2.4 Hypothesentests

2.4.1 Grundprinzipien statistischer Hypothesentests; Motivationsbeispiel

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

Statistischer Test: Überprüfung von Hypothesen über die Grundgesamtheit anhand einer Stichprobe

Statistische Testtheorie: Schließe von Stichprobe auf Grundgesamtheit
Vorgehen:

- inhaltliche Hypothese aufstellen
- Operationalisierung
- inhaltliche Hypothese in statistische Hypothese „übersetzen“
- statistischer Test

Bsp. 2.21.

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

Wir betrachten eine Teilstudie: Kooperationsbereitschaft in der Befragung.

1. „Theorie“: Aktive Stellung im öffentlichen Leben beeinflusst Kooperationsbereitschaft positiv.

$$\begin{array}{lll} Aktiv & \leftrightarrow & \text{Altruismus} \\ & \leftrightarrow & \text{Interesse an öffentlichen Angelegenheiten} \\ & & \Rightarrow \text{eher bereit, die Rolle des Befragten einzunehmen} \end{array}$$

2. Hypothese: „Die Koorperationsbereitschaft der aktiven Personen unterscheidet sich vom Rest der Bevölkerung?“

3. Operationalisierung:

   - Aktiv im öffentlichen Leben

     $\rightarrow$ Verbandsmitgliedschaft ja/nein $=$ Variable X

- Kooperationsbereitschaft

  $\rightarrow$ Von den überhaupt nicht Kooperierenden hat man keine Daten, deshalb vergleicht man die sofort zur Antwort bereiten mit den „Primärverweigerern“, die sich erst nach langer Zeit zur Auskunft überreden ließen.
  $\rightarrow$ direkt kooperativ ja/nein $\rightarrow$Variable Y

4. Statistische Hypothesen: „Besteht ein Zusammenhang zwischen X und Y?“

   Statistisches Vorgehen:
   Kann die sog. $Nullhypothese$ „Es besteht kein Zusammenhang zwischen X und Y“ abgelehnt werden?

Herleitung / Motivation eines geeigneten Prüfverfahrens, basierend auf χ^2 (vgl. Statistik I)

Gegebene Daten (relative und absolute Häufigkeiten):

		kooperativ: ja	kooperativ: nein	
aktiv	ja	0.27 (95)	0.05 (17)	0.32 (112)
	nein.	0.53 (186)	0.15 (54)	0.68 (240)
		0.8 (281)	0.2 (71)	1 (352)

Vergleiche gegebene Tafel mit der „Unabhängigkeitstafel“
Wie würde denn die Tafel aussehen, wenn kein Zusammenhang bestünde?
Genauer: wie würde das Innere der Tabelle aussehen, wenn Unabhängigkeit (und die

gleichen Randverteilungen) herrschen würde, also die Nullhypothese zutreffen würde?

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

Die Häufigkeiten in der Unabhängigkeitstafel weichen von den tatsächlichen Daten ab. Vgl. Statistik I: Je stärker die Abweichung, desto stärker ist der Zusammenhang.

Kardinalfrage der Testtheorie:

Wann ist die Abweichung „groß“, d.h. „überzufällig“?

- Bestimme eine Zufallsvariable T, die in geeigneter Weise den Unterschied einer zufälligen Stichprobe zur Situation der Nullhypothese misst (hier: der χ^2-Abstand zwischen einer Stichprobe und der Unabhängigkeitstafel, vgl. Statistik I).
- Bestimme einen Bereich KR, der sehr unwahrscheinlich ist, falls H_0 gilt („$Ablehnungsbereich$“, „$kritische\ Region$“)
- Bestimme die Realisation t von T anhand der konkreten Daten (hier: χ^2=2.11). Ist $t \in$ KR, dann Nullhypothese ablehnen, sonst nicht.

Dabei kann man zwei Arten von Fehlern machen:

Wahrheit \ Aktion	H_0 beibehalten	H_0 ablehnen
H_0 wahr	$\surd$	Fehler 1.Art
H_0 falsch	Fehler 2. Art	$\surd$

Fehler 1. Art („α-Fehler “):

Die Nullhypothese wird abgelehnt, obwohl sie in Wirklichkeit richtig ist.

z.B.: Man behauptet, es bestünde ein Zusammenhang, obwohl in Wirklichkeit kein Zusammenhang besteht.

Der Fehler 1. Art soll klein sein (üblich sind 5% oder 10%). Allerdings kann man nicht fordern, dass der Fehler 1. Art bei 0% liegen soll, sonst würde man die Nullhypothese nie ablehnen können.

Fehler 2. Art („β-Fehler “):

Die Nullhypothese wird beibehalten, obwohl sie in Wirklichkeit falsch ist.

Dabei wird so vorgegenagen, dass die **Wahrscheinlichkeit**, einen **Fehler 1.Art** zu begehen, **beschränkt** ist durch eine inhaltliche vorgegebene Schranke α („***Signifikanzniveau***“)
Ein guter statistischer Test garantiert bei einem vergegebenen niedrigen Signifikanzniveau (als Schranke für den Fehler 1. Art) auch einen möglichst geringen Fehler 2. Art.

2.4.2 Präzisierung: Konstruktion eines parametrischen statistischen Tests

1. *Aufstellen der substanzwissenschaftlichen Hypothese / inhaltliche Fragestellung*

   (z.B. Rot/Grün bekommt die absolute Mehrheit, der mittlere Intelligenzscore der Gruppe $\cdots$ beträgt mindestens 130 Einheiten)

2. *Formulieren eines geeigneten statistischen Modells*

   Im Folgenden stets $X_1, \ldots, X_n$ i.i.d. Stichprobe sowie parametrisches Modell mit unbekanntem Parameter ϑ.

