Statistische Tests

Vitaly Belik

Institut für Veterinär-Epidemiologie und Biometrie, FU Berlin

10/27/2019

Was haben wir letztes mal gelernt?

In der letzen Vorlesung haben wir über die über *Schliessende Statistik* gersprochen.

Test-Teorie

Punkt- und Intervallschätzer

t-Test (Vergleich zwei Mittelwerte)

Hypothesen H_0 und H_1

Nullhypothese ist normalerweise die Behauptung, dass eine Behandlung oder Maßnahme in einem Versuch *keine Auswirkungen* hat und jegliche Unterschiede zwischen den Messwerten nur durch *Zufall* entstanden sind.

- Die Test-Theorie stellt eine Verbindung zwischen Stichtproben und Grundgesamtheit.
- Es wird geprüft, aufgrund von Stichprobenwerten, ob gewisse *Hypothesen* über die Grundgesamtheit wahr sind oder nicht.
- Es soll entschieden werden ob eine Hypothese beizubehalten oder zu verwerfen ist.

Fehler I. und II. Art (α- und β-Fehler)

α-Fehler (1. Art)

Durch den Test wird die *Nullhypothese verworfen, obwohl sie in Wirklichkeit richtig* ist (*false negatives*).

β-Fehler (2. Art)

Die *Nullhypothese wird beibehalten, obwohl sie in Wirklichkeit falsch* ist (*false positives*).

Klassifikation der statistischen Tests

THE TABLE: Systematic of statistical tests and guide lines on how to choose a test.

Choose adequate combination based on row 1, columns A and B/C.

⇨ If possible, use non-parametric approaches

⇨ Examples of other statistical approaches: cluster, discriminant, principal component, path, time series analysis, ...

	A	B: explanatory variables		C: explanatory variables		
			Nonparametric tests		**Parametric tests**	
1			Outcome variable: ordinal scale		Outcome variable: interval or ratio scale possibly after transformation	Outcome variable: other distributions
			Residuals „symmetrical“		Residuals normally distributed	Residuals follow other distribution (e.g. Poisson, Binomial)
					linear models	**generalised linear models**
	Non-paired, independent fixed effects only	1 factor 2 levels	Mann-Whitney-U-test (Wilcoxon, rank sum test)	1 factor 2 levels	t-Test for independent data	🡻
		1 factor >2 levels	Kruskal-Wallis-test	≥1 factor >2 levels	ANOVA (analysis of variance, F-test)	🡻
		1 factor >2 ordered levels	Jonkheere-trend-test	≥1 factor >2 ordered levels	Analysis of variance with ordered factors or corresponding contrasts	🡻
		1 ordinally scaled	Spearman- , Kendall correlation	1 continuous	Pearson-correlation	–
				≥1 continuous	Regression	🡻
				≥1 any type in combination	Linear models[1]	Poisson-regression Logistic regression
					linear mixed-effects models	**Generalised linear mixed-effects models**
	dependent, repeated, nested additionally: random effects	1 factor 2 levels	Wilcoxon (signed rank test)	1 factor 2 levels	paired t-test	🡻
		1 factor >2 levels	Friedman-test	≥1 factor >2 levels	repeated measures, nested ANOVA	🡻
		1 factor >2 ordered levels	Page-Trend-test	≥1 factor >2 ordered levels	🡻	🡻
		1 ordinally scaled	No test if all data dependent	1 continuous	No test if all data dependent	🡻
				≥1 continuous	🡻	🡻
				≥1 any type in combination	Linear mixed-effects models[1]	Generalised linear mixed-Effects models
	Occurrences		χ^2-test, contingency table			Loglinear models

[1] all models listed above can be considered special cases, 🡻: choose model listed below

Figure 1: Test-Systematik [Quelle: L. Gygax]

Einseitige und zweiseitige Fragestellungen

Einseitige Fragestellung

Wenn schon *vor dem Versuch* feststeht, dass die Abweichung der Messgröße nur in *eine Richtung* möglich oder von Interesse ist. Dann prüft der Test nur, ob eine signifikante Abweichung in diese Richtung nachweisbar ist oder nicht.

- Beispiel. Überlebungsrate von Vieren nach der Bestralung mit Röntgenstrahlen verglichen mit der Kontrolle.

Zweiseitige Fragestellung

Kann in keine Richtung eine Veränderung ausgeschlossen werden, liegt eine *zweiseitige Fragestellung* vor.

- Beispiel. Ertrag nach der Bestralung der Pflanzensamen mit niederen Dosen Röntgenstrahlen verglichen mit der Kontrolle.

Einseitige und zweiseitige Fragestellungen (1)

Beim einseitigen Testen kann man bessere Signifikanzen nachweisen. Der einseitige Test darf aber nur dann angewandt werden, wenn aus theoretischen Erwägungen vor dem Versuch nur eine einseitige Veränderung von Interesse ist.

