

2.4.3 Typische Tests I: Tests auf Lageparameter

Hier werden exemplarisch nur wenige, ausgewählte Tests, die typisch sind, besprochen. Das Grundprinzip ist bei anderen Tests analog.

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

Wir betrachten hier 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, \dots, 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*:

$$\text{Fall 1: } H_0 : \mu \leq \mu_0 \quad H_1 : \mu > \mu_0$$

$$\text{Fall 2: } H_0 : \mu \geq \mu_0 \quad H_1 : \mu < \mu_0$$

$$\text{Fall 3: } H_0 : \mu = \mu_0 \quad H_1 : \mu \neq \mu_0$$

Gleichheitszeichen immer bei H_0 !

4. *Festlegen des Signifikanzniveaus*: Wir rechnen im Folgenden allgemein. Übliche Werte sind, wie gesagt: 10%, 5%, 1%.

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. 2.26.

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 : Seien also X_1, \dots, X_n eine *i.i.d.* Stichprobe, wobei X_i , $i = 1, \dots, n$ jeweils normalverteilt sei mit unbekanntem Mittelwert μ und *unbekannter* Varianz σ^2 .

Analoges Vorgehen zur Konstruktion des Tests, aber

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

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, \dots, 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:*

$$1 \quad H_0 : \pi \leq \pi_0 \quad H_1 : \pi > \pi_0$$

$$2 \quad H_0 : \pi \geq \pi_0 \quad H_1 : \pi < \pi_0$$

$$3 \quad H_0 : \pi = \pi_0 \quad H_1 : \pi \neq \pi_0$$

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*: Für großen Stichprobenumfang n und wahren Anteil π kann man mit der Normalapproximation arbeiten (vgl. Kap.1.7 und Kap. 2.3). Es gilt

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

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

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

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, \dots, X_n i.i.d. Stichprobe aus Gruppe A ,

Y_1, \dots, Y_m i.i.d. Stichprobe aus Gruppe B ,

$$X_i \sim \mathcal{N}(\mu_X; \sigma_X^2) \quad \text{und} \quad Y_i \sim \mathcal{N}(\mu_Y; \sigma_Y^2), i = 1, \dots, n.$$

X_i und Y_j stochastisch unabhängig für beliebige i und j , $i = 1, \dots, n$; $j = 1, \dots, m$. Zunächst seien die Varianzen σ_X^2 und σ_Y^2 als bekannt angenommen.

X : Vorurteilsindex aus Fragebatterie mit Statements (1, ..., 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:*

$$1 \quad H_0 : \mu_X \leq \mu_Y \quad H_1 : \mu_X > \mu_Y$$

$$2 \quad H_0 : \mu_X \geq \mu_Y \quad H_1 : \mu_X < \mu_Y$$

$$3 \quad H_0 : \mu_X = \mu_Y \quad H_1 : \mu_X \neq \mu_Y$$

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

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 unbekanntem Varianzen:

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

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:

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

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

nicht X und Y , sondern X und \tilde{Y} vergleicht, wobei $\tilde{Y}_i = Y_i + \delta$, $i = 1, \dots, m$.

Es gilt $\tilde{Y} \sim \mathcal{N}(\mu_{\tilde{Y}}, \sigma_{\tilde{Y}}^2)$, wobei $\mu_{\tilde{Y}} = \mu_Y + \delta$.

Obige Hypothese ist also äquivalent zur Hypothese

$$\mu_X \leq \mu_{\tilde{Y}}$$

die mit den behandelten Methoden getestet werden kann.

Eine weitere Möglichkeit ergibt sich letztendlich über die Dualität zu Konfidenzintervallen (vgl. Abschnitt 2.4.9).

Varianzanalyse:

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

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:

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

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

Zum Testen von Hypothesen der Form

- 1 $H_0 : \mu_X \leq \mu_Y$ gegen $H_1 : \mu_X > \mu_Y$
- 2 $H_0 : \mu_X \geq \mu_Y$ gegen $H_1 : \mu_X < \mu_Y$
- 3 $H_0 : \mu_X = \mu_Y$ gegen $H_1 : \mu_X \neq \mu_Y$

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

$$\mu_D = E(D_i) =$$

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

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

also

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

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

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

und man kann unmittelbar die Tests aus Kapitel 2.4.3 anwenden. Sind die Varianzen unbekannt, so kann man σ_D^2 aus den Differenzen $D_i, i = 1, \dots, n$ 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* Abweichung von einer Verteilung z.B. der diskreten 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

