2.4 Grundprinzipien statistischer Hypothesentests

2.4.1 Motivation und Hinführung

Hypothese: „Behauptung einer Tatsache, deren Überprüfung noch aussteht“ (Leutner in: Endruweit, Trommsdorff: Wörterbuch der Soziologie (1989))

Statistik: Überprüfung von Hypothesen über die Grundgesamtheit anhand einer Stichprobe

Idealtypische Vorgehensweise:

Bsp. 2.20

- Studie zur Einstellung der Münchner Bevölkerung zu psychisch Kranken (1989)

- Teilstudie: Kooperationsbereitschaft in der Befragung

1. *"Theorie":* Aktive Stellung im öffentlichen Leben beeinflusst Kooperationsbereitschaft positiv
   Aktiv $\leftrightarrow$ Altruismus
   $\leftrightarrow$ Interesse an öffentlichen Angelegenheiten
   $\leftrightarrow$ häufiger affektiv-neutrale Rollenbeziehung
   $\Rightarrow$ eher bereit, affektiv-neutrale Rolle des Befragten einzunehmen
   $\leftrightarrow$ häufiger in spezifischen Rollen
   $\leftrightarrow$ häufiger elaborierter Code?

2. *Hypothese:* Befragte, die aktiv am öffentlichen Leben teilnehmen, sind kooperativer.

3. *Operationalisierung:*

   - Aktiv im öffentlichen Leben
     $\rightarrow$ Verbandsmitgliedschaft ja/nein Variable X
   - *Kooperationsbereitschaft* $\rightarrow$ Variable Y
     * Kooperative: ließen sich interviewen
     * Primärverweigerer: auf „sanften Druck“ zu schriftlicher Befragung bereit
     * Totalverweigerer: keine Info, aber Annahme: Primärverweigerer sind ihnen ähnlicher

4. *statistische Hypothesen:*
   Es besteht (k)ein Zusammenhang zwischen X und Y

   Kann die sog. *Nullhypothese* (H_0)
   „Es besteht kein Zusammenhang zwischen X und Y“ abgelehnt werden?

Zur Herleitung/Motivation eines geeigneten Prüfverfahrens

Daten (echt, relative und absolute Häufigkeiten)

	ja	nein	
kooperativ	0.27	0.53	0.80
	(95)	(186)	(281)
Primärv.	0.05	0.15	0.20
	(17)	(54)	(71)
	0.32	0.68	1
	(112)	(240)	(352)

erster Schritt: relative Häufigkeit; „Schätzen der gemeinsamen Verteilung“

	ja	nein	
kooperativ	0.256	0.544	0.8
Primärv.	0.064	0.136	0.2
	0.32	0.68	1

Die Zahlen in der Unabhängigkeitstafel weichen von den tatsächlichen Daten ab ⇒

Kardinalsfrage der Testtheorie

Wann ist etwas überzufällig? - Testen mit Hilfe des sog. p-Wertes

- Bestimme eine Zufallsvariable T, die in geeigneter Weise den Unterschied einer zufälligen Stichprobe zur Situation der Nullhypothese mißt (hier insb. der χ^2-Abstand zwischen einer Stichprobe und der Unabhängigkeitstafel (siehe auch Statistik I und später))
- Bestimme die Realisation t_0 von T anhand der konkreten Daten (hier: χ^2=2.11)
- Berechne die Wahrscheinlichkeit, einen mindestens so extremen Wert von T zu beobachten, falls H_0 richtig ist. D.h. hier

$$p\text{-value} \quad := \quad P(T \geq t_0 || H_0)$$

  (Hier: p-value=0.15)
- Falls p-value klein (kleiner als eine substanzwissenschaftlich vorgegebene Fehlerwahrscheinlichkeit), dann H_0 ablehnen, sonst beibehalten.
  (Hier: p-value zu groß: Die Nullhypothese kann nicht abgelehnt werden.)

