Mathematik für Biologen

Prof. Dr. Rüdiger W. Braun

Heinrich-Heine-Universität Düsseldorf

19. Januar 2011

Ordinalskalierte Daten

- Beispiel: Tomaten sind “delikat”, “schmackhaft”, “genießbar” oder “ekelhaft”.
- Zahlenkodierung

delikat	schmackhaft	genießbar	ekelhaft
1	2	3	4

- Es handelt sich um ein diskretes quantitatives Merkmal.
- Je 12 Tester beurteilen die Qualität von zwei Tomatensorten, um herauszubekommen, ob die Neuzüchtung besser schmeckt.
- Wir machen einen U-Test. Da nur vier Werte möglich sind, sind Bindungen unvermeidlich.

U-Test: Wiederholung

- R_1 sei die Summe sämtlicher Ränge in der ersten Gruppe und R_2 sei die Summe sämtlicher Ränge in der zweiten Gruppe
- Die Zufallsvariable U wird definiert als

$$U = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1 + 1)}{2} - R_1$$

  U gibt die Anzahl der Rangplatzüberschreitungen an
- r_1, r_2, u sind die Realisierungen
- Die Teststatistik ist

$$t = \frac{u - \mu}{\sigma}$$

  wobei

$$\mu = \frac{n_1 \cdot n_2}{2}$$

$$\sigma = \sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}}$$

U-Test, Fortsetzung

- Das Signifikanzniveau sei α
- Entscheidung:

  $H_0 = \{\mu_1 = \mu_2\}$: Die Nullhypothese H_0 wird abgelehnt, wenn
  $$|t| > q_{1-\alpha/2}$$
  $H_0 = \{\mu_1 \leq \mu_2\}$: Die Nullhypothese H_0 wird abgelehnt, wenn
  $$t < -q_{1-\alpha}$$
  $H_0 = \{\mu_1 \geq \mu_2\}$: Die Nullhypothese H_0 wird abgelehnt, wenn
  $$t > q_{1-\alpha}$$

U-Test mit Bindungen

- Wenn mehrere Werte den gleichen Rang aufweisen, spricht man von Bindungen
- Dann wird jedem Messwert als Rangplatz das arithmetische Mittel der fortlaufend vergebenen Rangplätze zugewiesen
- und σ muss durch die korrigierte Größe

$$\sigma_{\text{korr}} = \sqrt{\frac{n_1 \cdot n_2}{12n \cdot (n-1)} \left(n^3 - n - \sum_{k=1}^{K} (t_k^3 - t_k) \right)}$$

  ersetzt werden
- Dabei

$$n = n_1 + n_2$$

K = Anzahl der Messwertgruppen mit gleichen Werten

t_k = Anzahl der Bindungen in der k-ten Gruppe

Beispieldaten

Bewertung der Tomaten

Urteil	Sorte A Anzahl	Sorte B Anzahl
delikat	2	1
schmackhaft	3	7
genießbar	6	4
ekelig	1	0

Kann man zum Signifikanzniveau $\alpha = 0.05$ sagen, dass Sorte B bessere Bewertungen bekommt als Sorte A?

Beispiel, Fortsetzung

Sorte A			Sorte B		
Urteil	Rang	gemittelt	Urteil	Rang	gemittelt
1	1	2	1	2	2
1	3	2	2	4	8.5
2	5	8.5	2	6	8.5
2	7	8.5	2	8	8.5
2	9	8.5	2	10	8.5
3	14	18.5	2	11	8.5
3	15	18.5	2	12	8.5
3	16	18.5	2	13	8.5
3	17	18.5	3	18	18.5
3	19	18.5	3	20	18.5
3	21	18.5	3	22	18.5
4	24	24	3	23	18.5
Rangsumme r_1:		164.5	Rangsumme r_2:		135.5

Beispiel, Fortsetzung

Damit kann man u ausrechnen

$$u = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1 + 1)}{2} - r_1 = 57.5$$

$$\mu = \frac{n_1 \cdot n_2}{2} = 72$$

$$K = 3$$

Die Bindungszahlen sind

$$t_1 = 3 \qquad t_2 = 10 \qquad t_3 = 10$$

Beispiel, Fortsetzung

- Berechnung des korrigierten σ

$$\sigma_{\text{korr}} = \sqrt{\frac{n_1 \cdot n_2}{12 \cdot n \cdot (n-1)} \left(n^3 - n - \sum_{k=1}^{K} (t_k^3 - t_k) \right)}$$

$$= \sqrt{\frac{12 \cdot 12}{12 \cdot 24 \cdot 23} \left(24^3 - 24 - (3^3 - 3 + 10^3 - 10 + 10^3 - 10)\right)}$$

$$= \sqrt{256.4} = 16.01$$

- Das unkorrigierte σ ist $\sqrt{300} = 17.31$
- Damit

$$t = \frac{u - \mu}{\sigma_{\text{korr}}} = \frac{57.5 - 72}{16.01} = -0.906$$

- Das Quantil ist $q_{0.95} = 1.645$. Daher kann die Nullhypothese nicht abgelehnt werden

