Biostatistik, WS 2017/18

Kontingenztafeln und Chi-Quadrat-Test

Matthias Birkner

http://www.staff.uni-mainz.de/birkner/Biostatistik1718/

12.1.2018

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

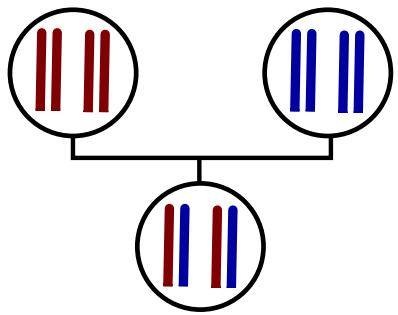
Mendels Erbsenexperiment

Betrachte zwei Merkmale:
Farbe: grün (rezessiv) vs. gelb (dominant)

Form: rund (dominant) vs. runzlig (rezessiv)

Beim Kreuzen von Doppelhybriden erwarten wir folgende Phänotypwahrscheinlichkeiten unter Mendelscher Segregation:

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



Im Experiment beobachtet ($n = 556$ Versuche):

	grün	gelb
runzlig	32	101
rund	108	315

Frage:
Passen die Beobachtungen zu den theoretischen Erwartungen?

Relative Häufigkeiten:

	grün/runzlig	gelb/runzlig	grün/rund	gelb/rund
erwartet	0,0625	0,1875	0,1875	0,5625
beobachtet	0,0576	0,1942	0,1816	0,5665

bzw. in absoluten Häufigkeiten ($n = 556$):

	grün/runzlig	gelb/runzlig	grün/rund	gelb/rund
erwartet	34,75	104,25	104,25	312,75
beobachtet	32	101	108	315

Können diese Abweichungen plausibel durch Zufallsschwankungen erklärt werden?

Wir messen die Abweichungen durch die χ^2-Statistik:

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

wobei $E_i =$ erwartete Anzahl in Klasse i und $O_i =$ beobachtete (engl. *observed*) Anzahl in Klasse i.

(im Beispiel durchläuft i die vier möglichen Klassen grün/runzlig, gelb/runzlig, grün/rund, gelb/rund.)

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:

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

Also

$$\mathbb{E}\, O_i = n\, p_i, \quad \mathbb{E}(O_i - E_i)^2 = \mathrm{Var}(O_i) = n\, p_i\, (1 - p_i)$$

und

$$\mathbb{E}\left[\frac{(O_i - E_i)^2}{E_i}\right] = \frac{\mathrm{Var}(O_i)}{\mathbb{E}O_i} = 1 - p_i$$

(was gar nicht von n abhängt).

Für das Erbsenbeispiel finden wir:

	gr/runz	ge/runz	gr/rund	ge/rund	Summe
theor. Ant.	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

$$\chi^2 = 0.47$$

Ist ein Wert von $\chi^2 = 0.47$ ungewöhnlich?

Die (asymptotische) Verteilung von χ^2 hängt ab von der sog. Anzahl der Freiheitsgrade **df** (eng. *degrees of freedom*), anschaulich gesprochen die Anzahl der Dimensionen, in denen man von der Erwartung abweichen kann.

In diesem Fall: Es gibt vier Klassen, die Summe der Beobachtungen muss die Gesamtzahl $n = 556$ ergeben.
$\rightsquigarrow$ wenn die ersten Zahlen 32, 101, 108 gegeben sind, ist die letzte bestimmt durch

