Lineare Regression

JiTT Ziele

- Ihr könnt Aussagen als Hypothesen formulieren und sie testen.
- Ihr könnt bei der linearen Regression bestimmen, ob die Steigung und Achsenabschnitt signifikant sind.
- Ihr kennt Fehler 1. und 2. Art.

Hypothesentests

Sinn: Man will Aussagen prüfen.

Beispiele für Aussagen:

- Es gibt mehr Geburten von Knaben als von Mädchen.
- Männer sind grösser als Frauen.
- Babies erkennen Farben.
- Dieser Fonds verspricht mindestens 5% Rendite.
- Der Koeffizient a in ax+b ist signifikant.

Vorgehen bei statistischen Tests

Das Prinzip eines statistischen Tests ist:

- Man stellt eine Nullhypothese H_0 auf und die Alternativhypothese H_1 und gibt das Signifikanzniveau α vor. α hängt vom Problem ab und ist häufig 0.01, 0.05 oder 0.1.
- Man berechnet die Wahrscheinlichkeit des Ereignisses unter der Voraussetzung von H_0.
- Ist die berechnete Wahrscheinlichkeit kleiner als das Signifikanzniveau α, dann wird H_0 abgelehnt, sonst wird H_0 angekommen.

Vorgehen bei statistischen Tests

dies will man widerlegen!

Nullhypothese H_0: Wird als wahr angenommen.

"Babies sehen keine Farben."

"Der Angeklagte ist unschuldig."

„Die Maschine produziert 2% schlechte Schrauben."

Alternativhypothese H1: Das Gegenteil von H_0

"Babies sehen Farben."

"Der Angeklagte ist schuldig."

„Die Maschine arbeitet momentan schlechter."

dies will man beweisen!

Vorgehen bei statistischen Tests

Achtung: H_0 kann man nie beweisen, nur widerlegen!
Also aufpassen, wie man die Hypothesen formuliert.

100%ig kann man keine Hypothese widerlegen, da man von Stichproben auf die Grundgesamtheit schliesst.

H_0: Always contains a '$=$', '$\leq$' or '$\geq$' sign

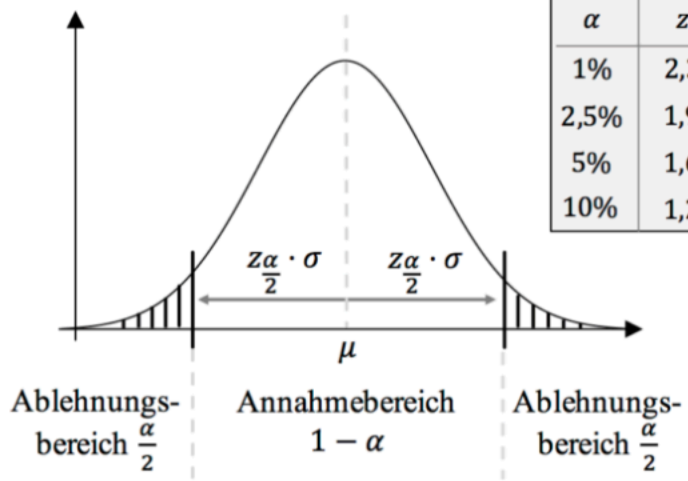
In der H_1 Hypothese steht niemals ein =, ≤ oder ≥.

Einseitig oder Zweiseitig?

Einseitiger Signifikanztest:	$H_0 : p \leq p_0$ gegen $H_1 : p > p_0$ $H_0 : p \geq p_0$ gegen $H_1 : p < p_0$
Zweiseitiger Signifikanztest:	$H_0 : p = p_0$ gegen $H_1 : p \neq p_0$

Einseitig oder Zweiseitig?