Diese Vorgehensweise sollte einem Einblick verschaffen; statistische Test können *nicht* immer genau so durchgeführt werden, aber die Kardinalsfrage bleibt die entscheidende Grundidee. Das allgemeine Vorgehen lässt sich am besten zunächst bei sog. parametrischen Test verdeutlichen:

2.4.2 Die prinzipielle Vorgehensweise bei einem parametrischen statistischen Test

1.) *Aufstellen der substanzwissenschaftlichen Hypothese / inhaltliche Fragestellung*
(z.B Rot/Grün bekommt die absolute Mehrheit, das Einkommen beträgt mindestens 3000 Euro)

2.) *Formulieren eines geeigneten statistischen Modells*
Hier: Stets $X_1, \ldots, X_n$ i.i.d. Stichprobe,
zunächst: parametrisches Modell mit unbekanntem Parameter ϑ

$$\begin{pmatrix} \pi & \text{Anteil Rot/Grün} \\ \mu & \text{Durchschnittseinkommen} \end{pmatrix}$$

3.) *Formulierung der statistischen Hypothesen*
Umformulieren der substantiellen Hypothesen als Hypothesen über ϑ
Verglichen wird immer eine sog. *Nullhypothese* (H_0) mit einer sog. *Alternativhypothese* (H_1)
Bei parametrischen Fragestellungen:

Bem: Da nur die Fehlerwahrscheinlichkeit 1. Art kontrolliert werden kann,

- kann H_0 nicht mit einer a priori kontrollierten Fehlerwahrscheinlichkeit angenommen werden; ($P(T \in KR||H_1)$ ist nur implizit gegeben)
  $\Rightarrow$ „Ablehnung von H_0“ oder „Nichtablehnung von H_0“
- setzt man bei einseitigen Test das, was man inhaltlich zeigen will, in die Alternativhypothese. (Bei zweiseitigen Tests geht dies – aus hier nicht zu besprechenden, mathematischen Gründen – nicht (so einfach).)

2.4.3 Typische Tests I: Ein-Stichproben-Gauss-Test, t-Test und Test auf einen Anteilswert

Aufgabe: Konstruiere Test für eine Hypothese über die Lage einer Verteilung

hier: ausschließlich Mittelwert μ eines normalverteilten Merkmals, z.B. IQ der Studierenden der Soziologie, Autoritarismusscore

a) Gauss-Test

1.) [*inhaltliche Hypothese*]

2.) *Statistisches Modell*
$X_1, \ldots, X_n$ *i.i.d.* Stichprobe, wobei X_i jeweils normalverteilt sei mit unbekanntem Mittelwert μ und bekannter Varianz σ^2

3.) *Formulierung der statistischen Hypothesen*

Fall I.	$H_0:$	$\mu \leq \mu_0$	$H_1:$	$\mu > \mu_0$
Fall II.	$H_0:$	$\mu \geq \mu_0$	$H_1:$	$\mu < \mu_0$
Fall III.	$H_0:$	$\mu = \mu_0$	$H_1:$	$\mu \neq \mu_0$

4. *Festlegen des Signifikanzniveaus*

   Ein Wert α; hier allgemein durchgerechnet, üblich:

   10% : tendenziell signifikant
   5% : signifikant
   1% : hoch signifikant

5. *Testgröße*

$$T := \frac{\bar{X} - \mu_0}{\sigma} \sqrt{n}$$

   T ist empfindlich gegenüber Abweichungen von H_0, also geeignet, falls $\mu = \mu_0$ gilt.

$$T = \frac{\bar{X} - \mu_0}{\sigma} \sqrt{n} = \left(\frac{\bar{X} - \mu_0}{\sigma} \sqrt{n} \right) \sim N(0, 1)$$

Bsp. 2.21:
Der IQ in einer gewissen Population sei normalverteilt mit unbekanntem Mittelwert μ und Varianz $\sigma^2 = 225$. Es wird vermutet, dass $\mu > 120$ gilt. Kann diese Vermutung mit einer Fehlerwahrscheinlichkeit von 5% bestätigt werden, wenn eine Stichprobe mit $n = 100$ den Wert $\bar{x} = 125$ ergab?

b) t-Test

Ist in Punkt 2 die Varianz σ^2 unbekannt, so kann man völlig analog vorgehen, wenn man