Test für ein Merkmal mit nur zwei Ausprägungen

- Beispielaufgabe:

  An der HHU sind 60.6% der Studierenden weiblich. Im Fach Biologie sind 530 von 906 Studierenden weiblich. Das sind 58.5%. Ist der Unterschied beim Anteil weiblicher Studierender signifikant?

- Für solche Fragestellungen verwendet man einen Chi-Quadrat-Anpassungstest. Diese Tests dienen zur Überprüfung der Gleichheit zweier Verteilungen.
- Die Beispielaufgabe ist aber untypisch einfach; wir können sie als Binomialtest mit Normalapproximation rechnen.

Beispielaufgabe

- Die Zufallsvariable X ist $B_{906,\,p}$-verteilt mit unbekanntem p
- Die Nullhypothese ist $H_0 = \{p = p_0\}$ für $p_0 = 0.606$
- Wir machen einen zweiseitigen Binomialtest zum Signifikanzniveau $\alpha = 0.05$

Beispiel, Fortsetzung

Der kritischen Wert c_1 und c_2 sind so zu wählen, dass

$$\sum_{k=0}^{c_1-1} \binom{n}{k} \cdot p_0^k \cdot (1-p_0)^{n-k} \leq \frac{\alpha}{2}$$

$$\sum_{k=0}^{c_1} \binom{n}{k} \cdot p_0^k \cdot (1-p_0)^{n-k} > \frac{\alpha}{2}$$

$$\sum_{k=0}^{c_2} \binom{n}{k} \cdot p_0^k \cdot (1-p_0)^{n-k} \geq 1 - \frac{\alpha}{2}$$

$$\sum_{k=0}^{c_2-1} \binom{n}{k} \cdot p_0^k \cdot (1-p_0)^{n-k} < 1 - \frac{\alpha}{2}$$

Beispiel, Fortsetzung

- Wir approximieren die erste und die dritte Formel mit der Normalapproximation

$$P(X \leq b) \cong \Phi\left(\frac{b + \frac{1}{2} - n \cdot p_0}{\sqrt{n \cdot p_0 \cdot (1 - p_0)}}\right)$$

- Hier $n \cdot p_0 = 906 \cdot 0.606 = 549.0$ und $n \cdot p_0 \cdot (1 - p_0) = 216.3$

Beispiel, Fortsetzung

- Also löst c_1 die Gleichung

$$\Phi\left(\frac{c_1 - 1 + 1/2 - 549.0}{\sqrt{216.3}}\right) = \frac{\alpha}{2}$$

- Die Gleichung ist äquivalent zu

$$\Phi\left(\frac{c_1 - 549.5}{14.71}\right) = 0.025$$

- $q_{0.025}$ ist das Quantil der Standardnormalverteilung

$$\frac{c_1 - 549.5}{14.71} = q_{0.025} = -q_{0.975} = -1.960$$

- Daher $c_1 = 549.5 - 14.71 \cdot 1.960 = 520.7$
- Gerundet $c_1 = 521$

Beispiel, Fortsetzung

- Der kritische Bereich ist

$$K_1 = \left\{ (x_1, \ldots, x_{906}) \;\middle|\; \sum_{j=1}^{906} x_j < c_1 \text{ oder } \sum_{j=1}^{906} x_j > c_2 \right\}$$

- $c_1 = 521$ und c_2 ist auf jeden Fall größer als 549.5
- Bei $\sum_{j=1}^{906} x_j = 530$ kann die Nullhypothese nicht abgelehnt werden
- Der Frauenanteil in der Biologie entspricht dem Durchschnitt über alle Studierenden der HHU

Mendelsche Erbregeln

- Bei den Mendelschen Erbversuchen tritt das Merkmal *Blütenfarbe* in drei Ausprägungen auf, nämlich weiß, rosa und rot
- weiß und rot haben dieselbe Wahrscheinlichkeit, rosa die doppelte
- 4 Blüten werden beobachtet, alle sind rosa
- Widerspricht diese Beobachtung den Mendelschen Regeln?