$(X_1, Y_1), \dots, (X_n, Y_n)$ 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			
		y_1	\dots	y_m	
X	x_1	h_{11}	\dots	h_{1m}	$h_{1\bullet}$
	\vdots		h_{ij}		
	x_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			
		y_1	\dots	y_m	
X	x_1	$\frac{h_{\bullet 1}h_{1\bullet}}{n}$	\dots		$h_{1\bullet}$
	\vdots		$\frac{h_{i\bullet}h_{\bullet j}}{n}$		$h_{i\bullet}$
	x_k		\dots		$h_{k\bullet}$
		$h_{\bullet 1}$	$h_{\bullet j}$	$h_{\bullet m}$	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			
		y_1	\dots	y_m	
X	x_1	$\frac{h_{11}}{n}$	\dots	$\frac{h_{1m}}{n}$	$\frac{h_{1\bullet}}{n}$
	\vdots		$\frac{h_{ij}}{n}$		
	x_k	$\frac{h_{k1}}{n}$	\dots	$\frac{h_{km}}{n}$	$\frac{h_{k\bullet}}{n}$
		$\frac{h_{\bullet 1}}{n}$	\dots	$\frac{h_{\bullet m}}{n}$	1

X/Y		Y			
		y_1	\dots	y_m	
X	x_1	$\frac{h_{\bullet 1} h_{1 \bullet}}{n^2}$	\dots		$\frac{h_{1 \bullet}}{n}$
	\vdots		$\frac{h_{i \bullet} h_{\bullet j}}{n^2}$		$\frac{h_{i \bullet}}{n}$
	x_k		\dots		$\frac{h_{k \bullet}}{n}$
		$\frac{h_{\bullet 1}}{n}$	$\frac{h_{\bullet j}}{n}$	$\frac{h_{\bullet m}}{n}$	1

Teststatistik:

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

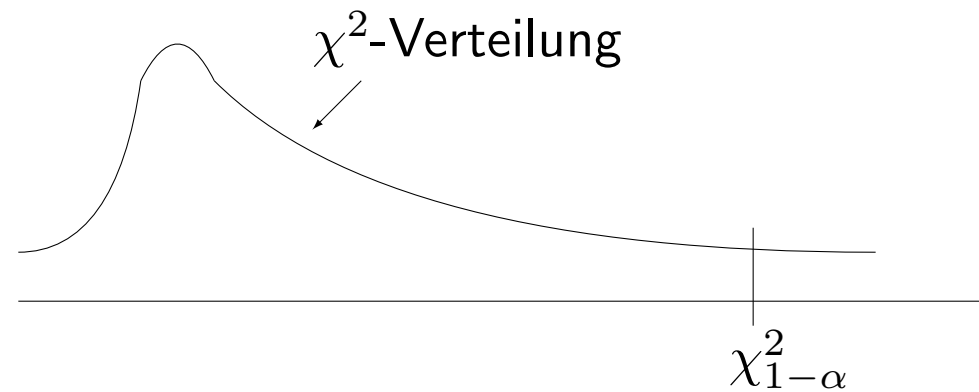
$$T = \sum_{\text{alle Zellen}} \frac{\text{„(beobachtet – erwartet)}^2\text{“}}{\text{Normierung}}$$

Unter H_0 gehorcht T approximativ einer sogenannten χ^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

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



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

Beachte: „Gleichheitsnullhypothese“, aber einseitige kritische Region

Bsp. 2.27.

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 $\chi_{1-\alpha}^2(1) = 2.7055$, also $KR = [2.7055, \infty)$.

Die Teststatistik T hat hier den Wert ⁵

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

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.

⁵Hier ist das Ergebnis stark rundungsabhängig. Dies wäre ein Argument, mit absoluten Häufigkeiten zu rechnen! (Bei Berechnung am Computer sollten hier Rundungsfehler praktisch keine Rolle mehr spielen.)

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

Statistik-Software löst Test-Probleme nicht direkt über die kritische Region, sondern berechnet meist den sogenannten p -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 Signifikanzniveau ist.

