Statistische Tests: Beispiele und Berichte

3./5 Februar 2014

ein Stichproben-t-Test: Modell

Annahme: n u.i.v. Beobachtungen $X_1, \ldots, X_n$, $X_i \sim \mathcal{N}_{\mu,\sigma^2}$ mit unbekanntem $\mu \in \mathbb{R}$ und unbekanntem $\sigma^2 > 0$
$M := \frac{1}{n}\sum_{i=1}^{n} X_i$, $S^2 := \frac{1}{n-1}\sum_{i=1}^{n-1}(X_i - M)^2$
Wir haben haben gesehen:
Für jedes $\mu \in \mathbb{R}$, $\sigma^2 > 0$ ist unter $P_{(\mu_0,\sigma^2)}$

$$T := \sqrt{n}\frac{M - \mu_0}{\sqrt{S^2}} \sim \text{Student-}(n-1)$$

Signifikanzniveau $\alpha \in (0, 1)$:
Zweiseitiger Test von $H_0 : \{\mu = \mu_0\}$ gegen $H_1 : \{\mu \neq \mu_0\}$:
Lehne H_0 ab, wenn $|T| > q_{n-1,1-\alpha/2}$, wobei $q_{n-1,1-\alpha/2} =$ $(1 - \alpha/2)$-Quantil der Student-$(n-1)$-Vert.

Einseitiger Test von $H_0 : \{\mu \leq \mu_0\}$ gegen $H_1 : \{\mu > \mu_0\}$:
Lehne H_0 ab, wenn $T > q_{n-1,1-\alpha}$, wobei $q_{n-1,1-\alpha} =$ $(1 - \alpha)$-Quantil der Student-$(n-1)$-Vert.

Beispiel:
Die Wirksamkeit eines gewissen Schlafmittels soll geprüft werden.

10 Patienten erhalten das Schlafmittel, die Anzahl zusätzlicher Stunden Schlaf wird in einer Nacht beobachtet.

Wir nehmen an, die Beob. sind u.i.v. $\sim \mathcal{N}_{\mu,\sigma^2}$ und wir möchten die Nullhypothese $\mu = 0$, sagen wir, zum Niveau $\alpha = 0.05$ testen.

Die Daten*

Patient i	1	2	3	4	5	6	7	8	9	10
zus. Schl.	0.7	-1.6	-0.2	-1.2	-0.1	3.4	3.7	0.8	0.0	2.0

Es ist $n = 10$, $\overline{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 0.75$,
$s^2 = \frac{1}{9}\sum_{i=1}^{10}(x_i - \overline{x})^2 \approx 1.79$, $t = \frac{\overline{x}-0}{s/\sqrt{10}} \approx 1.326$

Das 0.975-Quantil der Student-9-Verteilung ist ≈ 2.262, demnach können wir die Nullhypothese nicht ablehnen.

(Für ein Student-9-verteiltes T ist $P(|T| \geq 1.326) \approx 0.2176$, dies ist der p-Wert des Tests.)

---

* Aus Student (William S. Gosset), The Probable Error of a Mean, Biometrika 6:1–25 (1908)

Man kann unseren Befund folgendermaßen formulieren:

„Die Beobachtungen sind mit der Nullhypothese $\mu = 0$ (im statistischen Sinne) verträglich.“

oder

„Die beobachtete Abweichung $\overline{x} = 0.75$ ist nicht signifikant von 0 verschieden (t-Test, $\alpha = 0.05$).“

Das Beispiel in R:

```
> schlaf <- c(0.7, -1.6, -0.2, -1.2, -0.1, 3.4,
              3.7, 0.8, 0.0, 2.0)
> t.test(schlaf)

One Sample t-test

data:  schlaf
t = 1.3257, df = 9, p-value = 0.2176
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.5297804  2.0297804
sample estimates:
mean of x
     0.75
```

Beispiel:
Die Wirksamkeit eines Schlafmittels soll mit der eines anderen verglichen werden

10 Patienten erhalten Schlafmittel *A*, die Anzahl zusätzlicher Stunden Schlaf wird in einer Nacht beobachtet.
Dann erhalten dieselben 10 Patienten Schlafmittel *B*, wieder wird die Anzahl zusätzlicher Stunden Schlaf wird in einer Nacht beobachtet.

Da dieselben Patienten untersucht werden, können (und sollten) wir die Messungen paaren:
Wir interessieren uns bei jedem Patienten für die Differenz des (zusätzlichen) Schlafs bei Mittel 2 und bei Mittel 1.

Wir nehmen an, die beobachteten Differenzen sind Realisierungen von u.i.v. ZVn mit Vert. $\mathcal{N}_{\mu,\sigma^2}$ und wir möchten die Nullhypothese $\mu \leq 0$ gegen die Alternative $\mu > 0$, sagen wir, zum Niveau $\alpha = 0.05$ testen.

Dies wäre beispielsweise in folgender Situation angemessen: Wir möchten darlegen, dass Mittel *B* wirksamer ist als Mittel *A*, indem wir die Nullhypothese „$\mu \leq 0$“ entkräften.