Interpretation als Vergleich zweier Verteilungen

- Modellannahme: Die Mendelschen Regeln gelten für die untersuchte Situation
- Das entspricht der Verteilung

Nummer	Ausprägung	Wahrscheinlichkeit
1	weiß	25%
2	rosa	50%
3	rot	25%

- Zu vergleichen mit der tatsächlichen Verteilung der Blütenfarben in dem Kollektiv
- Der Stichprobenumfang ist 4
- Das ist für praktische Zwecke zu wenig

Mendelsche Erbregeln, Fortsetzung

- Strategie: Ordne die möglichen Ergebnisse mit aufsteigender Wahrscheinlichkeit an
- Der kritische Bereich besteht dann aus den unwahrscheinlichsten Ergebnissen
- Dabei werden aus der Liste die obersten Ereignisse genommen, bis die erlaubte Fehlerwahrscheinlichkeit erster Art ausgeschöpft ist

Test auf Übereinstimmung zweier Verteilungen

- Unabhängige Zufallsvariable $X_1, \ldots, X_n$, die alle mit Wahrscheinlichkeit p_1 den Wert w_1, mit Wahrscheinlichkeit p_2 den Wert w_2, ..., mit Wahrscheinlichkeit p_s den Wert w_s annehmen
- Vergleichswahrscheinlichkeiten $\pi_1, \pi_2, \ldots, \pi_s$ mit $\pi_1 + \pi_2 + \cdots + \pi_s = 1$
- Nullhypothese und Alternative:

$$\begin{aligned} H_0 &: \quad p_1 = \pi_1, p_2 = \pi_2, \ldots, p_s = \pi_s \\ H_1 &: \quad \text{mindestens ein } p_j \neq \pi_j \end{aligned}$$

Test auf Übereinstimmung zweier Verteilungen: Summenvariable

- Summenvariable

$$Y_1 = \text{Anzahl aller } X_j \text{ mit } X_j = w_1$$
$$Y_2 = \text{Anzahl aller } X_j \text{ mit } X_j = w_2$$
$$\vdots$$
$$Y_s = \text{Anzahl aller } X_j \text{ mit } X_j = w_s$$

- Erwartungswerte unter H_0

$$E(Y_1) = n \cdot \pi_1$$
$$E(Y_2) = n \cdot \pi_2$$
$$\vdots$$
$$E(Y_s) = n \cdot \pi_s$$

Test auf Übereinstimmung für kleine Stichproben

- Bestimme für jede mögliche Kombination von Werten von $Y_1, \ldots, Y_s$ deren Wahrscheinlichkeit
- Ordne diese Wahrscheinlichkeiten aufsteigend in einer Liste
- Der kritische Bereich besteht aus den obersten Zeilen dieser Liste
- Man nimmt genau so viele Zeilen, dass die erlaubte Fehlerwahrscheinlichkeit erster Art nicht überschritten, aber möglichst gut ausgeschöpft wird

Beispiel Mendel: Formalisierung

- $s = 3$
- X_1 ist die (der Zahlencode der) Blütenfarbe der ersten Blüte, X_2 dasselbe für die zweite Blüte, ...
- Y_1 bezeichnet die Anzahl der weißen, Y_2 die der rosafarbenen und Y_3 die der roten Blüten
- Dann $Y_1 + Y_2 + Y_3 = 4$
- Im Beispiel $Y_1 = 0,\ Y_2 = 4,\ Y_3 = 0$
- Rechne sämtliche Einzelwahrscheinlichkeiten aus

