

## 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:** SchlieÙe von Stichprobe auf Zutreffen bestimmter Tatsachen in der Grundgesamtheit

Vorgehen:

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

**Bsp. 2.24.**

Studie zur Einstellung der Münchner Bevölkerung zu psychisch Kranken (1989, Methodenkurs Siegfried Lamnek).

Wir betrachten eine Teilstudie: Kooperationsbereitschaft in der Befragung.

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

*Aktiv* ↔ Altruismus  
↔ Interesse an öffentlichen Angelegenheiten  
⇒ eher bereit, die Rolle des Befragten einzunehmen

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

### 3. Operationalisierung:

- Aktiv im öffentlichen Leben

→ Verbandsmitgliedschaft ja/nein = Variable  $X$

- Kooperationsbereitschaft

→ Von den überhaupt nicht Kooperierenden hat man keine Daten, deshalb vergleicht man die sofort zur Antwort bereiten Befragten mit den „Primärverweigerern“, die sich erst nach langer Zeit zur Auskunft überreden ließen.

→ direkt kooperativ ja/nein → 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 $\chi^2$ (vgl. Statistik I)

Gegebene Daten (relative und absolute Häufigkeiten):

		kooperativ		
		ja	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.)

- Wie groß muss die Abweichung sein, um die Nullhypothese abzulehnen?
- Beachte: Die Daten entstammen einer Stichprobe, die mit einem Zufallsfehler behaftet ist. Selbst bei tatsächlich vorliegender Unabhängigkeit ist die Wahrscheinlichkeit, genau die Unabhängigkeitstafel zu beobachten, sehr gering.

## Kardinalfrage der Testtheorie:

Weichen die tatsächlichen Daten von der bei Gültigkeit der Nullhypothese zu erwartenden Situation „überzufällig“ stark ab, d.h. so stark, dass man die Abweichung nicht mehr nur der Zufallsstreuung zuschreiben kann?

Nur in diesem Fall ist die Nullhypothese abzulehnen.

## 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  $\chi^2$ -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:  $\chi^2=2.11$ )!  
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	✓
$H_0$ falsch	Fehler 2. Art	✓

## **Fehler 1. Art („ $\alpha$ -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 („ $\beta$ -Fehler“):

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

Dabei wird so vorgegeben, dass die **Wahrscheinlichkeit**, einen **Fehler 1.Art** zu begehen, **beschränkt** ist durch eine inhaltliche vorgegebene Schranke  $\alpha$  („**Signifikanzniveau**“)

Ein guter statistischer Test garantiert bei einem vorgegebenen niedrigen Signifikanzniveau (als Schranke für die Fehlerwahrscheinlichkeit 1. Art) auch eine möglichst geringe Fehlerwahrscheinlichkeit 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 ... beträgt mindestens 130 Einheiten)

### 2. *Formulieren eines geeigneten statistischen Modells*

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

Anteil:  $B(1, \pi)$

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

### 3. Formulierung der statistischen Hypothesen

- Umformulieren der substanzwissenschaftlichen Hypothesen als Hypothesen über  $\vartheta$ .
- 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:*

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

$\vartheta_0$  ist ein fester, vorgegebener Wert, der von inhaltlichem Interesse ist; zu unterscheiden von wahrem Wert  $\vartheta$ .

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 $\alpha$*

##### Wiederholung aus der Einleitung

Beim Testen sind folgende Entscheidungen möglich:

$H_0$ : ablehnen    oder     $H_0$ : beibehalten

Damit sind zwei verschiedene Arten von Fehlern möglich:

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

Man kann nicht beide Fehlerwahrscheinlichkeiten gleichzeitig kontrollieren! (Trade-off!)

⇒ asymmetrische Vorgehensweise:

Die Fehlerwahrscheinlichkeit 1. Art wird kontrolliert durch die Angabe einer Obergrenze  $\alpha$  („Signifikanzniveau“)

Typische Werte: üblich

$$\alpha = 0.1,$$

„marginal signifikant“

$$\alpha = 0.05,$$

„signifikant“

$$\alpha = 0.01,$$

„hoch signifikant“

$$\alpha = 0.001$$

„höchst signifikant“

## 5. Festlegen einer Testgröße und einer kritischen Region

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

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

Damit der Fehler 1. Art durch  $\alpha$  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  $\alpha$  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. 2.25.**

- Die wesentlichen 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 „gute(r) Forscher(in)“ sollte man deshalb immer das, was man zeigen will, in die Alternativhypothese schreiben.

z.B. Forscher(in) 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  $\alpha$  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 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  $\mu$  eines normalverteilten Merkmals, bzw. den Erwartungswert  $\pi$  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  $\mu$  und bekannter Varianz  $\sigma^2$ .

