Willkommen zur Vorlesung Statistik (Master)

Thema dieser Vorlesung:
Verteilungsfreie Verfahren

Prof. Dr. Wolfgang Ludwig-Mayerhofer

Universität Siegen – Philosophische Fakultät, Seminar für Sozialwissenschaften

Nicht-parametrische Verfahren: Warum?

Viele statistische Tests beziehen sich auf die Parameter der Grundgesamtheit – z. B. das arithmetische Mittel. Diese Tests sind meist an Voraussetzungen gebunden, nicht zuletzt die Annahme der Normalverteilung der gemessenen Größen.

Ist diese Annahme nicht gegeben, kann man Tests verwenden, die sich nicht auf bestimmte Parameter beziehen. Da sie auch keine Annahme hinsichtlich der Verteilung der untersuchten Variablen machen (etwa Normalverteilungsannahme), heißen Sie auch „verteilungsfreie Verfahren“.

Nicht-parametrische Verfahren: Welche

Die Zahl nichtparametrischer Testverfahren ist groß.

Wir besprechen hier zunächst zwei einfache Verfahren, die in folgenden Fällen angewendet werden können:

- Unabhängige Stichproben
- Die abhängige Variable ist metrisch, aber nicht normalverteilt, oder sie ist rangskaliert (mit echten Rängen – was einige Bindungen nicht ausschließt).
- Die unabhängige Variable bezeichnet eine Gruppenzugehörigkeit.

Außerdem wird der χ^2-Test aus Statistik I wiederholt.

Zwei Gruppen: Wilcoxons Rangsummen-Test

Statt der Originaldatenwerte werden die Rangplätze der Messwerte betrachtet. Geprüft wird die Abweichung der Summe der Rangplätze von jener Summe, die zu erwarten wäre, wenn keine Unterschiede zwischen den Gruppen bestünden.

Die Größe T_W (siehe übernächste Seite) folgt einer (Standard-) Normalverteilung, jedenfalls bei ausreichenden Gruppengrößen (manche Autoren: mindestens 8, andere: mindestens 20, wieder andere: mindestens 25 Fälle in einer Gruppe) und nicht zu vielen Bindungen (siehe später).

Bei sehr kleinen Fallzahlen gibt es Möglichkeiten der exakten Berechnung der Testgröße (hier nicht besprochen).

Wilcoxon-Test: Beispieldaten

Männer	Rang	Frauen	Rang
620	1	3580	2
3600	3	3840	5
3820	4	3850	6
4180	9	3920	7
4760	11	4160	8
5350	13	4520	10
6900	15	5300	12
7120	16	5690	14
7220	17		
8560	18		
9370	19		
10350	20		

Wilcoxon-Test: Die Teststatistik

$$T_W = \frac{W - W_0}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

mit

W = Summe der Ränge der ersten Gruppe (bzw. kleinere der beiden Rangsummen).

W_0: Bei Gültigkeit der Nullhypothese (gleiche durchschnittliche Rangsumme) zu erwartende Rangsumme für diese Gruppe.

Nenner des Ausdrucks: Standardfehler von $W - W_0$, mit n_1 und n_2 = Umfang der 1. bzw. 2. Gruppe.

Wilcoxon-Test: Berechungsbeispiel

Summe der Ränge in Gruppe 1, hier Gruppe mit niedrigerer Rangsumme (=Frauen):

$$W = \sum_{i=1}^{n_1} rg(X_i) = 64$$

$$W_0 = n_1 \frac{n_1 + n_2 + 1}{2} = 8\frac{21}{2} = 84$$

$$T_W = \frac{W - W_0}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} = \frac{-20}{\sqrt{\frac{8 \cdot 12 \cdot 21}{12}}} = -1{,}543$$

Zum Vergleich: t-Test für gleiche Varianzen: $-1{,}605$, für ungleiche Varianzen: $-1{,}925$.

Wilcoxon-Test: Ranggleichheiten (Ties)

Wenn mehrere Originalmesswerte gleich sind, so spricht man von Bindungen oder Ties. Treten solche Bindungen über beide Gruppen hinweg auf, werden Durchschnittsränge gebildet (z. B. 12,5 statt Rang 12 und 13).

