

Wahrscheinlichkeitsrechnung und  
Statistik für Biologen  
**5. Der zwei-Stichproben- $t$ -Test**  
( $t$ -Test für ungepaarte Stichproben)  
**und der Wilcoxon-Test**

Martin Hutzenthaler & Dirk Metzler

20. Mai 2010

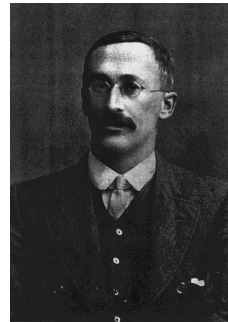
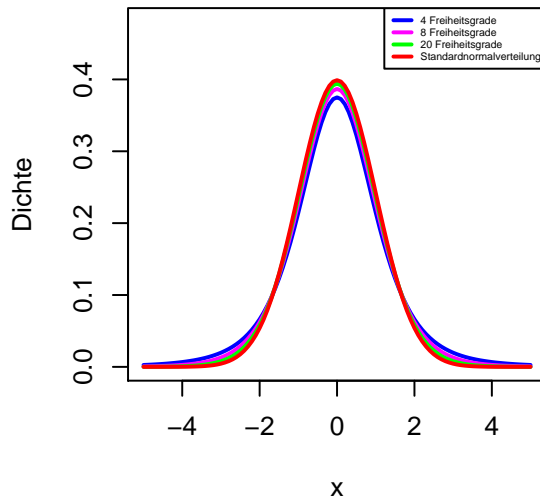
**Inhaltsverzeichnis**

<b>1</b>	<b>Wiederholung: Ein-Stichproben-<math>t</math>-Test</b>	<b>1</b>
<b>2</b>	<b><math>t</math>-Test für ungepaarte Stichproben</b>	<b>1</b>
2.1	Beispiel: Backenzähne von Hipparions . . . . .	1
2.2	Power eines Tests . . . . .	5
2.3	Vergleich: gepaarter $t$ -Test und ungepaarter $t$ -Test . . . . .	5
<b>3</b>	<b>Wilcoxons Rangsummentest</b>	<b>6</b>
3.1	Motivation . . . . .	6
3.2	Wilcoxon-Test für unabhängige Stichproben . . . . .	6
<b>4</b>	<b>Zusammenfassung</b>	<b>11</b>

# Inhaltsverzeichnis

## 1 Wiederholung: Ein-Stichproben-*t*-Test

„Student“ und seine Verteilung(en)



William S. Gosset,  
1876–1937

### Zusammenfassung gepaarter *t*-Test

Gegeben: gepaarte Beobachtungen

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

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

Berechne Differenz  $X := Y - Z$

Berechne Teststatistik

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

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

$$X_1, X_2, \dots, X_n$$

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

Berechne Teststatistik

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

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 Beispiel: Backenzähne von Hipparions