3. *Formulierung der statistischen Hypothesen*:

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

Gleichheitszeichen immer bei  $H_0$  !

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

10% : „marginal signifikant“  
5% : „signifikant“  
1% : „hoch signifikant“

5. *Testgröße*:

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

$T$  ist empfindlich gegenüber Abweichungen von  $H_0$ .

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  $\mu$  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  $\sigma^2$ : 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  $\mu$  und *unbekannter* Varianz  $\sigma^2$ .

Analoges Vorgehen zur Konstruktion des Tests, aber

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

$H_0$  ablehnen, falls

$$\begin{array}{ll} \text{Fall 1} & T \geq t_{1-\alpha}(n-1) \\ \text{Fall 2} & T \leq -t_{1-\alpha}(n-1) \\ \text{Fall 3} & T \leq -t_{1-\frac{\alpha}{2}}(n-1) \text{ oder } T \geq t_{1-\frac{\alpha}{2}}(n-1) \end{array}$$

### 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  $\pi$  der Anteil der Einheiten mit Ausprägung 1 in der Grundgesamtheit ist.

### 3. *Statistische Hypothesen:*

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

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*: Für großen Stichprobenumfang  $n$  und wahren Anteil  $\pi$  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  $\sigma_X^2$  und  $\sigma_Y^2$  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:

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

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

#### 4. Festlegen eines Signifikanzniveaus:

Allgemein  $\alpha$ , 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  $\sigma_X^2, \sigma_Y^2$  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:

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

**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.

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:

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

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

## Zum Testen von Hypothesen der Form

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

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

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

und für die Varianz  $\sigma_D^2$  (da ja  $X_i$  und  $Y_i$  nicht unabhängig sind)

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

also

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

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

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

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

## 2.4.6 $\chi^2$ -Tests am Beispiel des $\chi^2$ -Unabhängigkeitstests

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

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

- \*  $\chi^2$ -*Homogenitätstest*: Gilt in  $k$  Subpopulationen jeweils dieselbe Verteilung?

Hier nur ausführlicher:  $\chi^2$ -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:  $\chi^2$ -Abstand

*Beobachtete Tafel* der absoluten Häufigkeiten:

		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:

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

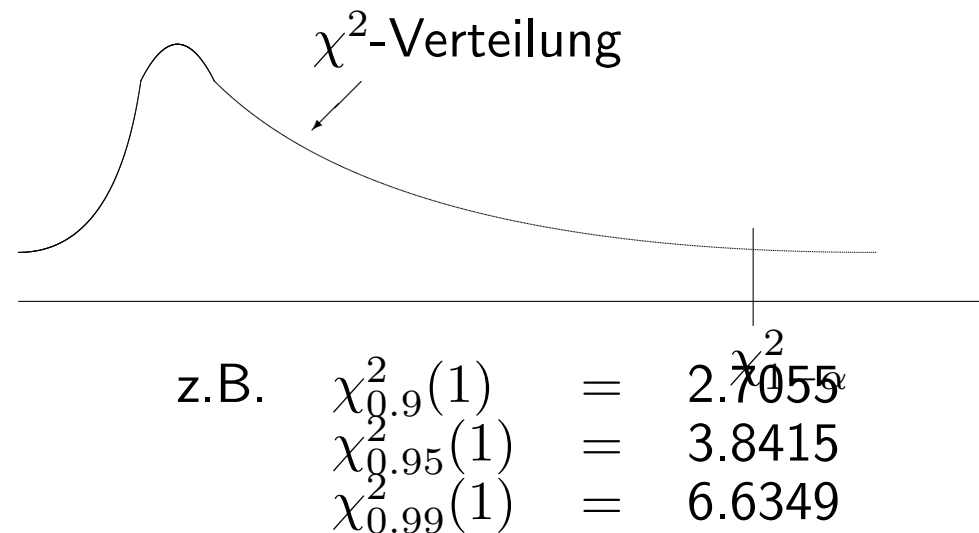
$$T = \sum_{\text{alle Zellen}} \frac{\text{„(beobachtet – erwartet)}^2\text{“}}{\text{Normierung}}$$

Unter  $H_0$  gehorcht  $T$  approximativ einer sogenannten  $\chi^2$ -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

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



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 <sup>4</sup>

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

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.

