

Wahrscheinlichkeitsrechnung und  
Statistik für Biologen  
**6. Chi-Quadrat-Test und  
Fishers exakter Test**

Dirk Metzler & Noémie Becker

9. Mai 2014

## Inhaltsverzeichnis

1	$X^2$ -Anpassungstest für eine vorgegebene Verteilung	1
2	$X^2$ -Test auf Homogenität bzw. Unabhängigkeit	4
3	Fisher's exakter Test	6
4	$X^2$ -Test für Modelle mit angepassten Parametern	8

## 1 $X^2$ -Anpassungstest für eine vorgegebene Verteilung

### Mendels Erbsenexperiment

grün (rezessiv) vs. gelb (dominant)

rund (dominant) vs. runzlig (rezessiv)

Erwartete Häufigkeiten beim Kreuzen von Doppelhybriden:

	grün	gelb
runzlig	$\frac{1}{16}$	$\frac{3}{16}$
rund	$\frac{3}{16}$	$\frac{9}{16}$

Im Experiment beobachtet ( $n = 556$ ):

	grün	gelb
runzlig	32	101
rund	108	315

Passen die Beobachtungen zu den Erwartungen?

Relative Häufigkeiten:

	grün/runz.	gelb./runz.	grün/rund	gelb./rund
erwartet	0.0625	0.1875	0.1875	0.5625
beobachtet	0.0576	0.1942	0.1816	0.5665

Können diese Abweichungen plausibel mit Zufallsschwankungen erklärt werden?  
Wir messen die Abweichungen durch die  $X^2$ -Statistic:

$$X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

wobei  $E_i$  = erwartet Anzahl in Klasse  $i$  und  $O_i$  = beobachtete (engl. *observed*) Anzahl in Klasse  $i$ .  
Wieso teilen wir dabei  $(O_i - E_i)^2$  durch  $E_i = \mathbb{E}O_i$ ?

Sei  $n$  die Gesamtzahl und  $p_i$  die Wahrscheinlichkeit (unter der Nullhypothese) jeder Beobachtung, zu  $O_i$  beizutragen.

Unter der Nullhypothese ist  $O_i$  binomialverteilt:

$$\Pr(O_i = k) = \binom{n}{k} p_i^k \cdot (1 - p_i)^{n-k}.$$

Also

$$\mathbb{E}(O_i - E_i)^2 = \text{Var}(O_i) = n \cdot p \cdot (1 - p).$$

Wenn  $p$  klein ist, gilt  $n \cdot p \cdot (1 - p) \approx n \cdot p$  und

$$\mathbb{E} \frac{(O_i - E_i)^2}{E_i} = \frac{\text{Var}(O_i)}{\mathbb{E}O_i} = 1 - p \approx 1.$$

Anders ausgedrückt:

die Binomialverteilung mit kleinem  $p$  und großem  $n$  kann durch die Poissonverteilung mit Parameter  $\lambda = n \cdot p$  approximiert werden:

$$\binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k} \approx \frac{\lambda^k}{k!} \cdot e^{-\lambda} \quad \text{mit} \quad \lambda = n \cdot p.$$

Eine Zufallsvariable  $Y$ , die Werte in  $0, 1, 2, \dots$  annehmen kann, ist *poissonverteilt* mit Parameter  $\lambda$ , wenn

$$\Pr(Y = k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}.$$

Es gilt dann  $\mathbb{E}Y = \text{Var}(Y) = \lambda$ .

	gr/runz	ge/runz	gr/rund	ge/rund	sum
theorie	0.0625	0.1875	0.1875	0.5625	
erw. ( $E$ )	34.75	104.25	104.25	312.75	556
beob. ( $O$ )	32	101	108	315	556
$O - E$	-2.75	-3.25	3.75	2.25	
$(O - E)^2$	7.56	10.56	14.06	5.06	
$\frac{(O-E)^2}{E}$	0.22	0.10	0.13	0.02	0.47

$$X^2 = 0.47$$

Ist ein Wert von  $X^2 = 0.47$  ungewöhnlich?

