Wahrscheinlichkeitsrechnung und
Statistik für Biologen

5. Der zwei-Stichproben-t-Test

(t-Test für ungepaarte Stichproben)

und der Wilcoxon-Test

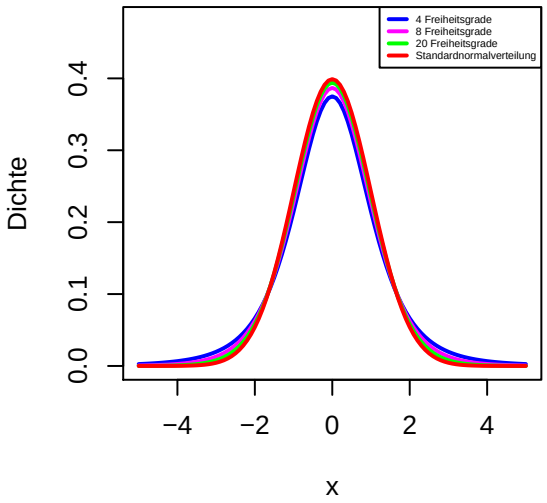
Dirk Metzler

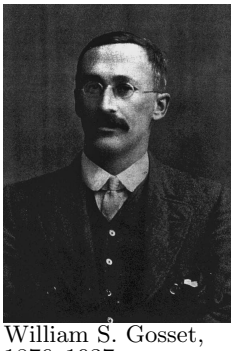
24. März 2026

Inhaltsverzeichnis

1 Wiederholung: t-Test für gepaarte Stichproben

„Student“ und seine Verteilung(en)





William S. Gosset, 1876–1937

Zusammenfassung gepaarter t-Test
Gegeben: gepaarte Beobachtungen

$$(Y_1, Z_1), (Y_2, Z_2), \ldots, (Y_n, Z_n)$$

Nullhypothese H_0: $\mu_Y = \mu_Z$ **Signifikanzniveau:** α (meist $\alpha = 5\%$) **Test**: gepaarter t-Test (genauer: zweiseitiger gepaarter t-Test)

Berechne Differenz $X := Y - Z$ Berechne Teststatistik

$$t := \frac{\overline{X}}{s(X)/\sqrt{n}}$$

p-Wert $= \Pr(|T_{n-1}| \geq |t|)$ ($n-1$ Freiheitsgrade) Verwirf Nullhypothese, falls p-Wert $\leq \alpha$

Zusammenfassung Ein-Stichproben t-Test
Gegeben: Beobachtungen

$$X_1, X_2, \ldots, X_n$$

Nullhypothese H_0: $\mu_X = c$ (Den Wert c testet man, oft $c = 0$) **Signifikanzniveau:** α (meist $\alpha = 5\%$) **Test**: t-Test

Berechne Teststatistik

$$t := \frac{\overline{X} - c}{s(X)/\sqrt{n}}$$

p-Wert $= \Pr(|T_{n-1}| \geq |t|)$ ($n-1$ Freiheitsgrade) Verwirf Nullhypothese, falls p-Wert $\leq \alpha$

2 t-Test für ungepaarte Stichproben

2.1 Angenommen, die Varianzen sind gleich

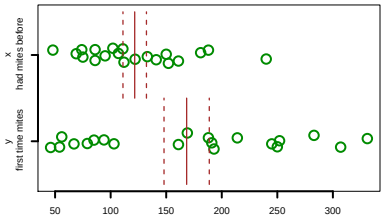
Beispiel: Bevorzugen Spinnmilben Pflanzen, die bisher nicht von Spinnmilben befallen waren?

Infiziere Baumwollsträucher mit Milben (*Tetranychus urticae*) und zähle die Milben auf Pflanzen, die schon mal befallen waren, und auf solchen, die zum ersten Mal befallen sind.

Die hier gezeigten Daten sind per Computersimulation erzeugt, aber echten Daten nachempfunden, siehe z.B.

