Stochastik

t-Verteilung, t-Test, Vertrauensintervalle, Wilcoxon-Test und gepaarte/ungepaarte Stichproben

Mirko Birbaumer

Hochschule Luzern Technik & Architektur

Problem in Praxis: σ_X ist unbekannt!

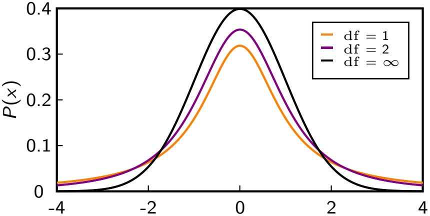
- Falls σ_X unbekannt ist, dann müssen wir die Varianz aus den Daten schätzen:

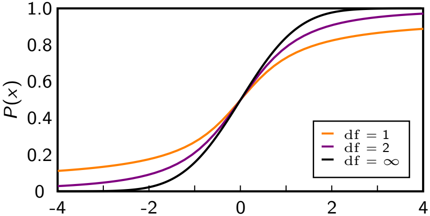
$$\widehat{\sigma_X^2} = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}_n\right)$$

- Neue Teststatistik: $T = \dfrac{\overline{X}_n - \mu_0}{\frac{\widehat{\sigma}_X}{\sqrt{n}}}$
- Verteilung von T, falls $H_0 \; : \; \mu = \mu_0$ stimmt: $T \sim t_{n-1}$
- t_{n-1} ist die sogenannte **t-Verteilung mit** $n-1$ **Freiheitsgraden**

"Student's" t-Verteilung – Zoo Teil 3

- Annahme: $X_1, X_2, \ldots, X_n \sim \mathcal{N}(\mu, \sigma_X^2)$ und unabhängig
- $\widehat{\sigma}_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}_n\right)^2$ ist geschätzte Varianz
- $T = \left(\overline{X}_n - \mu\right) / \left(\frac{\widehat{\sigma}_X}{\sqrt{n}}\right)$ folgt einer „t-Verteilung mit $n-1$ Freiheitsgraden", $T \sim t_{n-1}$
- Werte mit Computer ermittelbar
- Falls $n = \infty$: $t_n = \mathcal{N}(0,1)$



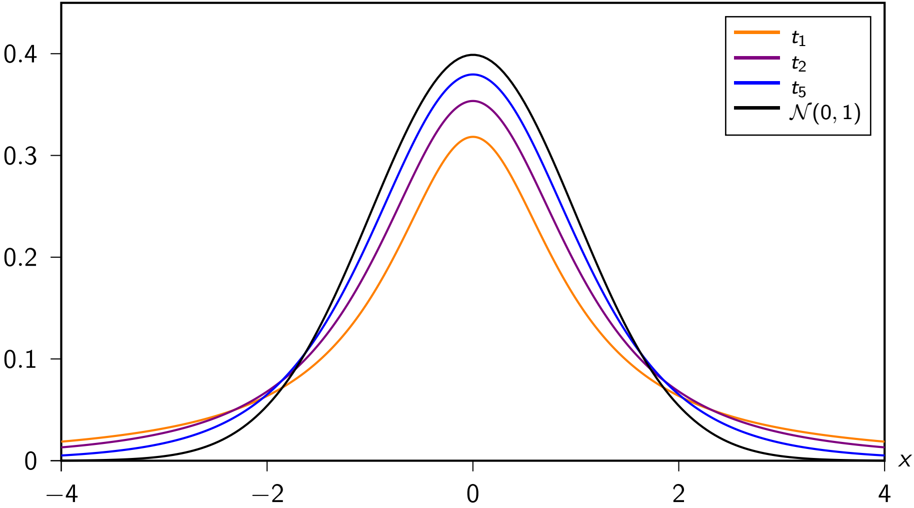


t-Verteilung: Eigenschaften

- Wie die Standardnormalverteilung ist die t-Verteilung symmetrisch um 0
- Sie ist jedoch „langschwänziger“, d.h. ihr Peak in der Mitte ist weniger hoch und „weit aussen“ ist die Dichte grösser (insbesondere falls die Anzahl Freiheitsgrade n klein ist)
- D.h. verglichen mit der Standardnormalverteilung liefert sie (betragsmässig) **eher grosse Werte**
- Es gilt: $t_n \to \mathcal{N}(0, 1)$ für $n \to \infty$ (siehe auch Plot mit Dichten)

t-Verteilung: Dichten

Dichte $f(x)$



t-Test: σ_X unbekannt

1. **Modell**: X_i ist eine kontinuierliche Messgrösse;

$$X_1, \ldots, X_n \text{ iid } \mathcal{N}(\mu, \sigma_X^2), \ \sigma_X \text{ wird durch } \widehat{\sigma_X} \text{ geschätzt}$$