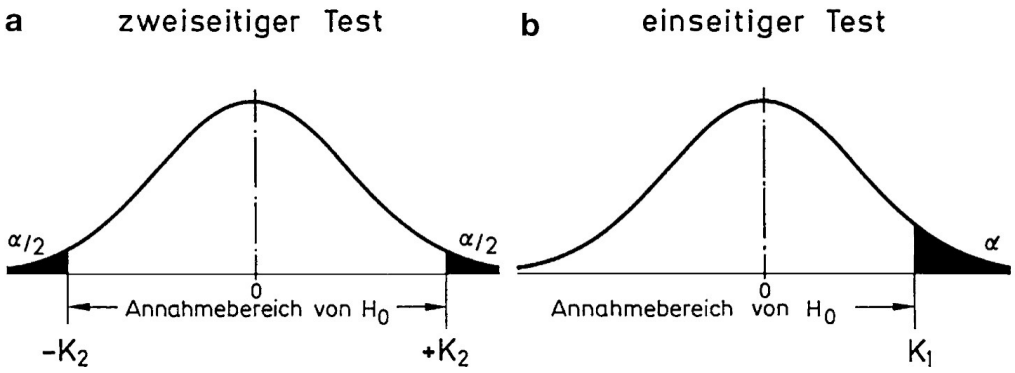

Figure 2: (a) Zweiseitiger Test. (b) einseitiger Test. [Quelle: Köhler et. al]

Teststatistiken

Zur Ermittlung von α und manchmal auch von β, hatten wir für unsere Testgröße basierend auf der Stichprobejeweils eine Wahrscheinlichkeitsverteilung verwendet. Man bezeichnet solche Testgrößen wie i (Äpfel) und $\bar{d}$ (Medikament) als *Prüfstatistiken* (Teststatistiken) und die zugehörigen Verteilungen als Prüfverteilungen.

Die Voraussetzung ist, dass die Stichprobe *zufällig* gezogen wird. Teststatistiken hängen von *Stichprobenumfang* ab. Daraus werden die *Freicheitsgrade* abgeleitet.

Die wichtigsten Verteilungen sind tabelliert oder werden mit Hilfe von Software berechnet.

Vorgehensweise bei statistischen Tests

1. Formulieren der zu überprüfenden Hypothese zu Grundgesamtheiten. Da wir beim Überprüfen der Hypothese nur auf die Stichproben zurückgreifen können, ist unsere Entscheidung fehlerhaft.

2. Wir legen die maximale Irrtumswahrscheinlichkeit in Form des *Signifikanzniveaus* α fest.

3. Wir wählen den *geeigneten Test*, der über unsere Hypothese entscheiden kann und zu den Gegebenheiten passt.

4. Wir berechnen die im Test *vorgeschriebenen* Teststatistiken und vergleichen sie mit geeigneten Tabellenwerten. Dabei entscheiden wir uns für die Beibehaltung oder Verwerfung der Nullhypothese.

5. Wenn wir den Test mit Hilfe der Software ausführen, bekommen wir den *P-Wert*. Falls

$P \geq \alpha \Rightarrow H_0$, die Nullhypothese wird beibehalten.

$P < \alpha \Rightarrow H_1$, die Alternativhypothese wird angenommen.

Unter der Bedinging, dass die Nullhypothese gilt, gibt der P-Wert due Überschreitungswahrscheinlichkeit P der aus den Daten berechneten Prüfstatistik.

Statistische Tests

Intervalldaten

Ordinaldaten

Nominale Daten

Test zu intervalskalierten (normalverteilten) Daten

Vergleich eines Mittelwertes mit dem theoretischen Wert (t-Test)

$$\hat{t} = \frac{|\bar{x} - \mu_T|}{s}\sqrt{n}$$

wird mit dem Tabellenwert t_{Tab} der t-Verteilung verglichen für gewünsches α und $FG = n - 1$.

$\hat{t} \leq t_{\text{Tab}} \Rightarrow H_0 \, (\mu = \mu_T)$

$\hat{t} > t_{\text{Tab}} \Rightarrow H_1 \, (\mu \neq \mu_T)$

n ist der Stichprobenumfang

s ist die Standardabweichung der Stichprobe

μ_T ist der theoretischer Mittelwert

$\bar{x}$ ist der arithmetischer Mittelwert der Stichporbe

Vergleich eines Mittelwertes mit dem theoretischen Wert (t-Test) (1)

Beispiel

Stichprobe von $n = 20$, $\bar{x} = 42.0$, und $s = 5.0$.

$\mu_T = 45.0$

$\hat{t} = 2.28$

$t_{\text{Tab}}(FG = 19; \alpha = 0.05) = 2.09$

$\hat{t} > t_{\text{Tab}} \rightarrow H_1(\mu \neq \mu_T)$

$\bar{x}$ weicht von μ_t signifikant ab.