- σ durch $S = \sqrt{S^2}$ schätzt
- die Teststatistik T ersetzt durch

$$T = \frac{\bar{X} - \mu_0}{S} \sqrt{n}$$

H_0 ablehnen, falls

Fall I.	$T \geq t_{1-\alpha}(n-1)$	(2.4.6)
Fall II.	$T \leq -t_{1-\alpha}(n-1)$	(2.4.7)
Fall III.	$T \leq -t_{1-\frac{\alpha}{2}}(n-1)$ oder $T \geq t_{1-\frac{\alpha}{2}}(n-1)$	(2.4.8)

c) Approximative Tests für Hypothesen über Anteilswerte

Mit Hilfe der Binomialapproximation (vgl. Kapitel 1.7) ermöglichen die eben besprochenen Tests auch unmittelbar die Prüfung von Hypothesen über Anteilswerte.

Eingebettet in Beispiel:

1.) *Rot/Grün wird nicht die Mehrheit bekommen*

2.) *Statistisches Modell*

$X_1, \ldots, X_n$ *i.i.d.* Stichprobe von $X = \begin{cases} 1 \text{ Rot/Grün} \\ 0 \text{ sonst} \end{cases}$,

wobei π der Anteil der Einheiten mit Ausprägung 1 in der Grundgesamtheit ist.

3.) *Statistische Hypothesen*

I. $H_0: \ \pi \leq \pi_0 \quad H_1: \ \pi > \pi_0$
II. $H_0: \ \pi \geq \pi_0 \quad H_1: \ \pi < \pi_0$
III. $H_0: \ \pi = \pi_0 \quad H_1: \ \pi \neq \pi_0$

Hier: $\pi_0 = 0.5$, $H_0: \ \pi \geq 0.5 \quad H_1: \ \pi < 0.5$

4.) *Vorgabe des Signifikanzniveaus*

5.) *Testgröße und kritische Region*

Wie in Beispiel 2.19 ist

$$\frac{\overline{X} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim N(0,1) \ ,$$

also unter $\pi = \pi_0$

$$T = \frac{\overline{X} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \sim N(0,1) \qquad (2.4.9)$$

kritische Regionen

6.) *Berechnung der Realisationen t_0 von T*

An den Daten von Bsp. 2.19

Wahlumfrage: n=500, $\overline{X} = 46.5\%$ (Anteil Rot/Grün)

II. $H_0 : \pi \geq 0.5$ gegen $H_1 : \pi < 0.5$

$\alpha = 0.05$

Weitere Tests

2.4.4 Lagevergleiche aus unabhängigen Stichproben

a) Situation

- Ein stetiges Merkmal X
- Zwei Gruppen A, B

- *Ziel*: Vergleich der Mittelwerte in den beiden Gruppen

- Typische Fragestellungen, z.B.
  - Verdienen Männer mehr als Frauen?
    (X Erwerbseinkommen, A=Männer, B=Frauen)
  - Sind A-Wähler autoritärer als B-Wähler?
    (X Autoritarismusscore)
  - Konkret aus Studie: Bild des psychisch Kranken, Kooperationsbereitschaft und Vorurteile:

Schritt 1: Substanzwissenschaftliche Hypothese

Je weniger die Einstellung gegenüber psychisch Kranken durch Vorurteile und Stereotype gekennzeichnet ist, desto größer ist die Kooperationsbereitschaft im Interview.

Schritt 2: Statistisches Modell

$X_1, \ldots, X_n$ i.i.d. Stichprobe aus Gruppe A, $Y_1, \ldots, Y_n$ i.i.d. Stichprobe aus Gruppe B

$$X_i \sim N(\mu_X; \sigma_X^2)$$

$$Y_i \sim N(\mu_Y; \sigma_Y^2).$$

Zunächst seien die Varianzen als bekannt angenommen.

X: Vorurteilsindex (aus Fragebatterie mit Statements $(1, \ldots, 5)$ und anschließender Likert-Skalierung gewonnen); hier als normalverteilt angenommen. (Kleiner Wert entspricht großen Vorurteilen.)