2. **Nullhypothese**: $H_0: \ \mu = \mu_0$
   Alternative: $H_A: \ \mu \neq \mu_0$ (oder „$<$" oder „$>$")
3. **Teststatistik**:

$$T = \frac{(\overline{X}_n - \mu_0)}{\widehat{\sigma}_{\overline{X}_n}} = \frac{(\overline{X}_n - \mu_0)}{\widehat{\sigma}_X/\sqrt{n}} = \frac{\text{beobachtet - erwartet}}{\text{Standardfehler}}$$

   Verteilung der Teststatistik unter $H_0: \ T \sim t_{n-1}$
4. **Signifikanzniveau**: α
5. **Verwerfungsbereich für die Teststatistik**:

$$K = (-\infty, t_{n-1;\frac{\alpha}{2}}] \cup [t_{n-1;1-\frac{\alpha}{2}}, \infty) \quad \text{bei } H_A: \ \mu \neq \mu_0$$

$$K = (-\infty, t_{n-1;\alpha}] \quad \text{bei } H_A: \ \mu < \mu_0$$

$$K = [t_{n-1;1-\alpha}, \infty) \quad \text{bei } H_A: \ \mu > \mu_0$$

6. **Testentscheid**: Überprüfe, ob der beobachtete Wert der Teststatistik im Verwerfungsbereich liegt

Beispiel: t-Test

- Unsere Messreihe für die Körpergrösse von 150 Frauen in Luzern ergab:

$$\overline{x}_{150} = 168\mathrm{cm}$$

  Wir vermuten, dass die Durchschnittsgrösse der Schweizerinnen entweder grösser oder kleiner als 164 cm ist (zweiseitiger Test).
- σ_X wurde nun geschätzt, und beträgt $\hat{\sigma}_X = 10$cm.
- Wir berechnen zuerst den P-Wert für eine einseitige Alternativhypothese

$$P[\overline{X}_{150} \geq 168] = 1 - P[X \leq 168] = 1 - P\left[T \leq \frac{168 - 164}{10/\sqrt{150}}\right]$$

  mit **R**:

R-Befehl: pt()

```
> 1-pt((168-164)/(10/sqrt(150)),df=149) 1.241988e-06
```

- Der P-Wert wäre also $2 \cdot 1.241988 \cdot 10^{-6}$. Auch in diesem Fall können wir die Nullhypothese auf dem Signifikanzniveau $\alpha = 0.05$ verwerfen.

Vertrauensintervall für μ

- Vertrauensintervall: besteht aus denjenigen Werten μ, bei denen der entsprechende Test nicht verwirft.
- Bei einem zweiseitigen t-Test hat der Verwerfungsbereich die Form

$$K = (-\infty, t_{n-1;\frac{\alpha}{2}}] \cup [t_{n-1;1-\frac{\alpha}{2}}, \infty)$$

- Der t-Test verwirft H_0 nicht, wenn der Wert der Teststatistik nicht im Verwerfungsbereich der Teststatistik ist:

$$t_{n-1;\frac{\alpha}{2}} \leq \frac{\sqrt{n}(\overline{x}_n - \mu_0)}{\widehat{\sigma}_X}$$

und

$$t_{n-1;1-\frac{\alpha}{2}} \geq \frac{\sqrt{n}(\overline{x}_n - \mu_0)}{\widehat{\sigma}_X}$$