Beidseitiger Test



α	z_α
1%	2,33
2,5%	1,96
5%	1,64
10%	1,28

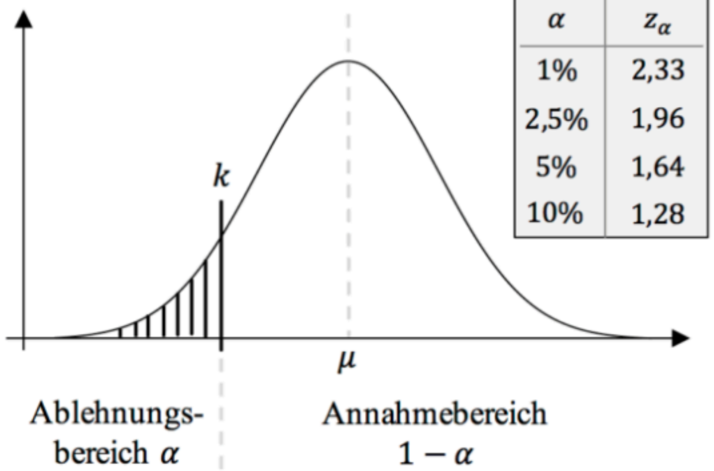
Wenn $\alpha = 5\%$ ist, verwenden wir nicht $z_{0,05} = 1{,}64$ sondern den Wert für $\alpha = 2{,}5\%$, also $z_{0,025} = 1{,}96$.

Warum? Weil sich der Ablehnungsbereich i.H.v. 5% links und rechts aufteilt! In Summe lehnen wir natürlich 2,5%+2,5%=5% ab.

Der Annahmebereich würde lauten:

$$A = [\mu - 1{,}96\sigma; \mu + 1{,}96\sigma]$$

Linksseitiger Test



α	z_α
1%	2,33
2,5%	1,96
5%	1,64
10%	1,28

Wenn $\alpha = 5\%$ ist, verwenden wir $z_{0,05} = 1{,}64$.

Warum? Weil sich der Ablehnungsbereich i.H.v. 5% nur auf den linken Bereich bezieht und nicht aufgeteilt werden muss.

Der Ablehnungsbereich würde lauten:

$$\bar{A} = [0; \mu - 1{,}64\sigma]$$

Abbildung: Quelle: `https://www.studyhelp.de/online-lernen/mathe/hypothesentests/`

Hypothesen - Übung

Was ist H_0 und H_1?

„Es gibt mehr Geburten von Knaben als von Mädchen.“

H_0: Knabengeburten ≤ Mädchengeburten
H_1: Knabengeburten > Mädchengeburten

Hypothesen - Übung

Was ist H_0 und H_1?

„Männer sind grösser als Frauen.“

H_0: Männergrösse ≤ Frauengrösse
H_1: Männergrösse > Frauengrösse

Hypothesen - Übung

Was ist H_0 und H_1?

„5% der Einwohner sind Handwerker.“

H_0: Handwerker = 5%
H_1: Handwerker ≠ 5 %

Hypothesen - Übung

Was ist H_0 und H_1?

„In Zürich ist es im November wärmer als in Bern.“

H_0: T in Zürich ≤ T in Bern
H_1: T in Zürich > T in Bern

Hypothesen - Fehlermöglichkeiten

Fehler 1. Art (Irrtumswahrscheinlichkeit α)
Obwohl H_0 richtig ist, wird H_0 auf Grund der Stichprobenfunktion abgelehnt.

Fehler 2. Art (Irrtumswahrscheinlichkeit β)
Obwohl H_0 falsch ist, wird H_0 auf Grund der Stichprobenfunktion angenommen.

Hypothesen - Fehlermöglichkeiten

Man kann die Optionen von Hypothesentests also folgendermassen betrachten:

	H_0 wird nicht abgelehnt	H_0 wird abgelehnt (=Entscheid für H_1)
H_0 richtig	richtige Entscheidung	Fehler 1. Art (α-Fehler)
H_1 richtig	Fehler 2. Art (β-Fehler)	richtige Entscheidung

Fehlentscheidungen sind in statistischen Tests immer möglich. Zudem kann man nach einem Test nicht beurteilen, ob die Entscheidung richtig oder falsch ist. Man wählt deswegen das Signifikanzniveau α genügend klein, sodass die Wahrscheinlichkeit für Fehler der 1. Art klein wird.