Um zu entscheiden, ob ein Wert von  $X^2 = 0.47$  signifikant ist, müssen wir etwas über die Verteilung von  $X^2$  unter der Nullhypothese wissen. (Die Nullhypothese lautet hier: Die erwarteten Häufigkeiten sind

durch Mendels Gesetze gegeben) Falls die Nullhypothese gilt und die Erwartungswerte  $E_i$  nicht zu klein sind (Faustregel: sie sollten alle  $\geq 5$  sein), ist  $X^2$  *ungefähr*  $\chi^2$ -verteilt. Die  $\chi^2$ -Verteilung hängt ab von der Anzahl der Freiheitsgrade **df**.

Die von  $X^2$  hängt ab von der Anzahl der Freiheitsgrade **df** (eng. *degrees of freedom*), d.h. die Anzahl der Dimensionen in denen man von der Erwartung abweichen kann.

In diesem Fall: Die Summe der Beobachtungen muss die Gesamtzahl  $n = 556$  ergeben.

↪ wenn die ersten Zahlen 32, 101, 108 gegeben sind, ist die letzte bestimmt durch

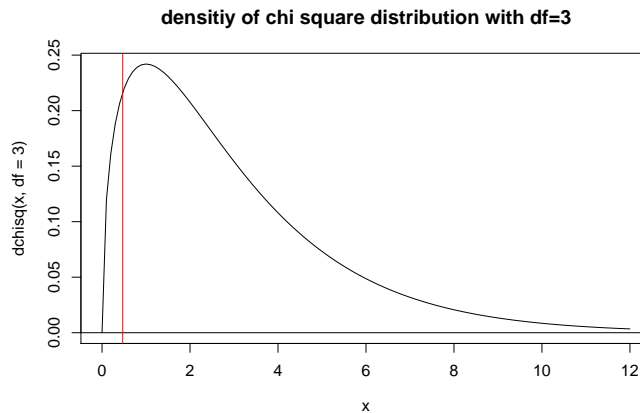
$$315 = 556 - 32 - 101 - 108.$$

$$\Rightarrow \text{df} = 3$$

**Merkregel 1.** Allgemein gilt beim Chi-Quadrat-Anpassungstest mit  $k$  Klassen

$$\text{df} = k - 1.$$

In unserem Beispiel können wir die Verteilung von  $X^2$  also durch die  $\chi^2$ -Verteilung mit  $\text{df}=4-1=3$  approximieren:



```
> pchisq(0.47,df=3) [0.2ex] [1] 0.07456892[0.2ex] > pchisq(0.47,df=3,lower.tail=FALSE) [0.2ex]
[1] 0.925431 ← p-Wert!!!
```

```
> prob <- c(0.0625,0.1875,0.1875,0.5625)
> obs <- c(32,101,108,315)
> (n <- sum(obs))
[1] 556
> (erw <- prob*n)
[1] 34.75 104.25 104.25 312.75
> erw-obs
[1] 2.75 3.25 -3.75 -2.25
> (erw-obs)^2
[1] 7.5625 10.5625 14.0625 5.0625
> (erw-obs)^2/erw
[1] 0.21762590 0.10131894 0.13489209 0.01618705
> sum((erw-obs)^2/erw)
[1] 0.470024
> pchisq(0.470024,df=3,lower.tail=FALSE)
[1] 0.9254259
```

```
> obs <- c(32,101,108,315)
> prob <- c(0.0625,0.1875,0.1875,0.5625)
> chisq.test(obs,p=prob)
```

Chi-squared test for given probabilities

```
data: obs
X-squared = 0.47, df = 3, p-value = 0.9254
```

Ergebnis dieses Beispiels: Die Abweichungen der beobachteten Häufigkeiten von den Vorhersagen der Mendelschen Regeln für zwei unabhängige Genloci sind nicht signifikant. Die Daten sind also in Bezug auf die durch die  $X^2$ -Statistik gemessenen Abweichungen mit der Theorie verträglich.

## 2 $X^2$ -Test auf Homogenität bzw. Unabhängigkeit

Der Kuhstärling ist ein Brutparasit des Oropendola.

### Literatur

[Smi68] N.G. Smith (1968) The advantage of being parasitized. *Nature*, **219(5155)**:690-4

- Kuhstärling-Eier sehen Oropendola-Eiern sehr ähnlich.
- Normalerweise entfernen Oropendolas alles aus ihrem Nest, was nicht genau nach ihren Eiern aussieht.
- In einigen Gegenden sind Kuhstärling-Eier gut von Oropendola-Eiern zu unterscheiden und werden trotzdem nicht aus den Nestern entfernt.
- Wieso?
- Mögliche Erklärung: Dasselfliegenlarven töten häufig junge Oropendolas.
- Nester mit Kuhstärling-Eier sind möglicherweise besser vor Dasselfliegenlarven geschützt.