Die Daten*

Patient *i*	1	2	3	4	5	6	7	8	9	10
Mittel *A*	0.7	-1.6	-0.2	-1.2	-0.1	3.4	3.7	0.8	0.0	2.0
Mittel *B*	1.9	0.8	1.1	0.1	-0.1	4.4	5.5	1.6	4.6	3.4
Diff.	1.2	2.4	1.3	1.3	0.0	1.0	1.8	0.8	4.6	1.4

* Aus Student (William S. Gosset), The Probable Error of a Mean, Biometrika 6:1–25 (1908)

Patient i	1	2	3	4	5	6	7	8	9	10
Diff.	1.2	2.4	1.3	1.3	0.0	1.0	1.8	0.8	4.6	1.4

Es ist $n = 10$, $\overline{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 1.58$,
$s^2 = \frac{1}{9}\sum_{i=1}^{10}(x_i - \overline{x})^2 \approx 1.23$, $t = \frac{\overline{x}-0}{s/\sqrt{10}} \approx 4.062$

Das 0.95-Quantil der Student-9-Verteilung ist ≈ 1.833, demnach können wir die Nullhypothese ablehnen.

(Für ein Student-9-verteiltes T ist $P(T > 4.062) \approx 0.0014$, dies ist der p-Wert des Tests.)

Mögliche knappe Formulierung:

„Die beobachtete Differenz $\overline{x} = 1.58$ ist signifikant größer als 0 (einseitiger t-Test, $\alpha = 0.05$).“

Das Beispiel in R:

```
> diff <- c(1.2, 2.4, 1.3, 1.3, 0.0, 1.0, 1.8,
            0.8, 4.6, 1.4)
> t.test(diff, alternative="greater")

One Sample t-test

data:  diff
t = 4.0621, df = 9, p-value = 0.001416
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 0.8669947       Inf
sample estimates:
mean of x
     1.58
```

ungepaarter t-Test: Modell

Annahme: m u.i.v. Beobachtungen $X_1, \ldots, X_m$ und davon unabhängig n u.i.v. Beobachtungen $Y_1, \ldots, Y_n$, unter P_ϑ
$X_i \sim \mathcal{N}_{\mu_1,\sigma^2}$, $Y_j \sim \mathcal{N}_{\mu_2,\sigma^2}$ mit $\vartheta = (\mu_1, \mu_2, \sigma^2) \in \mathbb{R} \times \mathbb{R} \times (0, \infty)$

Seien

$$M_X := \frac{1}{m}\sum_{i=1}^{m} X_i, \; M_Y := \frac{1}{n}\sum_{j=1}^{n} Y_j$$

$$S_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - M_X)^2, \; S_Y^2 = \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - M_Y)^2,$$

$$S^2 := \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2} \left(= \frac{1}{m+n-2}\left(\sum_{i=1}^{n}(X_i - M_X)^2 + \sum_{j=1}^{n}(Y_j - M_Y)^2 \right) \right),$$

(bem.: $\mathbb{E}_{(\mu_1,\mu_2,\sigma^2)}[S^2] = \sigma^2$)

$$T = \frac{M_X - M_Y}{S\sqrt{\frac{1}{m} + \frac{1}{n}}}.$$

Für $\mu_0 \in \mathbb{R}$, $\sigma^2 > 0$ ist unter $P_{(\mu_0,\mu_0,\sigma^2)}$

$$T = \frac{M_X - M_Y}{S\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim \text{Student-}(m+n-2)$$

Signifikanzniveau $\alpha \in (0,1)$:

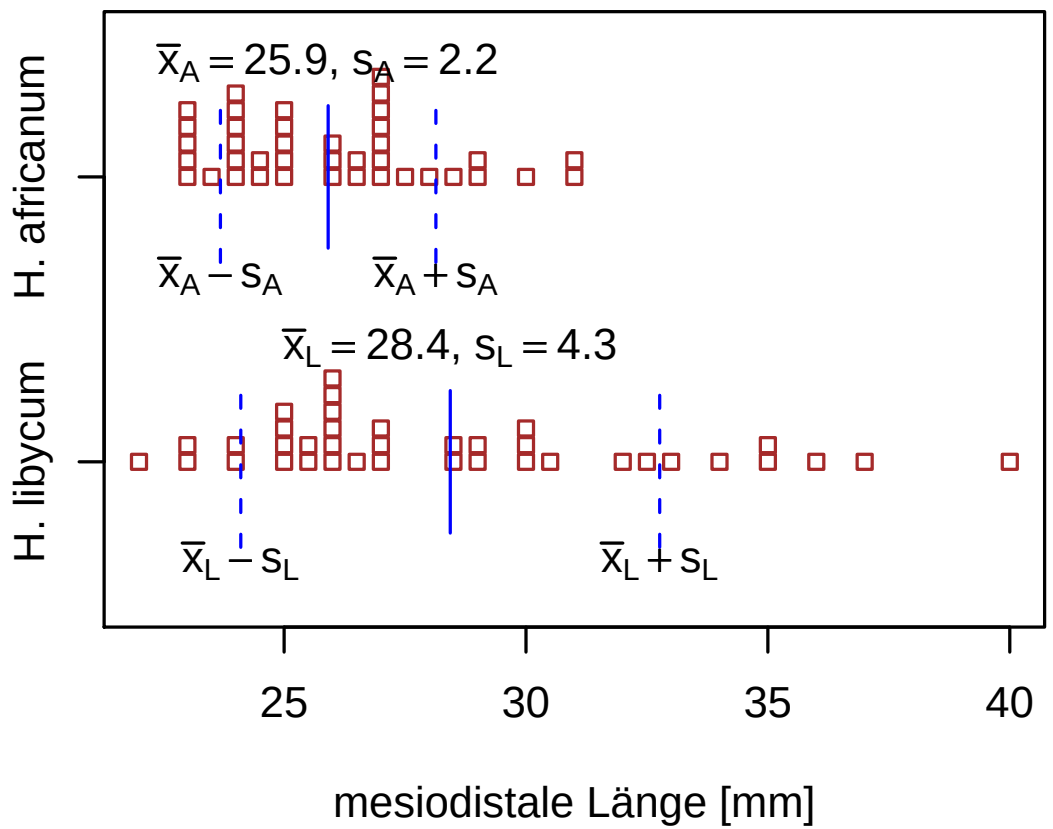
Zweiseitiger Test von $H_0 : \{\mu_1 = \mu_2\}$ gegen $H_1 : \{\mu_1 \neq \mu_2\}$:
Lehne H_0 ab, wenn $|T| > q_{m+n-2,1-\alpha/2}$, wobei $q_{m+n-2,1-\alpha/2} =$ $(1-\alpha/2)$-Quantil der Student-$(n-1)$-Vert.

