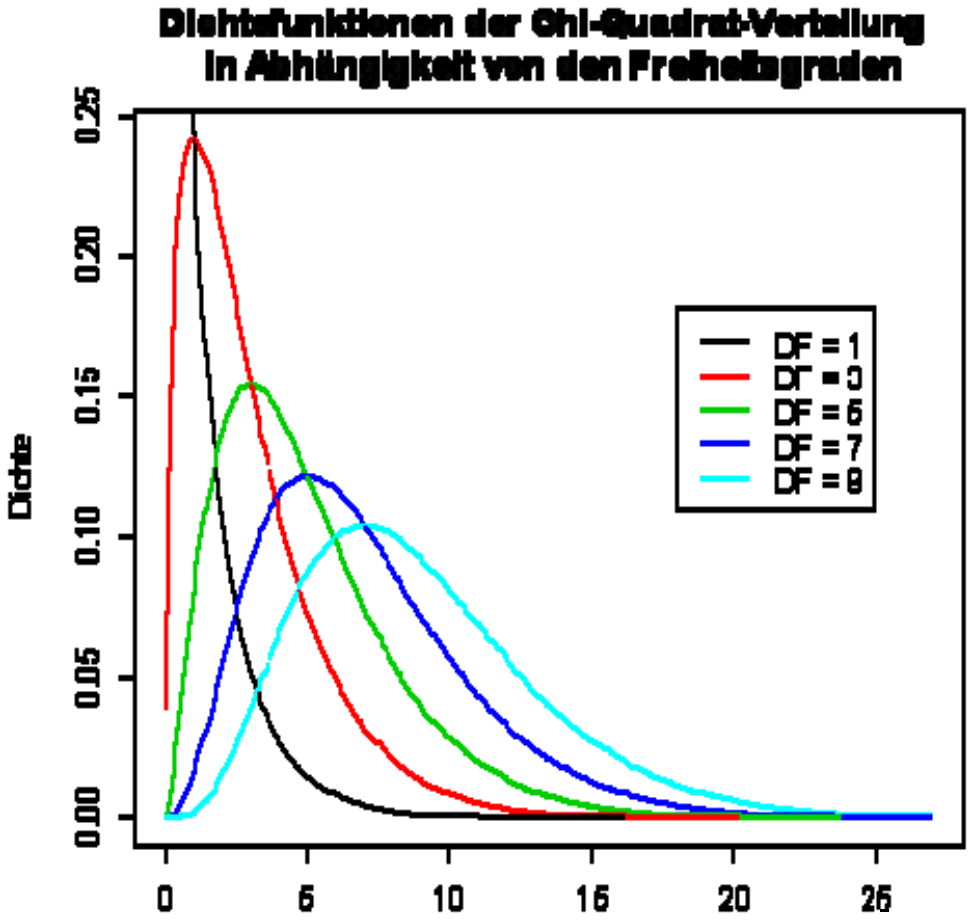
Berechnung des χ^2-Wertes

$$\chi^2 = \sum_{i=1}^{l} \sum_{j=1}^{m} \frac{\left(Beobachteter\ Wert_{ij} - Erwarteter\ Wert_{ij}\right)^2}{Erwarteter\ Wert_{ij}}$$

l	...	Anzahl der Zeilen
m	...	Anzahl der Spalten

Anzahl der Freiheitsgrade: (l-1)(m-1)

Form der Chi²-Verteilungsdichte



Die Chi²-Verteilung ergibt sich aus der Normalverteilung wie folgt:
Seien n Zufallsvariablen Z_i, die unabhängig und standardnormalverteilt sind, gegeben, so ist die Chi-Quadrat-Verteilung mit n Freiheitsgraden definiert als die Verteilung der Summe der quadrierten Zufallsvariablen $Z_1^2+\ldots+ Z_n^2$