Literatur

[1] S. Harrison, R. Karban: Behavioral response of spider mites (*Tetranychus urticae*) to induced resistance of cotton plants *Ecological Entomology* **11**:181-188, 1986.



$$\mu(y) = 168.4$$
$$sd(y) = 91.09763$$
$$sd(y)/\sqrt{20} = 20.37005$$
$$\mu(x) = 121.65$$
$$sd(x) = 47.24547$$
$$sd(x)/\sqrt{20} = 10.56441$$

Unsere Nullhypothese H_0: Alle Werte sind unabhängig aus der selben Normalverteilung gezogen.(Passt streng genommen nicht, da es hier um Anzahlen geht. Da es aber nicht sehr kleine Zahlen sind, approximativ okay.)

Diese Nullhypothese H_0 beinhaltet, dass die beiden Stichproben ("schon vorher infiziert" und "zum erste mal infiziert") aus Verteilungen stammen, die nicht nur den selben Mittelwert haben (was wir eigentlich testen wollen), sondern auch die selbe Varianz. Letzteres verwenden wir, wenn wir für die Berechnung der t-Statistik die Standardabweichung der Differenz der Stichprobenmittelwerte schätzen.

```
> t.test(y,x,var.equal=TRUE)

Two Sample t-test

data:  y and x
t = 2.0373, df = 38, p-value = 0.04862
alternative hypothesis: true difference in
 means is not equal to 0
95 percent confidence interval:
  0.2970719 93.2029281
```

```
sample estimates:
mean of x mean of y
   168.40    121.65
```

Theorem 1 (zwei-Stichproben t-Test, ungepaart mit gleichen Varianzen) *Seien $X_1, \ldots, X_n$ und $Y_1, \ldots, Y_m$ unabhängige normalverteilte Zufallsvariablen mit der selben Varianz σ^2. Als **gepoolte Stichprobenvarianz** definieren wir*

$$s_p^2 = \frac{(n-1) \cdot s_X^2 + (m-1) \cdot s_Y^2}{m+n-2}.$$

Unter der Nullhypothese gleicher Erwartungswerte $\mu_X = \mu_y$ folgt die Statistik

$$t = \frac{\overline{X} - \overline{Y}}{s_p \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

einer t-Verteilung mit $n + m - 2$ mit Freiheitsgraden.

2.2 Wenn die Varianzen ungleich sein könnten

Beispiel: Backenzähne von Hipparions

(c): public domain

Die Daten

77 Backenzähne

gefunden in den Chiwondo Beds, Malawi,

jetzt in den Sammlungen des Hessischen Landesmuseums, Darmstadt

http://en.wikipedia.org/wiki/File:LocationMalawi.svg (c): Rei-artur

Zuordnung

Die Zähne wurden zwei Arten zugeordnet:

Hipparion africanum[0.3ex] $\approx$ 4 Mio. Jahre

Hipparion libycum[0.3ex] $\approx$ 2,5 Mio. Jahre

Geologischer Hintergrund

Vor 2,8 Mio. Jahren kühlte sich das Klima weltweit ab.

Das Klima in Ostafrika:[0.5ex] warm-feucht $\longrightarrow$ kühl-trocken

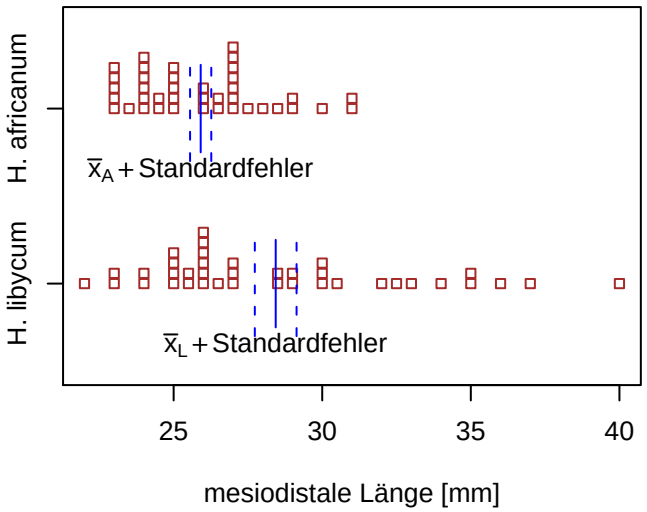
Hipparion:[0.5ex] Laubfresser $\longrightarrow$ Grasfresser

Frage

Hipparion:[0.5ex] Laubfresser $\longrightarrow$ Grasfresser

andere Nahrung $\longrightarrow$ andere Zähne?