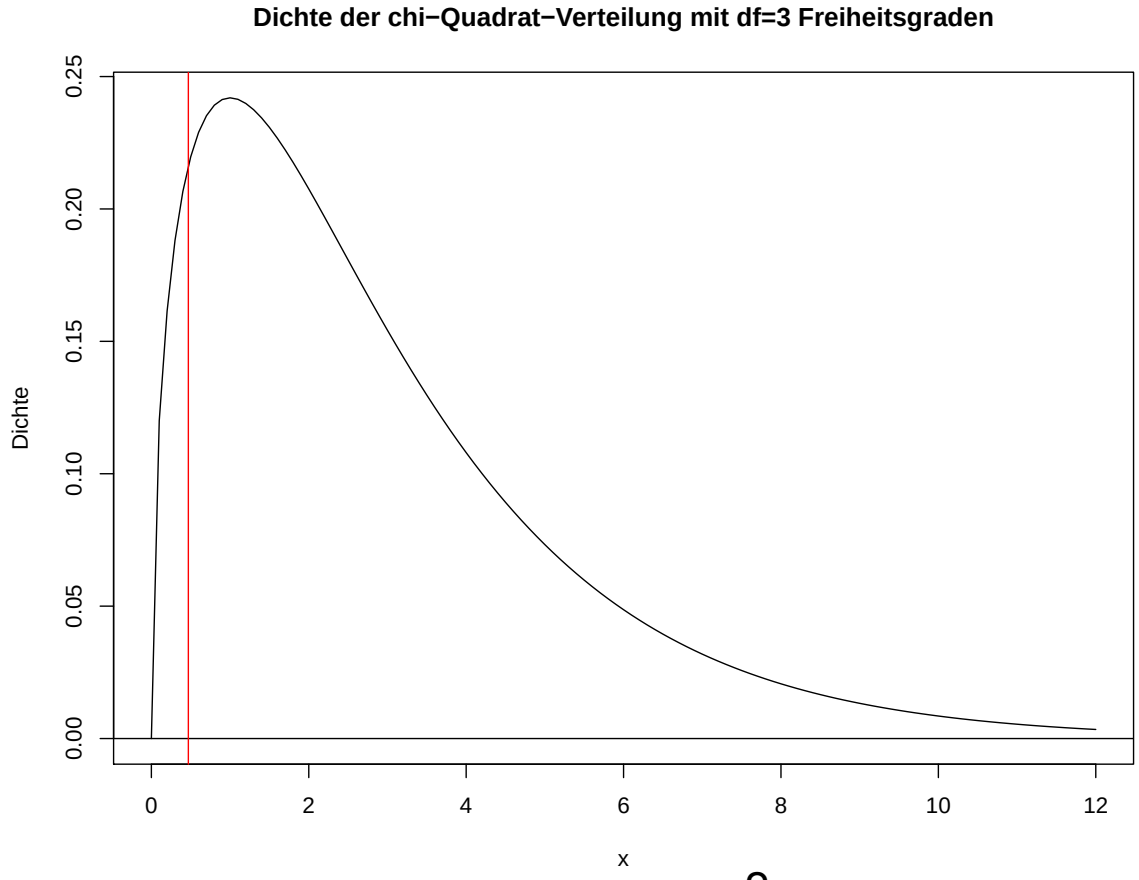
$$315 = 556 - 32 - 101 - 108.$$

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

Merkregel

Allgemein gilt beim Chi-Quadrat-Anpassungtest mit k Klassen (wenn das Modell voll spezifiziert ist, d.h. keine Parameter geschätzt werden)

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



Wir hatten im Erbsenbeispiel gesehen: $\chi^2 = 0.47$ mit df=3 Freiheitsgraden.

Für eine χ^2 mit 3 Freiheitsgraden-verteilte ZV X (man schreibt oft auch χ^2_3-verteilt) gilt

$\mathbb{P}(X \leq 0.47) \dot{=} 0{,}075$ (und somit ist der p-Wert $\mathbb{P}(X \geq 0.47) \dot{=} 0{,}925$), demnach zeigt der χ^2-Test keine signifikante Abweichung.

χ^2-Anpassungstest für eine vorgegebene Verteilung

Allgemeine Situation

Experiment mit k möglichen Ausgängen („Klassen“ oder „Ausprägungen“) wird n-Mal wiederholt,

theoretische Vorhersage („Nullhypothese“ H_0):
Ausgang i hat Wahrscheinlichkeit p_i (mit $p_1 + \cdots + p_k = 1$),

Wir erwarten demnach (unter H_0) etwa $E_i = np_i$ mal Ausgang i (für $i = 1, 2, \ldots, k$).

Wir möchten H_0 anhand von beobachteten Häufigkeiten $(O_1, O_2, \ldots, O_k)$ zum Signifikanzniveau $\alpha \in (0, 1)$ testen:

Unter H_0 ist

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

(approximativ) χ^2-verteilt mit $k - 1$ Freiheitsgraden.

χ^2-Anpassungstest für eine vorgegebene Verteilung

Allgemeine Situation

Wir erwarten unter H_0: $E_i = np_i$ mal Ausgang i (für $i = 1, 2, \ldots, k$)

und möchten H_0 anhand von beobachteten Häufigkeiten $(O_1, O_2, \ldots, O_k)$ zum Signifikanzniveau $\alpha \in (0, 1)$ testen:

Unter H_0 ist

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

(approximativ) χ^2-verteilt mit $k - 1$ Freiheitsgraden.