Gruppe A: Kooperative

Gruppe B: Primärverweigerer

b) Der Zwei-Stichproben-Gauss-Test, σ_x^2, σ_y^2 bekannt

Schritt 3: Formulieren der statistischen Hypothesen

I. $H_0: \ \mu_X \leq \mu_Y \quad H_1: \ \mu_X > \mu_Y$
II. $H_0: \ \mu_X \geq \mu_Y \quad H_1: \ \mu_X < \mu_Y$
III. $H_0: \ \mu_X = \mu_Y \quad H_1: \ \mu_X \neq \mu_Y$

In unserem Beispiel: vermuten wir, dass Gruppe A geringere Vorurteile, also einen größeren durchschnittlichen Score hat.

Deshalb:

Wir hoffen/erwarten

Schritt 4: Festlegen eines Signifikanzniveaus

Allgemein: α, hier z.B. 0.01

Schritt 5: Festlegen einer Testgröße und einer Kritischen Region

Testgröße: Vergleich der arithmetischen Mittel $\overline{X}$ und $\overline{Y}$, genauer betrachte:

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \qquad (2.4.10)$$

T ist $N(\mu_X - \mu_Y, 1)$ verteilt. Falls $\mu_X = \mu_Y$ ist, ist

$$T \sim N(0, 1)$$

Festlegen der Kritischen Region:

c) Der Zwei-Stichproben-t-Test

Abwandlung von Schritt 5 bei unbekannten Varianzen

$$\begin{aligned} X_i &\sim N(\mu_X, \sigma_X^2) \quad , \quad i = 1, \ldots, n \\ Y_i &\sim N(\mu_Y, \sigma_Y^2) \quad , \quad i = 1, \ldots, m \end{aligned}$$

wobei jetzt die Varianzen σ_X^2, σ_Y^2 unbekannt seien.

a) Ist bekannt, dass die Varianzen gleich sind, so schätzt man sie mittels S_X^2 und S_Y^2 und betrachtet

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right) \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}}} \qquad (2.4.14)$$

Bei $\mu_X = \mu_Y$ gehorcht T einer t-Verteilung mit $(n + m - 2)$ Freiheitsgraden.

Analoges Vorgehen bei der Konstruktion der Kritischen Region

Im Beispiel: *6. Realisationen* t_0 von T_0

$\bar{X}$	=	51.11	Kooperative	$n = 270$
$\bar{Y}$	=	48.76	Primärverweigerer	$m = 58$
S_X^2	=	40.2		
S_Y^2	=	35.5		

b) Sind die Varianzen unbekannt und können nicht als gleich angenommen werden, so kann man für ein großes n mit

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \qquad (2.4.16)$$

rechnen. T ist für $\mu_X = \mu_Y$ approximativ standardnormalverteilt und kann auch angewendet werden, wenn keine Normalverteilung vorliegt. (Ist aber eben nur approximativ, nicht exakt.)
Viele Software-Pakete rechnen beide Arten von t-Tests und geben oft auch das Ergebnis eines (in der Vorlesung nicht betrachteten F-)Tests auf Gleichheit der Varianzen an.
Die korrekte Teststatistik für kleines T ist außerordentlich kompliziert; sie wird in der Vorlesung nicht betrachtet, weshalb – aus Übungsgründen – im Rahmen der Veranstaltung bei ungleichen Varianzen stets mit (7.4.7) gearbeitet werden darf.

Erweiterung

Oft findet man Formeln für allgemeine Hypothesen der Form

$$\mu_X - \mu_Y \leq \delta \quad \Leftrightarrow \quad \mu_X \leq \mu_Y + \delta$$

Man kann dies (auch mit Software) direkt lösen, indem man

2.4.5 Gauss-Test und t-Test für verbundene Stichproben

Vorsicht: Leicht zu verwechseln mit vorheriger Fragestellung!
Verbundene Stichproben: Vergleich zweier Merkmale X und Y, die jetzt an *denselben Personen erhoben werden.*

- z.B. Evaluierung einer Schulungsmaßnahme:
  X Punktezahl *vor* der Schulung
  Y Punktezahl *nach* der Schulung
- Autoritarismusscore vor/nach Projekt
- klassisches Medizinbeispiel: rechts/links-Vergleiche: Test zweier Salben bei Ekzemen
- Split-Half Reliabilität von aus vielen Einzelfragen bestehenden Scores