	Anzahl Kuhstärling-Eier	0	1	2
Anzahlen von Nestern, die von Dasselfliegenlarven befallen sind	befallen	16	2	1
	nicht befallen	2	11	16

	Anzahl Kuhstärling-Eier	0	1	2
In Prozent:	befallen	89%	15%	6%
	nicht befallen	11%	85%	94%

- Anscheinend ist der Befall mit Dasselfliegenlarven reduziert, wenn die Nester Kuhstärlingeier enthalten.
- statistisch signifikant?
- Nullhypothese: Die Wahrscheinlichkeit eines Nests, mit Dasselfliegenlarven befallen zu sein hängt nicht davon ab, ob oder wieviele Kuhstärlingeier in dem Nest liegen.

	Anzahl Kuhstärling-Eier	0	1	2	$\Sigma$
Anzahlen der von Dasselfliegenlarven befallenen Nester	befallen	16	2	1	19
	nicht befallen	2	11	16	29
	$\Sigma$	18	13	17	48

Welche Anzahlen würden wir unter der Nullhypothese erwarten?

Das selbe Verhältnis  $19/48$  in jeder Gruppe.

Erwartete Anzahlen von Dasselfliegenlarven befallener Nester, bedingt auf die Zeilen- und Spalten-

	Anzahl Kuhstärling-Eier	0	1	2	$\Sigma$
summen:	befallen	7.3	5.2	6.5	19
	nicht befallen	10.7	7.8	10.5	29
	$\Sigma$	18	13	17	48

$$18 \cdot \frac{19}{48} = 7.3 \quad 13 \cdot \frac{19}{48} = 5.2$$

Alle anderen Werte sind nun festgelegt durch die **Summen**.

beobachtet (O, observed):	befallen	16	2	1	19
	nicht befallen	2	11	16	29
	$\Sigma$	18	13	17	48

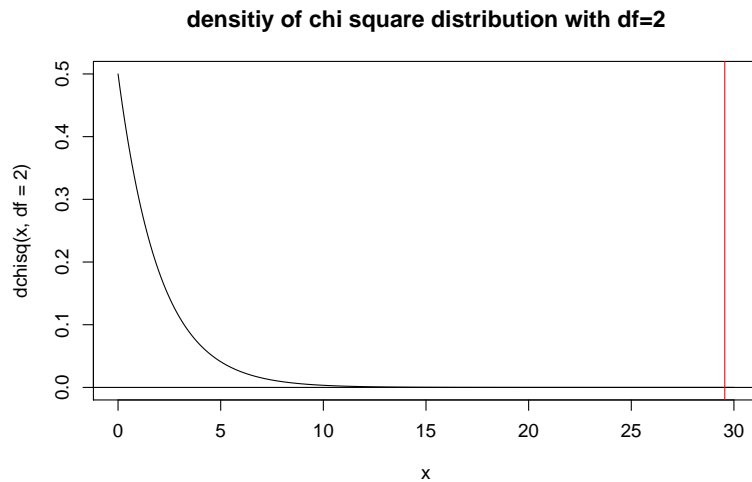
erwartet: (E):	befallen	7.3	5.2	6.5	19
	nicht befallen	10.7	7.8	10.5	29
	$\Sigma$	18	13	17	48

O-E:	befallen	8.7	-3.2	-5.5	0
	nicht befallen	-8.7	3.2	5.5	0
	$\Sigma$	0	0	0	0

$$X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} = 29.5544$$

- Wenn die Zeilen- und Spaltensummen gegeben sind, bestimmen bereits 2 Werte in der Tabelle alle anderen Werte
- $\Rightarrow df=2$  für Kontingenztafeln mit zwei Zeilen und drei Spalten.
- Allgemein gilt für  $n$  Zeilen und  $m$  Spalten:

$$df = (n - 1) \cdot (m - 1)$$



```
> M <- matrix(c(16,2,2,11,1,16),nrow=2)
> M
      [,1] [,2] [,3]
[1,]  16   2   1
[2,]   2  11  16
> chisq.test(M)

Pearson's Chi-squared test

data:  M
X-squared = 29.5544, df = 2, p-value = 3.823e-07
```

Ergebnis: Die Daten zeigen einen signifikanten Zusammenhang zwischen der Anzahl der Kuhstärling-Eier in einem Oropendola-Nest und dem Befall durch Dassenfliegenlarven ( $p < 10^{-6}$ ,  $\chi^2$ -Test,  $df=2$ ).

