

04

Statistische Tests

Dominic Schmitz & Janina Esser

Statistische Tests

- Einfachster Teil der **inferenziellen Statistik**:
wir nehmen unsere Daten und leiten etwa aus ihnen ab
- Geschieht meist anhand des „**Null-Hypothesis Significance Testing**“
- Resultat ist oftmals der berühmte ***p-Wert*** (*probability value*)

Tests für verschiedene Datentypen

Ziel	normalverteilte Daten	nicht normalverteilte Daten	kategoriale Daten
Beschreibung einer Gruppe	Durchschnitt, Standardabweichung	Median, Interquartial-Spannweite	Proportion
Testen auf Normalverteilung	Shapiro-Wilk Test		-
Vergleich zweier unabhängiger Gruppen	t-Test independent samples	Wilcoxon-Mann-Whitney Test	Fisher's Test
Vergleich zweier abhängiger Gruppen	t-Test dependent samples	Wilcoxon Signed-Rank Test	McNemar's Test
Vergleich dreier oder mehr unabhängiger Gruppen	ANOVA	Kruskal-Wallis Test	Chi-Quadrat Test
Vergleich dreier oder mehr abhängiger Gruppen	ANOVA	Friedman Test	Cochrane Q
Korrelation	Pearson	Spearman	

Tests für verschiedene Datentypen

Ziel	normalverteilte Daten	nicht normalverteilte Daten	kategoriale Daten
Beschreibung einer Gruppe	Durchschnitt, Standardabweichung	Median, Interquartial-Spannweite	Proportion
Testen auf Normalverteilung	Shapiro-Wilk Test		-
Vergleich zweier unabhängiger Gruppen	t-Test independent samples	Wilcoxon-Mann-Whitney Test	Fisher's Test
Vergleich zweier abhängiger Gruppen	t-Test dependent samples	Wilcoxon Signed-Rank Test	McNemar's Test
Vergleich dreier oder mehr unabhängiger Gruppen	ANOVA	Kruskal-Wallis Test	Chi-Quadrat Test
Vergleich dreier oder mehr abhängiger Gruppen	ANOVA	Friedman Test	Cochrane Q
Korrelation	Pearson	Spearman	

Tests für verschiedene Datentypen

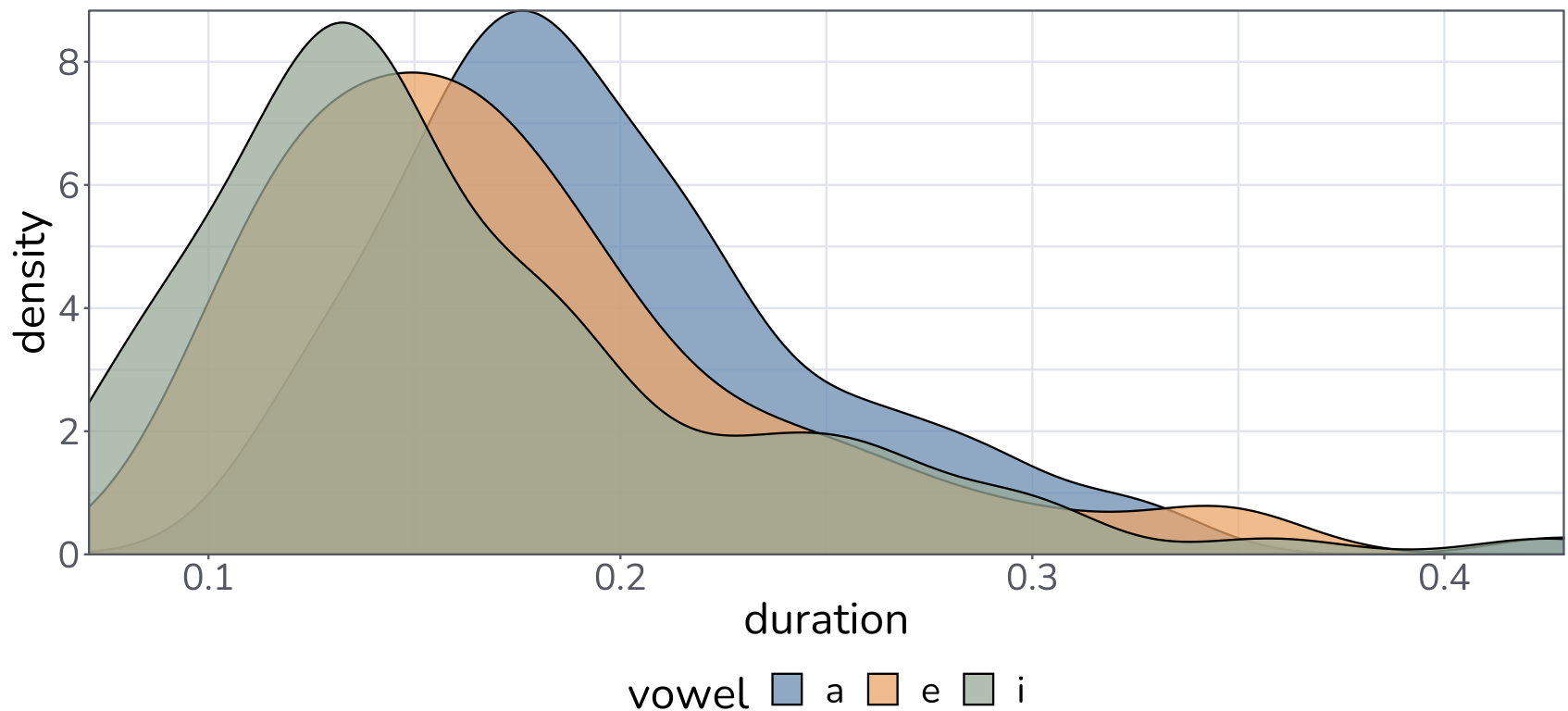
Ziel	normalverteilte Daten	nicht normalverteilte Daten	kategoriale Daten
Beschreibung einer Gruppe	Durchschnitt, Standardabweichung	Median, Interquartial-Spannweite	Proportion
Testen auf Normalverteilung	Shapiro-Wilk Test		-
Vergleich zweier unabhängiger Gruppen	t-Test independent samples	Wilcoxon-Mann-Whitney Test	Fisher's Test
Vergleich zweier abhängiger Gruppen	t-Test dependent samples	Wilcoxon Signed-Rank Test	McNemar's Test
Vergleich dreier oder mehr unabhängiger Gruppen	ANOVA	Kruskal-Wallis Test	Chi-Quadrat Test
Vergleich dreier oder mehr abhängiger Gruppen	ANOVA	Friedman Test	Cochrane Q
Korrelation	Pearson	Spearman	