Vergleich zweier Mittelwerte unabhängiger Stichproben (t-Test)

$$\hat{t} = \frac{|\bar{x} - \bar{y}|}{s_D} \sqrt{\frac{n_X \cdot n_Y}{n_X + n_Y}}$$

$$t_{\text{Tab}}(\alpha; FG = n_X + n_Y - 2)$$

Die gemeinsame Standardabweichung

$$s_D = \sqrt{\frac{(n_X - 1) \cdot s_x^2 + (n_Y - 1) \cdot s_y^2}{n_X + n_Y - 2}}$$

$\hat{t} \leq t_{\text{Tab}} \Rightarrow H_0 \, (\mu_X = \mu_Y)$

$\hat{t} > t_{\text{Tab}} \Rightarrow H_1 \, (\mu_X \neq \mu_Y)$

Vergleich zweier Mittelwerte unabhängiger Stichproben (t-Test) (1)

Beispiel

$n_X = 16,\ \bar{x} = 14.5,\ s_X^2 = 4$

$n_Y = 14,\ \bar{y} = 13.5,\ s_Y^2 = 3$

$s_D = 1.88$

$\hat{t} = 2.180$

$t_{\text{Tab}}(FG = 28; \alpha = 0.05) = 2.048$

$\hat{t} > t_{\text{Tab}} \Rightarrow H_1\,(\mu_X \neq \mu_Y)$

Die Mittelwerte der Stichproben sind signifikant verschieden

Vergleich zweier Mittelwerte verbundener Stichproben (t-Test)

Wenn die Annahme der Unabhängigkeit der Stichproben nicht erfüllt ist, spricht man über *verbundene Stichproben*. Z.B. wenn man dieselbe Gruppe von Individuen oder Objekten vor und nach einer Behandlung untersucht. Im Medikamenten-Beispiel haben wir uns gefragt, ob $\hat{d}$ signifikant *von null verschieden* ist.

Vergleich zweier Mittelwerte verbundener Stichproben (t-Test) (1)

$$\hat{t} = \frac{|\bar{d}|}{s_d}\sqrt{n}$$

$t_{\text{Tab}}(FG; \alpha)$

$\hat{t} \leq t_{\text{Tab}} \Rightarrow H_0(\delta = 0)$ oder $H_0\,(\mu_X = \mu_Y)$

$\hat{t} > t_{\text{Tab}} \Rightarrow H_1(\delta \neq 0)$ oder $H_1\,(\mu_X \neq \mu_Y)$

Der unverbundene t-Test bei verbundenen Stichproben seltener zu signifikanten Unterschieden führt als der verbundene (paarige or paired auf English) Test.

Vergelich zweier Mittelwerte verbundener Stichproben (t-Test) (2)

Beispiel

Baum	i	1	2	3	4	5	6	7	8
Jahr	X	36.0	31.5	34.0	32.5	35.0	31.5	31.0	35.5
Jahr	Y	34.0	35.5	33.5	36.0	39.0	35.0	33.0	39.5
Differenzen	d_i	2.0	−4.0	0.5	−3.5	−4.0	−3.5	−2.0	−4.0

Figure 3: Erträge in kg von acht Kirschbäumen in zwei Jahren. [Quelle: Köhler et. al]

Vergelich zweier Mittelwerte verbundener Stichproben (t-Test) (3)

$\bar{x} = 35.7$ $\bar{y} = 35.7$

$\bar{d} = -2.31$ $s_d = 2.33$ $n = 8$

Für verbundene Stichproben:

$\hat{t} = 2.8 > t_{\text{Tab}}(FG = 7; \alpha = 0.05) = 2.365 \Rightarrow H_1$

Es bestehen signifikante Mittelwertunterschiede.

Beim Ignorieren der Verbundenheit der Stichproben:

$\hat{t} = 2.07 < t_{\text{Tab}}(FG = 14; \alpha = 0.05) = 2.145 \Rightarrow H_0$

Vergleich zweier Varianzen (F-Test)

Der Test ist nach R.A. Fisher gennant und beschäftigt sich mit der Frage ob die Schätzwerte der Varianzen s_X^2 und s_Y^2 zwei Stichproben aus verschiedenen (normal verteilten) Grundgesamtheiten unterschiedlich sind.

Varianzquotient

$$\hat{F} = \frac{s_X^2}{s_Y^2}$$

dabei $s_X^2 > s_Y^2$.