Man könnte auf zweierlei Arten vorgehen:

a) Man bestimmt zufällig zwei Gruppen, in der *einen* erhebt man X, in der *anderen* Y.
Danach Vergleich der Mittelwerte wie in vorherigem Kapitel beschrieben.

b) Man erhebt an *jeder* Person *beide* Merkmale: „Matched pair design“

<u>Konstruktion der Tests</u>

Seien	$X_1, \ldots, X_n$	*i.i.d.*	$N(\mu_X, \sigma_X^2)$
und	$Y_1, \ldots, Y_n$	*i.i.d.*	$N(\mu_Y, \sigma_Y^2)$

Zum Testen von Hypothesen der Form

I.	$H_0:$	$\mu_X \leq \mu_Y$	gegen	$H_1:$	$\mu_X > \mu_Y$
II.	$H_0:$	$\mu_X \geq \mu_Y$	gegen	$H_1:$	$\mu_X < \mu_Y$
III.	$H_0:$	$\mu_X = \mu_Y$	gegen	$H_1:$	$\mu_X \neq \mu_Y$

betrachtet man die Differenz $D_i = X_i - Y_i$.
Für den Erwartungswert μ_D gilt:

$$\mathbb{E}\mu_D = \mathbb{E}(D_i) =$$

und für die Varianz σ_D^2 gilt:

$$\begin{aligned} \sigma_D^2 := Var(X_i - Y_i) &= \\ &= \end{aligned}$$

also

$$\sigma_D^2 = \sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY} \quad \text{mit } \sigma_{XY} = Cov(X, Y)$$

Im Folgenden sei immer angenommen, dass auch D_i normalverteilt ist. Wegen $D_i \sim N(\mu_D, \sigma_D^2)$ mit $\mu_D = \mu_X - \mu_Y$ und $\sigma_D^2 = \sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}$ sind obige Hypothesen äquivalent mit den Hypothesen

$$\begin{array}{lllll} \text{I'.} & H_0: & \mu_D \leq 0 & \text{gegen} & H_1: \quad \mu_D > 0 \\ \text{II'.} & H_0: & \mu_D \geq 0 & \text{gegen} & H_1: \quad \mu_D < 0 \\ \text{III'.} & H_0: & \mu_D = 0 & \text{gegen} & H_1: \quad \mu_D \neq 0 \quad , \end{array}$$

und man kann unmittelbar die Tests aus 2.4.2 anwenden.
Sind die Varianzen unbekannt, so kann man σ_D aus den Differenzen D_i, $i = 1, ..., n$ schätzen. Zur Prüfung ist dann die t-Verteilung heranzuziehen.

2.4.6 χ^2-Test(s)

- Tests basierend auf diskreten bzw. diskretisierenden Merkmalen
- Grob gesprochen eignen sich χ^2-Tests, um zu entscheiden, ob eine empirische Verteilung signifikant von einer Modellverteilung abweicht.
- *Haupttypen*
  - *Unabhängigkeitstest*
    vgl. Motivationsbeispiel in Abschnitt 2.4.1: weicht die empirische gemeinsame Verteilung von der unter Unabhängigkeit zu erwartenden signifikant ab?
  - *Anpassungstest* z.B. Abweichung von der Gleichverteilung

    $$H_0 : P(X = 1) = P(X = 2) = P(X = 3) = \frac{1}{3}$$

    (Abweichung von einer bestimmten stetigen Verteilung: durch Klassenbildung)
  - Homogenitätstest:
    Herrscht in k Subpopoulationen jeweils dieselbe Verteilung?

Hier nur ausführlicher: χ^2-Unabhängigkeitstest (vgl. Statistik I)

In Beispiel eingebettet:

1) Aktive Stellung im öffentlichen Leben beeinflußt Kooperationsbereitschaft im Interview (positiv)

2) *Statistische Modelle*

Zwei diskrete Merkmale X und Y

(hier Y Verbandsmitgliedschaft
X Kooperationsbereitschaft)

$(X_1, Y_1), \ldots, (X_n, Y_n)$ *i.i.d.* Stichprobe des zwei-dimensionalen Merkmals (X, Y)

3) *statistische Hypothesen*[2]