---

<sup>4</sup>Hier 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

### 2.4.7.10 Grundkonzept

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

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

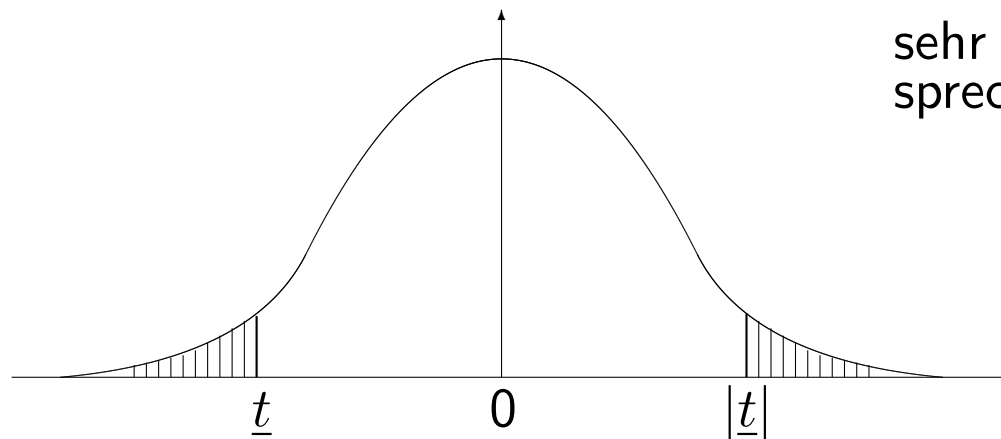
Üblicherweise: KR „extreme Werte von  $T$ , die gegen  $H_0$  sprechen“

### Jetzt Sicht wechseln

konkreter Wert  $t$  der Teststatistik

Berechne  $P(T$  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:  $H_0 : \mu = \mu_0$        $H_1 : \mu \neq \mu_0$



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:

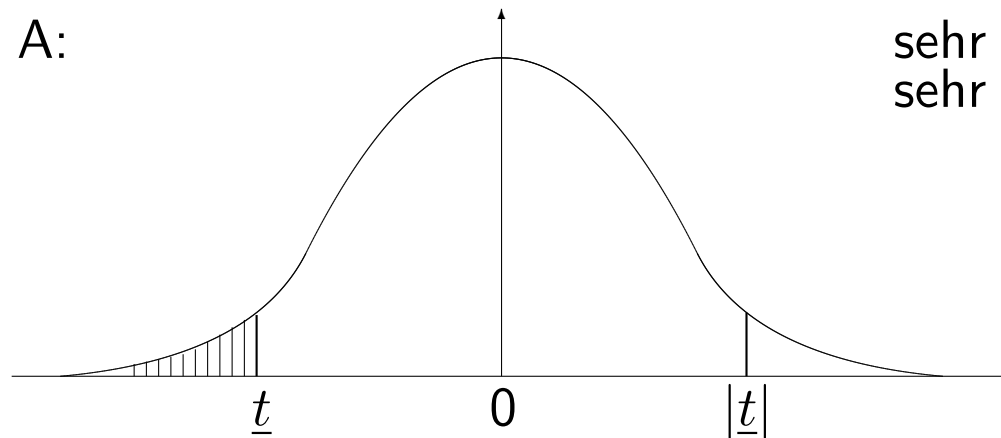
$$H_0 : \mu \leq \mu_0, \quad H_0 : \mu = \mu_0, \quad H_0 : \mu \geq \mu_0$$

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  $H_0$

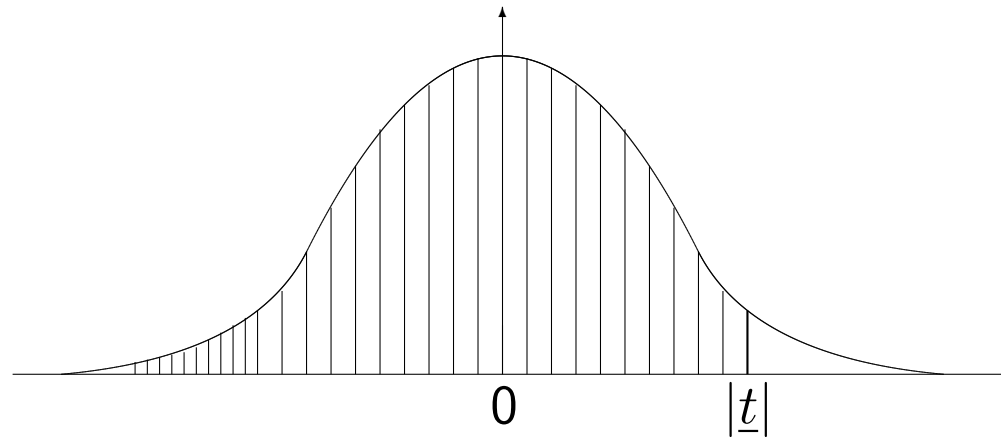
z.B. SPSS berechnet nach wie vor zweiseitigen p-Wert. Für den gesuchten p-Wert gilt