Shapiro-Wilk Test

- mit einem **Shapiro-Wilk Test** kann man feststellen, ob eine Stichprobe **normalverteilt** ist
- diese Info ist wichtig, da verschiedene andere Tests nur dann funktionieren, wenn Daten (annähernd) normalverteilt sind
- Daten müssen voneinander unabhängig sein; die Datenmenge sollte zwischen 3 und 5000 liegen
- als Beispiel nutzen wir das „Vowel Shortening in German“ Datenset aus dem SfL Package

Shapiro-Wilk Test

- Sind die Vokaldauern von /a/, /e/ und /i/ normalverteilt?



Shapiro-Wilk Test

- Sind die Vokaldauern von /a/, /e/ und /i/ normalverteilt?
- Der Shapiro-Wilk Test kommt zu folgenden Ergebnissen:

	p-Wert
/a/	$p < 0.001$
/e/	$p < 0.001$
/i/	$p < 0.001$

- Da die p -Werte kleiner 0.05 sind, sind die Daten **nicht normalverteilt**

Tests für verschiedene Datentypen

Ziel	normalverteilte Daten	nicht normalverteilte Daten	kategoriale Daten
Beschreibung einer Gruppe	Durchschnitt, Standardabweichung	Median, Interquartial-Spannweite	Proportion
Testen auf Normalverteilung	Shapiro-Wilk Test		-
Vergleich zweier unabhängiger Gruppen	t-Test independent samples	Wilcoxon-Mann-Whitney Test	Fisher's Test
Vergleich zweier abhängiger Gruppen	t-Test dependent samples	Wilcoxon Signed-Rank Test	McNemar's Test
Vergleich dreier oder mehr unabhängiger Gruppen	ANOVA	Kruskal-Wallis Test	Chi-Quadrat Test
Vergleich dreier oder mehr abhängiger Gruppen	ANOVA	Friedman Test	Cochrane Q
Korrelation	Pearson	Spearman	

t-Test

- Es gibt **verschiedene Arten** des t-Tests
- Wichtig dabei:
Stammen meine Daten aus dem gleichen Sample?
- Ja – z.B. falls zwei Experimente mit gleichen TN durchgeführt werden
→ **dependent samples t-test**
- Nein – z.B. falls zwei Experimente mit verschiedenen TN durchgeführt werden
→ **independent samples t-test**

t-Test – dependent samples

- ein Versuch wird n -mal durchgeführt
- ein Parameter wird geändert
- der Versuch wird mit den **gleichen TN** und dem geänderten Parameter erneut durchgeführt
- dann werden die Messergebnisse verglichen

t-Test – dependent samples

- unsere gemessene Variable sei in

Durchführung A: x

Durchführung B: y

- x und y wurden n -mal gemessen x_1, \dots, x_n und y_1, \dots, y_n
- der t-Test geht davon aus, dass x und y (annähernd) **normalverteilt** sind (wichtig!)
- **Frage:** Sind die Werte von x und y signifikant verschieden oder sind sie nur **zufällig** verschieden?