H_0: Es herrscht Unabhängigkeit
H_1: Es herrscht keine Unabhängigkeit

d.h. H_0 : $P(X = x_i, Y = y_i) = P(X = x_i) \cdot P(Y = y_i)$ für alle Paare i, j

gegen H_1 : $P(X = x_i, Y = y_i) \neq P(X = x_i) \cdot P(Y = y_i)$ für *ein* Paar i, j

4) *Festlegen des Signifikanzniveuas*

5) *Testgröße und kritische Region*

χ^2-Abstand:
beobachtete Tafel: relative Häufigkeiten

		Y			
	X/Y	1	$\ldots$	m	
	1	$\frac{h_{11}}{n}$	$\ldots$	$\frac{h_{1m}}{n}$	$\frac{h_{1\bullet}}{n}$
X	$\vdots$		$\frac{h_{ij}}{n}$		
	k	$\frac{h_{k1}}{n}$	$\ldots$	$\frac{h_{km}}{n}$	$\frac{h_{k\bullet}}{n}$
		$\frac{h_{\bullet 1}}{n}$	$\ldots$	$\frac{h_{\bullet m}}{n}$	

h_{ij} absolute Häufigkeit des Ereignisses $\{X = x_i\} \cap \{Y = y_i\}$ in der Stichprobe

$\frac{h_{ij}}{n}$ Schätzer für $P(X = x_i,\ Y = y_i)$

[2]Alternative Modellierung bei 2×2 Tafel über Vergleich von Anteilen

Zu vergleichen mit der Unabhängigkeitstafel: $\widetilde{h}_{ij}$

	X/Y	Y: 1	$\dots$	m	
	1	$\frac{h_{\bullet 1}h_{1\bullet}}{n^2}$	$\dots$		$\frac{h_{1\bullet}}{n}$
X	$\vdots$		$\frac{h_{i\bullet}h_{\bullet j}}{n^2}$		$\frac{h_{i\bullet}}{n}$
	k		$\dots$		$\frac{h_{k\bullet}}{n}$
		$\frac{h_{\bullet 1}}{n}$	$\frac{h_{\bullet j}}{n}$	$\frac{h_{\bullet m}}{n}$	1

Teststatistik[3]

$$\begin{aligned} T &= \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{\left(h_{ij} - \frac{h_{i\bullet}h_{\bullet j}}{n}\right)^2}{\frac{h_{i\bullet}h_{\bullet j}}{n}} = \sum_{i=1}^{k}\sum_{j=1}^{m} n \cdot \frac{\left(\frac{h_j}{n} - \frac{h_{i\bullet}h_{\bullet j}}{n^2}\right)^2}{\frac{h_{i\bullet}h_{\bullet j}}{n^2}} \\ &= \sum_{i=1}^{k}\sum_{j=1}^{m} n \cdot \frac{(f_{ij} - f_{i\bullet}f_{\bullet j})}{f_{i\bullet}f_{\bullet j}} \end{aligned} \tag{2.4.17}$$

Unter H_0 gehorcht T approximativ einer sog. χ^2 Verteilung mit $(k-1)\cdot(m-1)$ Freiheitsgraden.

$$T = \sum_{\text{alle Zellen}} \frac{(\text{beobachtet} - \text{erwartet})^2}{\text{Normierung}}$$

kritische Region

Je stärker H_0 verletzt ist, umso stärker weichen wohl die beobachteten Häufigkeiten $\frac{h_{ij}}{n}$ und die unter Unabhängigkeit zu erwartenden Häufigkeiten $\frac{h_{i\bullet}h_{\bullet j}}{n^2}$ voneinander ab, d.h desto größer ist T.
Also: kritische Region aus großen Werten von T
$KR = [z, \infty)$ wobei z so, dass $P(T \in KR||H_0) = P(T \geq z||H_0) \leq \alpha$

[3]Rechnet man die verschiedenen Arten mit dem Taschenrechner, so können die beiden Ergebnisse sich – wegen der Fortpflanzung von Rundungsfehlern – voneinander unterscheiden.

z.B. $\chi^2_{0.9}(1) = 2.7055$
$\chi^2_{0.95}(1) = 3.8415$
$\chi^2_{0.99}(1) = 6.6349$

z ist das $(1-\alpha)$-Quantil der entsprechenden χ^2 Verteilung: $\chi^2_{1-\alpha)}((k-1)\cdot(m-1))$.