Also: Lehne H_0 zum Niveau α ab, wenn

$$\chi^2 > q_{\chi^2_{k-1}, 1-\alpha} \quad \text{(das } (1-\alpha)\text{-Quantil der } \chi^2_{k-1}\text{-Verteilung)}$$

Die Quantile der χ^2-Verteilungen sind tabelliert, oder mit `R`:
`qchisq(1-`α`, df=`$k-1$`)`

χ^2-Anpassungstest für eine vorgegebene Verteilung

Allgemeine Situation

Wir erwarten unter H_0: $E_i = np_i$ mal Ausgang i (für $i = 1, 2, \ldots, k$) und beobachten Häufigkeiten $(O_1, O_2, \ldots, O_k)$.

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

Der p-Wert des Tests ist die Wahrscheinlichkeit (unter H_0), dass es χ^2_{k-1}-verteiltes X einen mindestens so großen Wert annimmt wie das aus den Daten bestimmte χ^2, d.h.

$$\mathbb{P}(X \geq \chi^2) = 1 - \mathbb{P}(X \leq \chi^2)$$

R kennt die Verteilungsfunktionen der χ^2-Verteilungen, z.B.

```
> pchisq(0.47,df=3)
[1] 0.07456892
```

Ein weiteres Beispiel

Wir vermuten, dass ein gegebener sechsseitiger Würfel unfair ist.

Bei 120-maligem Würfeln finden wir folgende Häufigkeiten:

i	1	2	3	4	5	6
O_i	13	12	20	18	26	31

Der χ^2-Test komplett in R:

```
> w <- c(13,12,20,18,26,31)
> chisq.test(w,p=c(1/6,1/6,1/6,1/6,1/6,1/6))

Chi-squared test for given probabilities

data:  w
X-squared = 13.7, df = 5, p-value = 0.01763
```

Oft zitierte „Faustregel“: Die χ^2-Approximation ist akzeptabel, wenn alle erwarteten Werte $np_i \geq 5$ erfüllen.

Lassen wir für das Würfel-Beispiel `R` den p-Wert via Simulation bestimmen:

```
> w <- c(13,12,20,18,26,31)
> chisq.test(w, p=c(1/6,1/6,1/6,1/6,1/6,1/6),
             simulate.p.value=TRUE)

Chi-squared test for given probabilities with
        simulated p-value (based on 2000 replicates)

data:  w
X-squared = 13.7, df = NA, p-value = 0.01799
```

(In diesem Beispiel ist $E_i = 120 \cdot \frac{1}{6} = 20$ (≥ 5).)

Der Kuhstärling ist ein Brutparasit des Oropendola.

photo (c) by J. Oldenettel

Wir verwenden simulierte Daten in Anlehnung an

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

- Kuhstärling-Eier sehen Oropendola-Eiern meist 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: Junge Oropendolas sterben häufig am Befall durch Dasselfliegenlarven.
- Nester mit Kuhstärling-Eier sind möglicherweise besser vor Dasselfliegenlarven geschützt.

Anzahlen von Nestern, die von Dasselfliegenlarven befallen sind

Anzahl Kuhstärling-Eier	0	1	2
befallen	16	2	1
nicht befallen	2	11	16

In Prozent:

Anzahl Kuhstärling-Eier	0	1	2
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.

Anzahlen der von Dasselfliegenlarven befallenen Nester

Anzahl Kuhstärling-Eier	0	1	2	$\sum$
befallen	16	2	1	19
nicht befallen	2	11	16	29
$\sum$	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 Spaltensummen:

Anzahl Kuhstärling-Eier	0	1	2	$\sum$
befallen	7.13	5.15	6.72	19
nicht befallen	10.87	7.85	10.28	29
$\sum$	18	13	17	48

$$18 \cdot \frac{19}{48} = 7.13 \qquad 13 \cdot \frac{19}{48} = 5.15$$

Alle anderen Werte sind nun festgelegt durch die Summen.

beobachtet (O, observed):

befallen	16	2	1	19
nicht befallen	2	11	16	29
$\sum$	18	13	17	48

erwartet: (E):

befallen	7.13	5.15	6.72	19
nicht befallen	10.87	7.85	10.28	29
$\sum$	18	13	17	48

O-E:

befallen	8.87	-3.15	-5.72	0
nicht befallen	-8.87	-3.15	5.72	0
$\sum$	0	0	0	0

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} = 29.5$$

- 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:

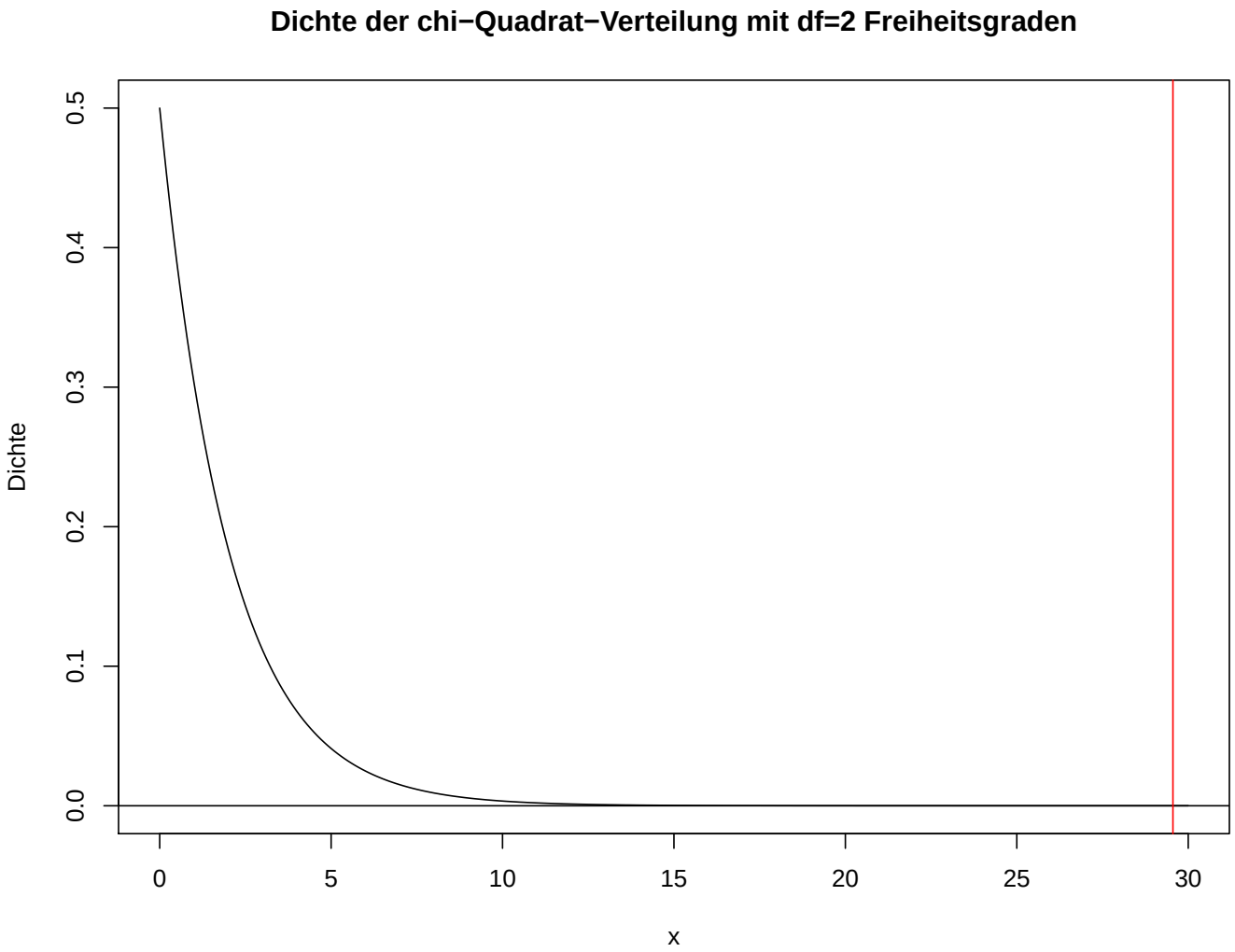
$$df = (n-1) \cdot (m-1)$$

Wir haben den Wert $\chi^2 = 29.5$ beobachtet.

Unter der Nullhypothese „die Wahrscheinlichkeit, mit der ein Nest von Dasselfliegenlarven befallen wird, hängt nicht von der Anzahl Kuhstärling-Eier ab“ ist die Teststatistik (approximativ) χ^2-verteilt mit $2 = (2-1)\cdot(3-1)$ Freiheitsgraden.

Das 99%-Quantil der χ^2-Verteilung mit df=2 ist 9.21 ($<$29.5), wir können also die Nullhypothese zum Signifikanzniveau 1% ablehnen.
(Denn wenn die Nullhypothese zutrifft, so würden wir in weniger als 1% der Fälle einen so extremen Wert der χ^2-Statistik beobachten.)

Faustregel: Die χ^2-Approximation ist akzeptabel, wenn alle Erwartungswerte $E_i \geq 5$ erfüllen, was in dem Beispiel erfüllt ist.
(Siehe die folgenden Folien für die mit dem Computer bestimmten exakten *p*-Werte.)