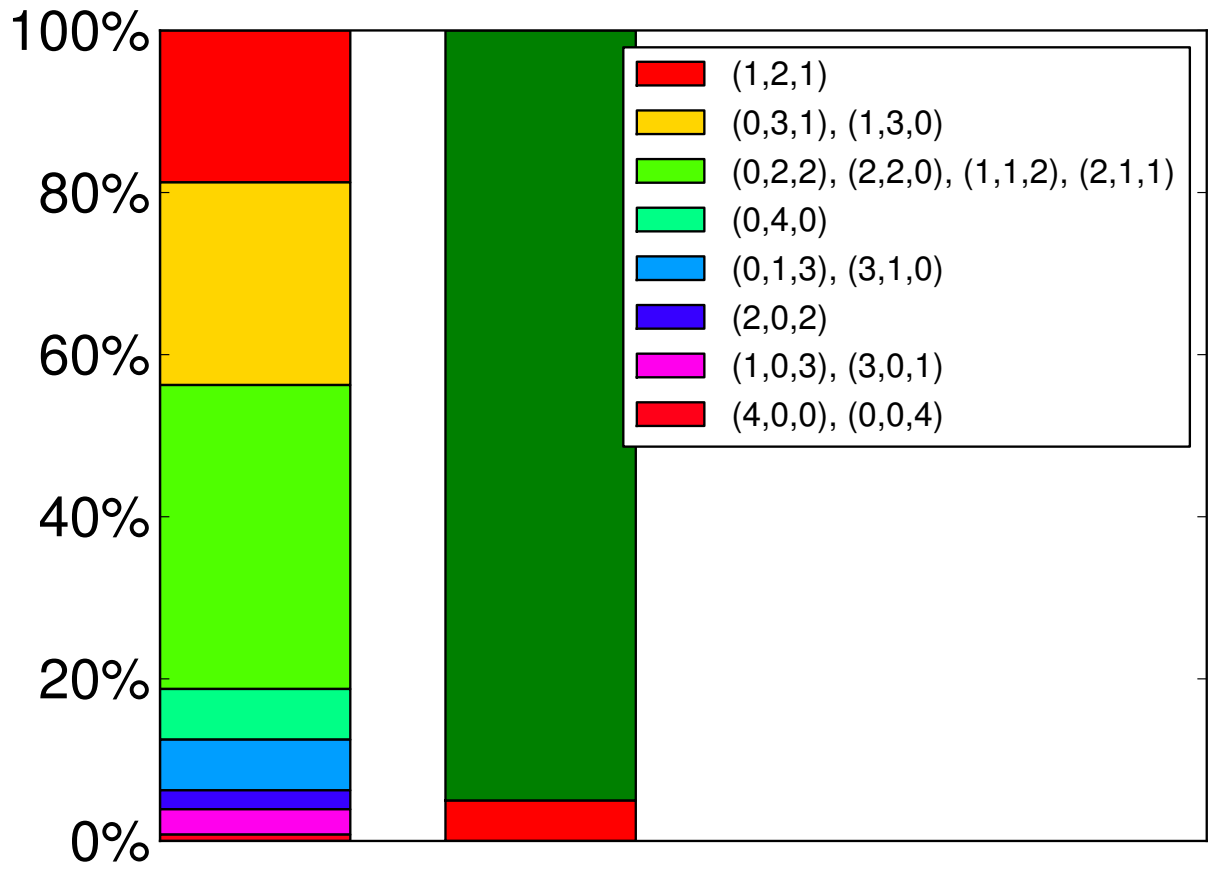
Beispiel Mendel: Wahrscheinlichkeiten der Einzelereignisse

$$
\begin{aligned}
&P(Y_1 = k_1, Y_2 = k_2, Y_3 = k_3) \\
&\qquad = \binom{4}{k_1} \cdot \binom{4-k_1}{k_2} \cdot \left(\frac{1}{4}\right)^{k_1} \cdot \left(\frac{1}{2}\right)^{k_2} \cdot \left(\frac{1}{4}\right)^{k_3} \\
&\qquad = \frac{4! \cdot (4-k_1)!}{k_1! \cdot (4-k_1)! \cdot (4-k_1-k_2)!} \cdot \left(\frac{1}{4}\right)^{k_1} \cdot \left(\frac{1}{2}\right)^{k_2} \cdot \left(\frac{1}{4}\right)^{k_3} \\
&\qquad = \frac{4!}{k_1! \cdot k_2! \cdot k_3!} \cdot \left(\frac{1}{4}\right)^{k_1} \cdot \left(\frac{1}{2}\right)^{k_2} \cdot \left(\frac{1}{4}\right)^{k_3}
\end{aligned}
$$

Beispiel Mendel: Tabelle der W'keiten der Einzelereignisse

k_1	k_2	k_3	$P(X_1 = k_1, X_2 = k_2, X_3 = k_3)$	kumulierte Summe
0	0	4	0.0039	0.0039
4	0	0	0.0039	0.0078
1	0	3	0.0156	0.0234
3	0	1	0.0156	0.0391
2	0	2	0.0234	0.0625
0	1	3	0.0312	0.0938
3	1	0	0.0312	0.1250
0	4	0	0.0625	0.1875
0	2	2	0.0938	0.2812
1	1	2	0.0938	0.3750
2	1	1	0.0938	0.4688
2	2	0	0.0938	0.5625
0	3	1	0.1250	0.6875
1	3	0	0.1250	0.8125
1	2	1	0.1875	1.0000