Fehlermöglichkeiten - Beispiel

Was ist H_0 und H_1?

„In Zürich ist es im November wärmer als in Bern.“

H_0: T in Zürich ≤ T in Bern
H_1: T in Zürich > T in Bern

Was ist der Fehler der 1. Art?

Fehlermöglichkeiten - Beispiel

Was ist H_0 und H_1?

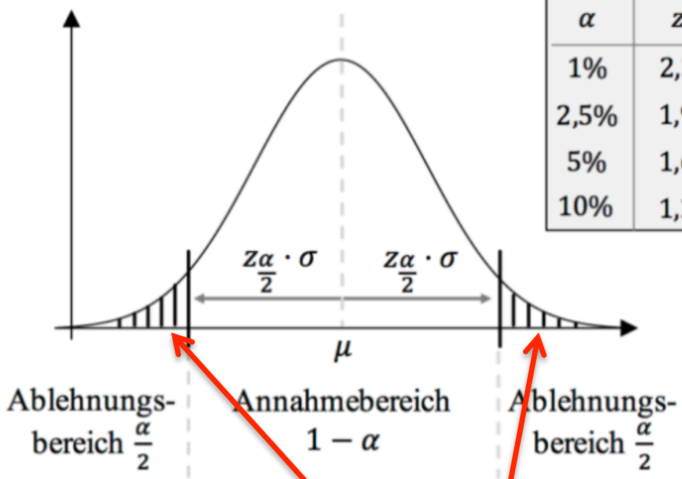
„In Zürich ist es im November wärmer als in Bern.“

H_0: T in Zürich ≤ T in Bern
H_1: T in Zürich > T in Bern

Was ist der Fehler der 1. Art?

H0 richtig, aber wird abgelehnt. D.h. T in Zürich ist in Realität kleiner oder identisch zu Bern, aber man denkt, dass T in Zürich höher sei.

Hypothesentests - Mathematik



α	z_α
1%	2,33
2,5%	1,96
5%	1,64
10%	1,28

Sehr ähnlich zu Konfidenzintervallen!

Nullhypothese H_0	Alternativhypothese H_1	Ablehnungsbereich von H_0
$\mu = \mu_0$	$\mu \neq \mu_0$	$\frac{\lvert\bar{x}-\mu_0\rvert}{\sigma_0}\sqrt{n} > z_{1-\alpha/2}$
$\mu \geq \mu_0$	$\mu < \mu_0$	$\frac{\lvert\bar{x}-\mu_0\rvert}{\sigma_0}\sqrt{n} < -z_{1-\alpha}$
$\mu \leq \mu_0$	$\mu > \mu_0$	$\frac{\lvert\bar{x}-\mu_0\rvert}{\sigma_0}\sqrt{n} > z_{1-\alpha}$

bei kleinen Stichproben t-Wert
und nicht z-Wert nehmen.

Hypothesentests - Beispiel

Eine Fabrik stellt Flaschen her mit einem Volumen von 5 Litern und Standard-abweichung 0.1 L. Seit kurzem wird neues Glas verwendet und der Manager vermutet, dass dies den Inhalt erhöht hat, wobei die Standardabweichung gleich geblieben ist.

Deswegen macht man einen Test:
Stichprobe: 16 Flaschen, mittleres Volumen 5.038 L.
Gewünschte Signifikanz: 5%

Hypothese?

Hypothesentests - Beispiel

Eine Fabrik stellt Flaschen her mit einem Volumen von 5 Litern und Standardabweichung 0.1 L. Seit kurzem wird neues Glas verwendet und der Manager vermutet, dass dies den Inhalt erhöht hat, wobei die Standardabweichung gleich geblieben ist.