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

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

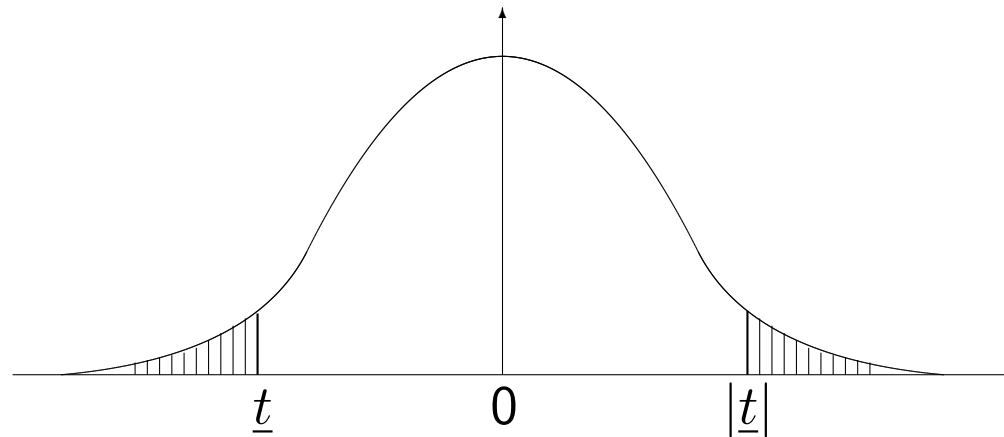
Dabei muss wie gesagt vor einer zu schematischen Vorgehensweise gewarnt werden. Wurde in derselben Situation  $H_0 : \mu \geq \mu_0$  gegen  $H_1 : \mu < \mu_0$  statt  $\underline{t}$  der Wert  $\bar{t} = |\underline{t}|$  beobachtet, so ist der korrekte p-Wert:  $P(T \text{ mindestens so extrem wie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

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  $\alpha$  und  $\gamma$  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, $\sigma$ 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  $\gamma$  genau aus all jenen Werten  $\vartheta_0$  eines Parameters  $\vartheta$ , 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  $\vartheta$ , 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.  $\sqrt{n} = 10 \iff n = 100 : \quad \varepsilon > 0.165$

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

## Mögliche Auswege:

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



### 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  $\alpha$  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  $\chi^2$ -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*.

## 2.5 Lineare Regressionsmodelle

### 2.5.1 Wiederholung aus Statistik I

Gegeben Datenpunkte  $(Y_i, X_i)$  schätze die beste Gerade  $Y_i = \beta_0 + \beta_1 X_i$ ,  $i = 1, \dots, n$ .

**Bsp. 2.30. [Kaffeeverkauf auf drei Flohmärkten]**

$X$  Anzahl verkaufter Tassen Kaffee

$Y$  zugehöriger Gewinn (Preis Verhandlungssache)

$i$	$x_i$	$y_i$	$y_i - \bar{y}$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	10	9	-1	0	0
2	15	21	11	5	25
3	5	0	-10	-5	25
	$\bar{x} = 10$	$\bar{y} = 10$			

Man bestimme die Regressionsgerade und interpretiere die erhaltenen KQ-Schätzungen!  
Welcher Gewinn ist bei zwölf verkauften Tassen zu erwarten?

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{0 \cdot (-1) + 5 \cdot 11 + (-5) \cdot (-10)}{0 + 25 + 25} = \frac{105}{50} = 2.1\end{aligned}$$

Mit der Erhöhung der Menge  $X$  um eine Einheit erhöht sich der Gewinn  $Y$  um 2.1 Einheiten, also ist  $\hat{b}$  so etwas wie der durchschnittliche Gewinn pro Tasse.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x} = 10 - 2.1 \cdot 10 = -11$$

„Grundlevel“, Gewinn bei 0 Tassen (Fixkosten).