Bemerkung 1: Genauere Rechnung ergibt: Für ein χ^2_2-verteiltes X gilt $\mathbb{P}(X \geq 29.6) \doteq 3.7 \cdot 10^{-7}$ (was hier wörtlich der p-Wert des χ^2-Tests auf Unabhängigkeit wäre, in dieser Genauigkeit für statistische Zwecke allerdings sinnlos ist).

χ^2-Test auf Homogenität komplett mit R

```
> 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
```

Bemerkung 2: Um die Gültigkeit der χ^2-Approximation (und der Faustregel) in diesem Beispiel einzuschätzen, könnten wir einen Computer beauftragen, durch vielfach wiederholte Simulation den *p*-Wert zu schätzen.
Mit R funktioniert das beispielsweise folgendermaßen:

```
> 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,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
```

Wir sehen: Der empirisch geschätze *p*-Wert $2 \cdot 10^{-5}$ stimmt zwar nicht mit dem aus der χ^2-Approximation überein, aber beide sind hochsignifikant klein (und in einem Bereich, in dem der exakte Wert sowieso statistisch „sinnlos“ ist). Insoweit ist die Faustregel hier bestätigt.

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

$\rightsquigarrow$ Genotyp-Häufigkeiten

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

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

beobachtet:	MM	MN	NN
	1787	3037	1305

Geschätzte Allelhäufigkeit von M:

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

(Dies ist der Anteil beobachteter M-Chromosomen in der Stichprobe,

denn in der Stichprobe sind 6129 diploide Individuen, also $2 \cdot 6129$ Chromosomenkopien,

davon sind $2 \cdot 1787 + 3037$ vom Typ M.)

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

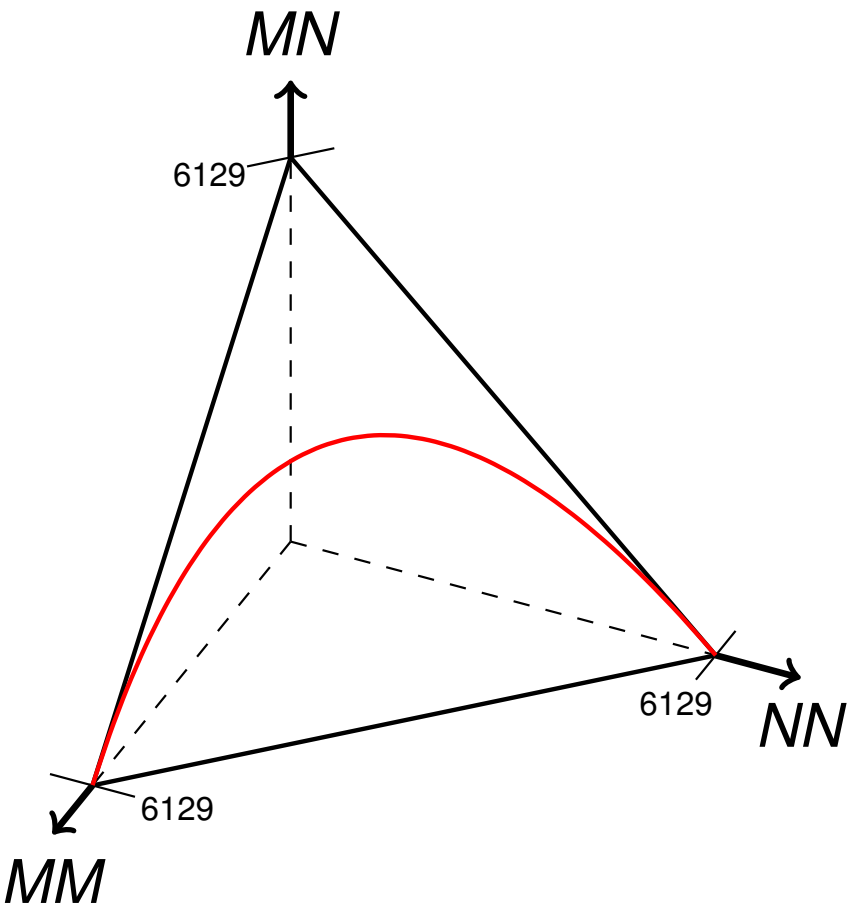
beobachtet:	MM	MN	NN
	1787	3037	1305

Geschätzte Allelhäufigkeit von M:

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

$\rightsquigarrow$ Erwartete Werte (anhand der Schätzung):