Deswegen macht man einen Test:
Stichprobe: 16 Flaschen, mittleres Volumen 5.038 L.
Gewünschte Signifikanz: 5%

Hypothese?
H_0: $\mu \leq 5$ L (ab und zu wird auch $\mu = 5$ geschrieben, math. korrekt ist aber $\leq$)
H_1: $\mu > 5$ L

Hypothesentests - Beispiel

Eine Fabrik stellt Flaschen her mit einem Volumen von 5 Litern und Standard-abweichung 0.1 L. Seit kurzem wird neues Glas verwendet und der Manager vermutet, dass dies den Inhalt erhöht hat, wobei die Standardabweichung gleich geblieben ist.

Deswegen macht man einen Test:
Stichprobe: 16 Flaschen, mittleres Volumen 5.038 L.
Gewünschte Signifikanz: 5%

Hypothese?
H_0: $\mu \leq 5$ L
H_1: $\mu > 5$ L

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{5.038 - 5}{0.1/\sqrt{16}} = 1.52$$

Hypothesentests – z-Tabelle

$\Phi_{0;1}(z) \rightarrow$

Gewünschte Signifikanz: 5%

z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,50000	0,50399	0,50798	0,51197	0,51595	0,51994	0,52392	0,52790	0,53188	0,53586
0,1	0,53983	0,54380	0,54776	0,55172	0,55567	0,55962	0,56356	0,56749	0,57142	0,57535
0,2	0,57926	0,58317	0,58706	0,59095	0,59483	0,59871	0,60257	0,60642	0,61026	0,61409
0,3	0,61791	0,62172	0,62552	0,62930	0,63307	0,63683	0,64058	0,64431	0,64803	0,65173
0,4	0,65542	0,65910	0,66276	0,66640	0,67003	0,67364	0,67724	0,68082	0,68439	0,68793
0,5	0,69146	0,69497	0,69847	0,70194	0,70540	0,70884	0,71226	0,71566	0,71904	0,72240
0,6	0,72575	0,72907	0,73237	0,73565	0,73891	0,74215	0,74537	0,74857	0,75175	0,75490
0,7	0,75804	0,76115	0,76424	0,76730	0,77035	0,77337	0,77637	0,77935	0,78230	0,78524
0,8	0,78814	0,79103	0,79389	0,79673	0,79955	0,80234	0,80511	0,80785	0,81057	0,81327
0,9	0,81594	0,81859	0,82121	0,82381	0,82639	0,82894	0,83147	0,83398	0,83646	0,83891
1,0	0,84134	0,84375	0,84614	0,84849	0,85083	0,85314	0,85543	0,85769	0,85993	0,86214
1,1	0,86433	0,86650	0,86864	0,87076	0,87286	0,87493	0,87698	0,87900	0,88100	0,88298
1,2	0,88493	0,88686	0,88877	0,89065	0,89251	0,89435	0,89617	0,89796	0,89973	0,90147
1,3	0,90320	0,90490	0,90658	0,90824	0,90988	0,91149	0,91309	0,91466	0,91621	0,91774
1,4	0,91924	0,92073	0,92220	0,92364	0,92507	0,92647	0,92785	0,92922	0,93056	0,93189
1,5	0,93319	0,93448	0,93574	0,93699	0,93822	0,93943	0,94062	0,94179	0,94295	0,94408
1,6	0,94520	0,94630	0,94738	0,94845	0,94950	0,95053	0,95154	0,95254	0,95352	0,95449
1,7	0,95543	0,95637	0,95728	0,95818	0,95907	0,95994	0,96080	0,96164	0,96246	0,96327

z=1.64 (oder 1.65)