Jetzt im Beispiel: vgl. Kapitel 2.4.1

Beobachtete Tabelle

	Mitglied		
	ja	nein	
Kooperativ	0.27	0.53	0.8
Primärv.	0.05	0.15	0.2
	0.32	0.68	1

2.4.7 Zur praktischen Anwendung statistischer Tests - Typische Fallen

Testentscheidungen und Statistik-Software

Statistik-Software berechnet meist einen sog. *p-Wert*,

Zur Hypothesenwahl

Es sei nochmal daran erinnert:

Dualitätsprinzip

Vorüberlegung: $H_0 : \mu = \mu_0$ gegen $H_1 : \mu \neq \mu_0$. H_0 ablehnen, wenn:

$$\begin{aligned} & \frac{\bar{X} - \mu_0}{\sigma} \cdot \sqrt{n} > z_{1-\frac{\alpha}{2}} \text{ oder } \frac{\bar{X} - \mu_0}{\sigma} \cdot \sqrt{n} < -z_{1-\frac{\alpha}{2}} \\ \Leftrightarrow \quad & \bar{X} - \mu_0 > z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \text{ oder } \bar{X} - \mu_0 < -z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \\ \Leftrightarrow \quad & \bar{X} > \mu_0 + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \text{ oder } \bar{X} < \mu_0 - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \end{aligned}$$

d.h. abgelehnt werden alle Nullhypothesen $\mu = \mu_0$ mit

$$\mu_0 < \bar{x} - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

oder

$$\mu_0 > \bar{x} + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Vgl. dies mit Konfidenzintervall!

$$\mu_0 \in \left[\bar{X} - z_{\frac{1+\gamma}{2}} \cdot \frac{\sigma}{\sqrt{n}} \; ; \; \bar{X} + z_{\frac{1+\gamma}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right]$$

Passen α und γ zusammen, gilt also $z_{1-\frac{\alpha}{2}} = z_{\frac{1+\gamma}{2}}$, so sind diese Ausdrücke gerade komplementär:

$$\begin{array}{lllllll} 1-\frac{\alpha}{2} & \stackrel{!}{=} & \frac{1+\gamma}{2} & \Leftrightarrow & 2-\alpha & = & 1+\gamma \\ & & & \Leftrightarrow & \gamma & = & 1-\alpha \quad \Leftrightarrow \quad \alpha = 1-\gamma \end{array}$$

Dieses Beispiel ist verallgemeinerbar. Es besteht generell ein sehr enger Zusammenhang zwischen Tests und Konfidenzintervallen: Gegeben eine Pivotgröße T, besteht ein Konfidenzintervall zum Vertrauensgrad γ genau aus all jenen Werten ϑ_0 eines Parameters ϑ, bei denen die Hypothese $H_0 : \vartheta = \vartheta_0$ zum Signifikanzniveau $\alpha = 1 - \gamma$ nicht abgelehnt wurde.

Eine praktische Konsequenz daraus: Gegeben ein Konfidenzintervall $[A(X_1, \ldots, X_n), B(X_1, \ldots, X_n)]$ für ϑ, kann man Hypothesen der Form $H_0 : \vartheta = \vartheta_0$ sofort testen!

Bsp. (vgl. Bsp. 2.3.7a)
Man interessiert sich, ob gewisse Gummibärchenpackungen genau die angegebene Füllmenge von 250 g enthalten, also
Testen: $H_0 : \mu = 250$ g $\quad H_1 : \mu \neq 250$ g $\quad$ zu $\alpha = 0.05$
Hat man zu $\gamma = 0.95$ folgendes – auf der t-Verteilung beruhendes – Konfidenzintervall

$$[239.675, 250.325]$$

erhalten, so kann obige Hypothese nicht abgelehnt werden, da der Wert 250 im Konfidenzintervall liegt.