MM	MN	NN	
$\widehat{p}^2$	$2 \cdot \widehat{p} \cdot (1 - \widehat{p})$	$(1 - \widehat{p})^2$	
0.291	0.497	0.212	(Anteile)
$n \cdot \widehat{p}^2$	$n \cdot 2 \cdot \widehat{p} \cdot (1 - \widehat{p})$	$n \cdot (1 - \widehat{p})^2$	
1782.7	3045.6	1300.7	(Häufigkeiten)



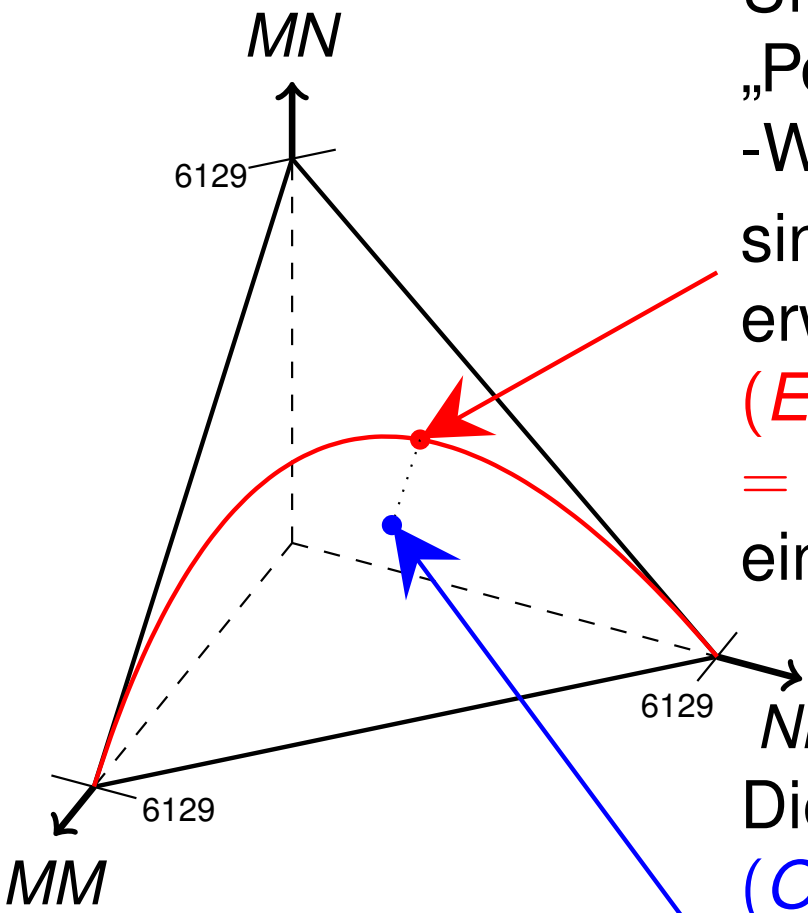
Die möglichen Beobachtungen (O_{MM}, O_{MN}, O_{NN}) liegen auf dem Dreieck („Simplex“) mit Eckpunkten $(6129, 0, 0)$, $(0, 6129, 0)$ und $(0, 0, 6129)$.

Wenn die Population im Hardy-Weinberg-Gleichgewicht ist, so liegen die erwarteten Häufigkeiten

$$(E_{MM}, E_{MN}, E_{NN}) = \left(n \cdot p^2, n \cdot 2p(1-p), n \cdot (1-p)^2\right)$$

auf der roten Kurve.

(Das wahre p könnte irgend ein Wert aus $[0, 1]$ sein.)



Unter der Nullhypothese „Population ist im Hardy -Weinberg-Gleichgewicht“ sind die tatsächlichen erwarteten Häufigkeiten