Hypothesentests – t-Tabelle

Gewünschte Signifikanz: 5%
16 Flaschen

Anzahl Freiheitsgrade *n*	***P* für zweiseitigen Vertrauensbereich**							
	0,5	**0,75**	**0,8**	**0,9**	**0,95**	**0,98**	**0,99**	**0,998**
	***P* für einseitigen Vertrauensbereich**							
	0,75	**0,875**	**0,90**	**0,95**	**0,975**	**0,99**	**0,995**	**0,999**
1	1,000	2,414	3,078	6,314	12,706	31,821	63,657	318,309
2	0,816	1,604	1,886	2,920	4,303	6,965	9,925	22,327
3	0,765	1,423	1,638	2,353	3,182	4,541	5,841	10,215
4	0,741	1,344	1,533	2,132	2,776	3,747	4,604	7,173
5	0,727	1,301	1,476	2,015	2,571	3,365	4,032	5,893
6	0,718	1,273	1,440	1,943	2,447	3,143	3,707	5,208
7	0,711	1,254	1,415	1,895	2,365	2,998	3,499	4,785
8	0,706	1,240	1,397	1,860	2,306	2,896	3,355	4,501
9	0,703	1,230	1,383	1,833	2,262	2,821	3,250	4,297
10	0,700	1,221	1,372	1,812	2,228	2,764	3,169	4,144
11	0,697	1,214	1,363	1,796	2,201	2,718	3,106	4,025
12	0,695	1,209	1,356	1,782	2,179	2,681	3,055	3,930
13	0,694	1,204	1,350	1,771	2,160	2,650	3,012	3,852
14	0,692	1,200	1,345	1,761	2,145	2,624	2,977	3,787
15	0,691	1,197	1,341	$\boxed{1,753}$	2,131	2,602	2,947	3,733
16	0,690	1,194	1,337	1,746	2,120	2,583	2,921	3,686
17	0,689	1,191	1,333	1,740	2,110	2,567	2,898	3,646
18	0,688	1,189	1,330	1,734	2,101	2,552	2,878	3,610

Hypothesentests - Beispiel

Da die Standardabweichung bekannt ist (=0.1), verwendet man hier eher den z-Test, trotz kleiner Stichprobe.

Das Resultat wäre hier bei beiden Tests identisch.

Hypothesentests - Beispiel

z-Wert (vorher berechnet): 1.52
Tabellenwert (abgelesen): 1.645

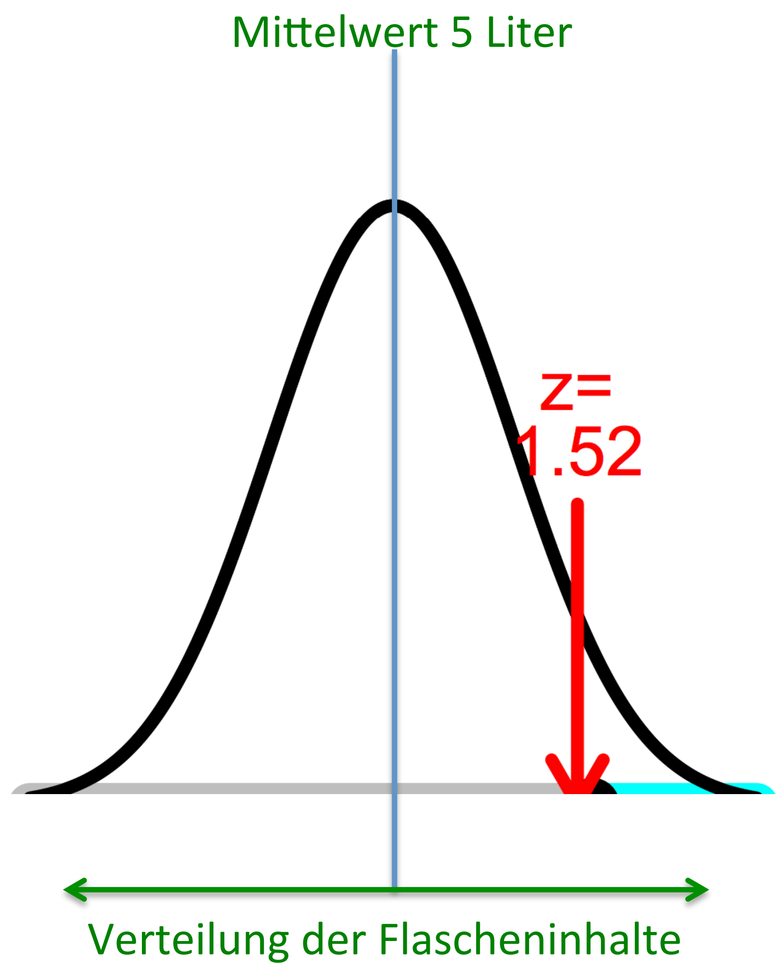
Weil 1.52 < 1.645 kann man H_0 bei 5% Signifikanz nicht verwerfen.