   Anteil Rot/Grün: $B(1, \pi)$

   Intelligenzscore: $\mathcal{N}(\mu; \sigma^2)$.

3. *Formulierung der statistischen Hypothesen*

   - Umformulieren der substantzwissenschaftlichen Hypothesen als Hypothesen über ϑ.

- Verglichen wird immer eine sog. *Nullhypothese* (H_0) mit einer sog. *Alternativhypothese* (H_1).
- Bei parametrischen Fragestellungen unterscheidet man:

a) *einseitige Testprobleme*:

$$H_0 : \vartheta \leq \vartheta_0 \text{ gegen } H_1 : \vartheta > \vartheta_0$$

$$H_0 : \vartheta \geq \vartheta_0 \text{ gegen } H_1 : \vartheta < \vartheta_0$$

b) *zweiseitiges Testproblem*:

$$H_0 : \vartheta = \vartheta_0 \text{ gegen } H_1 : \vartheta \neq \vartheta_0$$

ϑ_0 ist ein fester, vorgegebener Wert, der von inhaltlichem Interesse ist; zu unterscheiden von wahrem Wert ϑ.

Der Begriff einseitig/zweiseitig bezieht sich auf die Alternative, je nachdem ob die Alternative nur aus großen bzw. nur aus kleinen Werten besteht oder ob sowohl große als auch kleine Werte für die Alternative sprechen.

4. *Festlegung des Signifikanzniveaus* α

   Wiederholung aus der Einleitung

   Beim Testen sind folgende Entscheidungen möglich:

$$H_0\text{: ablehnen} \quad \text{oder} \quad H_0\text{: beibehalten}$$

Damit sind zwei verschiedene Arten von Fehlern möglich:

Wahrheit \ Aktion	H_0 beibehalten	H_0 ablehnen
H_0 wahr	$\surd$	Fehler 1.Art
H_0 falsch	Fehler 2. Art	$\surd$

Man kann nicht beide Fehlerwahrscheinlichkeiten gleichzeitig kontrollieren! (Trade-

off!)

$\Rightarrow$ asymmetrische Vorgehensweise:

Der Fehler 1. Art wird kontrolliert durch die Angabe einer Oberschranke α („Signifikanzniveau“)

Typische Werte: üblich

$\alpha = 0.1,$	$\alpha = 0.05,$	$\alpha = 0.01$	$\alpha = 0.001$
„marginal signifikant“	„signifikant“	„hoch signifikant“	„höchst signifikant“

5. *Festlegen einer Testgröße und einer kritischen Region*

Eine *Testgröße* T ist eine Zufallsgröße $T = g(X_1, \ldots, X_n)$, die „empfindlich gegenüber Abweichungen von H_0 ist“. Die *Kritische Region* KR („Ablehnungsbereich“) besteht aus potentiellen Werten von T, die gegen H_0 sprechen.

Liegt t (Realisation von T) in KR, wird man sich gegen H_0 entscheiden.

Damit der Fehler 1. Art durch α beschränkt bleibt, muss die kritische Region KR

also so gewählt werden, dass

$$P(T \in KR|H_0) \leq \alpha$$

gilt, d.h. die Wahrscheinlichkeit, dass T in der kritischen Region liegt und damit zur Ablehnung von H_0 führt, darf höchstens α sein, wenn H_0 stimmt.

Umgekehrt soll $P(T \in KR||H_1)$ möglichst groß sein, da dies die Wahrscheinlichkeit ist, die Nullhypothese H_0 abzulehnen, falls sie falsch ist. (Gegenwahrscheinlichkeit zur Wahrscheinlichkeit für den Fehler 2. Art, auch als $Power$ oder $Güte$ des Tests bezeichnet.)

6. *Auswerten der Stichprobe*

   Berechnung der Realisation t der Testgröße T basierend auf der konkret vorliegenden Stichprobe.

7. *Testentscheidung*

   Ist $t \in KR$, dann H_0 ablehnen, sonst nicht ablehnen.

Bem. 2.22.

- Die wesentliche Elemente des Tests (Signifikanzniveau, Testgröße, kritische Region) sind unabhängig von den Daten, also vor der Auswertung, zu bestimmen.

- Da nur die Fehlerwahrscheinlichkeit 1. Art kontrolliert werden kann, kann H_0 nicht mit einer a priori kontrollierten Fehlerwahrscheinlichkeit angenommen, sondern nur abgelehnt oder nicht abgelehnt werden.

- Als „guter Forscher“ sollte man deshalb immer das, was man zeigen will, in die Alternativhypothese schreiben.

  z.B. Forscher will zeigen, dass seine Interventionsmaßnahme besser wirkt als ein anderes.

  Nullhypothese: Sie wirkt schlechter oder gleich gut.
  Alternativhypothese: Sie wirkt besser.

Durch die Kontrolle des Fehlers 1. Art ist gewährleistet, dass die Wahrscheinlichkeit, der Maßnahme irrtümlich eine bessere Wirkung zuzuschreiben, höchstens α ist.