Beispiel: Backenzähne von Hipparions

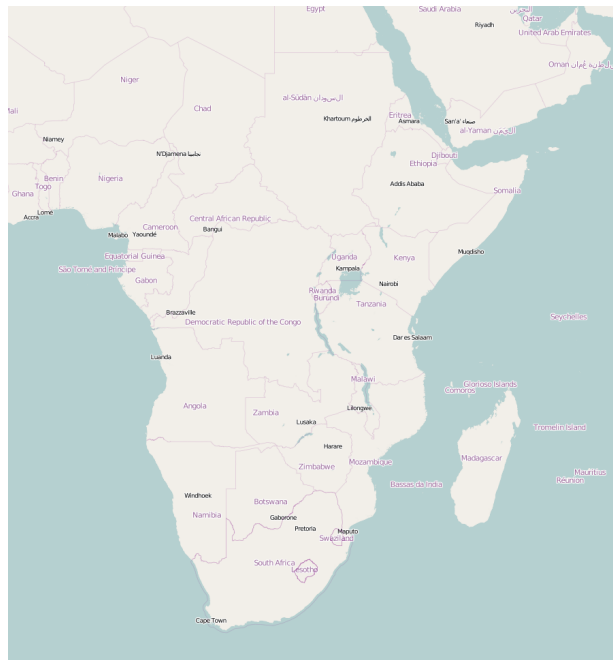


## Die Daten

77 Backenzähne

gefunden in den Chiwondo Beds, Malawi,

jetzt in den Sammlungen des Hessischen Landesmuseums, Darmstadt

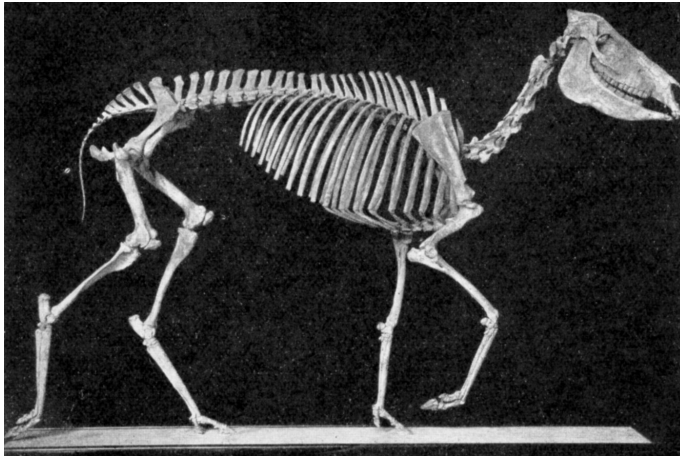


## Zuordnung

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 → kühl-trocken

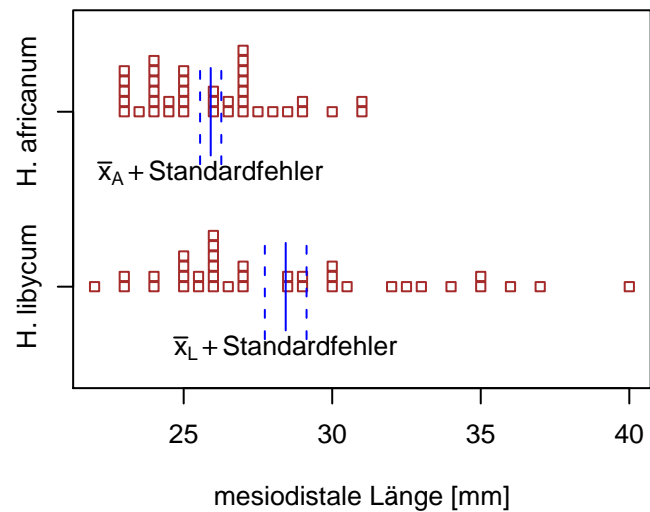
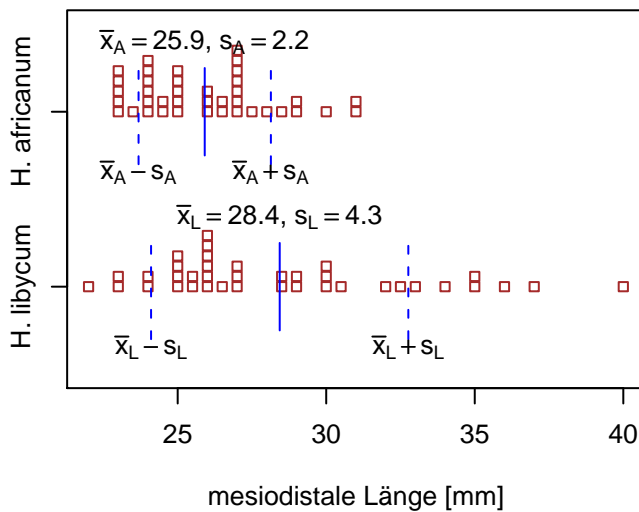
Hipparion: [0.5ex] Laubfresser → Grasfresser

### Frage

Hipparion: [0.5ex] Laubfresser → Grasfresser

andere Nahrung → andere Zähne?

Messungen: mesiodistale Länge  
distal = von der Mittellinie weg



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

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

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

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

### **t-Statistik**

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

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

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

$$\text{und bilden } t = \frac{\bar{x}_L - \bar{x}_A}{f}.$$

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

### **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,  $p = 0,002$ ).“

Wir haben keine Annahme an die Varianzen der beiden Stichproben gemacht. Es gibt auch eine Variante des  $t$ -Tests mit der zusätzlichen Annahme  $\sigma_A^2 = \sigma_L^2$ :

```
> t.test(L,A,var.equal=TRUE)
```

Two Sample t-test

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

## 2.2 Power eines Tests

### Testpower bzw. Testmacht

Salopp gesprochen ist die **Power** oder **Macht** eines Tests die Wahrscheinlichkeit, die Nullhypothese abzulehnen, falls die Alternative zutrifft.