Der  $p$ -Wert basiert wieder auf einer Approximation durch die  $\chi^2$ -Verteilung.

Faustregel: Die  $\chi^2$ -Approximation ist akzeptabel, wenn alle Erwartungswerte  $E_i \geq 5$  erfüllen.

Alternative: approximiere  $p$ -Werte durch Simulation:

```
> chisq.test(M,simulate.p.value=TRUE,B=50000)

Pearson's Chi-squared test with simulated p-value
(based on 50000 replicates)

data:  M
X-squared = 29.5544, df = NA, p-value = 2e-05
```

### 3 Fisher's exakter Test

#### Literatur

[McK91] J.H. McDonald, M. Kreitman (1991) Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* **351**:652-654.

	synonym	verändernd
polymorph	43	2
fixiert	17	7

```
> McK <- matrix(c(43,17,2,7),2,
                dimnames=list(c("polymorph","fixed"),
                              c("synon","replace")))
```

```
> McK
      synonym replace
polymorph  43      2
fixed     17      7
> chisq.test(McK)
```

Pearson's Chi-squared test  
with Yates' continuity correction

```
data:  McK
X-squared = 6.3955, df = 1, p-value = 0.01144
```

Warning message: In chisq.test(McK) :  
Chi-Square-Approximation may be incorrect

Yates' Stetigkeitskorrektur: Wegen der kleinen erwarteten Werte wird  $\sum_i \frac{(O_i - E_i - 0.5)^2}{E_i}$  verwendet.

```
> chisq.test(McK,simulate.p.value=TRUE,B=100000)
```

Pearson's Chi-squared test with simulated p-value  
(based on 1e+05 replicates)

```
data:  McK
X-squared = 8.4344, df = NA, p-value = 0.00649
```

### Fishers exakter Test

A	B
C	D

- Nullhypothese:  $\frac{EA/EC}{EB/ED} = 1$
- Für  $2 \times 2$ -Tabellen können die  $p$ -Werte exakt berechnet werden. (keine Approximation, keine Simulation).

```
> fisher.test(McK)
```

Fisher's Exact Test for Count Data

```
data:  McK
p-value = 0.006653
alternative hypothesis: true odds ratio
                    is not equal to 1
95 percent confidence interval:
 1.437432 92.388001
sample estimates:
odds ratio
 8.540913
```

43	2	Σ	45
17	7		24
Σ	60	9	69

a	b	Σ	K
c	d		M
Σ	U	V	N

Unter der Annahme, dass die Zeilen und Spalten unabhängig sind, ist die Wahrscheinlichkeit, dass links oben in der Tabelle der Wert  $a$  bzw. oben rechts ein  $b = K - a$  steht:

$$\Pr(a \text{ oben links}) = \frac{\binom{K}{a} \binom{M}{c}}{\binom{N}{U}} = \Pr(b \text{ oben rechts}) = \frac{\binom{K}{b} \binom{M}{d}}{\binom{N}{V}}$$

“hypergeometrische Verteilung”

a	b	Σ	45
c	d		24
Σ	60	9	69

**Einseitiger Fisher-Test:**

für  $b = 2$ :

$p$ -Wert =  $\Pr(0) + \Pr(1) + \Pr(2) = 0.00665313$

für  $b = 3$ :

$p$ -Wert =  $\Pr(0) + \Pr(1) + \Pr(2) + \Pr(3) = 0.04035434$

**Zweiseitiger Fisher-Test:**

Addiere alle Wahrscheinlichkeiten, die kleiner oder gleich  $\Pr(b)$  sind.

für  $b = 2$ :

$p$ -Wert =  $\Pr(0) + \Pr(1) + \Pr(2) = 0.00665313$

für  $b = 3$ :

$p$ -Wert =  $\Pr(0) + \Pr(1) + \Pr(2) + \Pr(3) + \Pr(9) = 0.05599102$

b	Pr(b)
0	0.000023
1	0.00058
2	0.00604
3	0.0337
4	0.1117
5	0.2291
6	0.2909
7	0.2210
8	0.0913
9	0.0156

Bitte beachten: beim der zweiseitigen Version von Fishers exaktem Test werden nur die Wahrscheinlichkeiten aufsummiert, die kleiner oder gleich der Wahrscheinlichkeit des beobachteten Ergebnisses sind. Im zuvor betrachteten Beispiel mit  $b = 2$  führen aus diesem Grund der einseitige und der zweiseitige Test zum selben  $p$ -Wert.