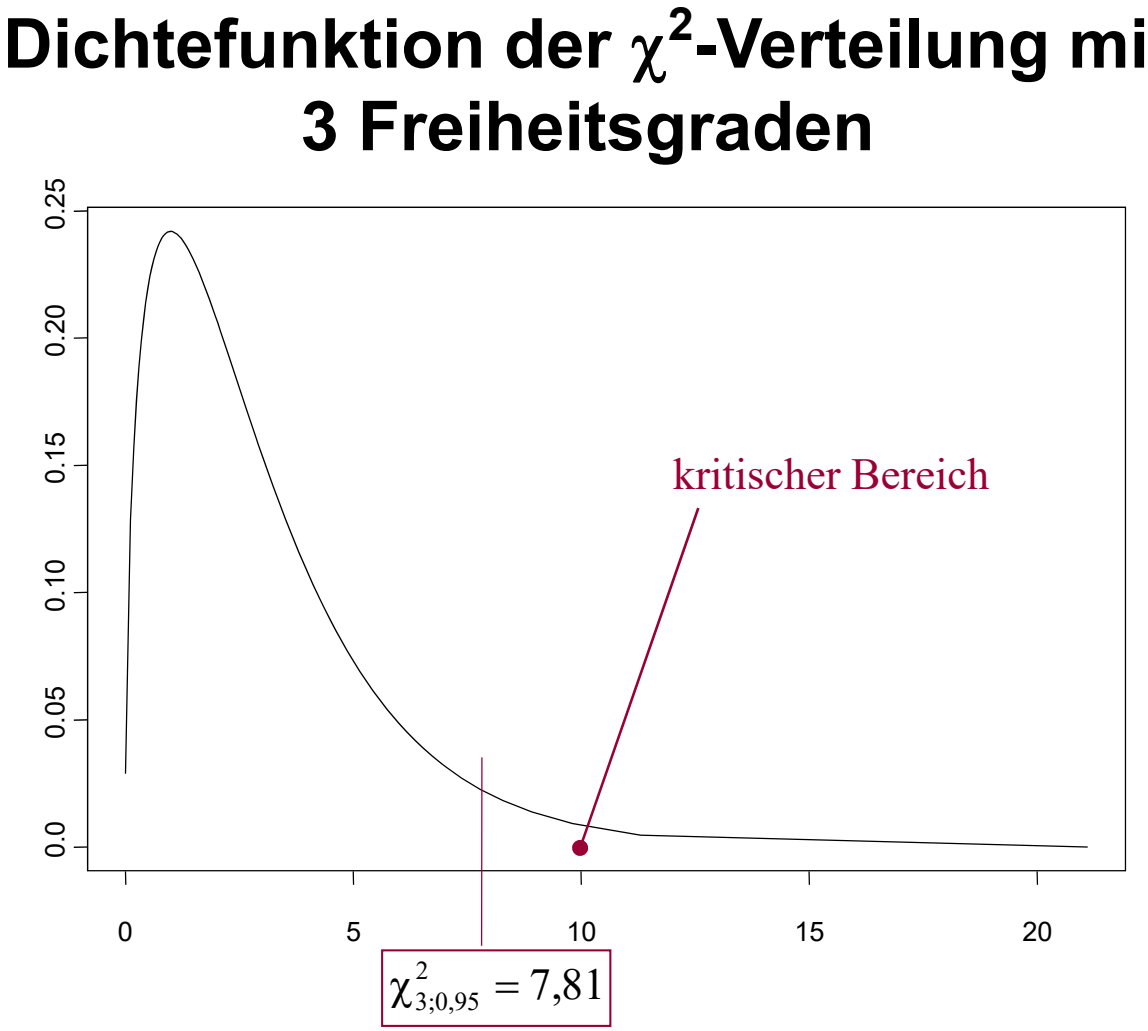
χ^2-Statistik

$$\frac{(o_{11} - e_{11})^2}{e_{11}}$$

		überlebt		Zeilen-summe
		NEIN	JA	
Klasse	1st Class	43,7	91,5	135,2
	2nd Class	3,5	7,3	10,8
	3rd Class	5,2	11,0	16,2
	Crew	9,1	19,1	28,2
Spaltensumme		61,5	128,9	190,4

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

Dichtefunktion der χ^2-Verteilung



χ^2-Statistik

$$\chi^2_{\text{Klasse nach Überleben}} = 190{,}4 \quad > \quad \chi^2_{kritisch} = 7{,}81$$

$\Longrightarrow$ Hoch signifikantes Ergebnis;
Unterschiede zwischen den Überlebenschancen in den verschiedenen Klassen können wohl nicht zufällig sein