- Allerdings gibt es keineswegs immer einen (einfachen) statistischen Test für jede Nullhypothese.

  z.B. ist es technisch viel einfacher als Nullhypothese $\theta = \theta_0$ zu verwenden, als $\theta \neq \theta_0$. Verwendet man deshalb einen Test mit $H_0 : \theta = \theta_0$, möchte man inhaltlich aber genau dies zeigen, kehren sich die Rollen des Fehlers 1. Art und des Fehlers 2. Art um. Um in diesem Fall einen geringeren Fehler 2. Art zu erzielen, sollte das Signifikanzniveau höher als üblich gewählt werden. Dies ist aber nur ein erster Versuch, diesem Problem beizukommen. Eine saubere technische Behandlung führt auf sogenannte „Äquivalenztests“.

2.4.3 Typische Tests I: Tests auf Lageparameter

Hier werden exemplarisch nur wenige, ausgewählte Test, die typisch sind, besprochen. Das Grundprinzip ist bei anderen Tests analog.
Aufgabe: Konstruiere Test für eine Hypothese über die Lage einer Verteilung.

Wir betrachten ausschließlich den Erwartungswert μ eines normalverteilten Merkmals, bzw. den Erwartungswert π einer binären Zufallsgröße.

2.4.3.1 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*:

$$\begin{array}{lllll} \text{Fall 1:} & H_0: & \mu \leq \mu_0 & H_1: & \mu > \mu_0 \\ \text{Fall 2:} & H_0: & \mu \geq \mu_0 & H_1: & \mu < \mu_0 \\ \text{Fall 3:} & H_0: & \mu = \mu_0 & H_1: & \mu \neq \mu_0 \end{array}$$

Gleichheitszeichen immer bei H!

4. *Festlegen des Signifikanzniveaus*: Wir rechnen im Folgenden allgemein. Übliche Werte sind, wie gesagt:

$$\begin{aligned} 10\%: &\quad \text{„marginal signifikant“} \\ 5\%: &\quad \text{„signifikant“} \\ 1\%: &\quad \text{„hoch signifikant“} \end{aligned}$$

5. *Testgröße*:

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

T ist empfindlich gegenüber Abweichungen von H_0.

Falls $\mu = \mu_0$ (falls also die Nullhypothese zutrifft) gilt

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

Kritische Regionen:

Bsp. 2.23.

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 $\alpha = 5\%$ bestätigt werden, wenn eine Stichprobe mit $n = 100$ den Wert $\bar{x} = 125$ ergab?

2.4.3.2 t-Test

Situation wie beim Gauß-Test, aber mit unbekannter Varianz σ^2: $X_1, \ldots, X_n$ $i.i.d.$ Stichprobe, wobei X_i jeweils normalverteilt sei mit unbekanntem Mittelwert μ und $unbekannter$ Varianz σ^2.

Analoges Vorgehen zur Konstruktion des Tests, aber

H_0 ablehnen, falls

Fall 1	$T \geq t_{1-\alpha}^{(n-1)}$
Fall 2	$T \leq -t_{1-\alpha}^{(n-1)}$
Fall 3	$T \leq -t_{1-\frac{\alpha}{2}}^{(n-1)}$ oder $T \geq t_{1-\frac{\alpha}{2}}^{(n-1)}$

2.4.3.3 Approximative Tests für Hypothesen über Anteilswerte

Mit Hilfe der Normalapproximation der Binomialverteilung (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

$$X_i = \begin{cases} 1 & \text{Rot/Grün} \\ 0 & \text{sonst} \end{cases}$$

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

3. *Statistische Hypothesen*:

$$\begin{array}{llll} 1 & H_0: & \pi \leq \pi_0 & H_1: \quad \pi > \pi_0 \\ 2 & H_0: & \pi \geq \pi_0 & H_1: \quad \pi < \pi_0 \\ 3 & H_0: & \pi = \pi_0 & H_1: \quad \pi \neq \pi_0 \end{array}$$

Hier: $\pi_0 = 0.5$ und

$$H_0: \pi \geq 0.5 \quad H_1: \pi < 0.5$$

4. *Vorgabe des Signifikanzniveaus*: $\alpha = 0.05$

5. *Testgröße und kritische Region*: Approximativ gilt für großen Stichprobenumfang n und wahren Anteil π (vgl. Kap. 1.7 und 2.3)

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

also speziell für $\pi = \pi_0$ (unter H_0)

$$T = \frac{\bar{X} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \sim \mathcal{N}(0,1)$$

Damit ergeben sich die kritischen Regionen

6. *Berechnung der Realisationen t_0 von T*

Wahlumfrage: $n = 500$, $\bar{x} = 46.5\%$ (Anteil Rot/Grün), $\alpha = 0.05$

2.4.4 Typische Tests II: Lagevergleiche aus unabhängigen Stichproben

2.4.4.4 Allgemeine Situation:

- Ein stetiges Merkmal, erhoben in zwei Gruppen A und B.
- Ziel: Vergleich der Erwartungswerte in den beiden Gruppen.
- Typische Fragestellungen, z.B.
  * Verdienen Männer mehr als Frauen?
  * Sind Anhänger von A autoritärer als Anhänger von B?
  * Konkret aus Studie (Bild des psychisch Kranken): Kooperationsbereitschaft und Vorurteile.

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

2. *Statistisches Modell*: $X_1, \ldots, X_n$ i.i.d. Stichprobe aus Gruppe A, $Y_1, \ldots, Y_m$ i.i.d. Stichprobe aus Gruppe B

$$X_i \sim \mathcal{N}(\mu_X; \sigma_X^2) \qquad Y_i \sim \mathcal{N}(\mu_Y; \sigma_Y^2).$$

Zunächst seien die Varianzen σ_X^2 und σ_Y^2 als bekannt angenommen.