$F_{\text{Tab}} = F_{n_Y-1}^{n_X-1}(\alpha)$

$\hat{F} \leq F_{\text{Tab}} \Rightarrow H_0\,(\sigma_X^2 = \sigma_Y^2)$

$\hat{F} > F_{\text{Tab}} \Rightarrow H_1\,(\sigma_X^2 \neq \sigma_Y^2)$

Vergleich zweier Varianzen (F-Test) (1)

Beispiel

$n_X = 16,\ \bar{x} = 14.5,\ s_X^2 = 4$

$n_Y = 14,\ \bar{y} = 13.5,\ s_Y^2 = 3$

$\hat{F} = 1.33$

$F_{13}^{15}(0.05) = 3.05$ (zweiseitiger Test)

$\hat{F} < F_{\text{Tab}} \Rightarrow H_0(\sigma_X^2 = \sigma_Y^2)$

Test zu ordinalskalierten Daten (nicht-parametrische Tests)

U-Test von Mann und Whitney (Wilcoxon-Rangsummen-Test)

Lagevergleich zweier unabhängiger Stichproben

- *Voraussetzung*: Die beiden Grundgesamtheiten sollen stetige Verteilungen von *gleichen Form* haben, die Stichproben seien unabhängig und die Daten mindestens ordinalskaliert.

Es wird eine gemeinsame Rangfolge der $(n_X + n_Y)$ Stichprobenwerte gebildet und daraus werden die Rangsummen R_X und R_Y berechnet.

Danach berechnet man

$$U_X = n_X \cdot n_Y + \frac{n_X(n_X+1)}{2} - R_X$$

$$U_Y = n_X \cdot n_Y + \frac{n_Y(n_Y+1)}{2} - R_Y$$

$\hat{U} = \min(U_X, U_Y)$ und $U_{\text{Tab}}(n_X, n_Y; \alpha)$

$\hat{U} \geq U_{\text{Tab}} \Rightarrow H_0$(Mediane gleich)

$\hat{U} < U_{\text{Tab}} \Rightarrow H_1$(Mediane verschieden)

U-Test von Mann und Whitney (Wilcoxon-Rangsummen-Test) (1)

Der *U*-Test hat geringere Voraussetzungen als *t*-Test und hat die *Effizienz* von 95. Effizienz ist das Verhältnis der Stichprobenumfänge, die in zwei verglichenen Tests zur selben *Güte* (auf English *power* $1 - \beta$ Wahrscheinlichkeit die Nullhypothese zu verwerfen, wenn H_1 erfüllt ist) führen. Also muss man den Stichprobenumfang von *U*-Test erhöhen um gleiche α- und β-Fehler wie im *t*-Tet zu haben.

Bei der Vergabe der Rangzahlen wird bei gleichen Werten (Bindungen oder *ties*) das arithmetische Mittel der zugehörigen Rangplätze vergeben. Bei zu vielen Bindungen benötigt $\hat{U}$ Korrekturen.

Falls man die Annahme der gleichen Verteilung fallen lässt vergleicht man mit *U*-Test die Unterschiede in den Verteilungen feststellen.

U-Test von Mann und Whitney (Wilcoxon-Rangsummen-Test) (2)

Beispiel

1955	Rangzahl						6		8	9	10	11		13		15	16	$R_1 = 88$
	Note X						2.2		2.5	2.9	3.0	3.2		3.8		4.2	4.5	
1975	Note Y	1.3	1.7	2.0	2.0	2.1		2.4					3.6		4.1			
	Rangzahl	1	2	3.5	3.5	5		7					12		14			$R_2 = 48$

Figure 4: Durchschnittsnoten und Ränge zweier Abitur-Jaurgänge. [Quelle: Köhler et. al]

$U_X = 12$ $U_Y = 52 \Rightarrow \hat{U} = 12$

$U_{\text{Tab}}(8, 8; 0.05) = 13$

$\hat{U} < U_{\text{Tab}} \Rightarrow H_1$

Die Unterschiede sind signifikant.

Wilcoxon-Test für Paardifferenz (Wilcoxon's signed-ranks test)

Legevergleich zweier verbundener Stichproben

Fragestellung: sind die Mediane zweier vernbundener Stichproben X und Y signifikant verschieden?

- *Voraussetzung*: Die beiden Grundgesamtheiten sollen stetige Verteilungen von *gleichen Form* haben, die Stichproben seien *verbunden* und die Daten mindestens ordinalskaliert.

Es werden die n Messwertdifferenzen $d_i = x_i - y_i$ $d_i \neq 0$ und die Ränge $r(|d_i|)$ gebildet.

Es werden die die Summen der Ränge der positiven Differenzen (W^+) und negativen differenzen (W^-) gebildet.

$\hat{W} = \min(W^+, W^-)$ und $W_{\text{tab}}(n; \alpha)$

$\hat{W} \geq W_{\text{Tab}} \Rightarrow H_0(\text{Mediane gleich})$ $\hat{W} < W_{\text{Tab}} \Rightarrow H_1(\text{Mediane verschieden})$

Test zu nominallskalierten Daten (wird später behandelt)