Bei einer einelementigen Alternative ist dies leicht zu formulieren:  $H_0 : \mu = 0$   $H_1 : \mu = m_1$

Die Testpower (oder auch Testmacht) ist dann definiert als  $\Pr_{H_1}(\text{Nullhypothese wird abgelehnt})$

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.

## 2.3 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)

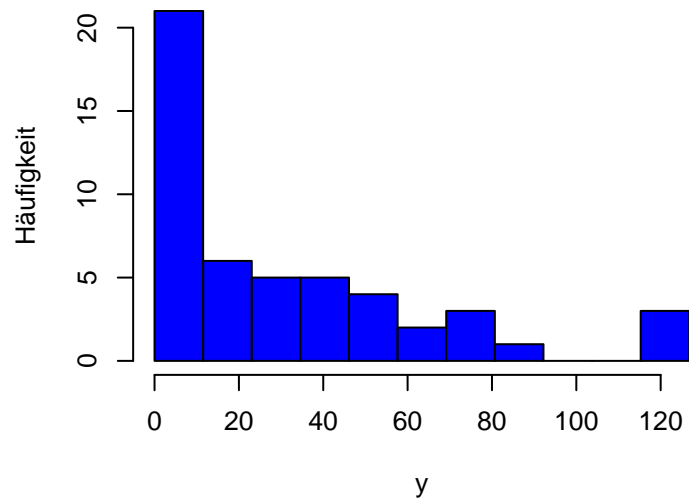
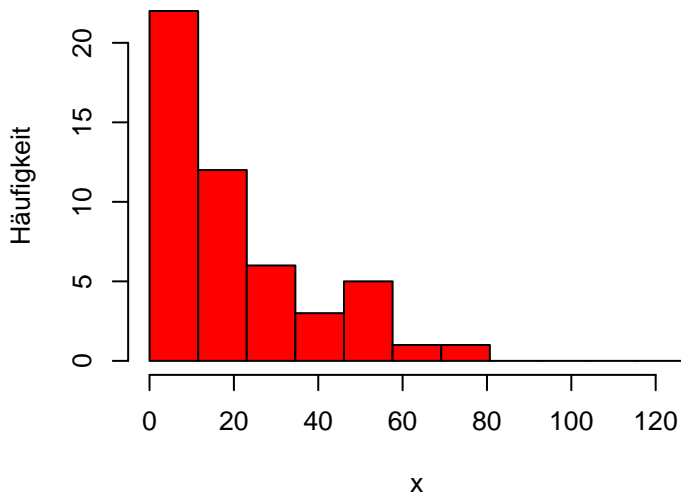
### 3 Wilcoxon's Rangsummentest

#### 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

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

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

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:

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

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

- 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  $\{1, 2, \dots, m + n\}$ .
- Berechne die Summe der  $X$ -Ränge, prüfe, ob dieser Wert untypisch groß oder klein.

## Wilcoxons Rangsummenstatistik

Beobachtungen:

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

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



Frank Wilcoxon,  
1892-1965

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

heißt

*Wilcoxons Rangsummenstatistik*

## Wilcoxons Rangsummenstatistik

Bemerkung:

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

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

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

Bemerkung

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



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

### Interpretation von $W$

$X$ -Population kleiner  $\implies W$  klein:

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

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

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

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

$X$ -Population größer  $\implies W$  groß:

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

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

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

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

### Signifikanz

Nullhypothese:

$X$ -Stichprobe und  $Y$ -Stichprobe stammen aus derselben Verteilung

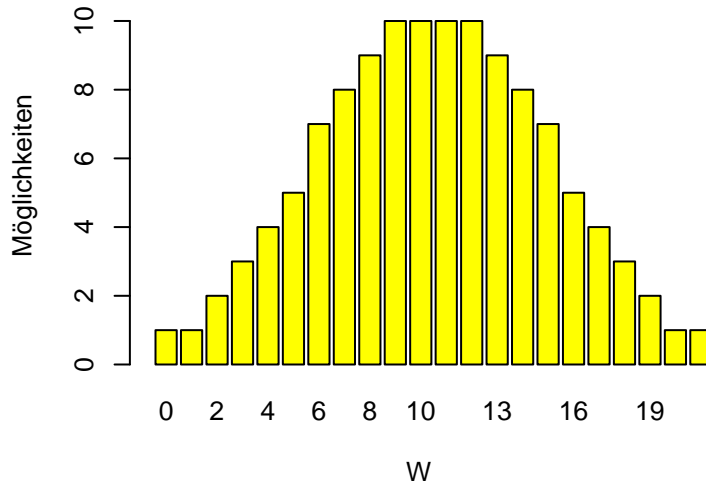
Die 3 Ränge der  $X$ -Stichprobe 1 2 3 4 5 6 7 8 9 10

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



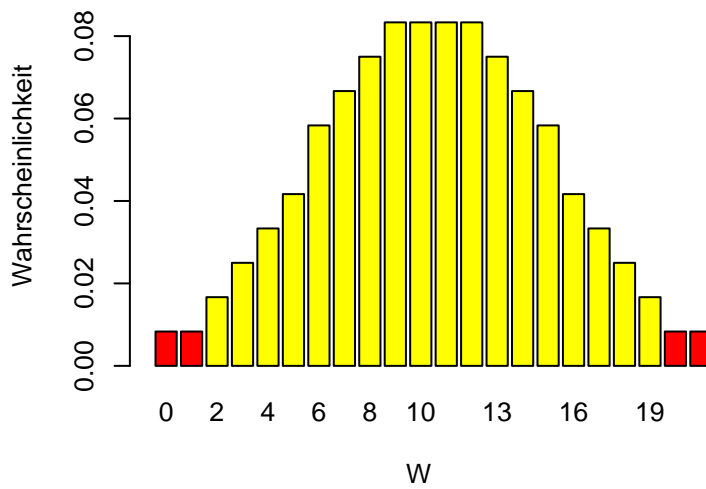
Unter der Nullhypothese sind alle Rangbelegungen gleich wahrscheinlich, also

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

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  $W = 1$

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

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 zu grunde liegenden Verteilungen verschiedene Formen haben.

Der Wilcoxon-Test kann beispielsweise Signifikanz anzeigen, **selbst wenn die Stichproben-Mittelwerte übereinstimmen!**

### 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.

**In den meisten Fällen liefern beide Tests dasselbe Ergebnis.** Im Allgemeinen empfehlen wir den  $t$ -Test, da er robuster ist.

In besonderen Fällen

- Verteilungen sind asymmetrisch
- Stichprobenlänge ist klein

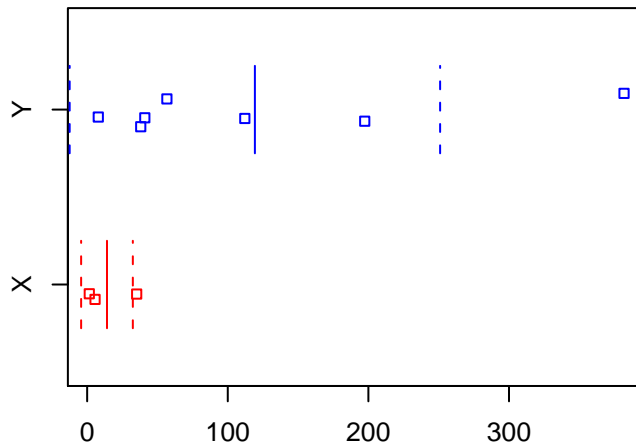
hat der Wilcoxon-Test eine höhere Testpower.

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	$\mu_1$	$\mu_2$

Nullhypothese:  $\mu_1 = \mu_2$

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

Um die Nullhypothese  $H_0$  zu prüfen, bilden wir die  $t$ -Statistik[1ex]

$$t = \frac{\bar{x}_1 - \bar{x}_2}{f} \quad \text{mit} \quad f^2 = \left( \frac{s_1}{\sqrt{n_1}} \right)^2 + \left( \frac{s_2}{\sqrt{n_2}} \right)^2$$

$p$ -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)

In besonderen Fällen können wir stattdessen den *Wilcoxon-Test* verwenden.