Bei Normalverteilung, d.h. bekannter Varianz σ ergibt sich

$$[240.100; 249.900]$$

250 g nicht mehr im Konfidenzintervall, H_0 kann abgelehnt werden.
Auch zur Abschätzung gut: zweiseitige Hypothese abgelehnt $\Rightarrow$ einseitige auch bei jedem $\alpha > 0.05$ ablehnbar, wenn der beobachtete Wert auf der richtigen Seite liegt (jedenfalls bei den hier betrachteten Tests)
Ferner oben: Zu $\alpha = 0.05$ nicht abgelehnt $\Rightarrow$ zu jedem $\alpha < 0.05$ nicht abgelehnt (jedenfalls bei Normalverteilung)

Signifikanz versus Relevanz[4]

Die üblichen Testgrößen hängen vom Stichprobenumfang n ab: Je größer n, umso leichter kann man eine Abweichung als signifikant nachweisen (ist auch sinnvoll).

Zweierlei praktische Konsquenzen

i) Aus der Nichtsignifikanz eines Unterschieds kann nicht notwendig geschlossen werden, dass kein inhaltlich relevanter Unterschied vorliegt. Vielleicht war nur der Stichprobenumfang zu klein, um einen durchaus vorhandenen Unterschied auch als signifikant nachweisen zu können.

ii) Andererseits kann es sein, dass bei großen Stichprobenumfängen selbst minimale Abweichungen signifikant sind.
Nicht jede statistisch signifikante Abweichung ist daher auch inhaltlich relevant, weswegen Vorsicht bei der inhaltlichen Interpretation gerade bei großen Datensätzen angebracht ist.

Mögliche Auswege:

- Ergebnisse kritisch betrachten
- Sog. Effektstärkemaße
- Untersuche statt der Hypothese „$\mu_A > \mu_B$“ die Hypothese „$\mu_A > \mu_B + \delta$“ mit relevantem Unterschied δ

[4](siehe z.B. Kriz: Statistik in den Sozialwissenschaften, Westdeutscher Verlag 1980, S. 116-121)

Multiples Testproblem

- Vorstellen: rein zufälliger Datensatz, 50 Variablen, ohne irgendeinen Zusammenhang
- Man testet alle Variablenpaare auf einen Zusammenhang
  $\Rightarrow \begin{pmatrix} 50 \\ 2 \end{pmatrix} = 1225$ Tests

  Irrtumswahrscheinlichkeit 5%.
  Für $X \sim B(1225, 0.05)$ gilt: $\mathbb{E}(X) = 61,25$.

  Im Durchschnitt wird also mehr als 61 mal die Nullhypothese, dass kein Zusammenhang besteht, verworfen

  $\Rightarrow$ wenige, sinnvolle Hypothesen ***vorher*** *inhaltlich* überlegen

  (In den Daten entdeckte „Zusammenhänge“ als statistisch signifikant nachzuweisen, ist (fast) zirkulär!)
- Es gibt Ansätze, wie man bei großen Hypothesensystemen diesem Problem entkommt

  $\rightarrow$ Theorie des multiplen Testens ; vorsichtige Adjustierung: Statt α betrachte man $\alpha/(Anzahl\ der\ Tests)$; ist aber meist überkonservativ.

Nichtparametrische Tests

Bis auf den χ^2-Unabhängigkeits-Test bauen alle Tests auf der (zumindestens approximativen Gültigkeit der) Normalverteilungsannahme auf.
Ist dies problematisch, z.B. bei kleinen Stichprobenumfängen, wie sie etwa bei der Vorbereitung von strukturierter Beobachtung, bei nicht reaktiven Verfahren oder in der Psychologie und Medizin oft auftreten, oder bei ordinalen Daten mit wenigen unterschiedlichen Ausprägungen, so kann die unreflektierte Anwendung der Standardtests zu krassen Fehlergebnissen führen.
Ein wichtiger Ausweg: nichtparametrische Tests = Verteilungsfreie Verfahren (Die Information in den Beobachtungen wird auf Ränge, bzw. großer/kleiner Vergleiche reduziert)
Bekannteste Beispiele: Wilcoxon-Test, Vorzeichentest

2.5 Inferenz bei der linearen Regression

2.5.1 Erinnerung an das Grundmodell

Statistik I

Modellbildung für jedes i