Sind viele Bindungen in den Daten, ist die Formel für den Standardfehler problematisch.

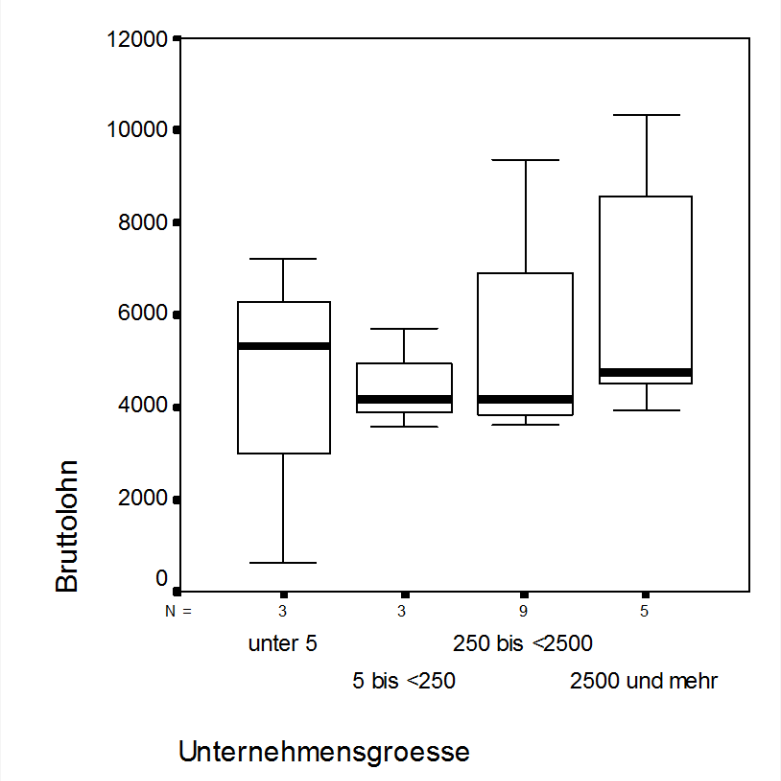
Hinweis: Der U-Test nach Mann und Whitney ist im Resultat identisch mit dem Wilcoxon-Test.

Mehr als zwei Gruppen: Der H-Test nach Kruskal und Wallis

Einkommen nach Unternehmensgröße (letzte Zeile: arithmetische Mittel)

unter 5	5 bis <250	250 bis <2500	2500+
620	3580	3600	3920
5350	4180	3820	4520
7220	5690	3840	4760
		3850	8560
		4160	10350
		5300	
		6900	
		7120	
		9370	
4397	4483	5329	6422

Die Beispiel-Daten graphisch



Kruskal-Wallis-Test: Die Teststatistik

Gegeben sind $i, i = 1 \ldots I$ Gruppen mit den Gruppengrößen n_i und den Summen der Rangplätze in den Gruppen R_i. Die Größe

$$H = \left(\frac{12}{n(n+1)} \sum_{i=1}^{I} \frac{R_i^2}{n_i} \right) - 3(n+1)$$

folgt einer χ^2-Verteilung mit $I - 1$ Freiheitsgraden (außer im Falle von nur drei sehr kleinen Gruppen, für diesen Fall gibt es einen hier nicht besprochenen exakten Text).

Beispieldaten

Einkommen nach Unternehmensgröße (letzte Zeile: arithmetische Mittel)

	unter 5	5 bis <250	250 bis <2500	2500+
	1	2	3	7
	13	9	4	10
	17	14	5	11
			6	18
			8	20
			12	
			15	
			16	
			19	
R	31	25	88	66
R^2	961	625	7744	4356
R^2/n_i	320,33	208,33	860,44	871,20

Der H-Test nach Kruskal und Wallis

Berechnung der Teststatistik:

$$\sum_{i=1}^{4} \frac{R_i^2}{n_i} = 2260{,}31$$

$$H = \left(\frac{12}{20 \cdot 21} \cdot 2260{,}31 \right) - 3 \cdot 21 = 1{,}580$$

Da wir wissen, dass der χ^2-Wert schon bei 1 Freiheitsgrad 3,841 beträgt (und somit der Wert bei 3 Freiheitsgraden größer sein muss), ist klar, dass die Teststatistik bei einem Signifikanzniveau von 5 Prozent nicht im Ablehnungsbereich liegt.