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

Gruppe A: Kooperative

Gruppe B: Primärverweigerer

2.4.4.5 Zwei-Stichproben-Gauss-Test

3. *Formulieren der statistischen Hypothesen*:

$$\begin{array}{llll} 1 & H_0: & \mu_X \leq \mu_Y & H_1: \quad \mu_X > \mu_Y \\ 2 & H_0: & \mu_X \geq \mu_Y & H_1: \quad \mu_X < \mu_Y \\ 3 & H_0: & \mu_X = \mu_Y & H_1: \quad \mu_X \neq \mu_Y \end{array}$$

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

4. *Festlegen eines Signifikanzniveaus*:

   Allgemein α, hier z.B. $\alpha = 0.01$.

5. *Festlegen einer Testgröße und einer Kritischen Region*:

Testgröße: Vergleich der arithmetischen Mittel $\bar{X}$ und $\bar{Y}$ basierend auf

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$$

T ist $\mathcal{N}(\mu_X - \mu_Y, 1)$ verteilt (vgl. Kap. 1.6). Falls $\mu_X = \mu_Y$ ist, gilt

$$T \sim \mathcal{N}(0, 1)$$

Festlegen der Kritischen Region:

2.4.4.6 Zwei-Stichproben-t-Test

Abwandlung von Schritt 5 bei unbekannten Varianzen:

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

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

Variante I: Ist bekannt, dass die Varianzen gleich sind, so schätzt man sie mittels S_X^2 und S_Y^2 und betrachtet

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

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

Vorgehen bei der Konstruktion der Kritischen Region analog zu vorher:

Im Beispiel:

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

Variante II Sind die Varianzen unbekannt und können nicht als gleich angenommen werden, so kann man für großes n und großes m mit

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}$$

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 n, m ist außerordentlich kompliziert; sie wird in der Vorlesung nicht betrachtet, weshalb – aus Übungsgründen – im Rahmen der Veranstaltung bei ungleichen Varianzen stets mit der approximativen Variante gearbeitet werden darf.

2.4.4.7 Erweiterungen

Erweiterung auf allgemeinere Hypothesen Oft benötigt man Formeln für allgemeine Hypothesen der Form

$$\mu_X - \mu_Y \leq \delta \iff \mu_X \leq \mu_Y + \delta, \text{ mit } \delta \text{ als „relevanten Unterschied“}$$

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

Varianzanalyse:

Sollen die Mittelwerte in mehr als zwei Gruppen verglichen werden, verwendet man die Varianzanalyse.

Dabei testet man zunächst, ob es einen signifikanten Unterschied zwischen mindestens zwei Mittelwerten gibt.

Danach wird in einem sogenannten „post-hoc-Test“ jeder Vergleich einzeln bzw. nach inhaltlichen Hypothesen überprüft.

Die Varianzanalyse lässt sich auch als lineares Modell auffassen (vgl. Statistik I und Ende der Vorlesung).

2.4.5 Gauss-Test und t-Test für verbundene Stichproben

2.4.5.8 Zum Begriff der verbundenen Stichprobe

Verbundene Stichproben: Vergleich zweier Merkmale X und Y, die jetzt an *denselben Einheiten erhoben werden*.

Vorsicht: Leicht zu verwechseln mit vorheriger Fragestellung!

Beispiele:

- 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:

1) Man bestimmt zufällig zwei Gruppen, in der $einen$ erhebt man X, in der $anderen$ Y.
   Danach Vergleich der Mittelwerte wie im vorherigen Kapitel beschrieben.

2) Man erhebt an $jeder$ Person $beide$ Merkmale.

Warum ist das zweite Vorgehen im Allgemeinen besser?

2.4.5.9 Konstruktion der Tests:

$$\begin{array}{lll} X_1, \ldots, X_n & i.i.d. & \mathcal{N}(\mu_X, \sigma_X^2) \\ Y_1, \ldots, Y_n & i.i.d. & \mathcal{N}(\mu_Y, \sigma_Y^2) \end{array}$$

(X_i, Y_i) unabhängig, $i = 1, \cdots, n$

Zum Testen von Hypothesen der Form

$$\begin{array}{llllll} 1 & H_0: & \mu_X \leq \mu_Y & \text{gegen} & H_1: & \mu_X > \mu_Y \\ 2 & H_0: & \mu_X \geq \mu_Y & \text{gegen} & H_1: & \mu_X < \mu_Y \\ 3 & H_0: & \mu_X = \mu_Y & \text{gegen} & H_1: & \mu_X \neq \mu_Y \end{array}$$

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

$$\mathrm{E}\,\mu_D = \mathrm{E}(D_i) =$$

und für die Varianz σ_D^2 (da ja X_i und Y_i nicht unabhängig sind)

$$\sigma_D^2 := \mathrm{Var}(X_i - Y_i) \quad =$$
$$=$$

also

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

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

$$\begin{array}{llllll} 1' & H_0: & \mu_D \leq 0 & \text{gegen} & H_1: & \mu_D > 0 \\ 2' & H_0: & \mu_D \geq 0 & \text{gegen} & H_1: & \mu_D < 0 \\ 3' & H_0: & \mu_D = 0 & \text{gegen} & H_1: & \mu_D \neq 0, \end{array}$$

und man kann unmittelbar die Tests aus 2.4.3 anwenden. Sind die Varianzen unbekannt, so kann man σ_D^2 aus den Differenzen D_i, $i = 1, \ldots, n$ schätzen. Zur Prüfung ist dann die t-Verteilung heranzuziehen.

