Lineare und logistische Regression

LE3 - Hypothesentests

Dr. Lucia Kleint

Fachhochschule Nordwestschweiz

Herbstsemester 2021

Inhaltsverzeichnis

Ziel der Kompetenz

Die Statistik ist die Grundlage zur zuverlässigen Auswertung von Daten. Allerdings sind Daten nie perfekt und somit muss ein geeignetes Modell gefunden werden. Das Hauptziel dieser Kompetenz ist zu verstehen, wie Daten in der Data Science durch lineare und logistische Regression praktisch ausgewertet werden können.

Jedes Teil-Skript enthält ein Lernergebnis, inklusive Verweisen auf detailliertere Quellen wie Bücher. Für die Prüfung werden die Themen dieses Skripts, aber so detailliert wie in den Verweisen vorausgesetzt.

last update: 12. Juli 2021

Verweise

Folgende Bücher/Unterlagen werden für das genauere Verständnis empfohlen:

- L. Fahrmeir et al., Statistik, Der Weg zur Datenanalyse, 8. Auflage, Springer (pdf via FHNW Netzwerk). Kapitel 10-12.
- wst Skript, Prof. Dr. M. Steiner-Curtis, 2015 (pdf via FHNW AD)
- Video D. Jung `https://www.youtube.com/watch?v=a1DlxpWhG40`
- Video D. Jung `https://www.youtube.com/watch?v=h0nzc-SApFk`

Section 1

Hypothesentests

Hypothesentests

Aussagen wie z.B. “Babies erkennen Farben”, “Ein Forscher hat eine neue Methode entwickelt und behauptet, er könne Krebs mit 99%iger Wahrscheinlichkeit 5 Jahre im Voraus voraussagen” muss man statistisch testen, um etwas über deren Richtigkeit aussagen zu können.
Hypothesentests sind statistische Tests, die solche Aussagen nach einem gewissen Prinzip untersuchen.

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 angenommen (bzw. kann nicht abgelehnt werden).

Hypothesentests

Die Nullhypothese ist immer die zu prüfende Hypothese und die Alternativhypothese ist ihre logische Verneinung, z.B. H_0 = "Babies sehen keine Farben", H_1="Babies sehen Farben". Selbst wenn der Test H_0 annimmt, kann man nie 100%ig sicher sein, dass H_0 wirklich wahr ist, da man aus einer Stichprobe auf die Grundgesamtheit schliesst.

Man kann deswegen zwei Arten von Fehlern machen:

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.

Hypothesentests

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.

Hypothesentests: Fehler

In der Praxis versucht man, die Fehler 1. und 2. Art möglichst klein zu halten. Allerdings hat man das Problem, dass eine Verkleinerung von α automatisch eine Vergrösserung von β zur Folge hat wenn man z.B. die Irrtumswahrscheinlichkeit anpasst. Eine Methode, wenn man β verringern will, aber α dabei nicht vergrössern will, ist die Stichprobengrösse zu erhöhen.

Wichtig: Man kann H_0 nie beweisen, sondern nur H_1. Aus diesem Grund ist es wichtig, dass man die Hypothesen richtig herum formuliert: Der Fall, den man nachweisen möchte, kommt in die Alternativhypothese.

H_0 beinhaltet immer ein '$=$', '$\leq$' oder '$\geq$'. H_1 nie!

Hypothesentests: Fehler

Die Berechnung von α und β folgt aus den Wahrscheinlichkeitsverteilungen, wie man sich hier veranschaulichen kann.

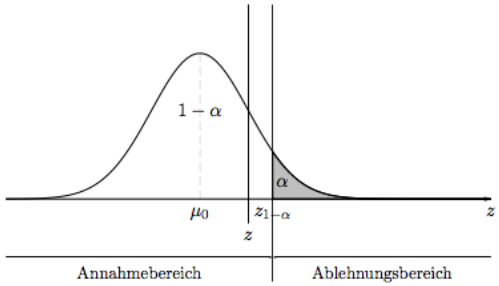


Abbildung 8.3.i: Es sei H_0 richtig: Da $z < z_{1-\alpha}$ wird die Nullhypothese angenommen. Dies ist die richtige Entscheidung, welche mit einer Wahrscheinlichkeit von $1-\alpha$ getroffen wird.

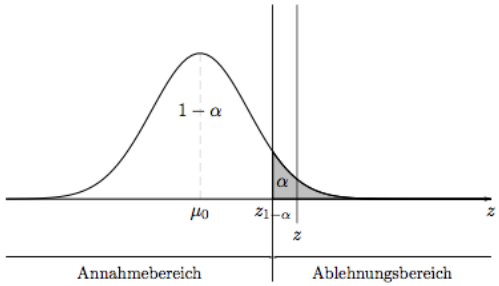


Abbildung 8.3.ii: Es sei H_0 richtig: Da $z \geq z_{1-\alpha}$ wird die Nullhypothese abgelehnt. Dies ist die falsche Entscheidung (**Fehler 1. Art**), welche mit einer Wahrscheinlichkeit von α getroffen wird.