Einseitiger Test von $H_0 : \{\mu_1 \leq \mu_2\}$ gegen $H_1 : \{\mu_1 > \mu_2\}$:
Lehne H_0 ab, wenn $T > q_{m+n-2,1-\alpha}$, wobei $q_{m+n-2,1-\alpha} =$ $(1-\alpha)$-Quantil der Student-$(n-1)$-Vert.

Beispiel: Es wurden fossile Backenzähne gefunden, die zwei Arten von Urpferden zugeordnet wurden, und jeweils die („mesiodistale“) Länge bestimmt.

Wir möchten die (Null-)Hypothese prüfen, ob die mittlere Zahnlänge bei den beiden Arten gleich ist.

Die Daten



Hipparion africanum

$n_A = 39$

$\bar{x}_A = 25.9$

$s_A = 2.2$

Hipparion lybicum

$n_L = 38$

$\bar{x}_L = 28.4$

$s_L = 4.3$

Wir verwenden Signifikanzniveau $\alpha = 0.01$, das 99,5%-Quantil der Student-Vert. mit 75 Freiheitsgraden ist ≈ 2.64.
Es ist

$$s^2 = \frac{(n_A - 1)s_A^2 + (n_L - 1)s_L^2}{n_A + n_L - 1} \approx 11.94, \quad t = \frac{\overline{x}_A - \overline{x}_L}{s\sqrt{\frac{1}{n_A} + \frac{1}{n_L}}} \approx -3.229,$$

Wir können die Nullhypothese „die mittlere mesiodistale Länge bei *H. lybicum* und bei *H. africanum* sind gleich“ zum Signifikanzniveau 1% ablehnen.

(Für ein Student-75-verteiltes T ist $P(|T| > 3.229) \approx 0.0018$, dies ist der p-Wert des Tests.)

Mögliche Formulierung unseres Befunds:

„Die mittlere mesiodistale Länge war signifikant größer (28.4 mm) bei *H. libycum* als bei *H. africanum* (25.9 mm) (t-Test, $\alpha = 0{,}01$).“

Das Beispiel in R:

```
> t.test(md[Art=="africanum"],md[Art=="libycum"],
         var.equal=TRUE)

Two Sample t-test

data:  md[Art == "africanum"] and md[Art == "libycum"]
t = -3.2289, df = 75, p-value = 0.001845
alternative hypothesis:
true difference in means is not equal to 0
95 percent confidence interval:
 -4.0811448 -0.9667634
sample estimates:
mean of x mean of y
 25.91026  28.43421
```

Betrachten wir (spaßeshalber) nochmal die Schlafmittel-Daten, diesmal ungepaart:

```
> medA <- c(0.7,-1.6,-0.2,-1.2,-0.1,3.4,3.7,0.8,0.0,2.0)
> medB <- c(1.9,0.8,1.1,0.1,-0.1,4.4,5.5,1.6,4.6,3.4)

> t.test(medB, medA, var.equal=TRUE)

Two Sample t-test

data:  medB and medA
t = 1.8608, df = 18, p-value = 0.07919
alternative hypothesis: true difference in means is not e
95 percent confidence interval:
 -0.203874  3.363874
sample estimates:
mean of x mean of y
     2.33      0.75
```

t-Statistik ohne Annahme gleicher Varianz

Es gibt auch eine Version des zwei-Stichproben-t-Tests, der die Annahme gleicher Varianzen nicht trifft (wir werden ihn im Verlauf der Vorlesung allerdings nicht verwenden):

Wäre eine beobachtete Abweichung $\overline{x} - \overline{y}$ mit der Nullhypothese verträglich, dass $\mu_X = \mu_Y$?