t-Test – dependent samples

- Schritt 1: Durchschnitt von z berechnen

$$\bar{z} = \bar{y} - \bar{x}$$

- Schritt 2: Standardabweichung von z berechnen

$$s := \sqrt{\frac{1}{n-1} \sum_{i=1}^n (z_i - \bar{z})^2}$$

- Schritt 3: t-Wert berechnen

$$t = \frac{\bar{z}}{s} * \sqrt{n}$$

t-Test – dependent samples

- mithilfe des t-Wertes und der Freiheitsgrade kann nun in einer Tabelle die t-Verteilung nachgeschlagen werden
- die Freiheitsgrade sind $df = n - 1$

t-Test – dependent samples

f	90%	95%	97.5%	99%	99.5%	99.9%
1	3.078	6.314	12.706	31.821	63.657	318.309
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.841	10.215
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.610
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
∞	1.282	1.645	1.960	2.326	2.576	3.090

t-Test – dependent samples

- t-Tests können **einseitig** oder **zweiseitig** sein
- es sind μ_1 und μ_2 die unbekannten wahren Erwartungswerte der beiden Stichproben
- bei **zweiseitigen** t-Tests ist die Nullhypothese von der Form

$$H_0 = \{\mu_1 \neq \mu_2\}$$

- bei **einseitigen** t-Tests ist die Nullhypothese von der Form

$$H_0 = \{\mu_1 > \mu_2\} \text{ oder } H_0 = \{\mu_1 < \mu_2\}$$

t-Test – dependent samples

- bei **zweiseitigen** t-Tests ist die Nullhypothese von der Form

$$H_0 = \{\mu_1 \neq \mu_2\}$$

- bei **zweiseitigen** t-Tests wissen wir nicht, ob x oder y im Durchschnitt größer ist; der Test ist **ungerichtet**

- bei **einseitigen** t-Tests ist die Nullhypothese von der Form

$$H_0 = \{\mu_1 > \mu_2\} \text{ oder } H_0 = \{\mu_1 < \mu_2\}$$


- bei **einseitigen** t-Tests wissen wir bereits, dass x größer/kleiner y ist; der Test ist **gerichtet**

t-Test – dependent samples

- das **Signifikanzniveau** sei $\alpha = 0.05$
- mit t-Wert, Freiheitsgraden und Signifikanzniveau können wir nun berechnen
- für **zweiseitige** t-Tests: $t_{n-1, 1-\frac{\alpha}{2}}$
- für **einseitige** t-Tests: $t_{n-1, 1-\alpha}$ bzw. $-t_{n-1, 1-\alpha}$

t-Test – dependent samples

Beispiel: Blutdruck

$$df = 10 - 1$$


Blutdruck	1	2	3	4	5	6	7	8	9	10
Placebo x	168	184	172	173	150	155	163	164	151	146
Medikament y	176	145	150	163	136	168	164	139	145	112
Differenz z	8	-39	-22	-10	-14	13	1	-25	-6	-34

- $\bar{z} = -12.8$
- $s = 17.36$
- $t = -2.332$

t-Test – dependent samples

- für **einseitige** t-Tests:

$$t_{n-1,1-\alpha} \text{ bzw. } -t_{n-1,1-\alpha}$$

- die Nullhypothese wird abgelehnt, wenn

$$t < -t_{n-1,1-\alpha}$$

- für unser Blutdruckbeispiel:

$$-t_{n-1,1-\alpha} = -t_{9,0.95}$$

t-Test – dependent samples

$$-t_{9,0.95}$$

f	90%	95%	97.5%	99%	99.5%	99.9%
1	3.078	6.314	12.706	31.821	63.657	318.309
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.841	10.215
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.610
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
∞	1.282	1.645	1.960	2.326	2.576	3.090

t-Test – dependent samples

- also, stimmt es nun, dass

$$t < -t_{n-1, 1-\alpha}$$

ist?

- ja, denn

$$-2.332 < -1.833$$

- damit ist die Wirksamkeit des Medikaments zum Signifikanzniveau $\alpha = 0.05$ nachgewiesen

t-Test – independent samples

- ein Versuch wird n -mal durchgeführt
- ein Parameter wird geändert
- der Versuch wird mit **anderen** TN und dem geänderten Parameter erneut durchgeführt
- da wir verschiedene Probandengruppen haben, kann $n_1 \neq n_2$ zutreffen
- dann werden die Messergebnisse verglichen

t-Test – independent samples

- unsere gemessene Variable sei in
Durchführung A: x
Durchführung B: y
- x und y wurden n -mal gemessen x_1, \dots, x_n und y_1, \dots, y_n
- der t-Test geht davon aus, dass x und y (annähernd) **normalverteilt** sind (wichtig!)
- **Frage:** Sind die Werte von x und y verschieden oder sind sie nur **zufällig** verschieden?