Vorhergesagte Werte und Residuen:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i, \quad \hat{\epsilon}_i = y_i - \hat{y}_i$$

$$\hat{y}_1 = -11 + 2.1 \cdot 10 = 10 \quad \Rightarrow \hat{\epsilon}_1 = -1$$

$$\hat{y}_2 = -11 + 2.1 \cdot 15 = 20.5 \quad \Rightarrow \hat{\epsilon}_2 = 0.5$$

$$\hat{y}_3 = -11 + 2.1 \cdot 5 = -0.5 \quad \Rightarrow \hat{\epsilon}_3 = 0.5$$

Zur Kontrolle:  $\hat{\epsilon}_1 + \hat{\epsilon}_2 + \hat{\epsilon}_3 = 0$

Prognose:  $x^* = 12 \quad \Rightarrow \hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 \cdot x^* = -11 + 2.1 \cdot 12 = 14.2$



## Bsp. 2.31. [Arbeitszeit und Einkommen]

Multiplres Regressionsmodell:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

mit

$$X_1 = \begin{cases} 1 & \text{männlich} \\ 0 & \text{weiblich} \end{cases}$$

$$X_2 = \text{(vertragliche) Arbeitszeit}$$

$$Y = \text{Einkommen}$$

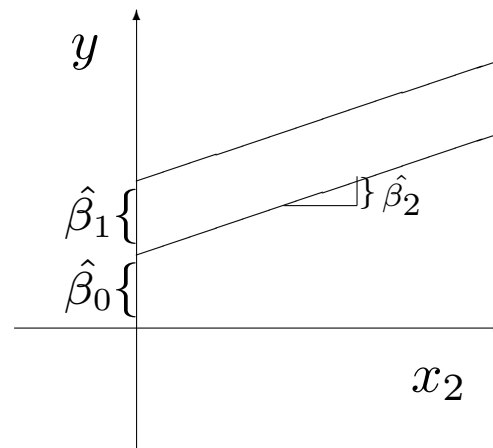
Interpretation:

Die geschätzte Gerade für die Männer lautet

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot x_{2i}$$

für die Frauen hingegen erhält man

$$\begin{aligned}\hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot x_{2i} \\ &= \hat{\beta}_0 + \hat{\beta}_2 \cdot x_{2i}\end{aligned}$$



$\beta_0$  Grundlevel

$\beta_2$  durchschnittlicher Stundenlohn

$\beta_1$  Zusatzeffekt des Geschlechts zum Grundlevel.

Die 0-1 Variable dient als Schalter, mit dem man den Männereffekt an/abschaltet.

**Bsp. 2.32. [Dummykodierung]**

Nominales Merkmal mit  $q$  Kategorien, z.B.  $X$  = Parteipräferenz mit

$$X = \begin{cases} 1 & \text{CDU/CSU oder FDP} \\ 2 & \text{SPD oder Grüne} \\ 3 & \text{Sonstige} \end{cases}$$

Man darf  $X$  nicht einfach mit Werten 1 bis 3 besetzen, da es sich um ein nominales Merkmal handelt.

Idee: Mache aus der einen Variable mit  $q$  (hier 3) Ausprägungen  $q - 1$  (hier 2) Variablen mit den Ausprägungen ja/nein ( $\hat{=}$ 0/1). Diese *Dummyvariablen* dürfen dann in der Regression verwendet werden.

$$X_1 = \begin{cases} 1 & \text{CDU/CSU oder FDP} \\ 0 & \text{andere} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{SPD, Grüne} \\ 0 & \text{andere} \end{cases}$$

Durch die Ausprägungen von  $X_1$  und  $X_2$  sind alle möglichen Ausprägungen von  $X$  vollständig beschrieben:

$X$	Text	$X_1$	$X_2$
1	CDU/CSU, FDP	1	0
2	SPD, Grüne	0	1
3	Sonstige	0	0

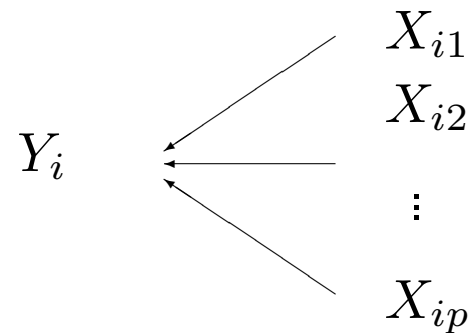
## Beispiel zur Interpretation:

- $Y$ : Score auf Autoritarismusskala
- $X$  bzw.  $X_1, X_2$ : Parteienpräferenz
- $X_3$ : Einkommen