2.4.6 χ^2-Tests am Beispiel des χ^2-Unabhängigkeitstests

- Tests basierend auf diskreten bzw. diskretisierten Merkmalen.
- Grob gesprochen eignen sich χ^2-Tests, um zu entscheiden, ob eine beobachtete Verteilung signifikant von einer Modellverteilung abweicht.
- Haupttypen:
  - χ^2-$Unabhängigkeitstest$: Weicht die beobachtete gemeinsame Verteilung von der unter Unabhängigkeit zu erwartenden signifikant ab?
  - χ^2-$Anpassungstest$ z.B. Abweichung von der Gleichverteilung

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

  - χ^2-$Homogenitätstest$: Gilt in k Subpopulationen jeweils dieselbe Verteilung?

Hier nur ausführlicher: χ^2-Unabhängigkeitstest

In Beispiel eingebettet (vgl. Anfang des Kapitels):

1. Aktive Stellung im öffentlichen Leben beeinflusst Kooperationsbereitschaft im Interview

2. *Statistische Modelle*: Zwei diskrete Merkmale X und Y

   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*: (jetzt, wie gewohnt, i, j Index für Ausprägungen)

   H_0: Es herrscht Unabhängigkeit
   H_1: Es herrscht keine Unabhängigkeit

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

gegen H_1 : $P(X = x_{i^*}, Y = y_{j^*}) \neq P(X = x_{i^*}) \cdot P(Y = y_{j^*})$ für mindestens *ein* Paar i^*, j^*

4. *Festlegen des Signifikanzniveaus*

5. *Testgröße und kritische Region*

   Sensitive Testgröße: χ^2-Abstand

Beobachtete Tafel der absoluten Häufigkeiten:

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

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

$$f_{ij} = \frac{h_{ij}}{n} \text{ Schätzer für } P(X = x_i,\ Y = y_j).$$

Zu vergleichen mit der Unabhängigkeitstafel: $\tilde{h}_{ij} = \frac{h_{i\bullet} h_{\bullet j}}{n}$, den unter Unabhängigkeit und gleichen Randverteilungen zu erwartenden Besetzungszahlen

	X/Y	Y 1	…	m	
	1	$\frac{h_{\bullet 1} h_{1\bullet}}{n}$	…		$h_{1\bullet}$
X	⋮		$\frac{h_{i\bullet} h_{\bullet j}}{n}$		$h_{i\bullet}$
	k		…		$h_{k\bullet}$
		$h_{\bullet 1}$	$h_{\bullet j}$	$h_{\bullet m}$	1

Analoges gilt für die relativen Häufigkeiten $f_{ij} = \frac{h_{ij}}{n}$; $\tilde{f_{ij}} = f_{j\bullet} \cdot f_{\bullet j}$

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

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

Teststatistik:

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

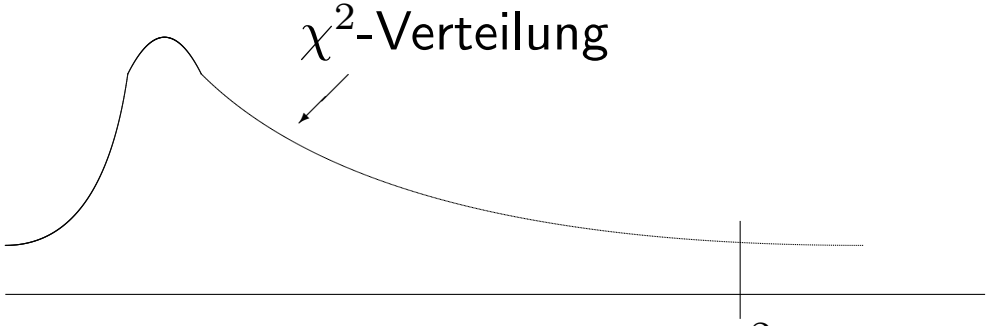
$$
T = \sum_{\text{alle Zellen}} \frac{\text{„(beobachtet} - \text{erwartet)}^2\text{“}}{\text{Normierung}}
$$

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

Kritische Region: Je stärker H_0 verletzt ist, umso stärker weichen 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

$$P(T \in KR|H_0) = P(T \geq z|H_0) \leq \alpha$$



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

Beachte: „Gleichheitsnullhypothese“, aber einseitige kritische Region

Bsp. 2.24.

Beobachtete Tabelle $f_{ij} = \left(\frac{h_{ij}}{n}\right)$:

		kooperativ		
		ja	nein	
Mitglied	ja	0.27	0.05	0.32
	nein	0.53	0.15	0.68
		0.8	0.2	1

Unabhängigkeitstabelle $\tilde{f}_{ij} = \left(\frac{\tilde{h}_{ij}}{n}\right)$:

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

Hier hat man 1 Freiheitsgrad, denn

$$(k-1)\cdot(m-1) = (2-1)\cdot(2-1) = 1$$

Bei $\alpha = 0.1$ erhält man $\chi^2_{1-\alpha}(1) = 2.7055$, also $KR = [2.7055, \infty)$.

Die Teststatistik T hat hier den Wert

$$t = 352 \cdot \left(\frac{(0.27 - 0.256)^2}{0.256} + \frac{(0.53 - 0.544)^2}{0.544} + \frac{(0.05 - 0.064)^2}{0.064} + \frac{(0.15 - 0.136)^2}{0.136} \right) = 1.98$$

Hier ist das Ergebnis stark rundungsabhängig. Dies wäre ein Argument, mit absoluten Häufigkeiten zu rechnen! (Bei Berechnung am Computer sollten Rundungsfehler praktisch keine Rolle mehr spielen.)