Der H-Test: Bindungen

Kommen in den Daten Bindungen vor, muss ein Korrekturfaktor berechnet werden:

1) Berechne für jede Gruppe von gebundenen Werten die Zahl der gebundenen Werte t_i sowie die Größe $T_i = t_i^3 - t_i$ (mit i = 1, 2 . . . m = erste, zweite . . . m-te Gruppe von Bindungen).

2) Die korrigierte Größe lautet:

$$H_{korr} = \frac{H}{1 - \sum_{i}^{m} T_i/(n^3 - n)}$$

mit m = Anzahl der Gruppen gebundener Werte

Der χ^2-Test für Kreuztabellen

Der χ^2-Test für Kreuztabellen ist eigentlich aus der Vorlesung Statistik (Bachelor) bekannt. Er wird hier wiederholt wegen seiner großen Bedeutung.

Das im folgenden gewählte Beispiel einer 2-mal-2-Felder-Tabelle lässt sich problemlos auf größere Tabellen verallgemeinern.

Anordnung von Merkmalen in Kreuztabellen

Wenn möglich: Unabhängige (erklärende) Variable als Spaltenvariable, abhängige Variable als Zeilenvariable. Prozentuierung in den Spalten! (Statt Prozenten auch rohe Anteilswerte [40 %=0,4] möglich, aber unüblich.)

Beispiel (Zahlen fiktiv, der Zusammenhang an sich ist nicht unrealistisch): Berufliche Stellung des Vaters (Arbeiter vs. Angestellter) und Schulbesuch des Kindes (Angaben in Spaltenprozent)

	Arbeiter	Angest.	n
Hauptschule	90	45	630
RS/Gymnasium	10	55	370
n	400	600	1000

Formale Notation zur Bezeichnung der Zellen

In Formeln werden die Zellen der Tabelle mit Indices versehen. Der erste Index bezieht sich auf die Zeile, der zweite auf die Spalte. Das Zeichen $\bullet$ steht für die gesamte Zeile bzw. Spalte.

Am Beispiel einer Vier-Felder-Kreuztabelle:

	X=1	X=2	$\sum$
Y=1	n_{11}	n_{12}	$n_{1\bullet}$
Y=2	n_{21}	n_{22}	$n_{2\bullet}$
$\sum$	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet\bullet}$

Lesebeispiele: n_{21} heißt: Häufigkeit in der zweiten Zeile und der ersten Spalte. $n_{\bullet 2}$ heißt: Häufigkeiten über alle Zeilen in der zweiten Spalte.

Formale Notation: Bedingte Anteilswerte

Um die Spaltenprozent (oder Zeilenprozent) kennzuzeichen, verwenden wir ein Symbol für „bedingte Anteilswerte“ (Anteils- bzw. Prozentwert unter der Bedingung . . .).

Die folgende Tabelle zeigt ein Beispiel für Spaltenprozente:

	X=1	X=2
Y=1	$p_{1\|X=1}$	$p_{1\|X=2}$
Y=2	$p_{2\|X=1}$	$p_{2\|X=2}$

Lesebeispiel: $p_{1|X=2}$ heißt: Anteilswert (oder Prozentwert) in der ersten Zeile unter der Bedingung $X = 2$ (X hat den Wert 2).

Eine kürzere Schreibweise ist : $p_{1|2}$ usw.

Der χ^2-Test: Schritt 1

Unsere Frage: Können wir annehmen, dass der Unterschied in den Anteilswerten in unserer Stichprobe auch in der Grundgesamtheit besteht?

Formulierung der Hypothesen

- Nullhypothese: Die Verteilung der abhängigen Variablen ist über alle Ausprägungen der unabhängigen Variablen identisch.
- Alternativhypothese: Die Verteilung der abhängigen Variablen unterscheidet sich je nach Ausprägung der unabhängigen Variablen.

Formal lautet die H_0:

$$\pi_{1|X=1} = \pi_{1|X=2} = \ldots = \pi_{1|X=k} = \pi_{1\bullet}$$

usw. für jede Zeile der Tabelle.