t-Test – independent samples

- Schritt 1: Durchschnitt von x und y berechnen
- Schritt 2: Standardabweichung von x und y berechnen
- Schritt 3: Standardabweichung von $x + y$ berechnen

$$s_p = \sqrt{\frac{(n_y - 1) * s_x^2 + (n_x - 1) * s_y^2}{n_y + n_x - 2}}$$

- Schritt 4: t-Wert berechnen

$$t = \frac{\bar{y} - \bar{x}}{s_p} * \sqrt{\frac{n_y * n_x}{n_y + n_x}}$$

t-Test – independent samples

- das **Signifikanzniveau** sei $\alpha = 0.05$
- mit t-Wert, Freiheitsgraden und Signifikanzniveau können wir nun berechnen
- für **zweiseitige** t-Tests: $t_{n_y+n_x-2, 1-\frac{\alpha}{2}}$
- für **einseitige** t-Tests: $t_{n_y+n_x-2, 1-\alpha}$ bzw. $-t_{n_y+n_x-2, 1-\alpha}$

t-Test – independent samples

Beispiel: f0 bei Männern

f0	1	2	3	4	5	6	7	8	9	10
Gruppe 1 x	55	69	64	70	75	70	83	69	75	69
Gruppe 2 y	61	60	62	58	75	63	52	66	59	

- $n_y = 10, n_x = 9$
- $\bar{x} = 69.00, \bar{y} = 61.78$
- $s_x = 7.972, s_y = 6.280$

$$s_p = 7.226$$

$$t = -2.175$$

t-Test – independent samples

- also, stimmt es nun, dass

$$t < -t_{17,0.95}$$

ist?

- ja, denn

$$-2.175 < -1.740$$

- damit ist die f0 der zweiten Gruppe zum Signifikanzniveau $\alpha = 0.05$ nachgewiesen tiefer

Tests für verschiedene Datentypen

Ziel	normalverteilte Daten	nicht normalverteilte Daten	kategoriale Daten
Beschreibung einer Gruppe	Durchschnitt, Standardabweichung	Median, Interquartial-Spannweite	Proportion
Testen auf Normalverteilung	Shapiro-Wilk Test		-
Vergleich zweier unabhängiger Gruppen	t-Test independent samples	Wilcoxon-Mann-Whitney Test	Fisher's Test
Vergleich zweier abhängiger Gruppen	t-Test dependent samples	Wilcoxon Signed-Rank Test	McNemar's Test
Vergleich dreier oder mehr unabhängiger Gruppen	ANOVA	Kruskal-Wallis Test	Chi-Quadrat Test
Vergleich dreier oder mehr abhängiger Gruppen	ANOVA	Friedman Test	Cochrane Q
Korrelation	Pearson	Spearman	

Chi-Quadrat-Test

- mit **Chi-Quadrat-Tests** können wir bestimmen, ob zwei kategoriale Variablen zusammenhängen
- als Beispiel nutzen wir das „Age and Looks“ Datenset aus dem SfL Package

	blue	brown	green
blonde	3	7	3
brunette	5	15	2
red	1	3	1

Chi-Quadrat-Test

- nun können wir mit einem **Chi-Quadrat-Test** testen, ob Haar- und Augenfarbe in unserem Sample zusammenhängen
- Ergebnis: $p = 0.84 > 0.05$, d.h. nein, kein Zusammenhang