Wir schätzen die Streuung von $M_X - M_Y$ durch

$$\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}} \quad \text{und bilden} \quad T = \frac{M_X - M_Y}{\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}}.$$

Unter $P_{(\mu_0,\mu_0,\sigma_1^2,\sigma_2^2)}$ ist T „approximativ Student-verteilt mit g Freiheitsgraden“

wobei g aus den Daten geschätzt wird, $g = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\frac{s_X^4}{n_X^2(n_X-1)} + \frac{s_Y^4}{n_Y^2(n_Y-1)}}$

Welchs* zwei Stichproben-t-Test

$$T = \frac{M_X - M_Y}{\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}}, \qquad g = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\frac{s_X^4}{n_X^2(n_X-1)} + \frac{s_Y^4}{n_Y^2(n_Y-1)}}$$

Man verwirft die Nullhypothese „$\mu_1 = \mu_2$“ (zum Niveau α), wenn

$$\texttt{pt}(|t|, \texttt{df=}g, \texttt{lower.tail=FALSE}) \leq \alpha/2$$

ist, d.h. wenn die Wahrscheinlichkeit, dass eine Student-verteilte Zufallsgröße mit g Freiheitsgraden einen betragsmäßig mindestens so großen Wert wie den beobachteten t-Wert annimmt, $\leq \alpha$ ist. (Analoges Vorgehen für einseitige Tests)

---

*B. L. Welch, The Significance of the Difference between Two Means When the Population Variances Are Unequal, Biometrika 29:350–362, (1938)

Zwei-Stichproben-*t*-Test mit R:

```
> A <- md[Art=="africanum"]
> L <- md[Art=="libycum"]
> t.test(L,A)

Welch Two Sample t-test

data:  L and A
t = 3.2043, df = 54.975, p-value = 0.002255
alternative hypothesis: true difference in means
is not equal to 0
95 percent confidence interval:
 0.9453745 4.1025338
sample estimates:
mean of x mean of y
 28.43421  25.91026
```

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
h_i	13	12	20	18	26	31

```
> 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 $n\rho_i \geq 5$ erfüllen.

Lassen wir für das 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
```

χ^2-Test auf Homogenität/Unabhängigkeit: Situation

In einem Experiment werden zwei „Merkmale“ beobachtet, wobei das erste Merkmal a und das zweite Merkmal b viele Ausprägungen besitzt (also insgesamt $s = a \cdot b$ mögliche Ausgänge).

Unter n u.a. Wiederholungen werde h_{ij} mal Ausgang (i, j) beobachtet ($i \in \{1, 2, \ldots, a\}$, $j \in \{1, 2, \ldots, b\}$), man fasst die Beobachtungen in einer $a \times b$-Kontingenztafel zusammen:

i \ j	1	2	3	
1	h_{11}	h_{12}	h_{13}	$h_{1\cdot}$
2	h_{21}	h_{22}	h_{23}	$h_{2\cdot}$
	$h_{\cdot 1}$	$h_{\cdot 2}$	$h_{\cdot 3}$	$h_{\cdot\cdot} = n$

mit Zeilensummen $h_{i\cdot} = \sum_{j=1}^{b} h_{ij}$,
Spaltensummen $h_{\cdot j} = \sum_{i=1}^{a} h_{ij}$
und Gesamtsumme $h_{\cdot\cdot} = \sum_{i=1}^{a} \sum_{j=1}^{b} h_{ij} = n$

i \ j	1	2	3	
1	h_{11}	h_{12}	h_{13}	$h_{1\cdot}$
2	h_{21}	h_{22}	h_{23}	$h_{2\cdot}$
	$h_{\cdot 1}$	$h_{\cdot 2}$	$h_{\cdot 3}$	$h_{\cdot\cdot} = n$

Wir fassen die beobachteten Häufigkeiten als Realisierungen einer multinomial$(n, (\vartheta_{ij})_{i=1,\ldots,a;j=1,\ldots,b})$-verteilten ZV $(H_{ij})_{i=1,\ldots,a;j=1,\ldots,b}$ auf, wobei

$(\vartheta_{ij})_{i=1,\ldots,a;j=1,\ldots,b}$ ein $a \cdot b$-dimensionaler Vektor von Wahrscheinlichkeitsgewichten ist.

Passen die Beobachtungen zur Nullhypothese, dass

$$\vartheta_{ij} = \eta_i \cdot \rho_j, \quad \text{für } i = 1, \ldots, a,\ j = 1, \ldots, b$$

mit $(\eta_i)_{i=1,\ldots,a}$, $(\rho_j)_{j=1,\ldots,b}$ gewissen a- bzw. b-dimensionalen Vektoren von Wahrscheinlichkeitsgewichten?

Wir bilden

$$\widehat{\vartheta}_{i\cdot} = \frac{H_{i\cdot}}{n}, \quad \widehat{\vartheta}_{\cdot j} = \frac{H_{\cdot j}}{n}$$

und die Teststatistik

$$D = \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{(H_{ij} - n\widehat{\vartheta}_{i\cdot}\widehat{\vartheta}_{\cdot j})^2}{n\widehat{\vartheta}_{i\cdot}\widehat{\vartheta}_{\cdot j}}$$

Unter H_0 : „$(\vartheta_{ij})_{i=1,\ldots,a;j=1,\ldots,b}$ hat Produktform“
ist D (approximativ) $\chi^2_{(a-1)(b-1)}$-verteilt.