## 4 $\chi^2$ -Test für Modelle mit angepassten Parametern

Gegeben sei eine Population im *Hardy-Weinberg-Gleichgewicht* und ein Gen-Locus mit zwei möglichen Allelen A und B mit Häufigkeiten  $p$  und  $1 - p$ .

↪ Genotyp-Häufigkeiten

AA	AB	BB
$p^2$	$2 \cdot p \cdot (1 - p)$	$(1 - p)^2$

Beispiel: M/N Blutgruppen; Stichprobe: 6129 Amerikaner europäischer Abstammung

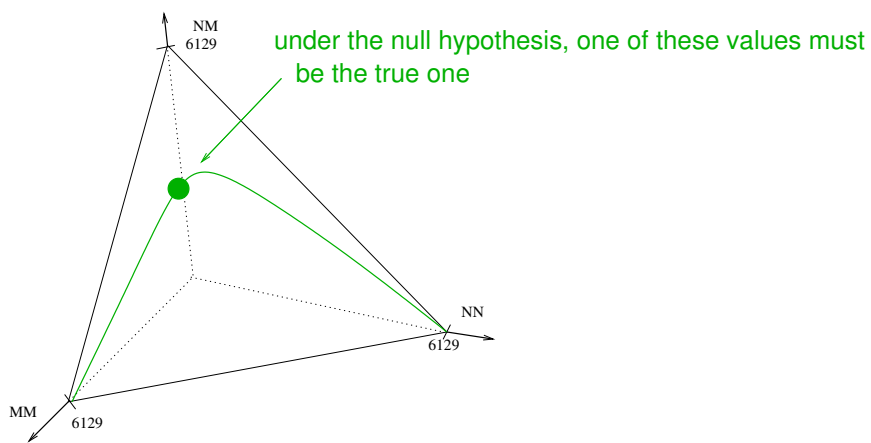
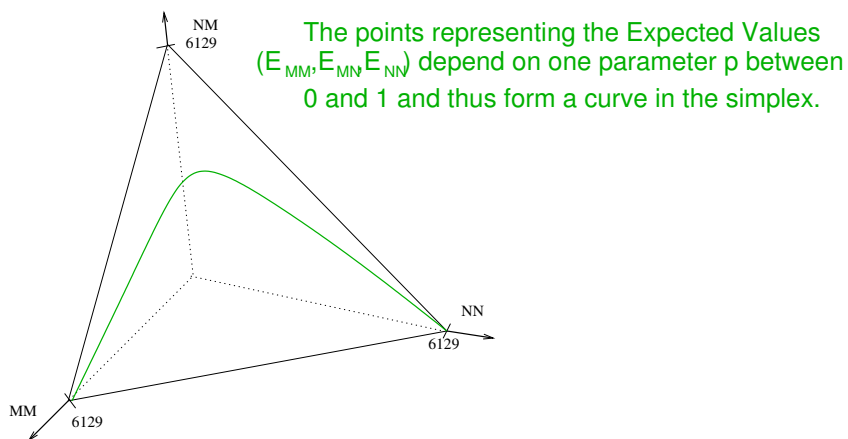
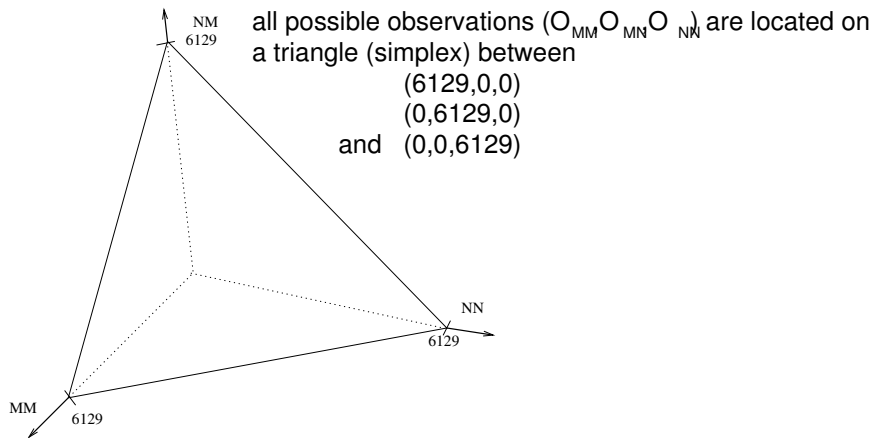
beobachtet:	MM	MN	NN
	1787	3037	1305