Testentscheidung: Da

$$t = 1.98 \notin KR$$

, kann die Nullhypothese nicht abgelehnt werden; ein Zusammenhang zwischen Aktivität im öffentlichen Leben und der Kooperationsbereitschaft konnte zum Signifikanzniveau von 10% nicht nachgewiesen werden.

2.4.7 Zur praktischen Anwendung statistischer Tests: Testentscheidungen und Statistik-Software, p-Wert

2.4.7.10 Grundkonzept

Statistik-Software löst Test-Probleme nicht direkt über die kritische Region, sondern berechnet meist den sogenannten $p\text{-}Wert$, also die Wahrscheinlichkeit unter H_0 mindestens einen so stark für die Alternative sprechenden Wert zu erhalten, wie den tatsächlich beobachteten Wert der Teststatistik. Dies ist die Wahrscheinlichkeit für den Fehler 1. Art, den man tatsächlich machen würde, wenn man die Nullhypothese aufgrund der konkreten Daten ablehnen würde. Man kann also sagen:

H_0 kann genau dann abgelehnt werden, wenn der p-Wert kleiner gleich dem vorgegebenen Signifikanzniveaus ist.

Also: das bisherige Konstruktionsprinzip lautete:

Nullhypothese ablehnen, wenn Wert t von $T \in$ kR, wobei

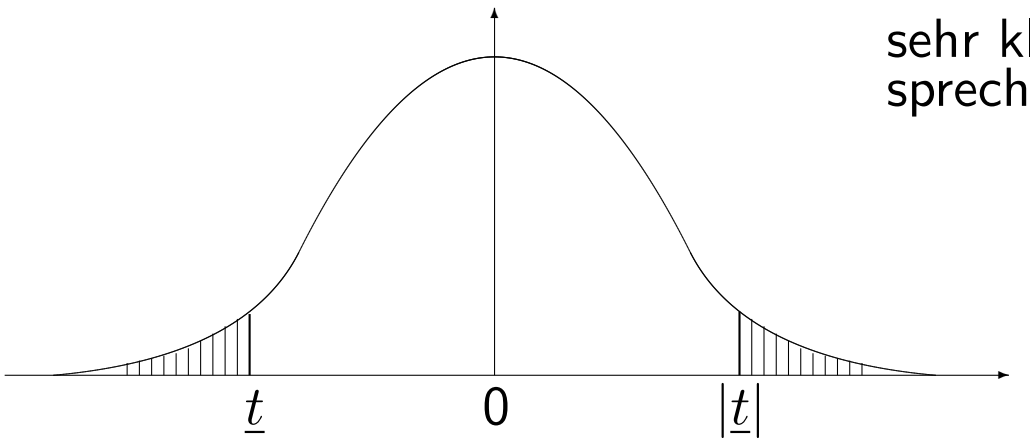
$$P(T \in \text{kR} || H_0) \leq \alpha$$

Üblicherweise: KR „extreme Werte von T, die gegen H_0 sprechen“

Jetzt Sicht wechseln

konkreter Wert t der Teststatistik
Berechne P(T mindestens so extrem gegen H_0 sprechend wie t) Ist diese Wahrscheinlichkeit $\leq \alpha$, so ist der Bereich „extremer als t“ als kritische Region geeignet.

Beispiel: zweiseitiger Test: $H_0 : \mu = \mu_0 \qquad H_1 : \mu \neq \mu_0$



sehr kleine und sehr große Werte sprechen gegen H_0

Vorsicht vor schematischer Anwendung

Bei vielen Tests ist hier aber Vorsicht geboten. Die vom Programm betrachtete Nullhypothese muss nicht die tatsächlich interessierende Nullhypothese sein!

Beim Gauss- und t-Test sind beispielsweise drei verschiedene Nullhypothesen möglich:

$$H_0 : \mu \leq \mu_0, \quad H_0 : \mu = \mu_0, \quad H_0 : \mu \geq \mu_0$$

SPSS gibt hier einen „zweiseitigen p-Wert“ (2-tailed significance) an, der zur Hypothese $H_0 : \mu = \mu_0$ gegen $H_1 : \mu \neq \mu_0$ und damit zur kritischen Region $(-\infty, -z_{1-\frac{\alpha}{2}}) \cup$

$(z_{1-\frac{\alpha}{2}}, \infty)$ gehört.

Möchte man dagegen $H_0 : \mu \geq \mu_0$ testen, so darf man H_0 ablehnen, falls

1. der Wert der Teststatistik kleiner als 0 ist (also „auf der richtigen Seite liegt“) und

2. falls gilt: p-Wert $\leq 2 \cdot$ Signifikanzniveau.

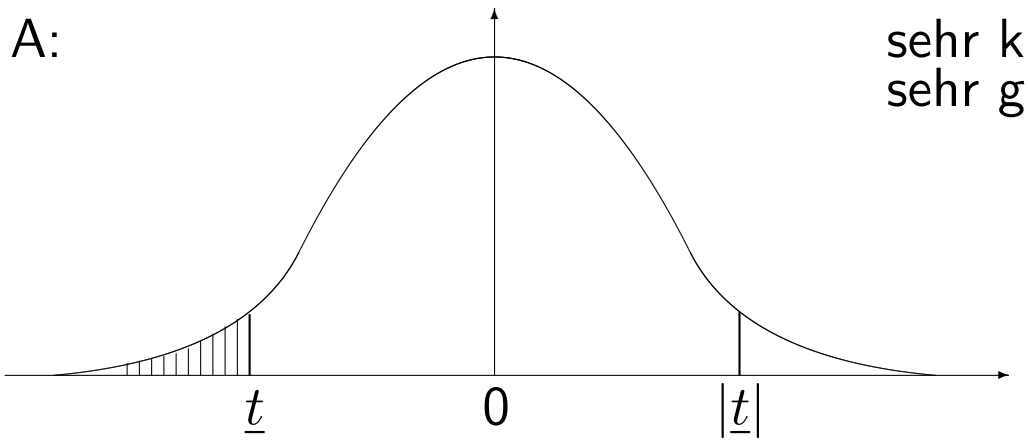
(analog für $H_0 : \mu \geq \mu_0$)