Wir würden also H_0 zum Niveau α ablehnen, falls der beobachtete Wert größer ist als das $(1-\alpha)$-Quantil der χ^2-Verteilung mit $(a-1)(b-1)$ Freiheitsgraden.

Beispiel

Der Kuhstärling ist ein Brutparasit des Oropendola.

photo (c) by J. Oldenettel

- 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.
- Mögliche Erklärung: Nester mit Kuhstärling-Eiern sind eventuell besser vor Befall durch Dasselfliegenlarven geschützt.

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

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

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 29.5$$

- 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 a Zeilen und b Spalten:

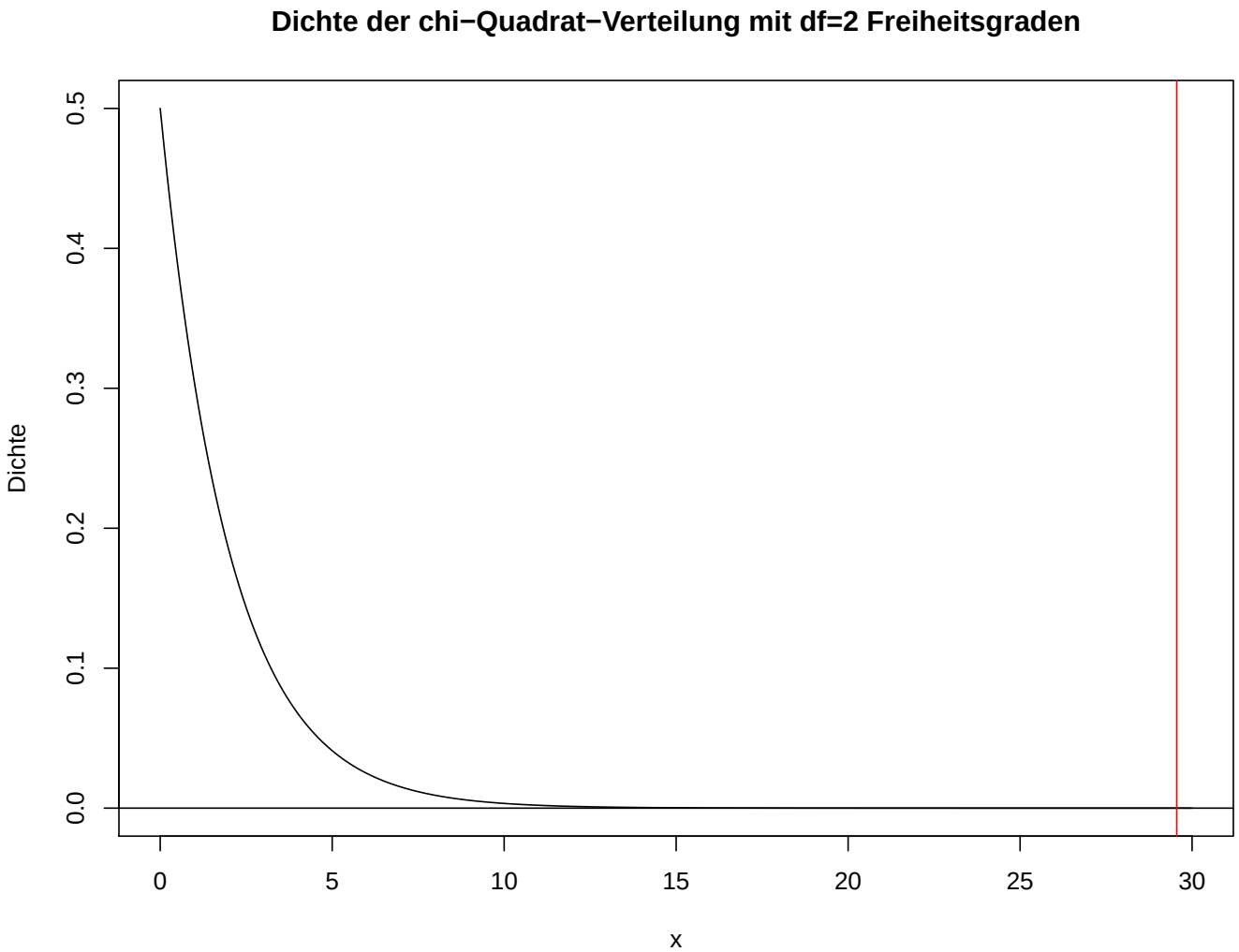
$$df = (a-1) \cdot (b-1)$$

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.)

(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 $\mathbb{P}(X \geq 29.6) = 3.74 \cdot 10^{-7}$ (was hier wörtlich der p-Wert des χ^2-Tests auf Unabhängigkeit wäre, in dieser Genauigkeit für statistische Zwecke allerdings sinnlos ist).

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** funktionierte 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.

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

Wilcoxons Rangsummen-Test

Ein „verteilungsfreier“ Test,
mit dem man die Lage zweier Verteilungen
zueinander testen kann.

Beobachtungen: Zwei Stichproben

$X : x_1, x_2, \ldots, x_m$ und $Y : y_1, y_2, \ldots, y_n$

Wir möchten die Nullhypothese:
X und Y haben diesselbe Verteilung
testen

gegen die Alternative:
Die beiden Verteilungen sind gegeneinander verschoben.