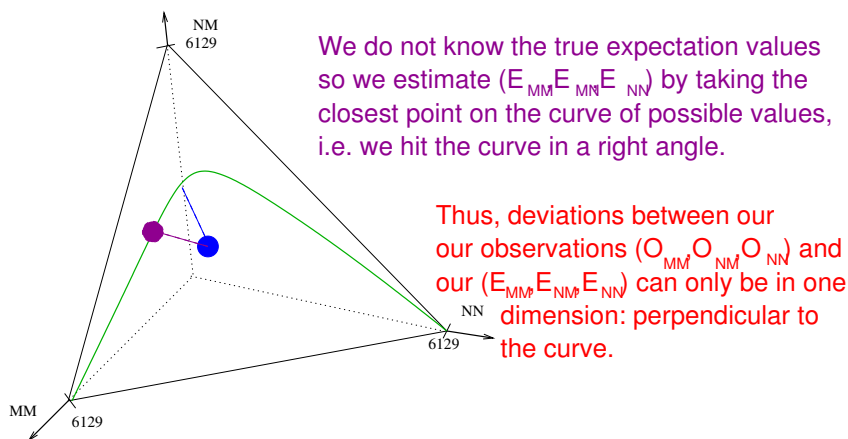
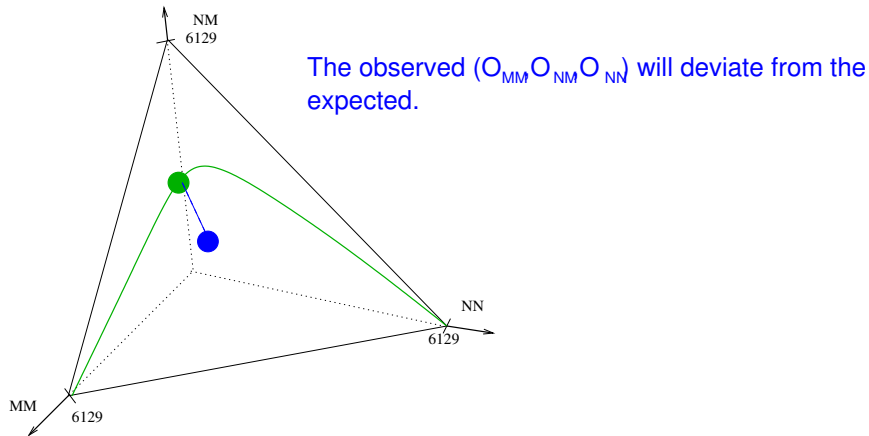
Geschätzte Allelhäufigkeit  $p$  von M:

$$\frac{2 \cdot 1787 + 3037}{2 \cdot 6129} = 0.5393$$

↪ Erwartungswerte:	MM	MN	NN
	$p^2$	$2 \cdot p \cdot (1 - p)$	$(1 - p)^2$
	0.291	0.497	0.212
	1782.7	3045.5	1300.7







$$df = k - 1 - m$$

$k$  = Anzahl Gruppen ( $k=3$  Genotypen)  $m$  = Anzahl Modellparameter ( $m=1$  Parameter  $p$ ) im Blutgruppenbeispiel:

$$df = 3 - 1 - 1 = 1$$

```
> obs <- c(1787,3037,1305)
> n <- sum(obs)
> p <- (2* 1787+3037)/(2* 6129)
> probs <- c(p^2,2*p*(1-p),(1-p)^2)
> erw <- probs*n
> (X2 <- sum((obs-erw)^2/erw))
[1] 0.04827274
> (p.value <- pchisq(X2,df=1,lower.tail=FALSE))
[1] 0.8260966
```