Nochmals detailierter:

Wenn jetzt ein einseitiger Test vorliegt, dann sind nicht mehr beide Seiten gegen H_0 sprechend, sondern nur noch eine.

Z.B. $H_0 : \mu \geq \mu_0$ gegen $H_1 : \mu < \mu_0$

Situation A:



sehr kleine, aber nicht mehr
sehr große Werte sprechen gegen H_0

z.B. SPSS berechnet nach wie vor zweiseitigen p-Wert. Für den gesuchten p-Wert gilt

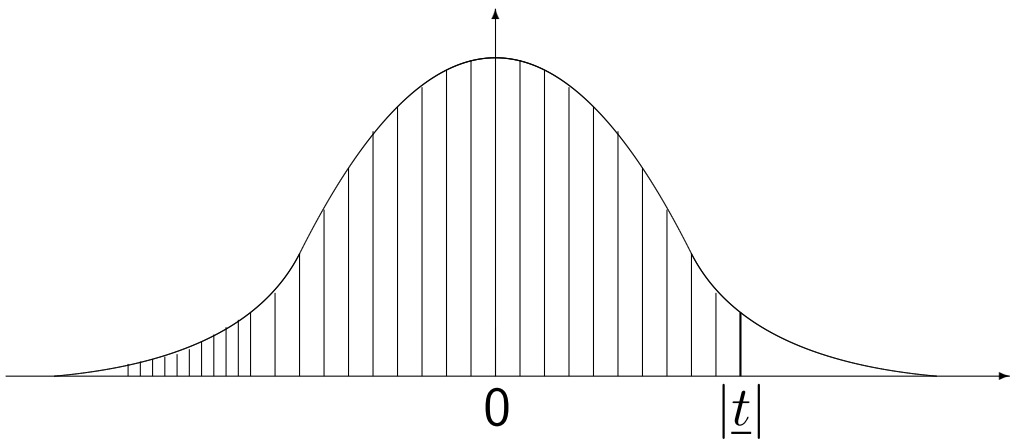
$$\text{p-Wert} = \frac{\text{zweiseitiger p-Wert}}{2}$$

Man kann in dieser Situation H_0 ablehnen, falls p-Wert $\leq \alpha$, also zweiseitiger p-Wert $\leq 2\alpha$.

Dabei muss vor einer zu schematischen Vorgehensweise gewarnt werden. Wurde in derselben Situation $H_0 : \mu \geq \mu_0$ gegen $H_1 : \mu < \mu_0$ statt $\underline{t}$ der Wert $\bar{t} = |\underline{t}|$ beobachtet,

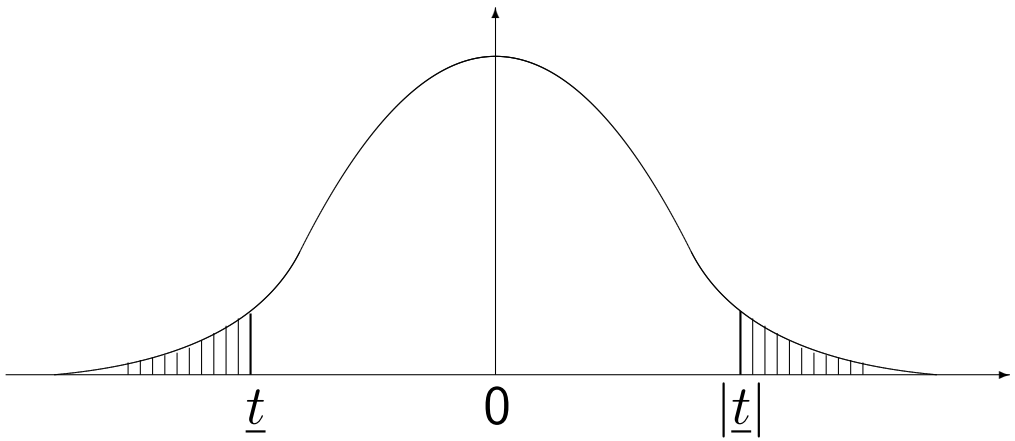
so ist der korrekte p-Wert: P(T mindestens so extrem wie $\bar{t}$):

Situation B:



und H_0 darf keinesfalls abgelehnt werden, Wahrscheinlichkeit von Fehlschluss deutlich größer als 50%.

SPSS berechnet aber den zweiseitigen p-Wert:



Man muss also, wenn rechnerisch gilt p-Wert (zweiseitig) $\leq 2\alpha$ noch sicherstellen, dass der beobachtete Wert von t auf der „richtigen Seite liegt“, d.h. Situation A und nicht Situation B vorliegt.

2.4.8 Zur Hypothesenwahl:

Es sei nochmal daran erinnert: Statistisch gesichert zur vorgegebenen Fehlerwahrscheinlichkeit ist nur die Ablehnung der Nullhypothese. Hat man die Wahl (bei einseitigen Tests), so setzt man das, was man zeigen will, in die Alternativhypothese.