	blue	brown	green
blonde	3	7	3
brunette	5	15	2
red	1	3	1

Tests für verschiedene Datentypen

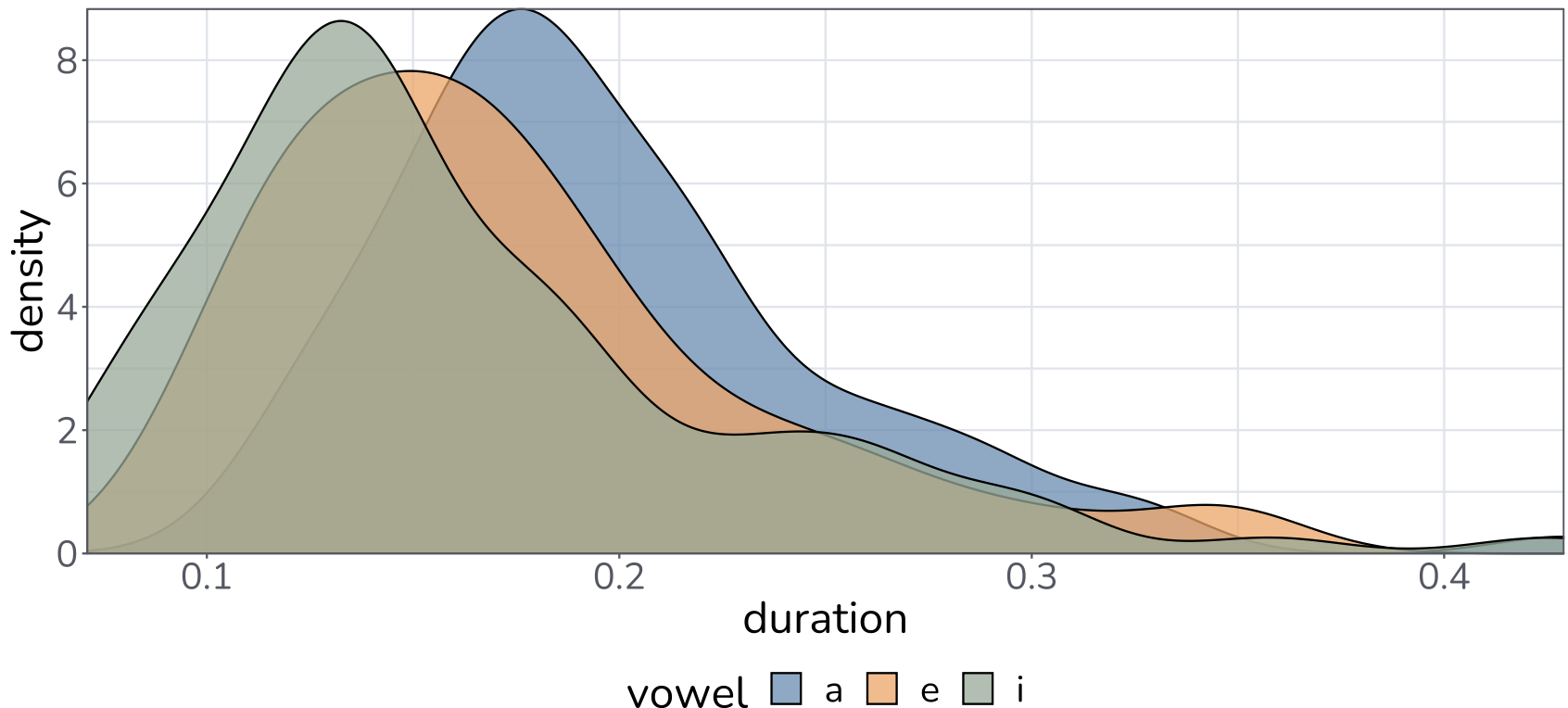
Ziel	normalverteilte Daten	nicht normalverteilte Daten	kategoriale Daten
Beschreibung einer Gruppe	Durchschnitt, Standardabweichung	Median, Interquartial-Spannweite	Proportion
Testen auf Normalverteilung	Shapiro-Wilk Test		-
Vergleich zweier unabhängiger Gruppen	t-Test independent samples	Wilcoxon-Mann-Whitney Test	Fisher's Test
Vergleich zweier abhängiger Gruppen	t-Test dependent samples	Wilcoxon Signed-Rank Test	McNemar's Test
Vergleich dreier oder mehr unabhängiger Gruppen	ANOVA	Kruskal-Wallis Test	Chi-Quadrat Test
Vergleich dreier oder mehr abhängiger Gruppen	ANOVA	Friedman Test	Cochrane Q
Korrelation	Pearson	Spearman	

Wilcoxon-Mann-Whitney Test

- Reminder: t-Tests setzen eine (annähernde) Normalverteilung der Daten voraus
- der **Wilcoxon-Mann-Whitney Test** kann auch mit **nicht-normalverteilten Daten** umgehen
- als Beispiel nutzen wir das „Vowel Shortening in German“ Datenset aus dem SfL Package

Wilcoxon-Mann-Whitney Test

- die Vokaldauern von /a/, /e/ und /i/ sind nicht normalverteilt (siehe Shapiro-Wilk Test)



Wilcoxon-Mann-Whitney Test

- Ergebnis:
ja, die Vokale haben unterschiedliche Dauern

	/a/ vs. /e/	/a/ vs. /i/	/e/ vs. /i/
t-Test	<0.001	<0.001	0.00568
WMW-Test	<0.001	<0.001	0.00241