Der χ^2-Test: Schritt 2

Die Teststatistik zur Prüfung der Nullhypothese bezieht sich auf den Vergleich der absoluten Häufigkeiten, die unter der Nullhypothese zu erwarten wären, mit den beobachteten Häufigkeiten.

Werte, die bei Gültigkeit der Nullhypothese (Unabhängigkeit der Merkmale) zu erwarten wären:

	Arbeiter	Angestellter	n
Hauptschule	252	378	630
	63 %	63 %	63 %
RS/Gymnasium	148	222	370
	37 %	37 %	37 %
n	400	600	1000

Die Verteilung der Werte in den Spalten entspricht der prozentualen Randverteilung über alle Gruppen.

Der χ^2-Test: Schritt 2

Einfache Berechnung der unter der H_0 erwarteten absoluten Häufigkeiten aus der Randverteilung:

$$e_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n} \quad \text{z. B.} \quad e_{21} = \frac{n_{2\bullet} \cdot n_{\bullet 1}}{n} = \frac{370 \cdot 400}{1000} = 148$$

	Arbeiter	Angestellter	n
Hauptschule	252	378	630
	63 %	63 %	63 %
RS/Gymnasium	148	222	370
	37 %	37 %	37 %
n	400	600	1000

Der χ^2-Test: Schritt 2

Anwendungsvoraussetzungen des χ^2-Tests:

- Der Test ist nur gültig, wenn gilt: $e_{ij} > 5$ für alle (oder: die meisten) e_{ij}.
- Außerdem soll evtl. bei $n < 60$ die sog. Kontinuitätskorrektur nach Yates verwendet werden (das ist aber unter Statistikern umstritten). Bei sehr kleinen Fallzahlen ($n < 30$) muss zu sog. exakten Testverfahren gegriffen werden (hier nicht behandelt).

Im vorliegenden Fall gibt es in beiden Hinsichten keine Probleme, wir können also den χ^2-Test bedenkenlos anwenden.

Der χ^2-Test: Schritt 3

Signifikanzniveau/Kritischer Wert/Ablehnungsbereich

Trifft die H_0 zu, beträgt der Wert der Teststatistik 0. Der kritische Wert der χ^2-Quadratverteilung liegt bei $1 - \alpha$ (bei $\alpha = 5$ % also: $1 - 0{,}05 = 0{,}95$) (in einer Vierfeldertabelle [1 Freiheitsgrad] beträgt er $\alpha = 5$ % 3,841). Erreicht oder übertrifft die Teststatistik diesen Wert, wird die Nullhypothese verworfen.

Es gibt hier also keine Unterscheidung zwischen ein- oder zweiseitigen Hypothesen – wir können nur prüfen: Zusammenhang ja oder nein.

Allgemein beträgt die Zahl der Freiheitsgrade (d. f.) für den χ^2-Test $(I - 1)(J - 1)$, wobei I die Zahl der Zeilen und J die Zahl der Spalten in der Tabelle ist (ohne Randverteilung!).

Der χ^2-Test: Schritt 4

Berechnung der Teststatistik und Entscheidung über H_0:

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

In Worten:

1. Berechne für jede einzelne Zelle der Tabelle die Differenz zwischen beobachtetem und unter der H_0 erwarteten Wert, quadriere diese Differenz und dividiere das Ergebnis durch den erwarteten Wert.
2. Summiere diese Werte über alle Spalten (j) und alle Zeilen (i).

Der χ^2-Test: Schritt 4

Berechnung der Teststatistik und Entscheidung über H_0 im Beispiel:

$$\begin{aligned}\chi^2 &= \frac{(360-252)^2}{252} + \frac{(40-148)^2}{148} \\ &\quad + \frac{(270-378)^2}{378} + \frac{(330-222)^2}{222} \\ &= 46{,}29 + 78{,}81 + 30{,}86 + 52{,}54 \\ &= 208{,}5\end{aligned}$$

$208{,}5 > 3{,}841 \rightarrow H_0$ wird verworfen (mit $\alpha = 0{,}05$).

Und zu guter Letzt . . .

Frohe Weihnachten und alles Gute für 2013!