(Die Situation ist ähnlich zum zwei-Stichproben-t-Test, aber wir möchten *nicht* die implizite Annahme treffen, dass es sich dabei (wenigstens ungefähr) um Normalverteilungen handelt.)

Idee

Beobachtungen:

$$X : x_1, x_2, \ldots, x_m \quad \text{und} \quad Y : y_1, y_2, \ldots, y_n$$

- Sortiere alle Beobachtungen der Größe nach.
- Bestimme die Ränge der m X-Werte unter allen $m + n$ Beobachtungen.
- Wenn die Nullhypothese zutrifft, sind die m X-Ränge eine rein zufällige Wahl aus $\{1, 2, \ldots, m + n\}$.
- Berechne die Summe der X-Ränge, prüfe, ob dieser Wert untypisch groß oder klein.

Wilcoxons Rangsummenstatistik

Beobachtungen:

$X : x_1, x_2, \ldots, x_m$

$Y : y_1, y_2, \ldots, y_n$

$W = \text{Summe der } X\text{-Ränge} - (1 + 2 + \cdots + m)$

heißt

Wilcoxons Rangsummenstatistik

Die Normierung ist so gewählt, dass $0 \leq W \leq m\,n$.

Wilcoxons Rangsummenstatistik

Bemerkung 1:

$$W = \text{Summe der } X\text{-Ränge} - (1 + 2 + \cdots + m)$$

Wir könnten auch die Summe der Y-Ränge benutzen, denn

$$\begin{aligned} &\text{Summe der } X\text{-Ränge} + \text{Summe der } Y\text{-Ränge} \\ = \;& \text{Summe aller Ränge} \\ = \;& 1 + 2 + \cdots + (m+n) = \frac{(m+n)(m+n+1)}{2} \end{aligned}$$

Bemerkung 2:

Der Wilcoxon-Test heißt auch Mann-Whitney-Test, die Rangsummenstatistik auch Mann-Whitney Statistik U, sie unterscheidet sich (je nach Definition) von W um eine Konstante.

(In der Literatur sind beide Bezeichnungen üblich, man prüfe vor Verwendung von Tabellen, etc. die verwendete Konvention.)

Ein kleines Beispiel

- Beobachtungen:

$$X : 1{,}5;\ 5{,}6;\ 35{,}2$$
$$Y : 7{,}9;\ 38{,}1;\ 41{,}0;\ 56{,}7;\ 112{,}1;\ 197{,}4;\ 381{,}8$$

- Lege Beobachtungen zusammen und sortiere:
  1,5; 5,6; 7,9; 35,2; 38,1; 41,0; 56,7; 112,1; 197,4; 381,8
- Bestimme Ränge:
  1, 2, 3, 4, 5, 6, 7, 8, 9, 10
- Rangsummenstatistik hier: $W = 1 + 2 + 4 - (1 + 2 + 3) = 1$

Interpretation von W

X-Population kleiner $\Longrightarrow$ W klein:

$1, 2, 3, 4, 5, 6, 7, 8, 9, 10 \quad W = 0$

$1, 2, 3, 4, 5, 6, 7, 8, 9, 10 \quad W = 1$

$1, 2, 3, 4, 5, 6, 7, 8, 9, 10 \quad W = 2$

$1, 2, 3, 4, 5, 6, 7, 8, 9, 10 \quad W = 2$

X-Population größer $\Longrightarrow$ W groß:

$1, 2, 3, 4, 5, 6, 7, 8, 9, 10 \quad W = 21$

$1, 2, 3, 4, 5, 6, 7, 8, 9, 10 \quad W = 20$

$1, 2, 3, 4, 5, 6, 7, 8, 9, 10 \quad W = 19$

$1, 2, 3, 4, 5, 6, 7, 8, 9, 10 \quad W = 19$

Signifikanz

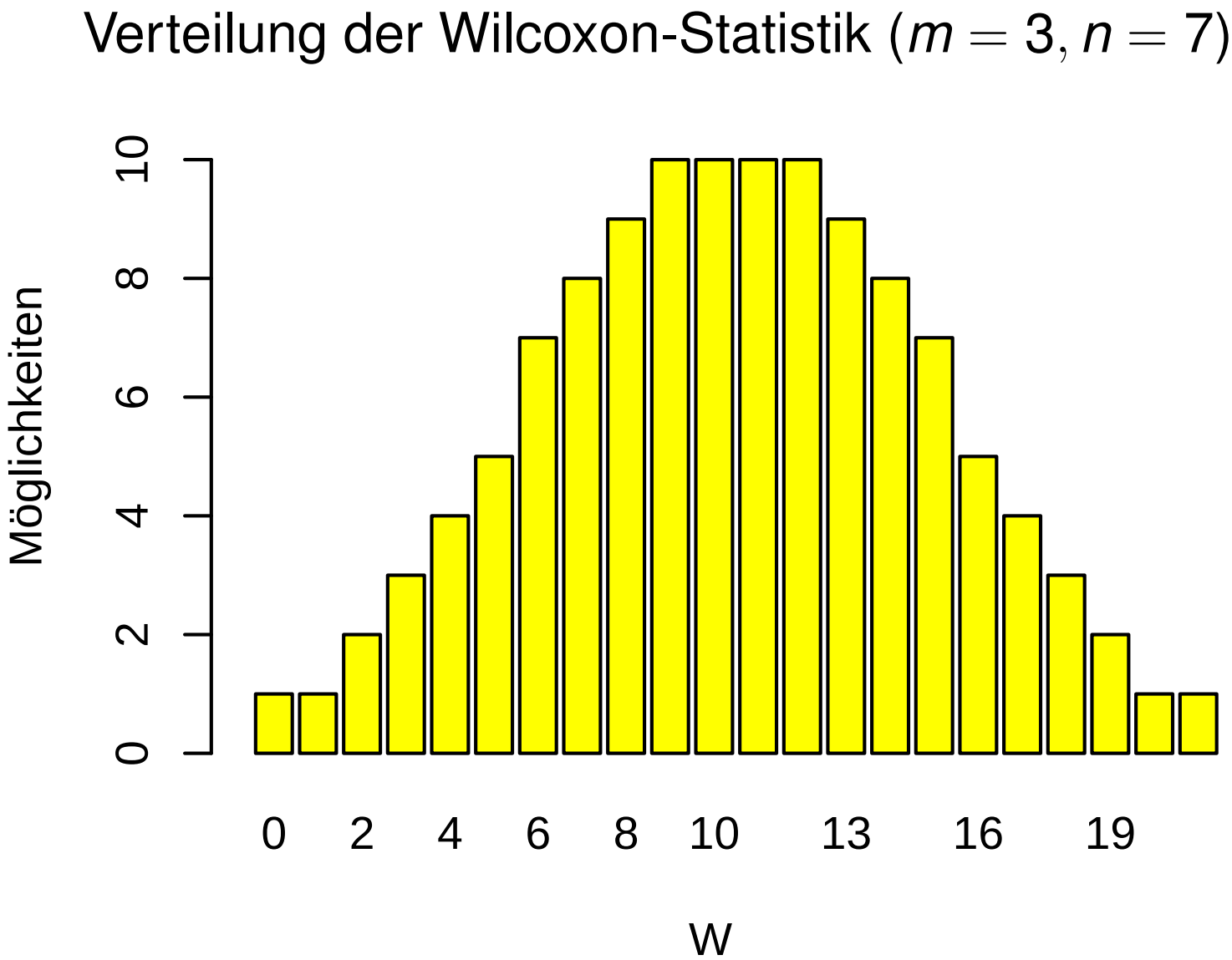
Nullhypothese:
X-Stichprobe und Y-Stichprobe
stammen aus
derselben Verteilung

Die 3 Ränge der X-Stichprobe
1 2 3 4 5 6 7 8 9 10

hätten genausogut irgendwelche 3 Ränge
1 2 3 4 5 6 7 8 9 10
sein können.

Es gibt $\frac{10 \cdot 9 \cdot 8}{3 \cdot 2 \cdot 1} = 120$ Möglichkeiten.

(Allgemein: $\frac{(m+n)(m+n-1)\cdots(n+1)}{m(m-1)\cdots 1}) = \frac{(m+n)!}{n!m!} = \binom{m+n}{m}$ Möglichkeiten)


Verteilung der Wilcoxon-Statistik ($m = 3, n = 7$)
Möglichkeiten
0
2
4
6
8
10
0
2
4
6
8
10
13
16
19
W

Unter der Nullhypothese sind alle Rangbelegungen gleich wahrscheinlich, also

$$\Pr(W = w) = \frac{\text{Anz. Möglichkeiten mit Rangsummenstatistik } w}{120}$$