(dieselbe Argumentation gilt, wenn man stattdessen den T-Wert verwendet: 1.52 < 1.75)

D.h. wir können nicht sagen, ob $\mu \leq 5$ L oder $\mu > 5$ L.

Hypothesentests - Beispiel

Grafische Darstellung:

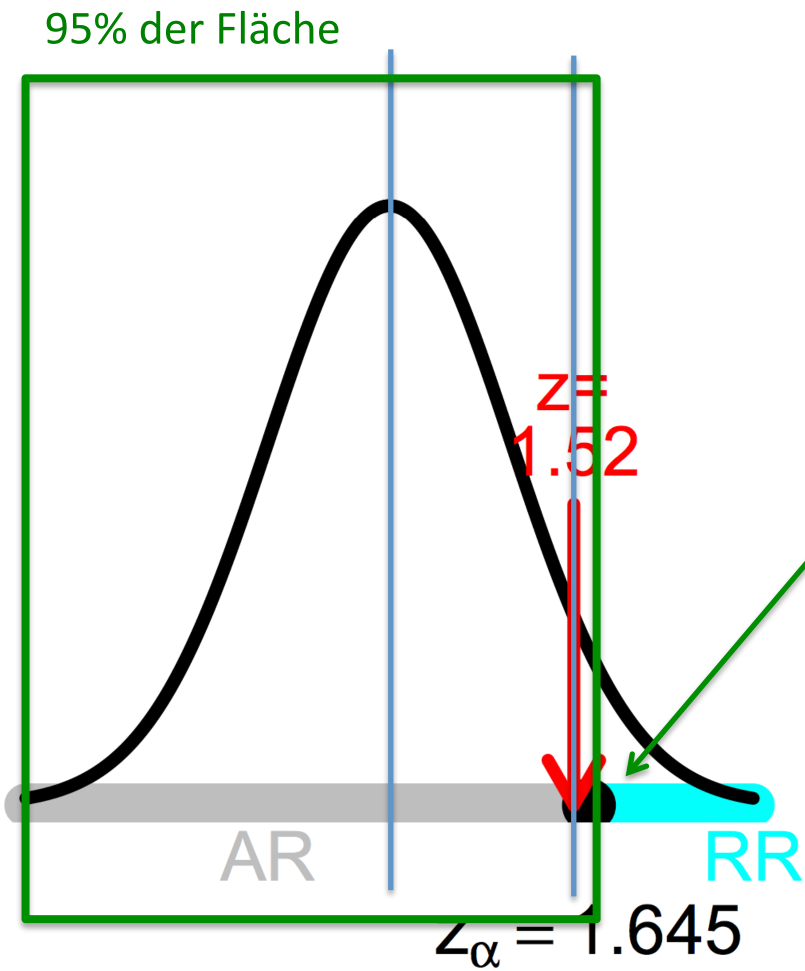


Hypothesentests - Beispiel

Wir hatten z = 1.645 in der Tabelle

Das entspricht der Grenze von 95% der Fläche

Unser berechnetes z (1.52) ist links von dieser Grenze!