Tests für verschiedene Datentypen

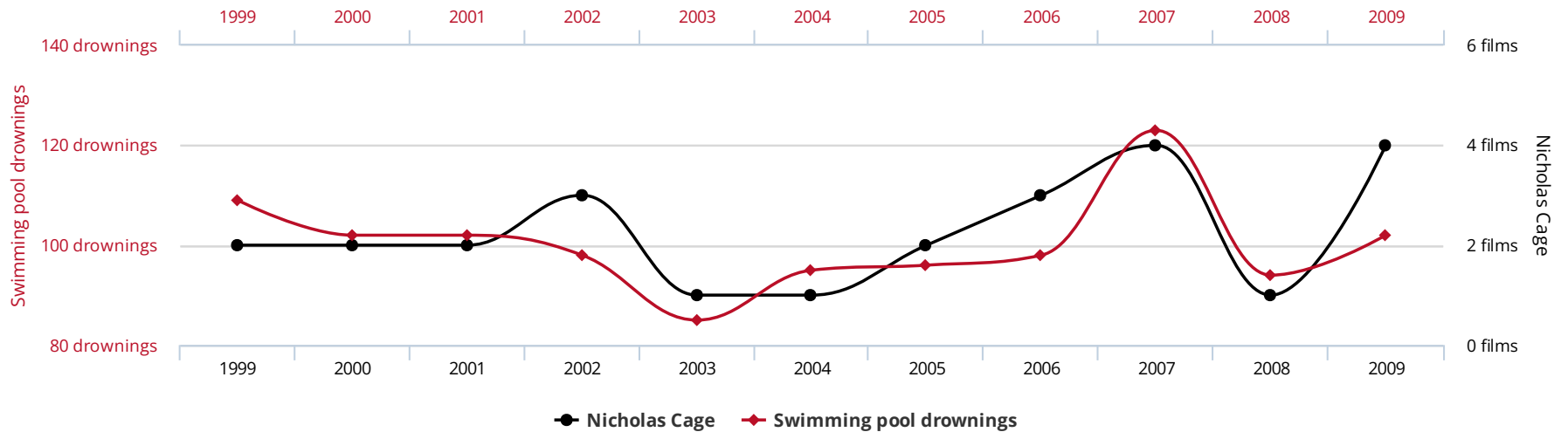
Ziel	normalverteilte Daten	nicht normalverteilte Daten	kategoriale Daten
Beschreibung einer Gruppe	Durchschnitt, Standardabweichung	Median, Interquartial-Spannweite	Proportion
Testen auf Normalverteilung	Shapiro-Wilk Test		-
Vergleich zweier unabhängiger Gruppen	t-Test independent samples	Wilcoxon-Mann-Whitney Test	Fisher's Test
Vergleich zweier abhängiger Gruppen	t-Test dependent samples	Wilcoxon Signed-Rank Test	McNemar's Test
Vergleich dreier oder mehr unabhängiger Gruppen	ANOVA	Kruskal-Wallis Test	Chi-Quadrat Test
Vergleich dreier oder mehr abhängiger Gruppen	ANOVA	Friedman Test	Cochrane Q
Korrelation	Pearson	Spearman	

Korrelation

- die **Korrelation** beschreibt eine Beziehung zwischen zwei oder mehr Variablen
- Korrelation **bedeutet nicht** Kausalität!
 - zwei Variablen können korreliert sein
 - ohne dabei in kausaler Verbindung zu stehen

Korrelation

Number of people who drowned by falling into a pool
correlates with
Films Nicolas Cage appeared in