Messungen: mesiodistale Länge
distal = von der Mittellinie weg



Wir beobachten ($n_A = 39$, $n_L = 38$):

$\overline{x}_A = 25{,}9$, $s_A = 2{,}2$, unser Schätzwert für die Streung von $\overline{x}_A$ ist also $f_A = s_A/\sqrt{n_A} = 2{,}2/\sqrt{n_A} = 0{,}36$ (Standardfehler),

$\overline{x}_L = 28{,}4$, $s_L = 4{,}3$, unser Schätzwert für die Streung von $\overline{x}_L$ ist also $f_L = s_L/\sqrt{n_L} = 4{,}3/\sqrt{n_L} = 0{,}70$.

Ist die beobachtete Abweichung $\overline{x}_L - \overline{x}_A = 2{,}5$ mit der *Nullhypothese* verträglich, dass $\mu_L = \mu_A$?

Da die Stichproben von zwei verschiedenen Arten kommen, beinhaltet unsere Nullhypothese diesmal nicht, dass beide aus der selben Verteilung kommen. Wir wollten also hier *nicht* voraussetzen, dass beide Arten die selbe Varianzen bei den Zahngrößen haben.

t-Statistik

Ist die beobachtete Abweichung $\overline{x}_L - \overline{x}_A = 2{,}5$ mit der *Nullhypothese* verträglich, dass $\mu_L = \mu_A$?

Wir schätzen die Streuung von $\overline{x}_L - \overline{x}_A$ durch f, wo

$$f^2 = f_L^2 + f_A^2$$

und bilden $t = \dfrac{\overline{x}_L - \overline{x}_A}{f}$.

Wenn die Nullhypothese zutrifft, ist t (approximativ) Student-verteilt mit g Freiheitsgraden (wobei g aus den Daten geschätzt wird.)

Theorem 2 (Welch-t-Test, die Varianzen dürfen ungleich sein) *Seien $X_1, \ldots, X_n$ und $Y_1, \ldots, Y_m$ unabhängige normalverteilte Zufallsvariablen mit (möglicherweise verschiedenen) Varianzen $VarX_i = \sigma_X^2$ und $VarY_i = \sigma_Y^2$. Seien s_X und s_Y die aus den Stichproben berechneten (korrigierten) Standardabweichungen und $f_X = \frac{s_X}{\sqrt{n}}$ und $f_Y = \frac{s_Y}{\sqrt{m}}$ die Standardfehler. Unter der Nullhypothese gleicher Mittelwerten $\mathbb{E}X_i = \mathbb{E}Y_j$ ist die Statistik*

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{f_X^2 + f_Y^2}}$$

ungefähr *t-verteilt mit* $\frac{\left(f_X^2+f_Y^2\right)^2}{\frac{f_X^4}{n-1}+\frac{f_Y^4}{m-1}}$ *Freiheitsgraden.*

Diese Approximation für die Freiheitsgrade brauchen Sie sich nicht auswendig zu wissen, denn R übernimmt das für Sie. (Aber natürlich gehört das auf's Formelblatt für die Klausur.)