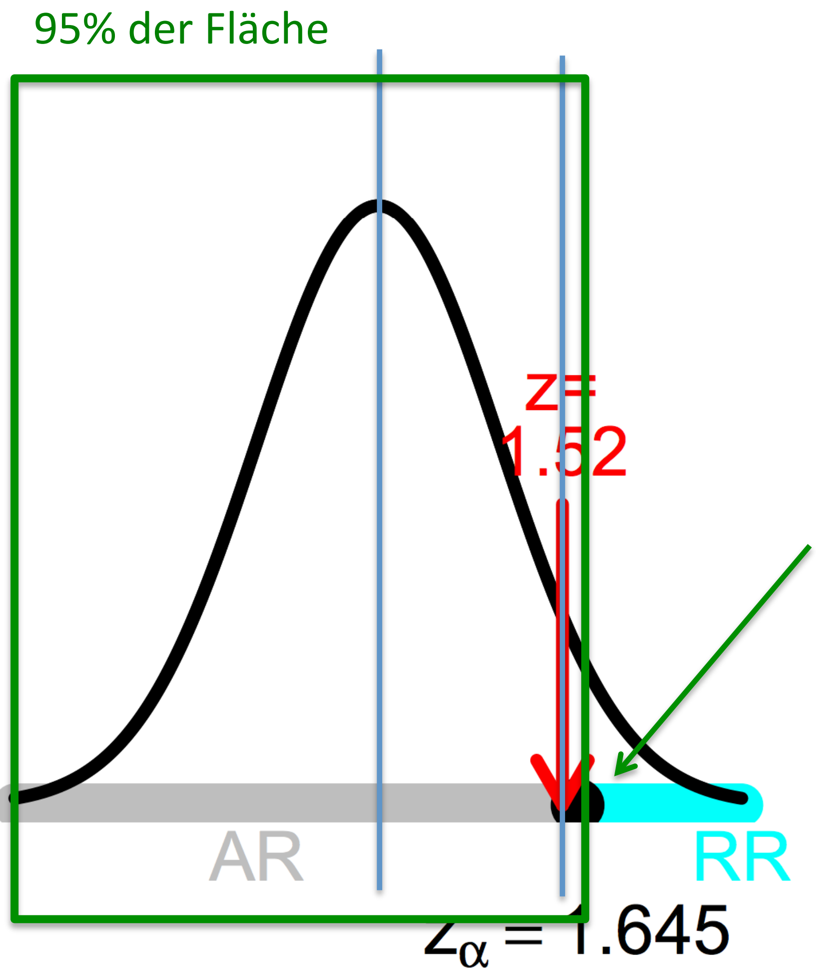
Hypothesentests - Beispiel

Wir hatten z = 1.645 in der Tabelle

Das entspricht der Grenze von 95% der Fläche

Unser berechnetes z (1.52) ist links von dieser Grenze!

Nur wenn wir (in der Grafik) rechts von z_α landen würden, dann wäre man im Ablehnbereich (RR = rejection region)



Hypothesentests - Beispiel

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{5.038 - 5}{0.1/\sqrt{16}} = 1.52$$

p-Wert: In der Z-Tabelle lesen wir bei 1.52 den Wert 0.93574 ab. Wir rechnen 1-0.93574 = 0.0643

Da $p>0.05$ kann man H_0 nicht verwerfen. Mit dem p-Wert sieht man sofort, bei welcher Signifikanz man verwerfen könnte!

Bei 7% Signifikanzniveau könnte man H_0 verwerfen. Warum wählt man das Signifikanzniveau generell nicht höher?

Signifikanzniveau

Warum wählt man das Signifikanzniveau generell nicht höher?

Bsp:
$\alpha = 0.25$
Falls H_0 wahr ist, wird es aber jedes 4. Mal abgelehnt (Fehler 1. Art).
Dies ist kein zuverlässiger Test!

Signifikanzniveau

Warum wählt man das Signifikanzniveau generell nicht höher?

Bsp:
$\alpha = 0.25$
Falls H_0 wahr ist, wird es aber jedes 4. Mal abgelehnt (Fehler 1. Art).
Dies ist kein zuverlässiger Test!

$\alpha = 0.05$
Falls H_0 wahr ist, wird es trotzdem jedes 20. Mal abgelehnt (Fehler 1. Art). Dies ist oft akzeptable Genauigkeit.

Signifikanzniveau

Warum wählt man das Signifikanzniveau generell nicht höher?

Bsp:
$\alpha = 0.25$
Falls H_0 wahr ist, wird es aber jedes 4. Mal abgelehnt (Fehler 1. Art).
Dies ist kein zuverlässiger Test!