Vertrauensintervall für μ

- Um das zweiseitige Vertrauensintervall von μ zu finden, müssen wir alle Werte von μ_0 finden, die obige Gleichungen erfüllen. Wir lösen nach μ_0 auf:

$$\mu_0 \leq \overline{x}_n - \frac{\widehat{\sigma}_X \cdot t_{n-1;\frac{\alpha}{2}}}{\sqrt{n}}$$

$$\mu_0 \geq \overline{x}_n - \frac{\widehat{\sigma}_X \cdot t_{n-1;1-\frac{\alpha}{2}}}{\sqrt{n}}$$

Zweiseitiges Vertrauensintervall

Dies führt dann auf die folgenden **zweiseitigen Vertrauensintervalle** (die dazugehörigen Tests sind zweiseitig mit Alternative $H_A : \mu \neq \mu_0$) zum Niveau $1 - \alpha$:

$$\left[\overline{x}_n - t_{n-1,1-\alpha/2} \cdot \frac{\widehat{\sigma}_X}{\sqrt{n}}, \overline{x}_n + t_{n-1,1-\alpha/2} \cdot \frac{\widehat{\sigma}_X}{\sqrt{n}}\right]$$

Beispiel: Vertrauensintervall

- Methode A zur Bestimmung der Schmelzwärme: Die **mittlere** mit Methode A gemessene Schmelzwärme ist $\overline{x}_{13} = 80.02$, die aus den Daten **geschätzte Standardabweichung** ist $\hat{\sigma}_X = 0.024$.

- Es wurden $n = 13$ Messungen ausgeführt. Also haben wir $n - 1 = 13 - 1 = 12$ **Freiheitsgrade**.

- Das Vertrauensintervall zum Niveau 95% ist gegeben durch

$$\left[80.02 - t_{12,1-0.025} \cdot \frac{0.024}{\sqrt{13}}, 80.02 + t_{12,1-0.025} \cdot \frac{0.024}{\sqrt{13}}\right]$$

- Berechnung von $t_{12,0.975}$ mit **R**:

R-Befehl: qt()

```
qt(0.975,df=12)=2.18
```

Beispiel: Vertrauensintervall

- Das Konfidenzintervall für die mit Methode A gemessene Schmelzwärme lautet also
$$I = 80.02 \pm 2.18 \cdot 0.024/\sqrt{13} = [80.01, 80.04].$$

- Alternative Berechnung mit **R**:

R-Befehl: t.test()

```
> x <- c(79.98, 80.04, 80.02, 80.04, 80.03, 80.03,
80.04, 79.97, 80.05,80.03, 80.02, 80.00, 80.02)
> t.test(x, altervative = "two.sided", mu = 80.00,
conf.level = 0.95)
One Sample t-test
data: x
t = 3.1246, df = 12, p-value = 0.008779
alternative hypothesis: true mean is not equal to 80
95 percent confidence interval:
80.00629 80.03525
```

Nicht-Normalverteilte Daten: Vorzeichentest

1. **Modell**: $X_1, \ldots, X_n$ iid, wobei X_i eine beliebige Verteilung hat
2. **Nullhypothese**: $H_0: \ \mu = \mu_0$, (μ ist der Median)
   Alternative: $H_A: \ \mu \neq \mu_0$ (oder einseitige Variante)
3. **Teststatistik**: V: Anzahl X_i's mit $(X_i > \mu_0)$
   Verteilung der Teststatistik unter H_0: $V \sim \text{Bin}(n, \pi_0)$ mit $\pi_0 = 0.5$
4. **Signifikanzniveau**: α
5. **Verwerfungsbereich für die Teststatistik**: $K = [0, c_u] \cup [c_o, n]$ falls $H_A: \ \mu \neq \mu_0$. Die Grenzen c_u und c_o müssen mit der Binomialverteilung oder der Normalapproximation berechnet werden
6. **Testentscheid**: Entscheide, ob der beobachtete Wert der Teststatistik im Verwerfungsbereich der Teststatistik liegt