Beispiel Mendel: Balkendiagramm



Der linke Balken zeigt die kumulierten Werte aus der Tabelle, der rechte die 5%-Schwelle

Beispiel Mendel: Ergebnis

- In den folgenden Fällen kann die Nullhypothese zum Signifikanzniveau $\alpha = 0.05$ abgelehnt werden
  - 4 weiße oder 4 rote Blüten
  - keine rosa, aber 3 weiße oder 3 rote Blüten
- Der p-Wert des beobachteten Ereignisses “4 rosa Blüten” beträgt 18.75%

Große Stichprobenumfänge

- Für große Stichprobenumfänge ist der soeben besprochene Test unpraktikabel
- Ziel: Zur Realisierung eine Teststatistik berechnen und dann mit einem passenden Quantil vergleichen

Teststatistik des Chi-Quadrat-Tests, Fortsetzung

Die Teststatistik misst die Abweichung der Realisierungen $y_1, y_2, \ldots, y_s$ von den Erwartungswerten

$$t = \sum_{j=1}^{s} \frac{(y_j - n \cdot \pi_j)^2}{n \cdot \pi_j}$$

Große Werte von t sprechen gegen H_0

Teststatistik für Beispiel Mendel

k_1	k_2	k_3	$P(X_1 = k_1, X_2 = k_2, X_3 = k_3)$	t
0	0	4	0.0039	12.0000
4	0	0	0.0039	12.0000
1	0	3	0.0156	6.0000
3	0	1	0.0156	6.0000
2	0	2	0.0234	4.0000
0	1	3	0.0312	5.5000
3	1	0	0.0312	5.5000
0	4	0	0.0625	4.0000
0	2	2	0.0938	2.0000
1	1	2	0.0938	1.5000
2	1	1	0.0938	1.5000
2	2	0	0.0938	2.0000
0	3	1	0.1250	1.5000
1	3	0	0.1250	1.5000
1	2	1	0.1875	0.0000

χ^2-Verteilung

- Die Quantile der χ^2-Verteilung sind die Referenzgröße beim Vergleich zweier Verteilungen für große Stichprobenumfänge
- Sprich: “Chi-Quadrat”
- Die χ^2-Verteilung mit n Freiheitsgraden besitzt die Dichte

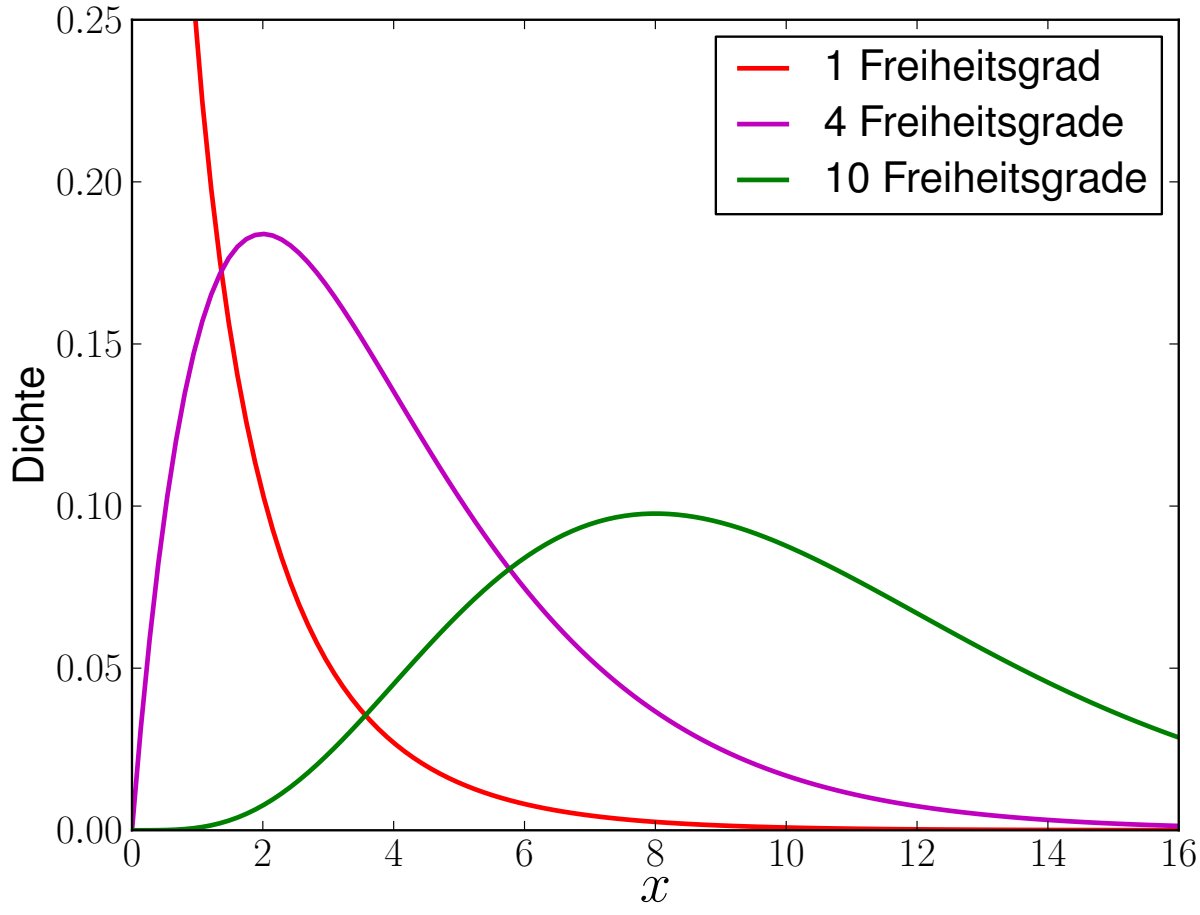
$$f_n(x) = c_n \cdot e^{-x/2} \cdot x^{n/2-1}, \qquad x > 0$$