$\alpha = 0.05$
Falls H_0 wahr ist, wird es trotzdem jedes 20. Mal abgelehnt (Fehler 1. Art). Dies ist oft akzeptable Genauigkeit.

$\alpha = 0.0005$
Vorteil: H_0 wahr, wird aber nur jedes 2000. Mal abgelehnt.
Problem: Falls H_0 nicht wahr ist, kann man die Nullhypothese oft nicht widerlegen. Fehler 2. Art wird gross.

Anwendung auf lineare Regression

Mittels Hypothesentests kann man verschiedene Annahmen prüfen. Beispiel: Für eine Vorlesung lernen 10 Studenten jeweils 1, 2, ..., 10 Stunden und alle erhalten am Ende die Note 5. In diesem Fall ist es offensichtlich, dass die Lerndauer keinen Einfluss auf die Note hat und ein linearer Fit würde y=5 ergeben und der Parameter β würde automatisch 0 ergeben. Aber was wenn ein Student nach 10 h Lernen eine 5.5 schreibt, ist ein lineares Modell dann sinnvoll?

Dazu stellt man die Hypothese auf

$$H_0 : \beta = 0$$

(d.h. eine Konstante erklärt die Werte besser) und die Alternativhypothese

$$H_1 : \beta \neq 0$$

Anwendung auf lineare Regression

Definiere die Daten

```
```{r}
note = c(5,5,5,5,5,5,5,5,5,5.5)
stunden = c(1,2,3,4,5,6,7,8,9,10)
```
```

Lineares Modell:

```
```{r}
model = lm(note ~ stunden)
summary(model)
#was wir fitten:    note = a * stunden + intercept
```
```

Anwendung auf lineare Regression

Definiere die Daten

```{r}
note = c(5,5,5,5,5,5,5,5,5,5.5)
stunden = c(1,2,3,4,5,6,7,8,9,10)
```

Lineares Modell:

```{r}
model = lm(note ~ stunden)
summary(model)
#was wir fitten:    note = a * stunden + intercept
#was wir erhalten: note = 0.02727 * stunden + 4.9
```

```
Call:
lm(formula = note ~ stunden)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.14545 -0.08409 -0.02273  0.03864  0.32727 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.90000    0.09770  50.153 2.77e-11 ***
stunden      0.02727    0.01575   1.732    0.122    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.143 on 8 degrees of freedom
Multiple R-squared:  0.2727,	Adjusted R-squared:  0.1818 
F-statistic:     3 on 1 and 8 DF,  p-value: 0.1215
```

Anwendung auf lineare Regression

Bei einem gegebenen Signifikanzlevel von 5% könnte man H0: a=0 nicht ablehnen, da 5% < 12.2%.

D.h. es ist möglich, dass ein Modell der Art Note = Konstante die Daten erklärt.

````
Definiere die Daten
```{r}
note = c(5,5,5,5,5,5,5,5,5,5.5)
stunden = c(1,2,3,4,5,6,7,8,9,10)
```

Lineares Modell:
```{r}
model = lm(note ~ stunden)
summary(model)
#was wir fitten:    note = a * stunden + intercept
#was wir erhalten: note = 0.02727 * stunden + 4.9
```
````

```
Call:
lm(formula = note ~ stunden)

Residuals:
     Min       1Q   Median       3Q      Max
-0.14545 -0.08409 -0.02273  0.03864  0.32727

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.90000    0.09770  50.153 2.77e-11 ***
stunden      0.02727    0.01575   1.732    0.122
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.143 on 8 degrees of freedom
Multiple R-squared:  0.2727,    Adjusted R-squared:  0.1818
F-statistic:     3 on 1 and 8 DF,  p-value: 0.1215
```

Lineare Regression

JiTT Ziele

- Ihr könnt Aussagen als Hypothesen formulieren und sie testen.

- Ihr könnt bei der linearen Regression bestimmen, ob die Steigung und Achsenabschnitt signifikant sind.

- Ihr kennt Fehler 1. und 2. Art.