2.4.9 Dualität von Test und Konfidenzintervall:

Man betrachte die Fragestellung

$H_0 : \mu = \mu_0$ gegen $H_1 : \mu \neq \mu_0$. H_0 wird abgelehnt, wenn

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

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

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

Vergleiche diese Ablehnbereiche mit dem Konfidenzintervall

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

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

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

Beispiel Normalverteilung: $\bar{X}$ unbekannt, σ bekannt

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

- Werte „in der Mitte“$\Rightarrow$ Konfidenzintervall

- extreme Werte $\Rightarrow$ Test

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

$$H_0 : \vartheta = \vartheta_0$$

unmittelbar testen:

Manche Softwarepakete geben deshalb bei bestimmten Prozeduren nur Konfidenzintervalle, aber keine Tests an.

Bsp. 2.25. [Beispiel Wahlumfrage (Fortsetzung von Bsp. 2.19)]

$n = 500$, $\bar{x} = 46.5\%$ Anteil Rot/Grün, $\gamma = 95\%$

Man erhielt das Konfidenzintervall $[0.421; 0.508]$. Da $\pi = 0.5$ im Konfidenzintervall liegt, kann die Hypothese $\pi = 0.5$ nicht abgelehnt werden.

Bsp. 2.26. [Fortsetzung von Bsp. 2.18]

Man interessiert sich, ob gewisse Gummibärchenpackungen genau die angegebene Füllmenge von $250g$ enthalten, möchte also $H_0 : \mu = 250g$ gegen $H_1 : \mu \neq 250g$ zu $\alpha = 0.05$ testen.

Hat man zu $\gamma = 0.95$ das – 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.

2.4.10 Signifikanz versus Relevanz:

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

1. 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.

2. Andererseits kann es sein, dass sich bei großen Stichprobenumfängen selbst minimale Abweichungen als signifikant erweisen. Nicht jede statistisch signifikante Abweichung ist daher auch inhaltlich relevant, weshalb Vorsicht bei der inhaltlichen Interpretation gerade bei großen Datensätzen angebracht ist. Insbesondere darf deshalb auch der p-Wert nicht als Maß für die Stärke einer Abweichung von der Nullhypothese interpretiert werden.

$X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ mit $\sigma^2 = 1$

z.B. $H_0 : \mu \leq 100 \qquad H_1 : \mu > 100$

$$\begin{aligned} \bar{X} &= 100 + \varepsilon \\ T &= \frac{\bar{X} - \mu_0}{\sigma}\sqrt{n} = \frac{\varepsilon}{\sigma}\sqrt{n} \end{aligned}$$

H_0 ablehnen $\iff$ $T \geq z_{1-\alpha} = 1.65$

$$\begin{aligned} \frac{\varepsilon}{\sigma}\sqrt{n} &> 1.65 \\ \varepsilon &> \frac{1.65}{\sqrt{n}}\sigma \end{aligned}$$

z.B. $\sqrt{n} = 10 \Leftrightarrow n = 100 :$ $\varepsilon > 0.165$

$\sqrt{n} = 100 \Leftrightarrow n = 10000 :$ $\varepsilon > 0.0165$

Mögliche Auswege:

- Ergebnisse kritisch betrachten.
- Betrachtung sogenannter *Effektstärkemaße*.
- Untersuche statt der Hypothese „$\mu_A > \mu_B$“ die Hypothese „$\mu_A > \mu_B + \delta$“ mit (inhaltlich) relevantem Unterschied δ.

2.4.11 Multiple Testprobleme:

- Gegeben sei ein rein zufälliger Datensatz mit 50 Variablen ohne irgendeinen Zusammenhang.
- Man testet alle Variablenpaare auf einen Zusammenhang

$$\Rightarrow \binom{50}{2} = 1225$$

Tests.

Bei vorgegebener Irrtumswahrscheinlichkeit von 5% gilt für die Anzahl fälschlich verworfener Nullhypothesen $X \sim B(1225, 0.05)$ und somit $\mathrm{E}(X) = 61,25$.

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

$\Rightarrow$ wenige, sinnvolle Hypothesen $vorher\ inhaltlich$ überlegen und nur diese testen!

- Es gibt Ansätze, wie man bei großen Hypothesensystemen diesem Problem entkommt:

  $\Rightarrow$ Theorie des multiplen Testens.

  Z.B. Bonferroni-Adjustierung des Irrtumswahrscheinlichkeit: Statt α betrachte man α/(Anzahl der durchzuführenden Tests). Diese spezielle Korrektur ist aber meist überkonservativ und kann durch bessere –aber komplexere– Korrekturen ersetzt werden.

2.4.12 Nichtparametrische Tests

- Bis auf den χ^2-Unabhängigkeits-Test bauen alle Tests auf der (zumindestens approximativen Gültigkeit der) Normalverteilungsannahme auf.
- Problematisch, z.B.
  - bei kleinen Stichprobenumfängen
  - oder bei ordinalen Daten mit wenigen unterschiedlichen Ausprägungen.
- Hier kann die unreflektierte Anwendung der Standardtests zu krassen Fehlergebnissen führen.
- Ein wichtiger Ausweg: nichtparametrische Tests = „Verteilungsfreie Verfahren“
- Hier wird die Information in den Beobachtungen auf Ränge, bzw. größer/kleiner Vergleiche reduziert.
- Bekannteste Beispiele: Wilcoxon-Test, Vorzeichentest.