Beispiel: Vorzeichentest

- Beobachtet: $x_1 = 13,\ x_2 = 9,\ x_3 = 17,\ x_4 = 8,\ x_5 = 14$
- Angenommen: $H_0:\ \mu = \mu_0 = 10$, $H_A:\ \mu \neq 10$
- Vorzeichen von $x_i - \mu_0:\ +,\ -,\ +,\ -,\ +$
- Führen Sie den Binomialtest durch mit

$$H_0:\ \pi = 0.5,\ H_A:\ \pi \neq 0.5,\ n = 5,\ x = 3 \text{ (Anzahl „+“)}$$

Beispiel: Vorzeichentest

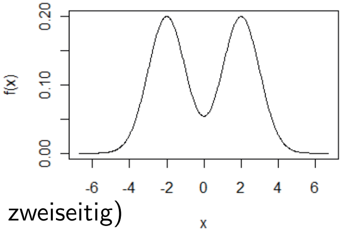
- Antwort: Binomialtest mit **R**

R-Befehl: binom.test()

```
> binom.test(x=3,n=5,p=0.5,alternative=''two.sided'')|
data: 3 and 5 number of successes = 3, number of trials = 5,
p-value = 1
```

- Die Nullhypothese beim Vorzeichentest wird nicht verworfen.
- Vorteil vom Vorzeichentest: Keine Annahme an Verteilung
- Nachteil vom Vorzeichentest: Kleinere Macht

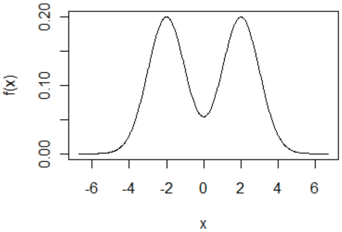
Nicht-normalverteilte Daten: Wilcoxon-Test



- Kompromiss zwischen Vorzeichen- und t-Test
- Annahme: $X_i \sim F$ iid, F ist symmetrisch
- Teste Median μ : H_0 : $\mu = \mu_0$ (einseitig oder zweiseitig)
- Intuition der Teststatistik
  - Rangiere $|x_i - \mu_0| \rightarrow r_i$
  - Gib Rängen ursprüngliches Vorzeichen von $(x_i - \mu_0)$ („signed ranks")
  - Teststatistik T: Summe aller Ränge, bei denen $(x_i - \mu_0)$ positiv ist
- Falls H_0 stimmt, sollte diese Rangsumme nicht zu gross und nicht zu klein sein

Beispiel: Wilcoxon-Test

- Bsp: H_0 : $\mu_0 = 0$
- Beobachte -1.9, 0.2, 2.9,-4.1, 3.9
- Absolutbeträge: 1.9, 0.2, 2.9, 4.1, 3.9
- Ränge der Absolutbeträge: 2,1,3,5,4
- Rangsumme der positiven Gruppe: 1+3+4=8
  Minimale Rangsumme: 0
  Maximale Rangsumme: 1+2+3+4+5 = 15



R-Befehl: wilcox.test()

```
> wilcox.test(c(-1.9, 0.2, 2.9, -4.1, 3.9), mu=0)|
wilcoxon singed rank test
data: c(-1.9, 0.2, 2.9, -4.1, 3.9) V=8, p-value = 1 alternative
hypothesis: true loaction is not equal to 0
```

Übersicht der Tests

Test	Annahme				$n_{\min}$ bei $\alpha = 0.05$	Macht für ein Beispiel (1)
	σ_X bekannt	$X_i \sim N$	Symm. Verteilung	iid		
z	×	×	×	×	1	89%
t		×	×	×	2	79%
Wilcoxon			×	×	6	79%
VZ				×	5	48%

(1): $X_i \sim \mathcal{N}(\mu, \sigma^2)$, $n = 10$; $H_0 : \mu = 0$; $H_A : \mu \neq 0$; $\alpha = 0.05$
Macht berechnet für konkrete Alternative: $X_i \sim \mathcal{N}(1, 1)$

Wilcoxon-Test versus t-Test

Wilcoxon-Test versus t-Test

Der **Wilcoxon-Test** ist in den allermeisten Fällen dem **t-Test** oder **Vorzeichen-Test** vorzuziehen: er hat in vielen Situationen oftmals wesentlich **grössere Macht**, und selbst in den ungünstigsten Fällen ist er nie viel schlechter.

Wenn man trotzdem den t-Test verwendet, dann sollte man die Daten auch graphisch ansehen, damit wenigstens grobe Abweichungen von der Normalverteilung entdeckt werden.

Insbesondere sollte der **Normal-Plot** angeschaut werden.