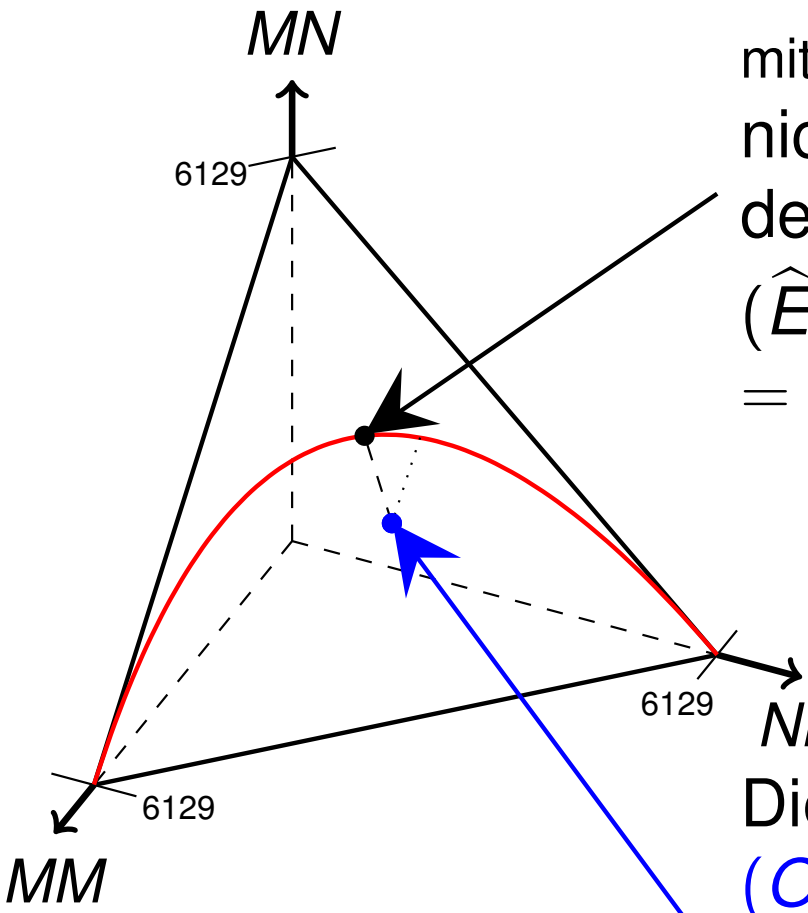
(E_{MM}, E_{MN}, E_{NN})
$= (n \cdot p^2, n \cdot 2p(1-p), n \cdot (1-p)^2)$

ein Punkt auf der roten Kurve.

Die beobachteten Häufigkeiten

(O_{MM}, O_{MN}, O_{NN})

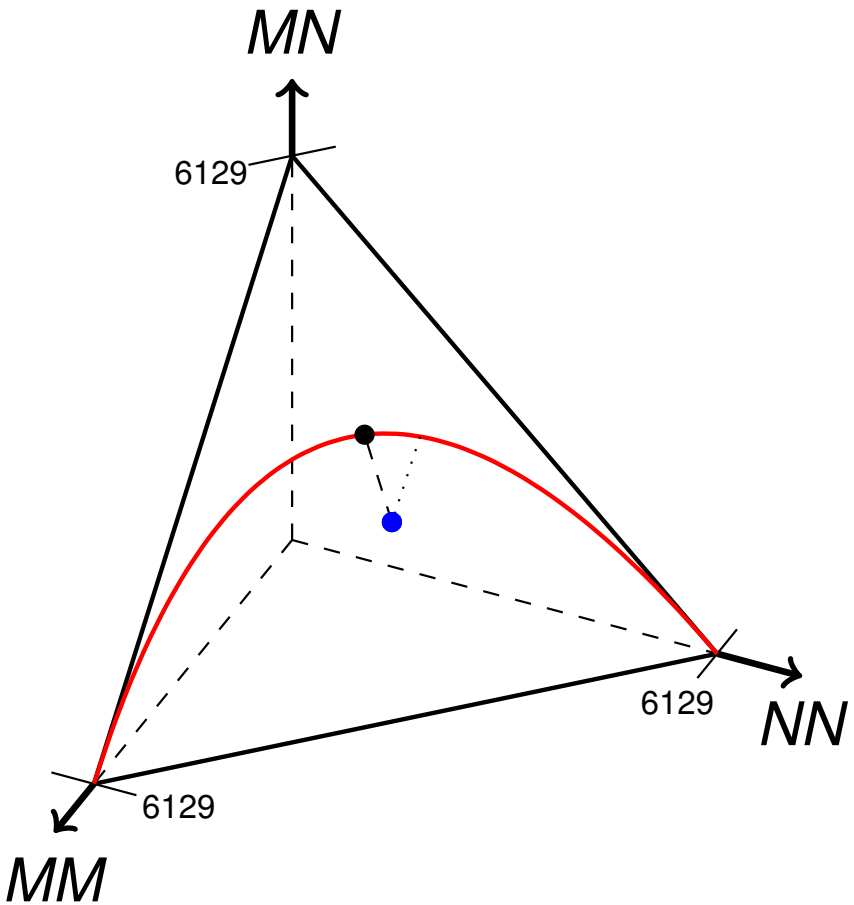
liegen (aufgrund von Zufallsschwankungen typischerweise) *nicht* auf der roten Kurve.