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

- $\beta_0$ : Grundniveau
- $\beta_1$ : ceteris paribus Effekt (Erhöhung des Grundniveaus) von CDU/CSU und FDP
- $\beta_2$ : ceteris paribus Effekt (Erhöhung des Grundniveaus) von SPD und Grünen
- $\beta_3$ : ceteris paribus Effekt des Einkommens

## Multiples Regressionsmodell:



abhängige Variable

metrisch/quasistetig

unabhängige Variablen

metrische/quasistetige oder dichotome (0/1) Variablen (kategoriale Variablen mit mehr Kategorien → Dummy-Kodierung)

## Ansatz:

- linearer Zusammenhang.
- Ermittle aus den Daten „Wirkungsstärke“ der einzelnen Variablen.
- Im Folgenden: Probabilistische Modelle in Analogie zu den deskriptiven Modellen aus Statistik I (damit Verallgemeinerung auf die Grundgesamtheit möglich).



## 2.5.2 Lineare Einfachregression

Zunächst Modelle mit nur *einer* unabhängigen Variable.

Statistische Sichtweise:

- Wahres Modell

$$y_i = \beta_0 + \beta_1 x_i$$

$\beta_0$  Grundniveau

$\beta_1$  „Elastizität“: Wirkung der Änderung von  $X_i$  um eine Einheit

- gestört durch zufällige Fehler  $\epsilon_i$  Man beobachtet Datenpaare,  $(X_i, Y_i)$ ,  $i = 1, \dots, n$  mit

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

wobei

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$\sigma^2$  für alle  $i$  gleich

$\epsilon_{i_1}, \epsilon_{i_2}$  stochastisch unabhängig für  $i_1 \neq i_2$

Nach den Modellannahmen gilt für die bedingte Verteilung von  $Y_i$  gegeben  $X_i = x_i$ :

$$Y_i | X_i = x_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2), \quad i = 1, \dots, n.$$

Interpretation: verschiedene Normalverteilungen jeweils mit verschobenem Mittelwert  $\mu_i = \beta_0 + \beta_1 x_i$ , aber gleicher Varianz.

Aufgabe: Schätze die Parameter  $\beta_0, \beta_1$  und  $\sigma^2$ . Die Schätzwerte und Schätzfunktionen werden üblicherweise mit  $\hat{\beta}_0, \hat{\beta}_1$  und  $\hat{\sigma}^2$  bezeichnet.

In der eben beschriebenen Situation gilt:

1. Die Maximum Likelihood Schätzer lauten:

$$\hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2},$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X},$$

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{\varepsilon}_i^2$$

mit den geschätzten Residuen

$$\hat{\varepsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i.$$

2. Mit

$$\hat{\sigma}_{\hat{\beta}_0} := \frac{\hat{\sigma} \sqrt{\sum_{i=1}^n X_i^2}}{\sqrt{n \sum_{i=1}^n (X_i - \bar{X})^2}}$$

gilt

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \sim t(n - 2)$$

und analog mit

$$\hat{\sigma}_{\hat{\beta}_1} := \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}}$$

gilt

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim t(n - 2).$$

**Bem. 2.33.**

- $\hat{\beta}_0$  und  $\hat{\beta}_1$  sind, wenn man die Realisationen  $(x_i, y_i)$  von  $(X_i, Y_i)$  einsetzt, die  $KQ$ -Schätzer aus Statistik I. Unter Normalverteilung fällt hier das  $ML$ - mit dem  $KQ$ -Prinzip zusammen.
- Man kann unmittelbar Tests und Konfidenzintervalle ermitteln (völlig analog zum Vorgehen in Kapitel 2.3 und 2.4).

Konfidenzintervalle zum Sicherheitsgrad  $\gamma$ :

$$\text{für } \beta_0 : \quad [\hat{\beta}_0 \pm \hat{\sigma}_{\hat{\beta}_0} \cdot t_{1+\frac{\gamma}{2}}(n-2)]$$

$$\text{für } \beta_1 : \quad [\hat{\beta}_1 \pm \hat{\sigma}_{\hat{\beta}_1} \cdot t_{1+\frac{\gamma}{2}}(n-2)]$$

Mit der Teststatistik

$$T_{\beta_1^*} = \frac{\hat{\beta}_1 - \beta_1^*}{\hat{\sigma}_{\hat{\beta}_1}}$$

ergibt sich

		Hypothesen		kritische Region
I.	$H_0 : \beta_1 \leq \beta_1^*$	gegen	$\beta_1 > \beta_1^*$	$T_{\beta_1^*} \geq t_{1-\alpha}(n-2)$
II.	$H_0 : \beta_1 \geq \beta_1^*$	gegen	$\beta_1 < \beta_1^*$	$T_{\beta_1^*} \leq t_{1-\alpha}(n-2)$
III.	$H_0 : \beta_1 = \beta_1^*$	gegen	$\beta_1 \neq \beta_1^*$	$ T_{\beta_1^*}  \geq t_{1-\frac{\alpha}{2}}(n-2)$

(analog für  $\hat{\beta}_0 \rightarrow T_{\beta_0^*}$ ).