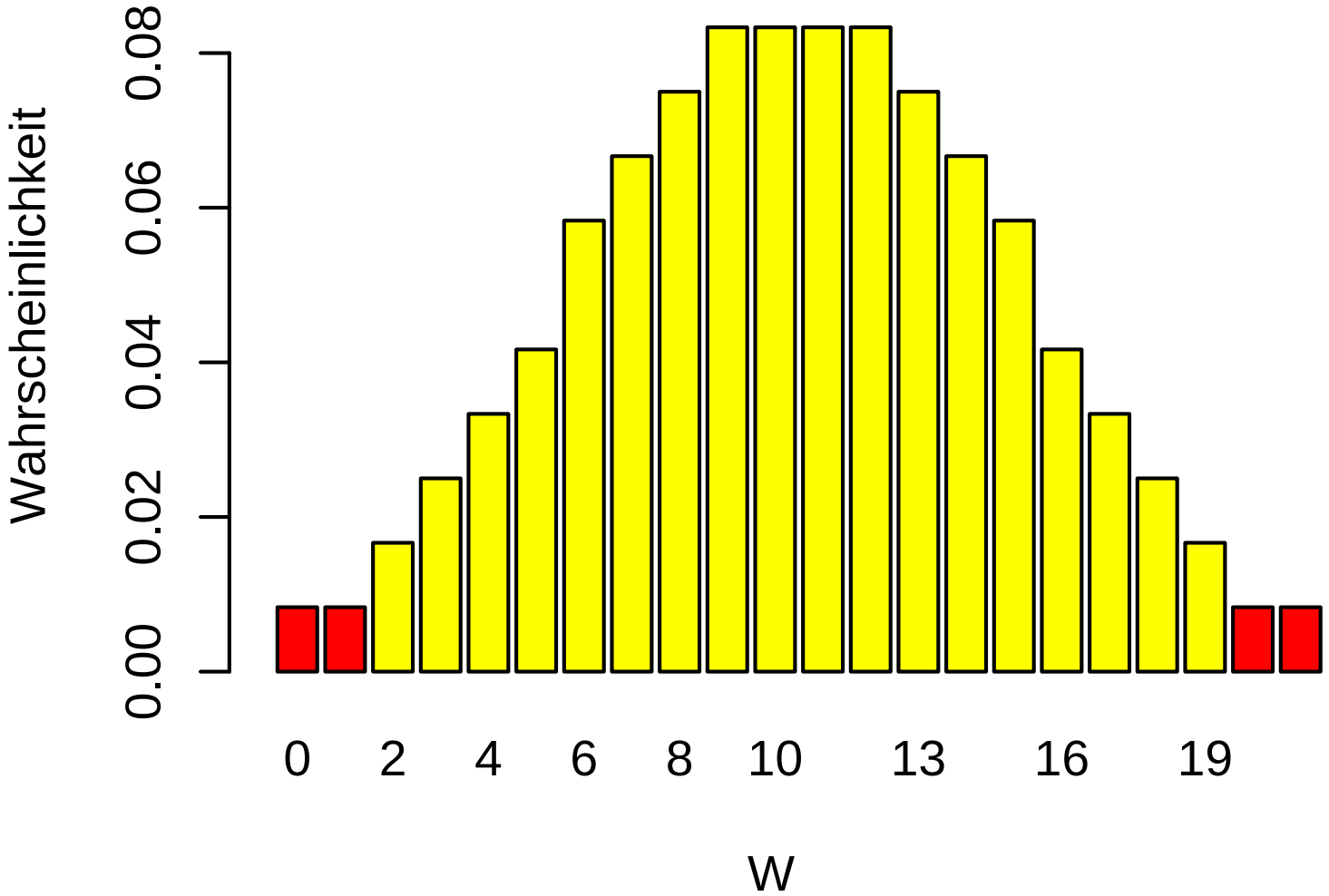
Wir beobachten in unserem Beispiel:

1,5, 5,6; 7,9; 35,2; 38,1; 41,0; 56,7; 112,1; 197,4; 381,8

somit $W = 1$

$$\begin{aligned} &\Pr(W \leq 1) + \Pr(W \geq 20) \\ &= \Pr(W = 0) + \Pr(W = 1) + \Pr(W = 20) + \Pr(W = 21) \\ &= \tfrac{1+1+1+1}{120} \doteq 0{,}033 \end{aligned}$$

Verteilung der Wilcoxon-Statistik ($m = 3, n = 7$)



Prüfen wir in unserem Beispiel die Nullhypothese, dass die Verteilungen von X und Y identisch sind, auf dem 5%-Niveau:

Wir haben $W = 1$ beobachtet, also

$$p\text{-Wert} = P(\text{ein so extremes } W) = 4/120 = 0{,}033$$

Wir lehnen die Nullhypothese auf dem 5%-Niveau ab.

Bem.: Die Verteilungsgewichte von W kann man mittels einer Rekursionsformel explizit bestimmen (was für „mittelgroße“ m und n praktikabel ist),
bei großem m und n verwendet man eine Normalapproximation.

R kennt den Wilcoxon-Test mittels `wilcox.test`:

```
> x
[1]  1.5  5.6 35.2
> y
[1]   7.9  38.1  41.0  56.7 112.1 197.4 381.8
> wilcox.test(x,y)

Wilcoxon rank sum test

data:  x and y
W = 1, p-value = 0.03333
alternative hypothesis: true location shift is
not equal to 0
```

Vergleich von *t*-Test und Wilcoxon-Test

Beachte:

Sowohl der *t*-Test als auch
der Wilcoxon-Test können verwendet werden,
um eine vermutete Verschiebung der Verteilung zu stützen.

Der *t*-Test testet „nur“ auf Gleichheit der Erwartungswerte.
Der Wilcoxon-Test dagegen testet auf Gleichheit
der gesamten Verteilungen.
(Der Wilcoxon-Test kann beispielsweise Signifikanz anzeigen,
selbst wenn die Stichproben-Mittelwerte übereinstimmen)

In besonderen Fällen

- Verteilungen sind asymmetrisch
- Stichprobenlänge ist klein

hat der Wilcoxon-Test eine höhere Testpower.

Vergleichen wir (spaßeshalber) mit dem *t*-Test (mit R ausgeführt):

```
> x
[1]  1.5  5.6 35.2
> y
[1]   7.9  38.1  41.0  56.7 112.1 197.4 381.8
> t.test(x,y,var.equal=TRUE)

Two Sample t-test

data:  x and y
t = -1.3319, df = 8, p-value = 0.2196
alternative hypothesis: true difference in means is not e
95 percent confidence interval:
 -287.30632   76.93489
sample estimates:
mean of x mean of y
  14.1000  119.2857
```