- Dabei ist c_n bestimmt durch das Erfordenis, dass

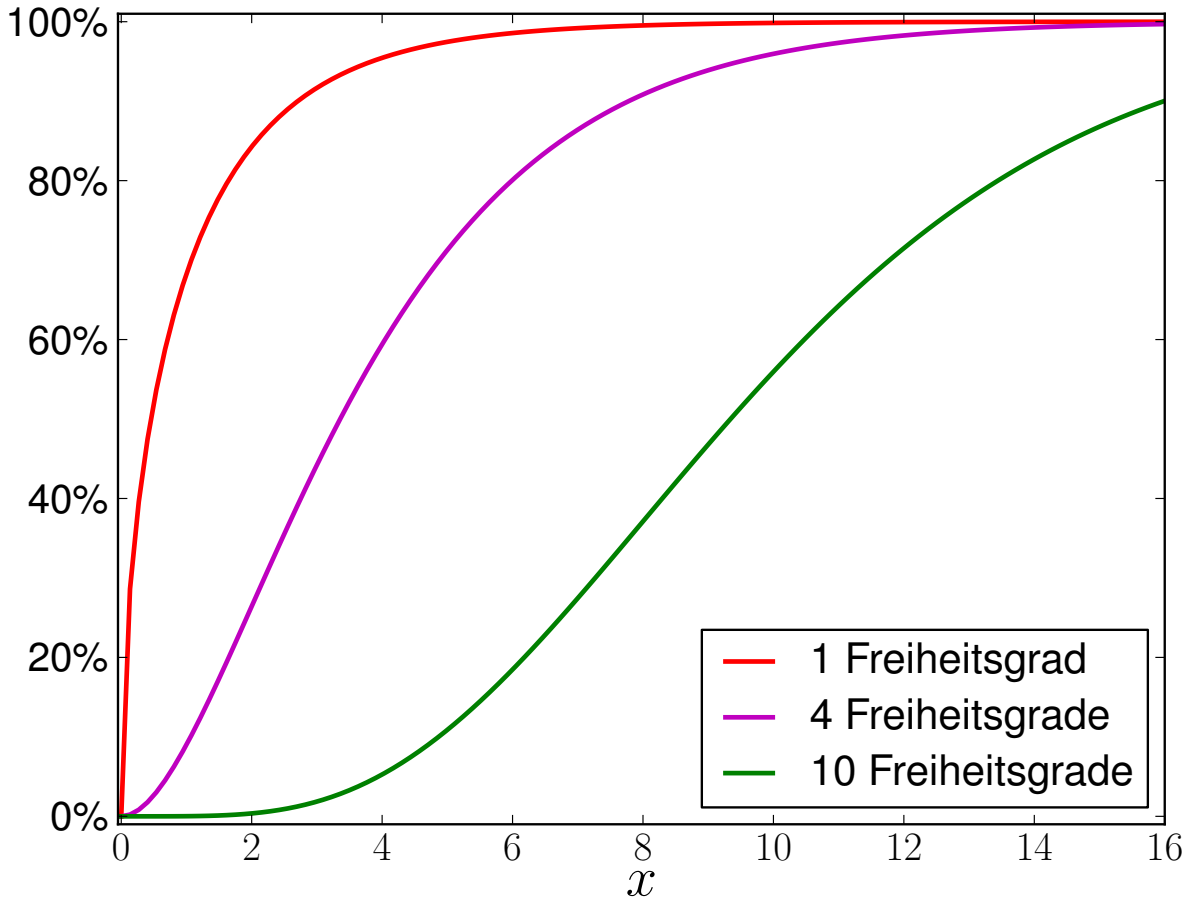
$$\int_0^\infty f_n(x)dx = 1$$

- Die Quantile der χ^2-Verteilung sind tabelliert

Graphen von Dichten von χ^2-Verteilungen



Graphen von Verteilungsfunktionen von χ^2-Verteilungen



Quantile der χ^2-Verteilung

f	90%	95%	97.5%	99%	99.5%	99.9%
1	2.71	3.84	5.02	6.63	7.88	10.83
2	4.61	5.99	7.38	9.21	10.60	13.82
3	6.25	7.81	9.35	11.34	12.84	16.27
4	7.78	9.49	11.14	13.28	14.86	18.47
5	9.24	11.07	12.83	15.09	16.75	20.52
6	10.64	12.59	14.45	16.81	18.55	22.46
7	12.02	14.07	16.01	18.48	20.28	24.32
8	13.36	15.51	17.53	20.09	21.95	26.12
9	14.68	16.92	19.02	21.67	23.59	27.88
10	15.99	18.31	20.48	23.21	25.19	29.59
11	17.28	19.68	21.92	24.72	26.76	31.26
12	18.55	21.03	23.34	26.22	28.30	32.91
13	19.81	22.36	24.74	27.69	29.82	34.53
14	21.06	23.68	26.12	29.14	31.32	36.12
15	22.31	25.00	27.49	30.58	32.80	37.70
16	23.54	26.30	28.85	32.00	34.27	39.25
17	24.77	27.59	30.19	33.41	35.72	40.79
18	25.99	28.87	31.53	34.81	37.16	42.31
19	27.20	30.14	32.85	36.19	38.58	43.82
20	28.41	31.41	34.17	37.57	40.00	45.31

Der Chi-Quadrat-Test zum Vergleich zweier Verteilungen

- Gegeben ein Signifikanzniveau α
- Berechne Teststatistik

$$t = \sum_{j=1}^{s} \frac{(y_j - n \cdot \pi_j)^2}{n \cdot \pi_j}$$

- Bestimme Quantil $\chi^2_{s-1,1-\alpha}$ der χ^2-Verteilung mit $s-1$ Freiheitsgraden
- Falls