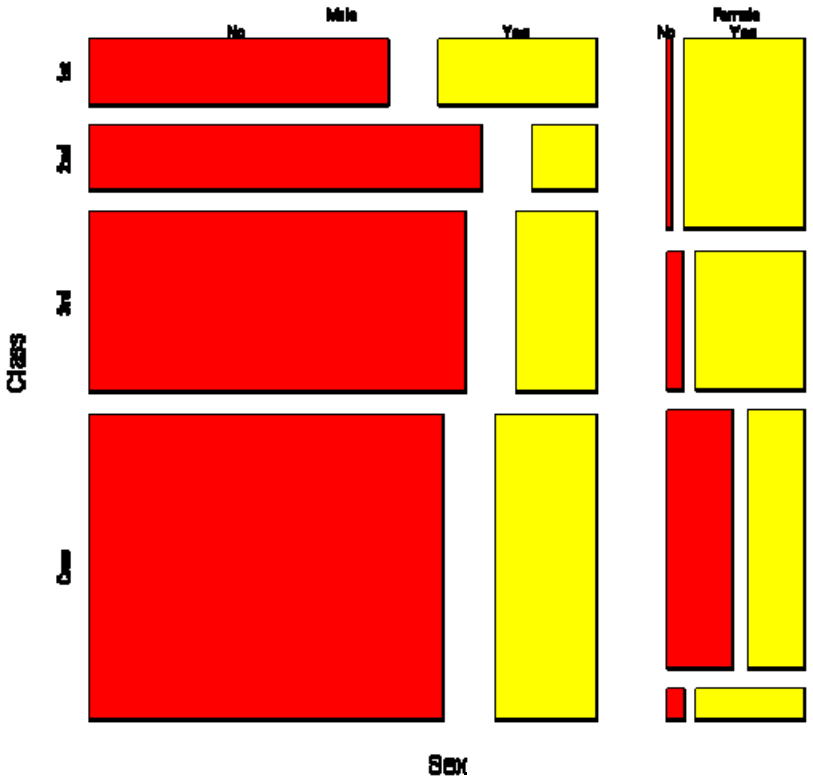
- Es bestehen signifikante Unterschiede zwischen den beobachteten und den unter Unabhängigkeit erwarteten Werten
- Überlebende Personen sind eher in den höherwertigen Unterkunftsklassen zu finden

Analyse von 3 Merkmalen

CLASS * SURVIVED * SEX Kreuztabelle

Anzahl

SEX			SURVIVED		
			No	Yes	Gesamt
Male	CLASS	1st Class	118	62	180
		2nd Class	154	25	179
		3rd Class	422	88	510
		Crew	670	192	862
	Gesamt		1364	367	1731
Female	CLASS	1st Class	4	141	145
		2nd Class	13	93	106
		3rd Class	106	90	196
		Crew	3	20	23
	Gesamt		126	344	470



<u>Log-lineare Modelle</u>:
Werkzeug zur Analyse höher-dimensionaler Häufigkeitstabellen

Äquivalenz zu 2-Stichprobenanteilstest

Beobachtete absolut

	Response	No Response	Total
DCF	43	68	111
CF	26	86	112
	69	154	223

Beobachtete Zeilenprozent

	Response	No Response	Total
DCF	38,7%	61,3%	111
CF	23,2%	76,8%	112
	30,9%	69,1%	223

Erwartete Zeilenprozent

	Response	No Response	Total
DCF	30,9%	69,1%	111
CF	30,9%	69,1%	112
	30,9%	69,1%	223

Erwartete absolut

	Response	No Response	Total
DCF	34,35	77	111
CF	35	77	112
	69	154	223

Kritischer Wert bei α=0,01 6,634897

Kritischer Wert bei α=0,05 3,841459

Abweichung

	Response	No Response	Total
DCF	8,655	-8,655	0,000
CF	-8,655	8,655	0,000
	0,000	0,000	0,000

CHI-WERT

	Response	No Response	Total
DCF	2,181	0,977	
CF	2,161	0,968	
			6,288

p-value 0,0122 6,288

Fallzahl und die Yates-Korrektur

- Die Approximation der Stichprobenverteilung mit der χ^2-Teststatistik darf nur angewendet werden, wenn alle erwarteten Häufigkeiten ≥ 5 sind.
- Andernfalls müssen Zeilen bzw. Spalten der Kreuztabelle zusammengefasst werden.
- Für den Fall der 4-Felder Tafel (Anzahl der Freiheits-grade = 1), wird in der Praxis häufig die sogenannte Yates-Korrektur herangezogen:

$$\chi^2_{\text{korr.}} = \sum_{i=1}^{l} \sum_{j=1}^{m} \frac{\left(\left|o_{ij} - e_{ij}\right| - 0{,}5\right)^2}{e_{ij}}$$

Beispiel zur Yates-Korrektur

Anhand eines Labortests (Digitalis-Konzentration im Blut) kann das Vorliegen einer bestimmten Krankheit nachgewiesen werden. 1975 wurde dazu folgende Statistik veröffentlicht:

T+	positiver Test
T-	negativer Test
D+	krank
D-	gesund

	D+	D-	Total
T+	25	14	39
T-	18	78	96
Total	43	92	135

$$\left.\begin{array}{l}\chi^2 = 26{,}28 \\ \chi^2_{\text{korr.}} = 24{,}23\end{array}\right\} > \chi^2_{1;0,95} = 3{,}84 \Longrightarrow \text{signifikantes Ergebnis}$$

Berechnungsschema in Excel

	D+	D-	
T+	25	14	39
T-	18	78	96
	43	92	135

Ohne Yates-Korrektur

12,7353	5,95236
5,17371	2,41815

26,28 p-Value 0,00000030

	D+	D-	
T+	12,42	26,58	39
T-	30,58	65,42	96
	43	92	135

Mit Yates-Korrektur

11,7429	5,48852
4,77055	2,22971

24,23 p-Value 0,00000085

krit. Wert: 11,3449

Funktion: CHITEST(Beobachtete Werte; Unter H0 erwartete Werte)

Simpson Paradoxon (1)

Clinical Center I

	Treatment		Sum
	A	B	
response	10	100	110
no response	100	730	830
Sum	110	830	940

Response

A:10 von 110 9%
B:100 von 830 12%

p-Value = 0,365

Clinical Center II

	Treatment		Sum
	A	B	
response	100	50	150
no response	50	20	70
Sum	150	70	220

Response

A:100 von 150 67%

B:50 von 70 71%

p-Value = 0,480

Simpson Paradoxon (2)

Data of Clinical Center I and II collapsed

	Treatment		Sum
	A	B	
response	110	150	260
no response	150	750	900
Sum	260	900	1160

p-Value = 0,0001

Response

A: 42% B:17%

Andere Problemstellung: Anpassungstest

Verteilung der Augenzahl x bei n = 235 Würfen mit einem antiken Würfel

x_i	1	2	3	4	5	6
n_i	37	17	49	59	28	45

Theoretische Wahrscheinlichkeit:

$P(X=x) = p_i = 1/6 = 0.167$

Berechnung der χ^2-Statistik

$$\chi^2 = \sum_{i=1}^{k} \frac{(\text{Beobachtet e Häufigkeit}_i - \text{Erwartete Häufigkeit}_i)^2}{\text{Erwartete Häufigkeit}_i}$$

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n \cdot p_i)^2}{n \cdot p_i} = n \sum_{i=1}^{k} \frac{(h_i - p_i)^2}{p_i}$$

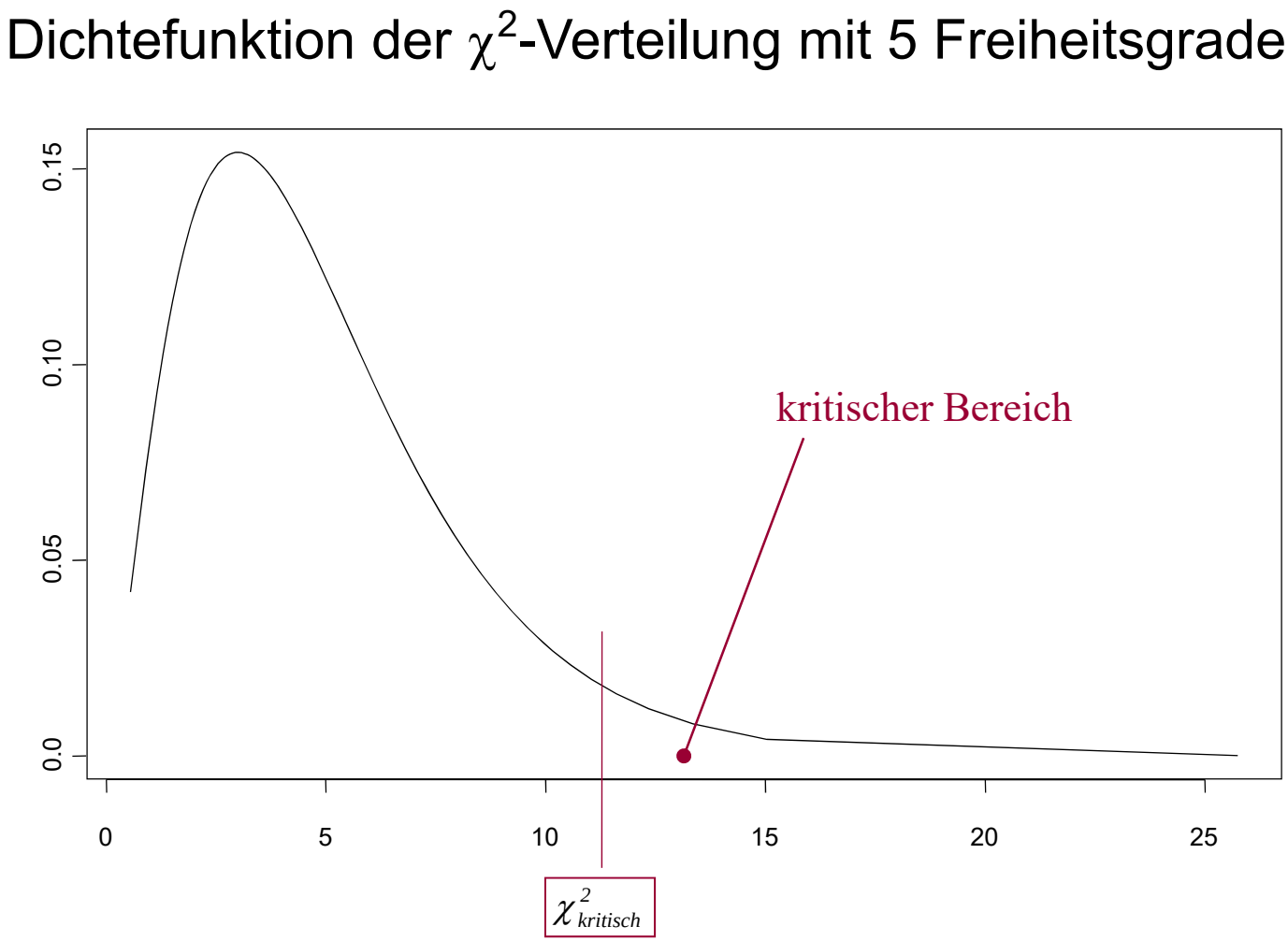
n	...	Stichprobenumfang
n_i	...	beobachtete Häufigkeit
p_i	...	theoretische Wahrscheinlichkeit
$h_i = n_i/n$...		relative Häufigkeit