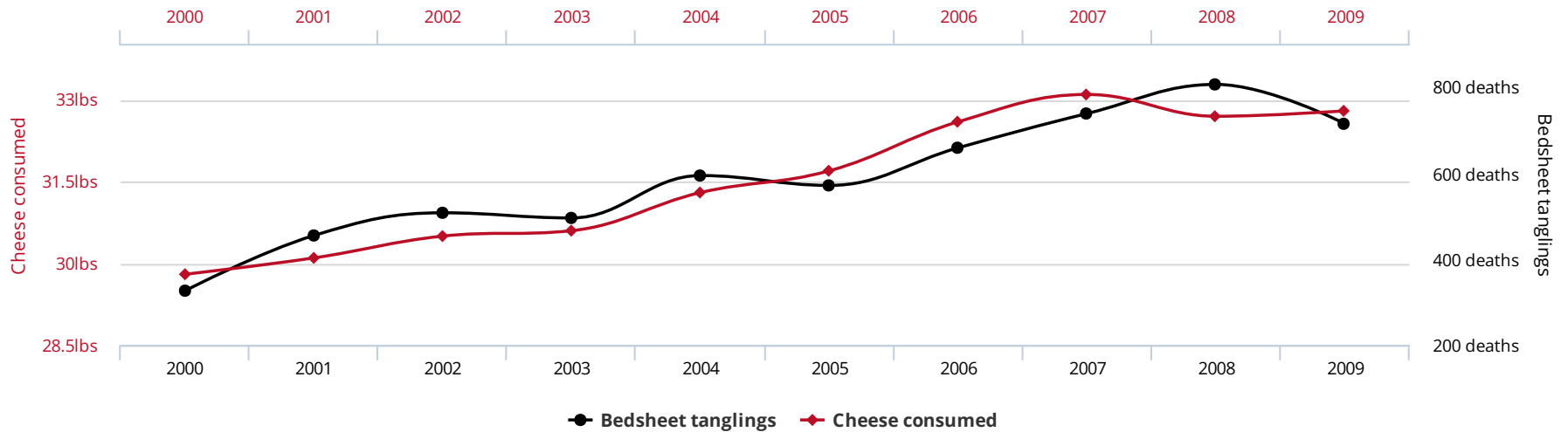
tylervigen.com

Korrelation

Per capita cheese consumption

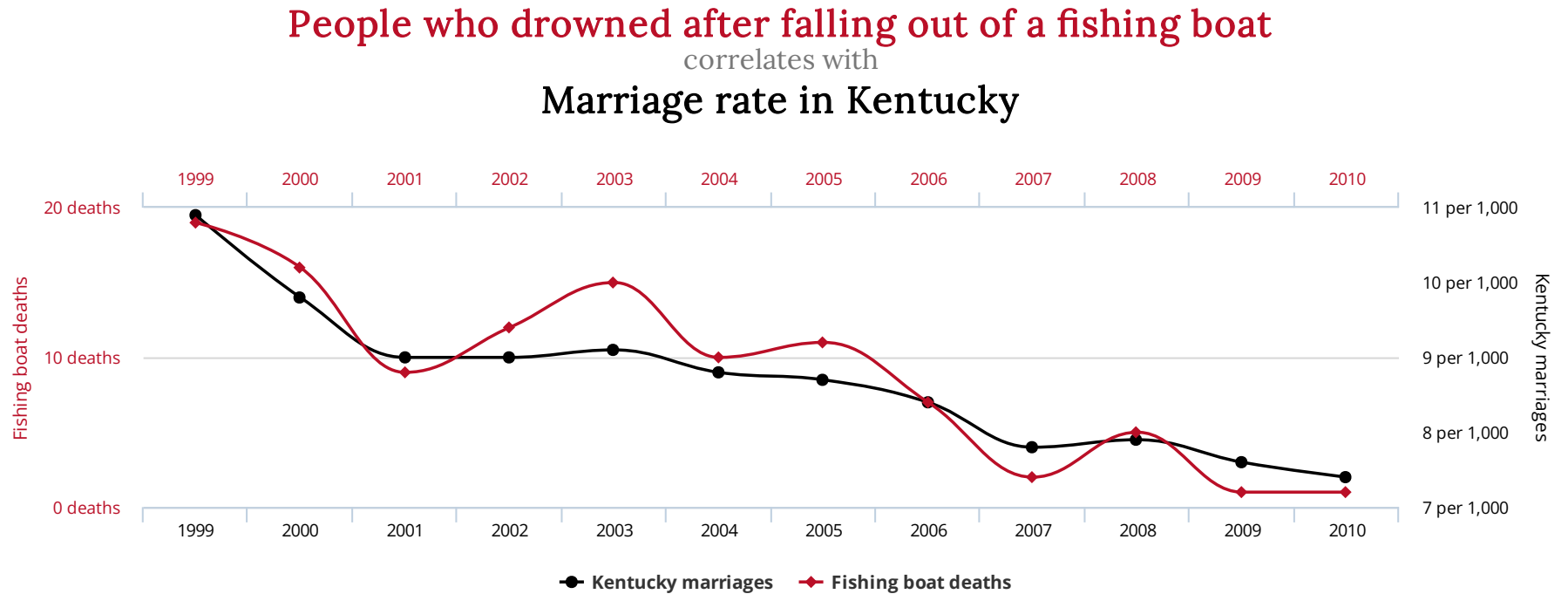
correlates with

Number of people who died by becoming tangled in their bedsheets



tylervigen.com

Korrelation



tylervigen.com

Korrelation

- sind die zu vergleichenden Daten normalverteilt und numerisch, nutzen wir **Pearson's r**
- sind die zu vergleichenden Daten nicht normalverteilt und/oder nicht numerisch, nutzen wir **Spearman's rho**
- als Beispiel nutzen wir das „Duration of word-final /s/ in English“ Datenset aus dem SfL Package

Korrelation

- Wann sprechen wir von Korrelation?

→ 4-stufige Version

Korrelationskoeffizient			Label	Richtung
0.7	< r ≤	1.0	sehr hoch	positive Korrelation
0.5	< r ≤	0.7	hoch	
0.2	< r ≤	0.5	mittel	
0.0	< r ≤	0.2	niedrig	
r ≈ 0			keine Korrelation	
0.0	> r ≥	-0.2	niedrig	negative Korrelation
-0.2	> r ≥	-0.5	mittel	
-0.5	> r ≥	-0.7	hoch	
-0.7	> r ≥	-1.0	sehr hoch	

Korrelation

- Wann sprechen wir von Korrelation?

→ 3-stufige Version

Korrelationskoeffizient			Label	Richtung
0.6	< r ≤	1.0	hoch	positive Korrelation
0.3	< r ≤	0.6	mittel	
0.0	< r ≤	0.3	niedrig	
r ≈ 0			keine Korrelation	
0.0	> r ≥	-0.3	niedrig	negative Korrelation
-0.3	> r ≥	-0.6	mittel	
-0.6	> r ≥	-1.0	hoch	

Korrelation

- Wann sprechen wir von Korrelation?