$$t \geq \chi^2_{s-1,1-\alpha}$$

dann lehne H_0 ab

Bemerkungen zum χ^2-Test

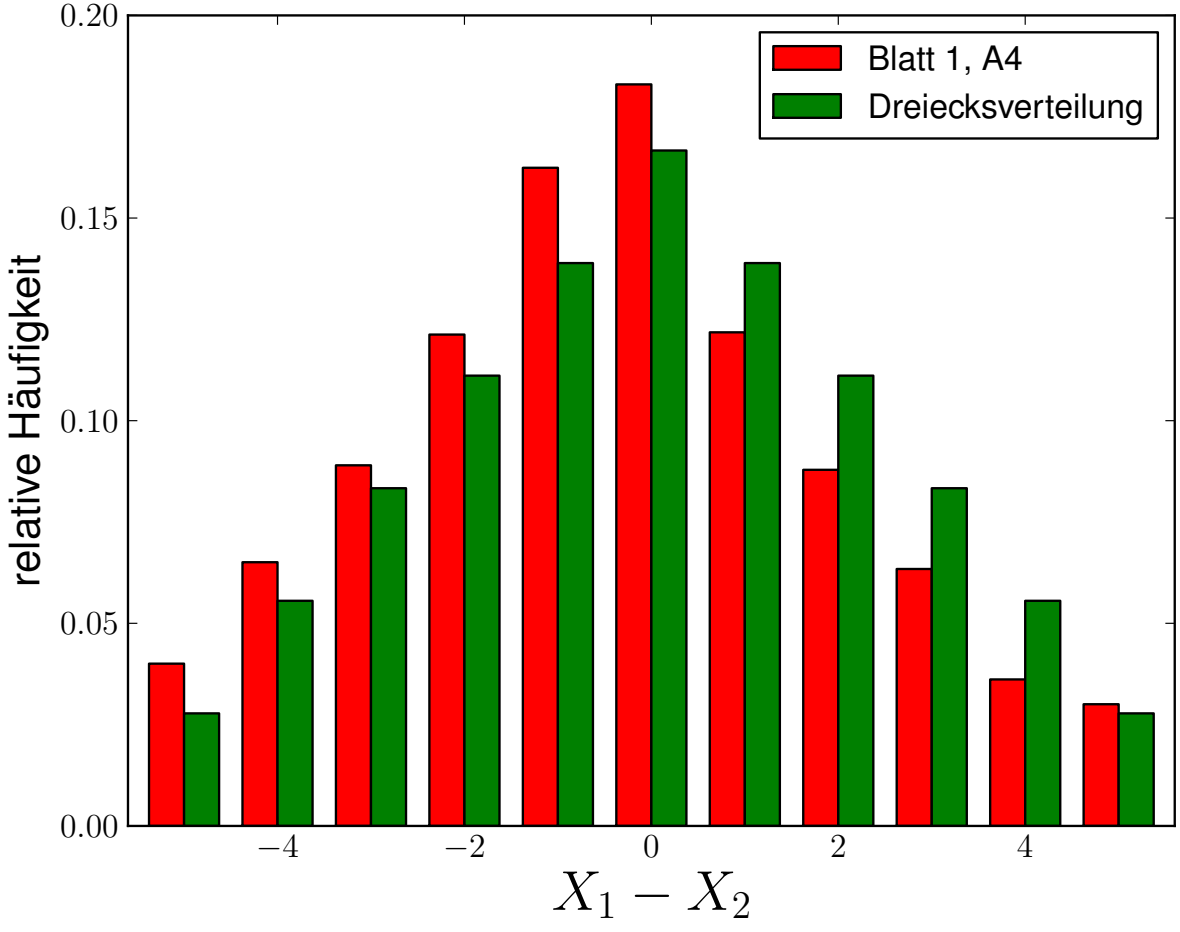
- Der χ^2-Test verwendet eine Approximation
- Er ist daher nur zulässig, wenn

$$
\begin{aligned}
n \cdot \pi_1 &\geq 5 \\
n \cdot \pi_2 &\geq 5 \\
&\vdots \\
n \cdot \pi_s &\geq 5
\end{aligned}
$$

- Die Zahl der Freiheitgrade beträgt $s - 1$

Beispiel zum χ^2-Test

Würfelexperiment aus Aufgabe 1 Blatt 4



Beispiel zum χ^2-Test, Fortsetzung

- Die Tabelle zeigt die empirische Häufigkeitsverteilung von $X_1 - X_2$
- Wir vergleichen mit der Dreiecksverteilung

j	-5	-4	-3	-2	-1	0
p_j	0.040	0.065	0.089	0.121	0.162	0.183
π_j	0.028	0.056	0.083	0.111	0.139	0.167

j	5	4	3	2	1
p_j	0.030	0.036	0.063	0.088	0.122
π_j	0.028	0.056	0.083	0.111	0.139

Beispiel zum χ^2-Test, Fortsetzung

- Ziel: Widerlege die Nullhypothese, dass die Daten gemäß der Dreiecksverteilung verteilt sind, zum Signifikanzniveau $\alpha = 0.1\%$
- Stichprobenumfang $n = 1798$

j	-5	-4	-3	-2	-1	0
y_j	72	117	160	218	292	329
$n \cdot \pi_j$	49.9	99.9	149.8	199.8	249.7	299.7

j	5	4	3	2	1
y_j	54	65	114	158	219
$n \cdot \pi_j$	49.9	99.9	149.8	199.8	249.7

- $s = 11$
- Der kleinste Wert von $n \cdot \pi_j$ ist 49.9
- Daher ist der χ^2-Test zulässig

Beispiel zum χ^2-Test, Fortsetzung

$$t = \sum_{j=1}^{s} \frac{(y_j - n \cdot \pi_j)^2}{n \cdot \pi_j} = 58.65$$

- Es gibt 10 Freiheitsgrade
- Das ▸ Quantil ist
$$\chi^2_{10,\,0.999} = 29.59$$
- Die Nullhypothese kann abgelehnt werden
- Die experimentell ermittelte Verteilung ist nicht die Dreiecksverteilung