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
```

Formulierung:

„Die mittlere mesiodistale Länge war signifikant größer (28,4 mm) bei *H. libycum* als bei *H. africanum* (25,9 mm) (t-Test, $p = 0{,}002$)."

2.3 Die Macht eines Tests

Testpower bzw. Testmacht

Die Power oder Macht eines Tests ist (vereinfacht gesagt) die Wahrscheinlichkeit, die Nullhypothese abzulehnen, falls die Alternative zutrifft.

Bei einer einelementigen Alternative ($H_0 : \mu = 0$, $H_1 : \mu = m_1$) kann man die Testmacht (oder Power) definieren als $\Pr_{H_1}$(Nullhypothese wird abgelehnt).

Meistens geht es aber um Alternativen wie $\mu > 0$ oder $\mu \neq 0$, und die Testmacht hängt dann vom tatsächlichen Wert von μ ab.

Warum interessiert uns die Testmacht?

Im Extremfall ist die Testmacht gleich 0, dann wird die Nullhypothese nie abgelehnt. Somit können wir unsere Vermutung nicht stützen.

Je größer die Testmacht, desto wahrscheinlicher wird die Nullhypothese abgelehnt. Beachte: Die Testmacht hängt stark von der Stichprobenlänge ab.

In der Praxis muss man sich bereits vor Versuchsbeginn Gedanken machen, wie groß die Stichprobenlänge sein muss, damit man die Vermutung stützen kann.

2.4 Vergleich: gepaarter t-Test und ungepaarter t-Test

Wann gepaarter t-Test (`paired=TRUE`) und wann ungepaarter t-Test (`paired=FALSE`)?

Wenn die Stichprobenlänge unterschiedlich ist, macht „gepaart“ keinen Sinn (`R` gibt Fehler aus).

Wenn die Stichprobenlänge gleich ist:

- Sind die Stichproben unabhängig voneinander? Falls ja, dann `paired=FALSE`, da wegen der höheren Zahl an Freiheitsgraden die Power größer ist.
- Sind die Stichproben voneinander abhängig? (z.B. Messungen von denselben Individuen bzw. Objekten) Falls ja, dann `paired=TRUE`. Bei starker Abhängigkeitsstruktur hat der gepaarte t-Test höhere Testpower (da der Test von Variabilität zwischen den Individuen bereinigt ist)

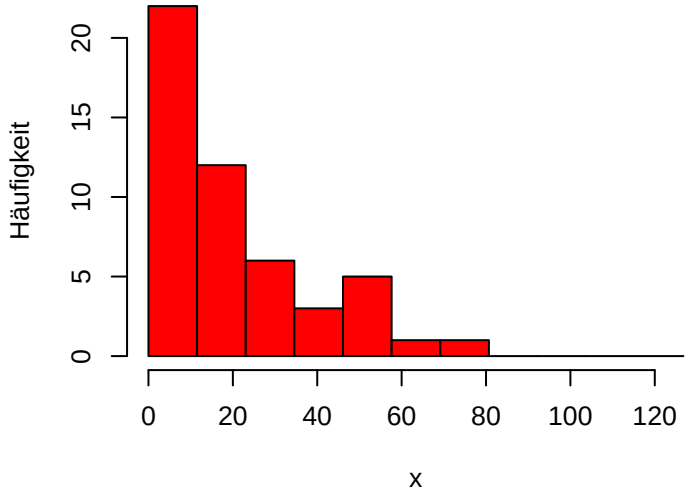
3 Wilcoxons Rangsummentest

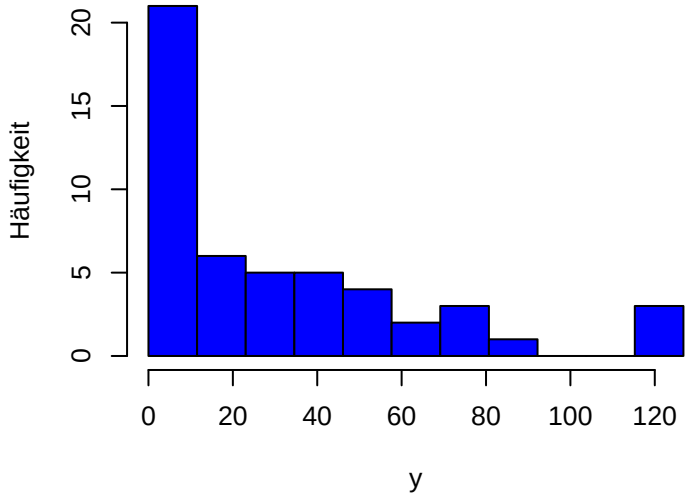
3.1 Motivation

Bei (ungefähr) glockenförmigen und symmetrisch verteilten Beobachtungen
oder wenn die Stichprobenumfänge genügend groß sind
können wir den t-Test benutzen, um die Nullhypothese $\mu_1 = \mu_2$ zu testen:
Die t-Statistik ist (annähernd) Student-verteilt.

Besonders bei sehr asymmetrischen und langschwänzigen Verteilungen kann das anders sein