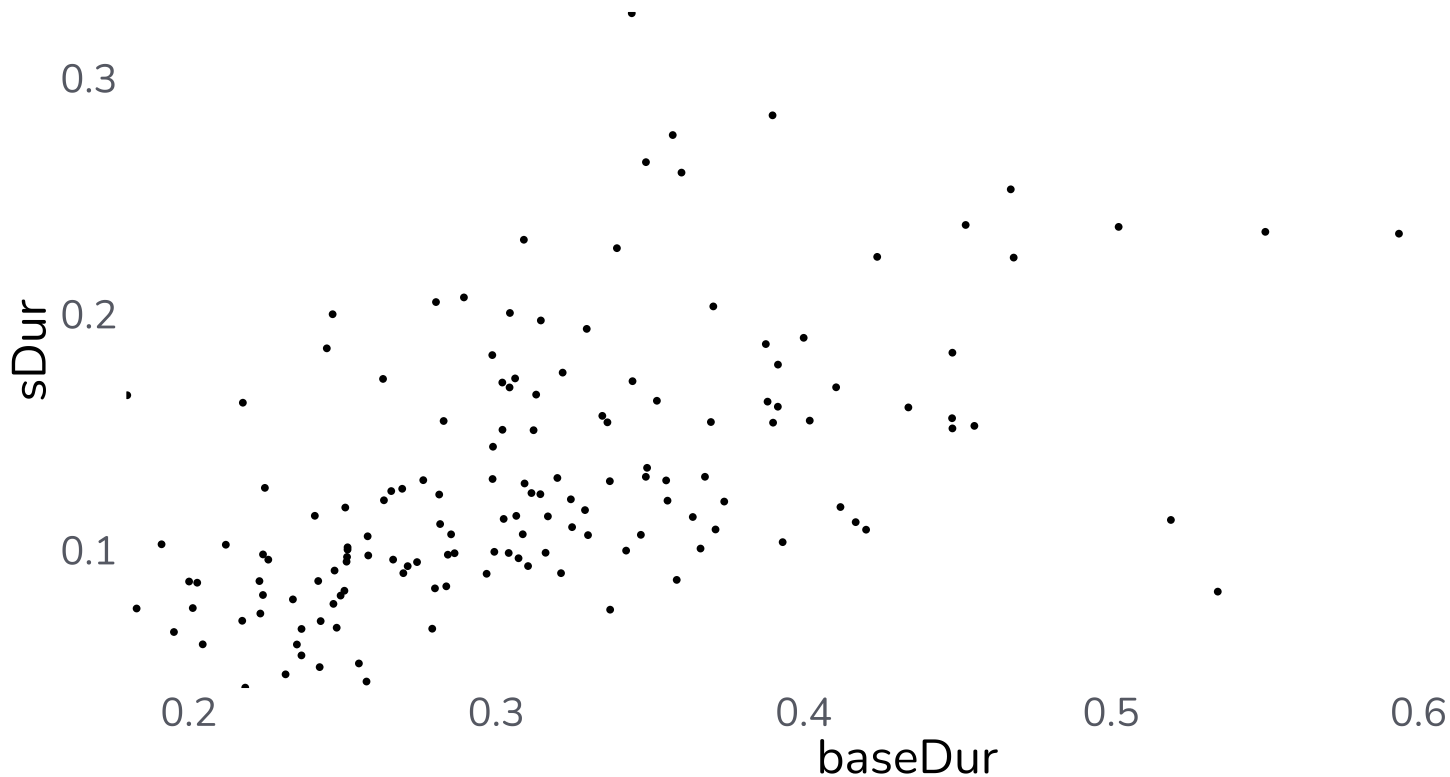
→ 2-stufige Version

Korrelationskoeffizient			Label	Richtung
0.6	< r ≤	1.0	hoch	positive Korrelation
0.2	< r ≤	0.5	mittel	
-0.2 ≤ r ≤ 0.2			niedrig bis keine Korrelation	
-0.2	> r ≥	-0.5	mittel	negative Korrelation
-0.5	> r ≥	-1.0	hoch	

- generell gilt: es gibt so viele Versionen wie wissenschaftliche Aufsätze

Korrelation

Frage: sind /s/-Dauer und base-Dauer korreliert?



Antwort: ja, da $r = 0.47$

Tests für verschiedene Datentypen

Ziel	normalverteilte Daten	nicht normalverteilte Daten	kategoriale Daten
Beschreibung einer Gruppe	Durchschnitt, Standardabweichung	Median, Interquartial-Spannweite	Proportion
Testen auf Normalverteilung	Shapiro-Wilk Test		-
Vergleich zweier unabhängiger Gruppen	t-Test independent samples	Wilcoxon-Mann-Whitney Test	Fisher's Test
Vergleich zweier abhängiger Gruppen	t-Test dependent samples	Wilcoxon Signed-Rank Test	McNemar's Test
Vergleich dreier oder mehr unabhängiger Gruppen	ANOVA	Kruskal-Wallis Test	Chi-Quadrat Test
Vergleich dreier oder mehr abhängiger Gruppen	ANOVA	Friedman Test	Cochrane Q
Korrelation	Pearson	Spearman	