Vergleich von zwei Stichproben

Mögliche Fragestellungen

- Vergleich von zwei Messverfahren (Messgerät A vs. Messgerät B): Gibt es einen signifikanten Unterschied?
- Vergleich von zwei Herstellungsverfahren (A vs. B): Welches hat die besseren Eigenschaften (z.B. bzgl. einer Festigkeitsgrösse)?
- Werden männliche Dozenten von weiblichen Studierenden besser als von männlichen Studierenden bewertet?
- Wir sammeln also jeweils Daten von zwei Gruppen

Gepaarte Stichproben

- Im Beispiel der Messgeräte messen wir jeden Prüfkörper mit **beiden** Messgeräten aus
- Wir haben also pro **Versuchseinheit** (hier: Prüfkörper) zwei Beobachtungen (einmal Gerät A und einmal Gerät B)
- Man spricht auch von **gepaarten Stichproben**
- Die beiden Beobachtungen sind **nicht** unabhängig, da wir an der **gleichen** Versuchseinheit zweimal messen!

Ungepaarte (unabhängige) Stichproben

- Im Beispiel der beiden Herstellungsverfahren nehmen wir eine Stichprobe von Verfahren A und eine andere Stichprobe von Verfahren B und messen jedes Objekt aus

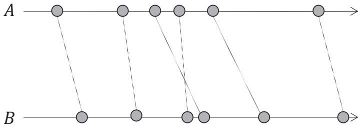
- Die Beobachtungen sind hier **unabhängig**; „es gibt **nichts**, was sie verbindet“

- Man spricht auch von **ungepaarten (oder unabhängigen) Stichproben**

Unterscheidung gepaart versus ungepaarte Stichproben

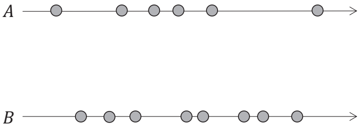
Gepaarte Stichproben

- Jede Beobachtung einer Gruppe kann eindeutig einer Beobachtung der anderen Gruppe zugeordnet werden
- Stichprobengrösse ist in beiden Gruppen zwangsläufig gleich



Ungepaarte Stichproben

- Keine Zuordnung von Beobachtungen möglich
- Stichprobengrössen können verschieden sein (müssen aber nicht!)
- Man kann die eine Gruppe vergrössern, ohne dass man die andere vergrössert

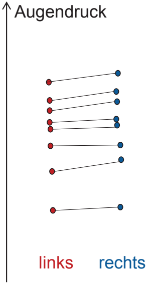


Gepaarte versus ungepaarte Stichproben

- Beispiel: Augeninnendruck; ein Auge behandelt, das andere nicht (gepaarter Test ist angebracht)

- Gemäss Vorraussetzungen dürfte auch ein ungepaarter Test angewendet werden

$H_0\!: \mu_X = \mu_Y$



Ungepaart:

Intuition Teststatistik: $T = \frac{\overline{X} - \overline{Y}}{\widehat{\sigma_{\overline{X}}}}$

Gepaart:

Differenz $D_i = X_i - Y_i$

Teststatistik $T = \frac{\overline{D}}{\widehat{\sigma_{\overline{D}}}}$

Statistischer Test für gepaarte Stichproben mit **R**

- **Gepaarte Stichproben:** Für $X_i \sim \mathcal{N}(\mu_X, \sigma_X^2)$ und $Y_i \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ betrachten wir die Differenzen $D_i = X_i - Y_i$
- Wir führen einen t-Test durch mit der Teststatistik D_i - normalerweise für die Nullhypothese $\mathrm{E}[D] = \mu_D = 0$, also kein Unterschied

R-Befehl: t.test(...,paired=TRUE)

```
> vorher <- c(25,25,27,44,30,67,53, 53,52,60,28)
> nachher <- c(27,29,37,56,46,82, 57,80,61,59,43)
> t.test(nachher, vorher, altervative = "two.sided",
mu = 0, paired = TRUE, conf.level = 0.95)
```

Statistischer Test für gepaarte Stichproben mit **R**

R-Befehl: t.test(....,paired=TRUE)

```
Paired t-test data: nachher and vorher
t = 4.2716, df = 10, p-value = 0.001633
alternative hypothesis: true difference in means
is not equal to 0
95 percent confidence interval:
4.91431 15.63114
sample estimates:
mean of the differences
10.27273
```