Nehmen wir an, wir sollten folgende Verteilungen vergleichen:





Beispiele

- Wartezeiten
- Ausbreitungsentfernungen
- Zelltypenhäufigkeiten

Gesucht:

ein „verteilungsfreier“ Test mit dem man die Lage zweier Verteilungen zueinander testen kann

3.2 Wilcoxon-Test für unabhängige Stichproben

Beobachtungen: Zwei Stichproben

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

Wir möchten die Nullhypothese: X und Y aus derselben Population (X und Y haben diesselbe Verteilung) testen.

Alternative: Die beiden Verteilungen sind gegeneinander verschoben.

Voraussetzung des Tests: Die beiden Verteilungen haben diesselbe Form, sind also bis auf eine Lageverschiebung (in etwa) identisch.

Idee

Beobachtungen:

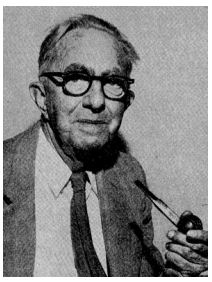
$$X : x_1, x_2, \ldots, x_m$$
$$Y : y_1, y_2, \ldots, y_n$$

- 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:

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



Frank Wilcoxon, 1892-–1965

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

Ein *kleines* Beispiel

- Beobachtungen:

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

- 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: $W = 1 + 2 + 4 - (1 + 2 + 3) = 1$

Interpretation von W

X-Werte tendenziell 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-Werte tendenziell 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 $\quad W = 19$

Wilcoxons Rangsummenstatistik

Bemerkung:

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

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

Der Wilcoxon Test heißt auch Mann-Whitney- Test. Die Mann-Whitney Statistik $U = W + \text{Konstante}$.

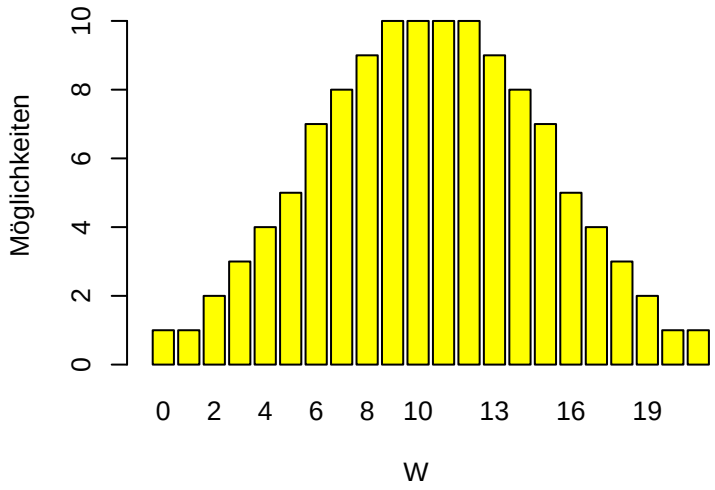
Signifikanz

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 $(m = 3, n = 7)$[1ex]



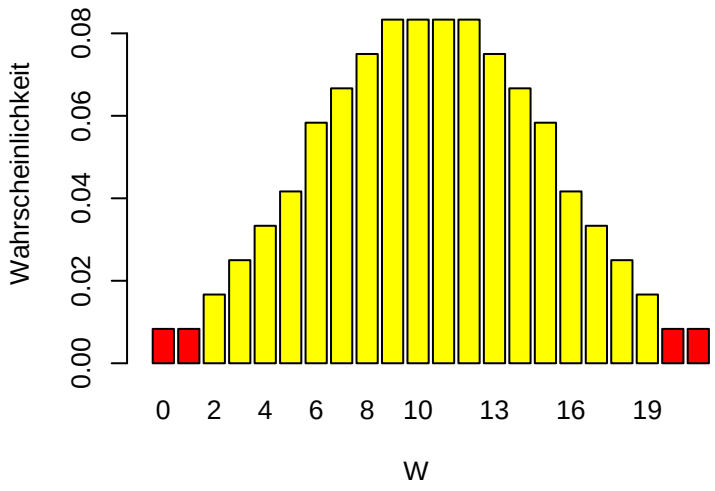
Unter der Nullhypothese sind alle Rangbelegungen gleich wahrscheinlich, also

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

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 $W = 1$

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

Verteilung der Wilcoxon-Statistik $(m = 3, n = 7)$[1ex]



Für unser Beispiel ($W = 1$) also:

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

Wir *lehnen* die *Nullhypothese*, dass die Verteilungen von X und Y identisch sind, auf dem 5%-Niveau *ab*.

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
```