Also: das bisherige Konstruktionsprinzip lautete:

Nullhypothese ablehnen, wenn Wert t von $T \in \text{KR}$, wobei

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

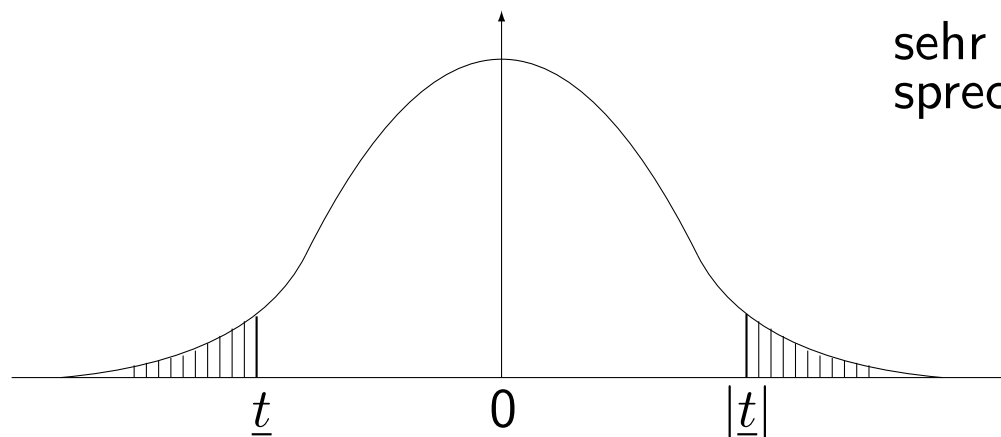
Üblicherweise: KR „extreme Werte von T , die gegen H_0 sprechen“

Jetzt Sicht wechseln

konkreter Wert t der Teststatistik

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

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



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:

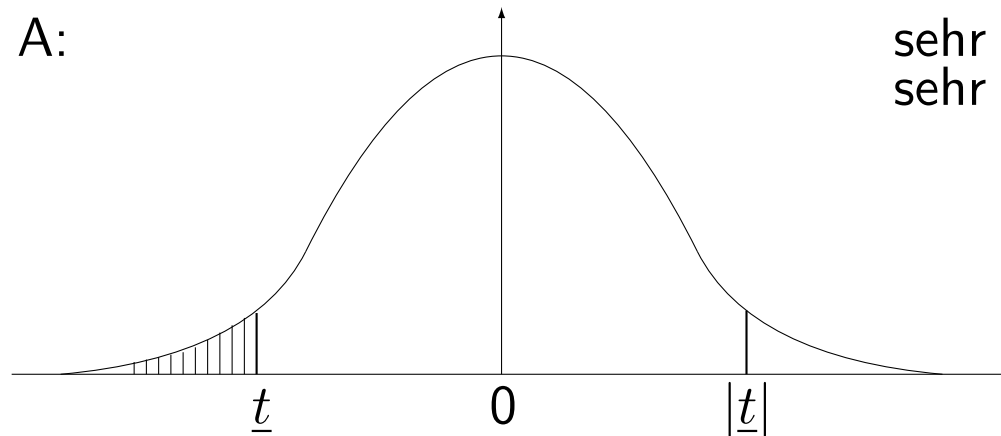
$$H_0 : \mu \leq \mu_0, \quad H_0 : \mu = \mu_0, \quad H_0 : \mu \geq \mu_0$$

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.

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

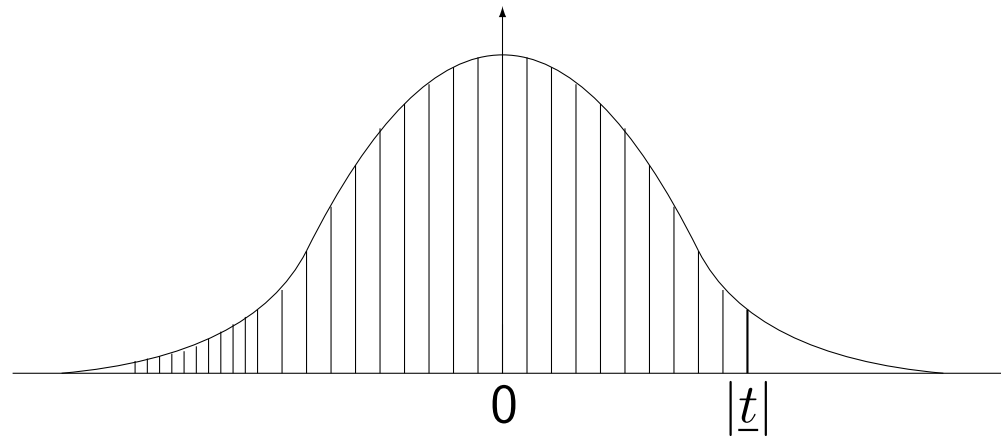
z.B. SPSS berechnet nach wie vor zweiseitigen p-Wert. Für den gesuchten p-Wert gilt