Von besonderem Interesse ist der Fall  $\beta_1^* = 0$ :

- Typischer SPSS-Output

Koeffizienten<sup>a</sup>

			Standardisierte Koeffizienten		
	$\beta$	Standardfehler	Beta	$T$	Signifikanz
Konstante	$\hat{\beta}_0$	$\hat{\sigma}_{\hat{\beta}_0}$	5)	1)	3)
Unabhängige Variable	$\hat{\beta}_1$	$\hat{\sigma}_{\hat{\beta}_1}$	6)	2)	4)

<sup>a</sup> abhängige Variable

1) Wert der Teststatistik

$$T_{\beta_0^*} = \frac{\hat{\beta}_0}{\hat{\sigma}_{\hat{\beta}_0}}.$$

zum Testen von  $H_0: \beta_0 = 0$  gegen  $H_1: \beta_0 \neq 0$ .



2) Analog: Wert von

$$T_{\beta_1^*} = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}$$

zum Testen von  $H_0: \beta_1 = 0$  gegen  $H_1: \beta_1 \neq 0$ .

3) p-Wert zu 1)

4) p-Wert zu 2)

5), 6) hier nicht von Interesse.

- Die Testentscheidung „ $\hat{\beta}_1$  signifikant von 0 verschieden“ entspricht dem statistischen Nachweis eines Einflusses von  $X$ .
- Man kann analog zu Kap. 2.4.7 auch einseitige Hypothesen testen

## 2.5.3 Multiple lineare Regression

- Analoger Modellierungsansatz, aber mit mehreren erklärenden Variablen:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} + \epsilon_i$$

- Schätzung von  $\beta_0, \beta_1, \dots, \beta_p$  und  $\sigma^2$  sinnvollerweise über Matrixrechnung bzw. Software.

Aus dem SPSS-Output sind  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$  sowie  $\hat{\sigma}_{\hat{\beta}_0}, \hat{\sigma}_{\hat{\beta}_1}, \dots, \hat{\sigma}_{\hat{\beta}_p}$  ablesbar.

(Outputs lesen können ist absolut klausurrelevant! Matrixrechnung wird nicht verlangt.)

- Es gilt für jedes  $j = 0, \dots, p$

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t(n - p - 1)$$

und man erhält wieder Konfidenzintervalle für  $\beta_j$ :

$$[\hat{\beta}_j \pm \hat{\sigma}_{\hat{\beta}_j} \cdot t_{1+\frac{\gamma}{2}}(n - p - 1)]$$

sowie entsprechende Tests.

Von besonderem Interesse ist wieder der Test

$$H_0 : \beta_j = 0, \quad H_1 : \beta_j \neq 0.$$

Der zugehörige p-Wert findet sich im SPSS-Ausdruck (Vorsicht mit Problematik des multiplen Testens!).

- Man kann auch simultan testen, z.B.

$$\beta_1 = \beta_2 = \dots = \beta_p = 0.$$

Dies führt zu einem sogenannten F-Test (→ Software).

## 2.5.4 Varianzanalyse (Analysis of Variance, ANOVA)

- Sind alle  $X_{ij}$  0/1-wertig, so erhält man die sogenannte *Varianzanalyse*, was dem Vergleich von mehreren Mittelwerten entspricht.

\* Für Befragte mit  $X_{ij} = 0$  für alle  $j$  gilt:

$$E(Y) = \beta_0$$

\* Ist  $X_{i1} = 1$  und  $X_{ij} = 0$  für  $j \geq 2$ , so gilt

$$E(Y) = \beta_0 + \beta_1$$

\* Ist  $X_{i1} = 1$  und  $X_{i2} = 1$ , sowie  $X_{ij} = 0$  für  $j \geq 3$ , so gilt

$$E(Y) = \beta_0 + \beta_1 + \beta_2$$

\* etc.