Achtung

Achtung!!!

Wenn der Wilcoxon-Test Signifikanz anzeigt, so kann das daran liegen, dass die zugrunde liegenden Verteilungen verschiedene Formen haben.

Der Wilcoxon-Test kann u. U. mit hoher Wahrscheinlichkeit Signifikanz anzeigen, selbst wenn die Stichproben-Mittelwerte übereinstimmen!

Vergleich von t-Test und Wilcoxon-Test

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

Der Welch-t-Test testet „nur“ auf Gleichheit der Erwartungswerte. Der Wilcoxon-Test dagegen testet auf Gleichheit der gesamten Verteilungen (so wie der 2-Stichproben-t-Test mit gleichen Varianzen).

In vielen Fällen liefern beide Tests dasselbe Ergebnis. Sofern die Verteilungen einigermaßen glockenförmig sind, empfehlen wir den Welch-t-Test.

In besonderen Fällen

- Verteilungen sind stark asymmetrisch oder haben “Ausreißer”
- Stichprobenlänge ist klein

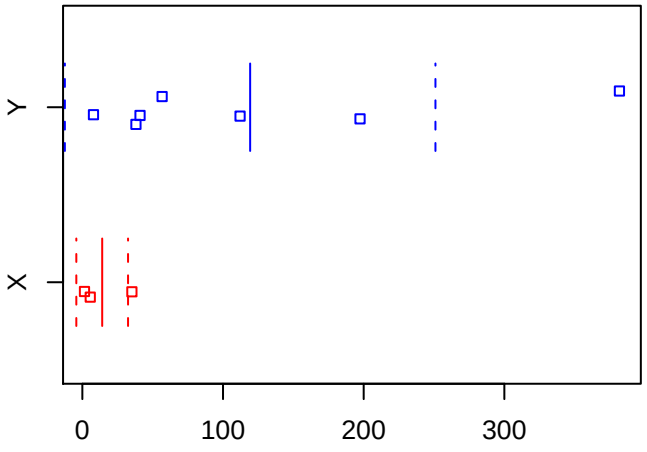
hat der Wilcoxon-Test eine höhere Testpower.

Vergleichen wir (spaßeshalber) mit dem t-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
> t.test(x,y)

Welch Two Sample t-test

data:  x and y
t = -2.0662, df = 6.518, p-value = 0.08061
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -227.39182   17.02039
sample estimates:
mean of x mean of y
  14.1000  119.2857
```



4 Zusammenfassung

Wir untersuchen ein Merkmal in zwei Populationen:

Population	1	2
Mittelwert	μ_1	μ_2

Nullhypothese: $\mu_1 = \mu_2$

Wir ziehen Stichproben aus den Populationen mit Stichproben-Mittelwerten $\overline{x}_1 \quad \overline{x}_2$

Um die Nullhypothese H_0 zu prüfen, bilden wir im Zweifelsfall die *Welch-t-Statistik* $t = \frac{\overline{x}_1 - \overline{x}_2}{f}$ mit $f = \sqrt{\left(\frac{s_1}{\sqrt{n_1}}\right)^2 + \left(\frac{s_2}{\sqrt{n_2}}\right)^2}$

p-Wert unter H_0: $p \approx \Pr(|T_g| \geq |t|)$ (g=(geschätzte) Anz. Freiheitsgrade, hängt von n_1, n_2, s_1, s_2 ab)

Wenn die Normalverteilungsannahmen offensichtlich grob verletzt ist und die Nullhypothese nicht nur ist, dass die beiden Mittelwerte gleich sind, sondern dass die Stichproben aus der selben Verteilung kommen, können wir stattdessen den *Wilcoxon-Test* verwenden.

Was Sie u.a. erklären können sollten

- Durchführung ungepaarter t-Test
- Wann welcher t-Test?
  - gepaart oder ungepaart?
  - gleiche oder ungleiche Varianzen?
  - einseitig oder zweiseitig?
- Wie und wann man den Wilcoxon-Rangsummentest anwendet