- Unterschied ist also auf dem 5% Signifikanzniveau signifikant, weil der P-Wert kleiner als 5% ist.
- 95%-Vertrauensintervall des Unterschieds in den Gruppenmittelwerten: Mit 95% Wahrscheinlichkeit ist der Gruppenmittelwert von $\mathtt{x}$ um eine Zahl im Bereich $[4.91431, 15.63114]$ grösser als der Gruppenmittelwert von $\mathtt{y}$

Statistischer Test für ungepaarte Stichproben mit **R**

- **Ungepaarte Stichproben:** Falls Daten X_i und Y_i normalverteilt sind, aber ungepaart

- Beispiel: Schmelzwärme von Eis: Wir berechnen den Zwei-Stichproben t-Test für ungepaarte Stichproben mit Nullhypothese $\mu_X = \mu_Y$:

R-Befehl: t.test(...,paired = FALSE)

```
> x <- c(79.98, 80.04, 80.02, 80.04, 80.03, 80.03, 80.04,
79.97, 80.05, 80.03, 80.02, 80.00, 80.02)

> y <- c(80.02, 79.94, 79.98, 79.97, 80.03, 79.95, 79.97)

> t.test(x, y, altervative = "two.sided", mu = 0, paired =
FALSE, , conf.level = 0.95)
```

Statistischer Test für ungepaarte Stichproben mit **R**

R-Befehl: T

```
wo Sample t-test
data: x and y
t = 3.4722, df = 19, p-value = 0.002551
alternative hypothesis: true difference in
means is not equal to 0
95 percent confidence interval:
0.01669058 0.06734788
sample estimates:
mean of x mean of y
80.02077 79.97875
```

- Unterschied ist also auf dem 5% Signifikanzniveau signifikant, weil der P-Wert kleiner als 5% ist.
- 95%-Vertrauensintervall des Unterschieds in den Gruppenmittelwerten: Mit 95% Wahrscheinlichkeit ist der Gruppenmittelwert von `x` um eine Zahl im Bereich $[0.0167, 0.0673]$ grösser als der Gruppenmittelwert von `y`

Mann-Whitney U-Test (aka Wilcoxon Rank-sum Test)

- Falls Daten nicht normalverteilt
- $X_i \sim F, \; i = 1, \ldots, n; \; Y_j \sim G, \; j = 1, \ldots m$
  $H_0 : \; F = G$
  $H_A : \; F = G + \delta \; (\delta \neq 0)$ (oder einseitig)
  (d.h., Verteilungen sind verschoben, haben aber gleiche From)

R-Befehl: wilcox.test(...,paired=FALSE)

```
> x <- c(79.98, 80.04, 80.02, 80.04, 80.03, 80.03, 80.04,
79.97, 80.05, 80.03, 80.02, 80.00, 80.02)

> y <- c(80.02, 79.94, 79.98, 79.97, 80.03, 79.95, 79.97)

> wilcox.test(x, y, altervative = "two.sided", mu = 0, paired =
FALSE, , conf.level = 0.95)
```

Übersicht: Tests für ungepaarte Stichproben

Test	**Annahme**				**$n_{\min}$ falls ($n = m$) bei $\alpha = 0.05$**	**Macht für ein Beispiel (1)**
	$\sigma_X = \sigma_Y$	$X_i \sim N$ $Y_i \sim N$	F, G haben gleiche Form	iid pro Gruppe		
t ($\sigma_X = \sigma_Y$)	x	x	x	x	2	57%
t ($\sigma_X \neq \sigma_Y$)		x		x	2	56%
MW U-Test	x		x	x	4	53%

(1): $X_i \sim \mathcal{N}(\mu_X, \sigma^2)$, $Y_i \sim \mathcal{N}(\mu_Y, \sigma^2)$ $n = m = 10$; $H_0 : \mu_X = \mu_Y$; $H_A : \mu_X \neq \mu_Y$; $\alpha = 0.05$
Macht berechnet für konkrete Alternative: $X_i \sim \mathcal{N}(0, 1)$, $Y_i \sim \mathcal{N}(1, 1)$

Übersicht Statistische Tests (stetige Verteilungen)