Arbeitstabelle zur Bestimmung der Prüfgröße

x_i	n_i	$n \cdot p_i$	$(n_i - n \cdot p_i)$	$\frac{(n_i - n \cdot p_i)^2}{n \cdot p_i}$
1	37	39.17	- 2.17	0.120
2	17	39.17	- 22.17	12.548
3	49	39.17	9.83	2.467
4	59	39.17	19.83	10.039
5	28	39.17	- 11.17	3.185
6	45	39.17	5.83	0.868
	n = 235			29.227

$$\chi^2_{\text{Würfel}} = 29{,}227$$

Dichtefunktion der χ^2 -Verteilung



χ^2-Statistik

Berechnung des χ^2-Wertes

$$\chi^2_{\text{Würfel}} = 29{,}227 \quad > \quad \chi^2_{\text{kritisch}} = \chi^2_{5;0.95} = 11{,}07$$

$\Longrightarrow$ signifikantes Ergebnis (Signifikanzniveau $\alpha = 0.05$);

Die beobachteten Häufigkeiten weichen signifikant von den unter der Annahme einer Gleichverteilung erwarteten Häufigkeiten ab.

2er oder 5er werden mit dem antiken Würfel seltener gewürfelt (2 Seiten die gegenüber liegen!)

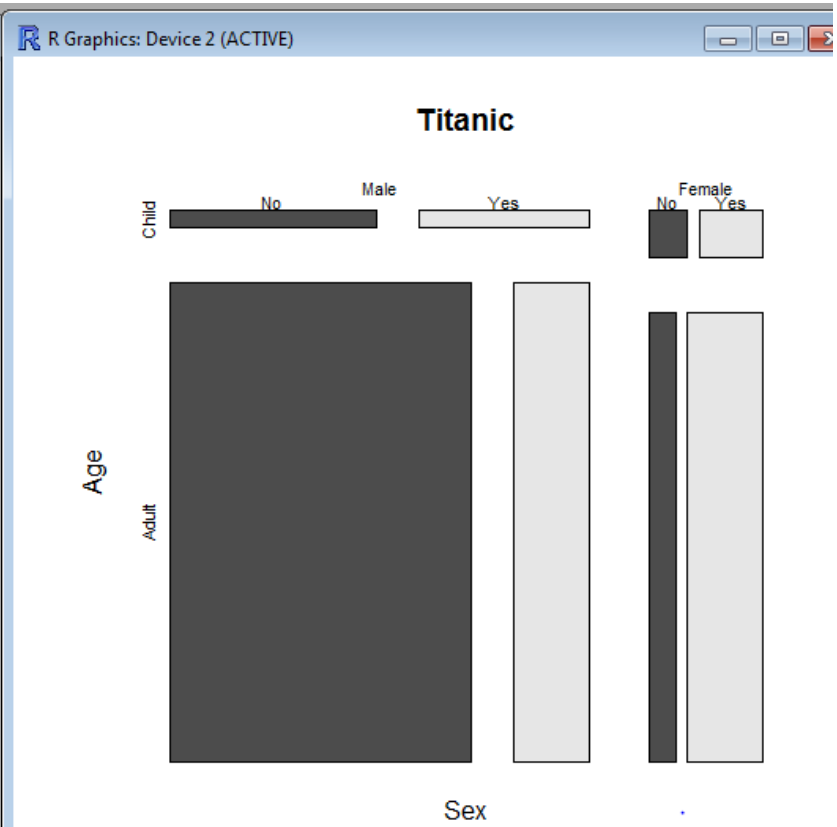
Der antike Würfel ist kein „fairer“ Würfel