- Vor allem in der angewandten Literatur, etwa in der Psychologie, wird die Varianzanalyse unabhängig vom Regressionsmodell entwickelt. Diese Sichtweise soll auch hier jetzt kurz besprochen werden.
- Ziel: Mittelwertvergleiche in mehreren Gruppen, häufig in (quasi-) experimentellen Situationen.
- Verallgemeinerung des t-Tests. Dort nur zwei Gruppen.
- Hier nur *einfaktorische Varianzanalyse* (*Eine* Gruppierungsvariable).

**Bsp. 2.34.**

Einstellung zu Atomkraft anhand eines Scores, nachdem ein Film gezeigt wurde.

3 Gruppen („Faktorstufen“):

- Pro-Atomkraft-Film
- Contra-Atomkraft-Film
- ausgewogener Film

Varianzanalyse: Vergleich der Variabilität in und zwischen den Gruppen

Beobachtungen:  $Y_{ij}$

$j = 1, \dots, J$     Faktorstufen

$i = 1, \dots, n_j$     Personenindex in der  $j$ -ten Faktorstufe



## Zwei äquivalente Modellformulierungen:

a) Modell in Mittelwertsdarstellung:

$$Y_{ij} = \mu_j + \epsilon_{ij} \quad j = 1, \dots, J, i = 1, \dots, n_j,$$

mit

$\mu_j$  factorspezifischer Mittelwert

$\epsilon_{ij}$  zufällige Störgröße

$\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ ,  $\epsilon_{11}, \epsilon_{12}, \dots, \epsilon_{Jn_J}$  unabhängig.

Testproblem:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_J$$

gegen

$$H_1 : \mu_\ell \neq \mu_q \quad \text{für mindestens ein Paar } (\ell, q)$$

b) Modell in Effektdarstellung:

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$$

wobei  $\alpha_j$  so, dass

$$\sum_{j=1}^J n_j \alpha_j = 0.$$

$\mu$  globaler Erwartungswert

$\alpha_j$  Effekt in der  $j$ -ten Faktorstufe, factorspezifische systematische Abweichung vom gemeinsamen Mittelwert  $\mu$

Testproblem:

$$H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_J = 0$$

gegen

$$H_1 : \alpha_j \neq 0 \text{ für mindestens ein } j$$

Die beiden Modelle sind äquivalent: setze  $\mu_j := \mu + \alpha_j$ .

## Streuungszerlegung

Mittelwerte:

$\bar{Y}_{\bullet\bullet}$  Gesamtmittelwert in der Stichprobe

$\bar{Y}_{\bullet j}$  Mittelwert in der  $j$ -ten Faktorstufe

Es gilt (vgl. Statistik I) die Streuungszerlegung:

$$\sum_{j=1}^J \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{\bullet\bullet})^2 = \sum_{j=1}^J \underbrace{n_j (\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet})^2}_{\text{Zwischenstufenstreuung}} + \sum_{j=1}^J \underbrace{\sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{\bullet j})^2}_{\text{Innerstufenstreuung}}$$

Die Testgröße

$$F = \frac{SQE/(J - 1)}{SQR/(n - J)}$$

ist geeignet zum Testen der Hypothesen

$$H_0 : \mu_1 = \mu_2 = \dots \mu_J$$

gegen

$$H_1 : \mu_\ell \neq \mu_q \text{ für mindestens ein Paar } (\ell, q)$$

beziehungsweise

$$H_0 : \alpha_1 = \alpha_2 = \dots \alpha_J = 0$$

gegen

$$H_1 : \alpha_j \neq 0 \text{ für mindestens ein } j$$

Sie besitzt eine sog. *F-Verteilung* mit  $(J - 1)$  und  $(n - J)$  Freiheitsgraden.

Die kritische Region besteht aus den *großen* Werten von  $F$  (Vorsicht: obwohl  $H_0$  von „Gleichheitsform“).

Also  $H_0$  ablehnen falls

$$T > F_{1-\alpha}(J - 1, n - J),$$

mit dem entsprechenden  $(1 - \alpha)$ -Quantil der  $F$ -Verteilung mit  $(J - 1)$  und  $(n - J)$  Freiheitsgraden.

(Je größer die Variabilität zwischen den Gruppen im Vergleich zu der Variabilität in den Gruppen, desto unplausibler ist die Nullhypothese, dass alle Gruppenmittelwerte gleich sind.)

Bei Ablehnung des globalen Tests ist dann oft von Interesse, welche Gruppen sich unterscheiden.

⇒ Testen spezifischer Hypothesen über die Effekte  $\alpha_j$  bzw. die Mittelwerte  $\mu_j$ . Dabei tritt allerdings wieder Problematik des multiplen Testens auf.