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

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

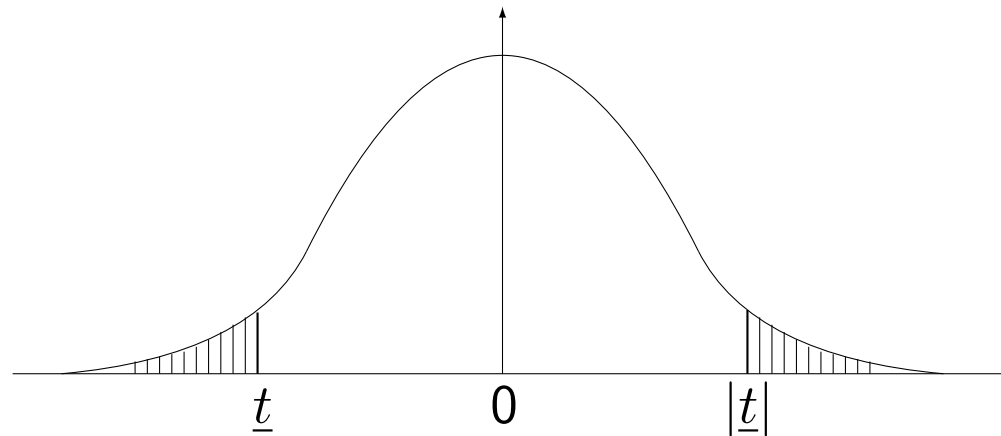
Dabei muss wie gesagt vor einer zu schematischen Vorgehensweise gewarnt werden. Wurde in derselben Situation $H_0 : \mu \geq \mu_0$ gegen $H_1 : \mu < \mu_0$ statt \underline{t} der Wert $\bar{t} = |\underline{t}|$ beobachtet, so ist der korrekte p-Wert: $P(T \text{ mindestens so extrem wie } \bar{t})$:

Situation B:



und H_0 darf keinesfalls abgelehnt werden; die Wahrscheinlichkeit eines Fehlschlusses ist deutlich größer als 50%.

SPSS berechnet aber den zweiseitigen p-Wert:



Man muss also, wenn rechnerisch gilt $p\text{-Wert (zweiseitig)} \leq 2\alpha$ 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

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

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:

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

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 abzulehnen ist.

Eine praktische Konsequenz daraus: Gegeben ein Konfidenzintervall

$$[A(X_1, \dots, X_n), B(X_1, \dots, X_n)]$$

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. 2.28. [Beispiel Wahlumfrage (Fortsetzung von Bsp. 2.22)]

$n = 500$, $\bar{x} = 46.5\%$ 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. 2.29. [Fortsetzung von Bsp. 2.3.3]

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.

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

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

$$\bar{X} = 100 + \varepsilon$$

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

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

$$\frac{\varepsilon}{\sigma} \sqrt{n} > 1.65$$

$$\varepsilon > \frac{1.65}{\sqrt{n}} \sigma$$

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

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

Mögliche Auswege:

- Ergebnisse kritisch betrachten (wie eigentlich immer).
- 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. Es ergeben sich

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

einzelne Tests.

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

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

⇒ wenige, sinnvolle Hypothesen *vorher inhaltlich* überlegen und nur diese testen!

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

⇒ Theorie des multiplen Testens.

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

2.4.12 Nichtparametrische Tests

- Bei großen Stichprobenumfängen kann man die besprochenen t-Tests auch anwenden, wenn die einzelnen Variablen X_1, \dots, X_n nicht normalverteilt sind. Unter den dort formulierten Regularitätsbedingungen folgt aus dem zentralen Grenzwertsatz, dass die betrachteten Teststatistiken dann approximativ normalverteilt sind. Auch der betrachtete χ^2 -Unabhängigkeitstest setzt große Stichproben voraus.
- Problematisch wäre die Anwendung dieser Tests
 - 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“:
 - * Vorzeichentest: Vergleiche die Anzahl der Beobachtungen größer als der vermutete Median mit der Anzahl kleiner als der Median
 - * Wilcoxon-Tests: Arbeitet mit Rängen
- Hier wird die Information in den Beobachtungen auf Ränge, bzw. größer/kleiner Vergleiche reduziert.
- Bekannteste Beispiele: *Wilcoxon-Tests*, *Vorzeichentest*.