Wir kennen das wahre p (und damit die wahren erwarteten Häufigkeiten) nicht und ersetzen es durch $\widehat{p}$, der geschätzte Punkt ist

$$(\widehat{E}_{MM}, \widehat{E}_{MN}, \widehat{E}_{NN}) = \left(n \cdot \widehat{p}^2, n \cdot 2\widehat{p}(1-\widehat{p}), n \cdot (1-\widehat{p})^2\right)$$

Die beobachteten Häufigkeiten (O_{MM}, O_{MN}, O_{NN}) liegen (aufgrund von Zufallsschwankungen typischerweise) *nicht* auf der roten Kurve.



Es gibt also „effektiv“ nur noch eine Richtung, in der die Beobachtungen von der Nullhypothese abweichen können:

Senkrecht zur roten Kurve.

Daher bleibt (hier)
nur 1 Freiheitsgrad übrig.

Für die Anzahl Freiheitsgrade im χ^2-Test mit angepassten Parametern gilt

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

mit

$k =$ Anzahl Gruppen (hier k=3 Genotypen)
$m =$ Anzahl Modellparameter (hier m=1, der Parameter p)

im Blutgruppenbeispiel also:

$$\text{df} = 3 - 1 - 1 = 1$$

Der Wert der χ^2-Statistik ist

$$\frac{(1787 - 1782.7)^2}{1782.7} + \frac{(3037 - 3045.6)^2}{3045.6} + \frac{(1305 - 1300.7)^2}{1300.7} = 0.049.$$

Dieser Wert gibt keinen Anlass, an der Nullhypothese „die Population ist bezüglich des M/N-Blutgruppensystems im HW-Gleichgewicht“ zu zweifeln:

0.049 liegt zwischen dem 10%- und dem 30%-Quantil der χ^2-Vert. mit einem Freiheitsgrad, wir könnten also eine solche oder noch größere Abweichung zwischen Beobachtung und Erwartung in ca. 80% der Fälle erwarten (der *p*-Wert ist 0.83).

Simpson-Paradoxon

Durch Zusammenfassen von Gruppen können sich (scheinbare) statistische Trends in ihr Gegenteil verkehren.

Dieses Phänomen heißt Simpson-Paradoxon oder Yule-Simpson-Effekt.

(nach Edward H. Simpson, *1922 und George Udny Yule, 1871–1951)

Simpson-Paradoxon

Beispiel: Zulassungsstatistik der UC Berkeley 1973

Im Herbst 1973 haben sich an der Universität Berkeley 12763 Kandidaten für ein Studium beworben, davon 8442 Männer und 4321 Frauen. Es kam zu folgenden Zulassungszahlen:

	Aufgenommen	Abgelehnt
Männer	3738	4704
Frauen	1494	2827

Demnach betrug die Zulassungsquote
bei den Männern $\frac{3738}{8442} \approx 44\%$, bei den Frauen nur $\frac{1494}{4321} \approx 35\%$.

Ein χ^2-Test auf Homogenität (z.B. mit R) zeigt, dass eine solche Unverhältnismäßigkeit nur mit verschwindend kleiner Wahrscheinlichkeit durch „reinen Zufall“ entsteht:

```
> berkeley <- matrix(c(3738,1494,4704,2827),nrow=2)
> berkeley
     [,1] [,2]
[1,] 3738 4704
[2,] 1494 2827
> chisq.test(berkeley,correct=FALSE)

Pearson's Chi-squared test

data:  berkeley
X-squared = 111.2497, df = 1, p-value < 2.2e-16
```

Dieser Fall hat einiges Aufsehen erregt, s.a. P.J. Bickel, E.A. Hammel, J.W. O'Connell, Sex Bias in Graduate Admissions: Data from Berkeley, *Science*, **187**, no. 4175, 398–404 (1975).

Das Ungleichgewicht verschwindet, wenn man die Zulassungszahlen nach Departments aufspaltet:

Es stellt sich heraus, dass innerhalb der Departments die Aufnahmewahrscheinlichkeiten nicht signifikant vom Geschlecht abhängen, aber sich Frauen häufiger bei Departments mit (absolut) niedriger Aufnahmequote beworben haben als Männer – dies ist ein Beispiel für das *Simpson-Paradox*.

Die genauen nach Departments aufgeschlüsselten Bewerber- und Zulassungszahlen sind leider nicht öffentlich zugänglich (siehe aber Abb. 1 in Bickel et. al, loc. cit., für eine grafische Aufbereitung der Daten, die den Simpson-Effekt zeigt).

Bickel et. al demonstrieren das Phänomen mittels eines hypothetischen Beispiels:

	Aufgenommen	Abgelehnt
	Department of machismathics	
Männer	200	200
Frauen	100	100
	Department of social warfare	
Männer	50	100
Frauen	150	300
	Gesamt	
Männer	250	300
Frauen	250	400