Chi²-Test mit R

R Console

```
> mosaicplot(~ Sex + Age + Survived, data = Titanic, color = TRUE)
> apply(Titanic, c(1,4), sum)
      Survived
Class   No Yes
  1st  122 203
  2nd  167 118
  3rd  528 178
  Crew 673 212
> chisq.test(apply(Titanic, c(1,4), sum))

        Pearson's Chi-squared test

data:  apply(Titanic, c(1, 4), sum)
X-squared = 190.4011, df = 3, p-value < 2.2e-16

>
```

R Graphics: Device 2 (ACTIVE)



Beispiel

Mendel überprüfte seine Theorien über die Vererbungsgesetze durch Kreuzung verschiedener Erbsensorten. Gemäß seiner Theorie sollte das Vorkommen von 4 Sorten im Verhältnis 9:3:3:1 stehen.

Eine Stichprobe von 556 Erbsen ergab: 315:108:101:32
Sind die beobachteten Abweichungen signifikant?

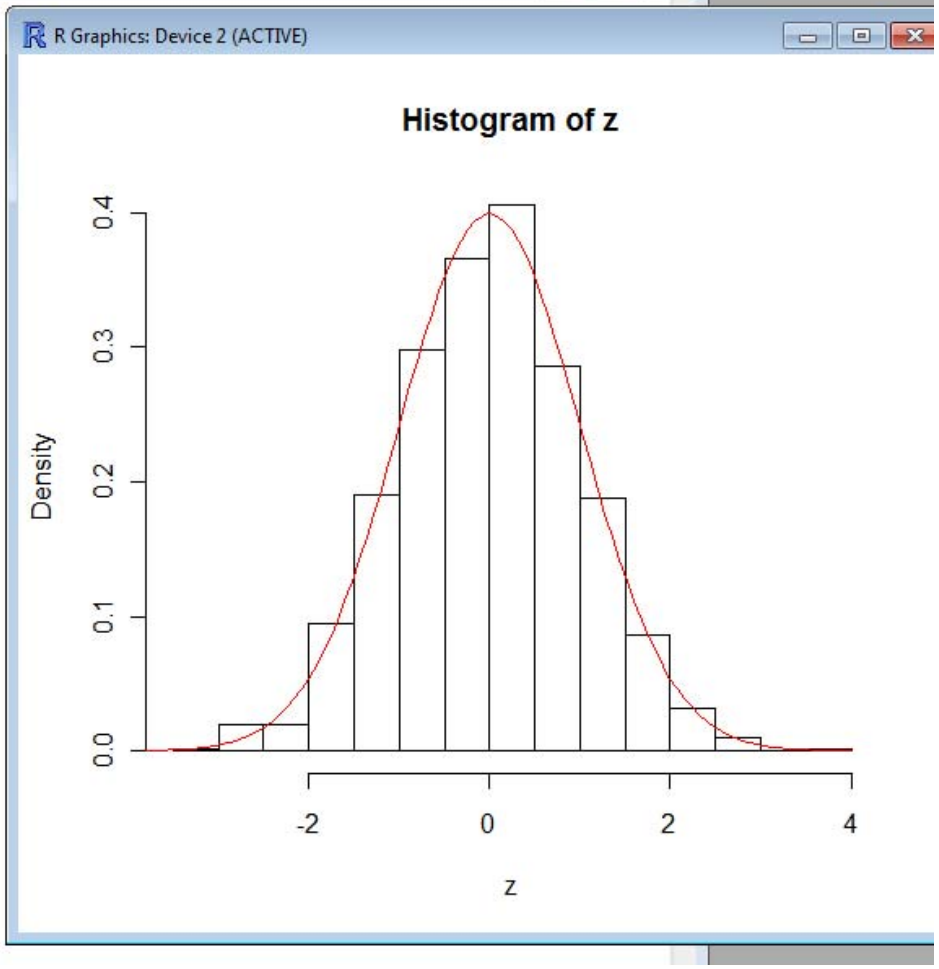
Observed	Soll-Odds	Soll-Rel	Expected	(Obs-Exp)^2	(O-E)^2/E
315	9	0,5625	312,75	5,0625	0,0162
108	3	0,1875	104,25	14,0625	0,1349
101	3	0,1875	104,25	10,5625	0,1013
32	1	0,0625	34,75	7,5625	0,2176
556	16	1	556		0,4700

CHI(3;0,99) 11,345

Anpassungstest auf Normalverteilung

R Console

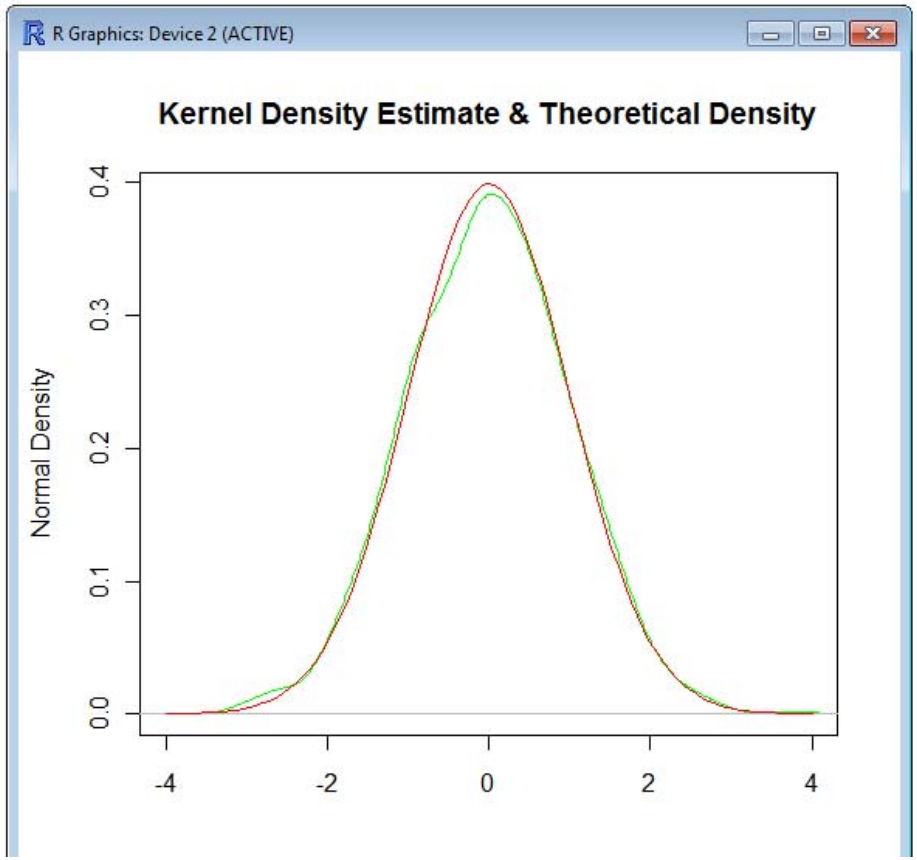
```
> # Chi-Square Test
> interv <- c(-Inf, -3:3, Inf)
> gz <- cut(z, breaks=interv)
> observed <- table(gz)
> expected  <- diff(c(0, pnorm(interv[-1], 0, 1)))
> cbind(observed, expected*length(z))
          observed
(-Inf,-3]        1   1.349898
(-3,-2]         20  21.400234
(-2,-1]        142 135.905122
(-1,0]         332 341.344746
(0,1]          346 341.344746
(1,2]          137 135.905122
(2,3]           21  21.400234
(3, Inf]         1   1.349898
> sum((observed-expected*length(z))^2 / (expected*length(z)))
[1] 0.8819606
> chisq.test(observed, p=expected)

        Chi-squared test for given probabilities

data:  observed
X-squared = 0.882, df = 7, p-value = 0.9965

Warnmeldung:
In chisq.test(observed, p = expected) :
  Chi-Quadrat-Approximation kann inkorrekt sein
>
> # Histogram with Normal Density superimposed
> hist(z, freq = F)
> curve(dnorm, -4, 4, add = T,col = "red")
>
```



Diagnostischer Plot (1)

R Console

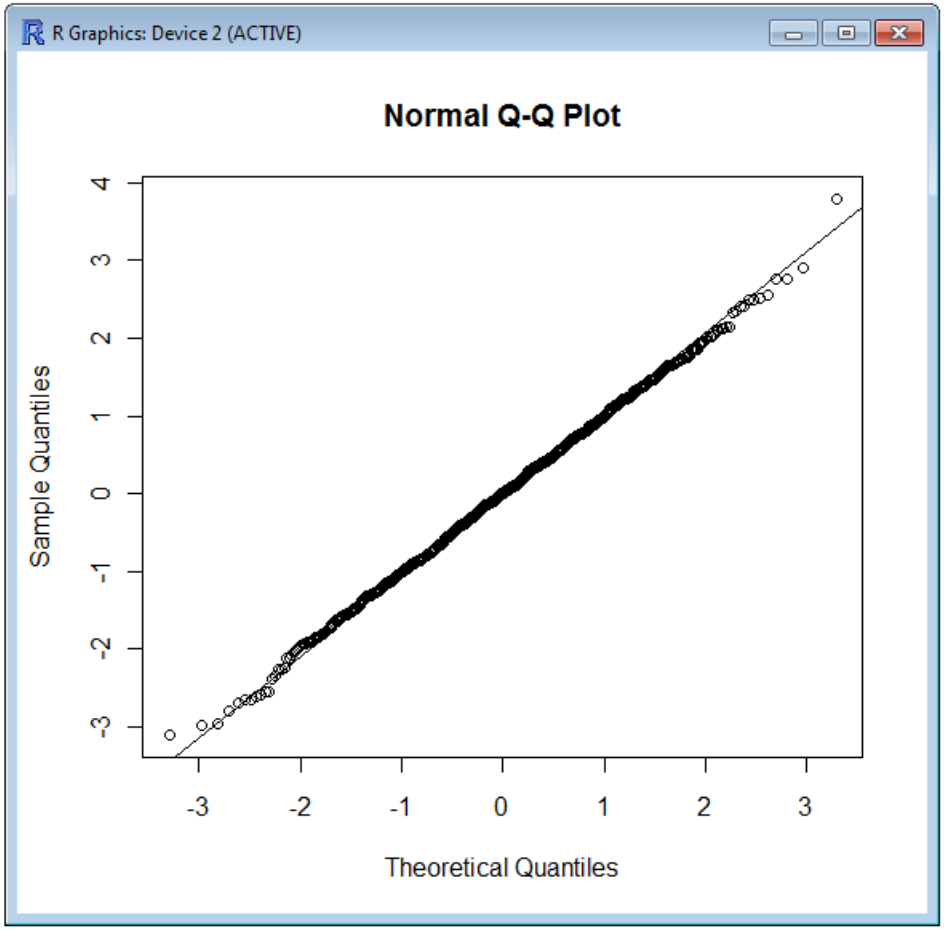
```
> # Kernel Density Estimate with Normal Density superimposed
> plot(density(z), col = "green", xlim=c(-4,4), xlab="", ylab="Normal Density",
+      main="Kernel Density Estimate & Theoretical Density")
> curve(dnorm, -4, 4, add = T,col = "red")
>
>
```



Diagnostischer Plot (2)

R Console

```
> # QQ-Plot
> qqnorm(z)
> qqline(z)
>
```



Q-Q plot

- A quantile-quantile (Q-Q) plot is used to see if a given set of data follows some specified distribution. It should be approximately linear if the specified distribution is the correct model.
- The quantile-quantile (Q-Q) plot is constructed using the theoretical cumulative distribution function F(x) of the specified model and making use of parameter estimates calculated.
- The values in the sample of data, in order from smallest to largest, are denoted $x_{(1)}, x_{(2)}, ..., x_{(n)}$.
- For i = 1, 2, ..., n, $x(_i)$ is plotted against $F^{-1}((i-0.5)/n)$.

Hinweise

- Der Chi-Quadrat Wert liefert eine summarische Beurteilung der Abweichung einer empirischen Verteilung von einer theoretisch erwarteten Verteilung.
- Damit die Verteilung der Teststatistik approximativ Chi-Quadrat verteilt ist, müssen die erwarteten Häufigkeiten in jeder Klasse größer 5 sein. Ist dies nicht der Fall müssen einzelne Klassen aggregiert werden.
- Die Anzahl der Freiheitsgrade ist die Anzahl der Klassen minus eins.
- Falls zur Bestimmung der erwarteten Häufigkeiten auch Parameter geschätzt werden müssen, so sind die Freiheitsgrade zusätzlich um die Anzahl der Parameter zu reduzieren.