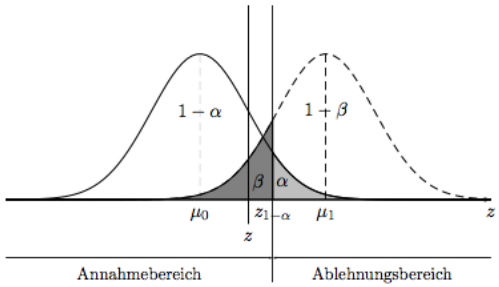


Abbildung 8.3.iii: Es sei H_0 falsch, H_1 richtig, d.h., die gestrichelte Dichte ist die richtige: Da $z < z_{1-\alpha}$ wird die Nullhypothese angenommen. Dies ist die falsche Entscheidung (**Fehler 2. Art**), welche mit einer Wahrscheinlichkeit von β getroffen wird.

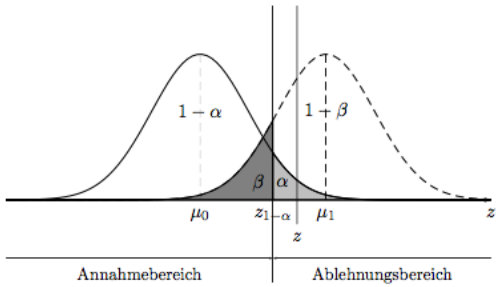


Abbildung 8.3.iv: Es sei H_0 falsch, H_1 richtig, d.h., die gestrichelte Dichte ist die richtige: Da $z \geq z_{1-\alpha}$ wird die Nullhypothese abgelehnt. Dies ist die richtige Entscheidung, welche mit einer Wahrscheinlichkeit (so genannte **Trennschärfe**) von $1-\beta$ getroffen wird.

Bildquelle: wst Vorlesung, Prof. Steiner-Curtis, 2015

Hypothesen aufstellen

Beispiel

Man möchte durch einen Test nachweisen, dass Berufseinsteiger mit Masterabschluss im Durchschnitt mehr verdienen als Berufseinsteiger mit einem Bachelorabschluss. Dazu befragt man 100 Berufseinsteiger nach ihrem Abschluss und Einstiegsgehalt. Wie lautet die Null- bzw. Alternativhypothese in diesem Fall?

Da wir nachweisen wollen, dass Berufseinsteiger mit Masterabschluss ein höheres Einstiegsgehalt haben, muss diese Behauptung in die Alternativhypothese. Die Nullhypothese ist das genaue Gegenteil davon. Solange wir keinen Unterschied im Einkommen nachweisen, müssen wir annehmen, dass beide Gruppen dasselbe verdienen oder sogar Bachelors mehr verdienen als Masters:
H_0: Bachelor- und Masterabsolventen bekommen das gleiche Einstiegsgehalt.
H_1: Masterabsolventen bekommen ein höheres Einstiegsgehalt als Bachelorabsolv.
Mathematisch formuliert: $H_0 : \mu_M \leq \mu_B$ und $H_1 : \mu_M > \mu_B$, wobei μ das durchschnittliche Einstiegsgehalt bezeichnet für Bachelor (B) und Master (M).[1]

1) Beispiel aus `https://www.studyhelp.de/online-lernen/mathe/hypothesentests/`

Hypothesentests

Man unterscheidet einseitige und zweiseitige Hypothesentests.

- Zweiseitige Hypothesentests testen nur auf Ungleichheit
- Einseitige Hypothesentests testen auf einen Unterschied in eine bestimmte Richtung (grösser oder kleiner)

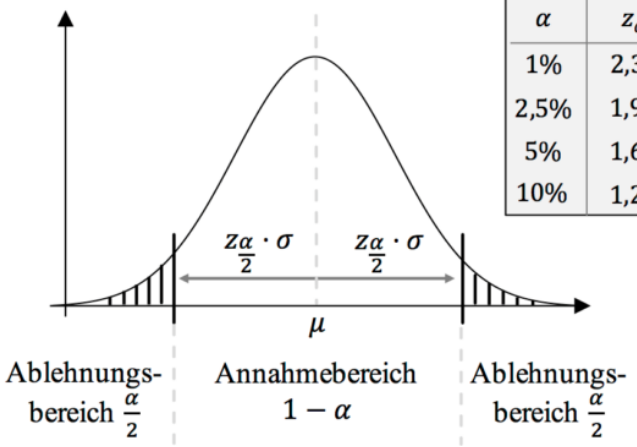
Ein Beispiel für einen zweiseitigen Test ist, wenn man testen will, ob Studenten gleich viel lesen wie Nicht-Studenten.
Die Nullhypothese ist H_0: $\mu_S = \mu_{NS}$.
Die Alternativhypothese ist H_1: $\mu_S \neq \mu_{NS}$.

Ein Beispiel für einen einseitigen Test ist, wenn man testen will, ob Studenten mehr für Bücher ausgeben als Nicht-Studenten.
Die Nullhypothese ist H_0: $\mu_S \leq \mu_{NS}$.
Die Alternativhypothese ist H_1: $\mu_S > \mu_{NS}$.

Hypothesentests

Der Unterschied der beiden Varianten ist hier bildlich dargestellt.

Beidseitiger Test



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

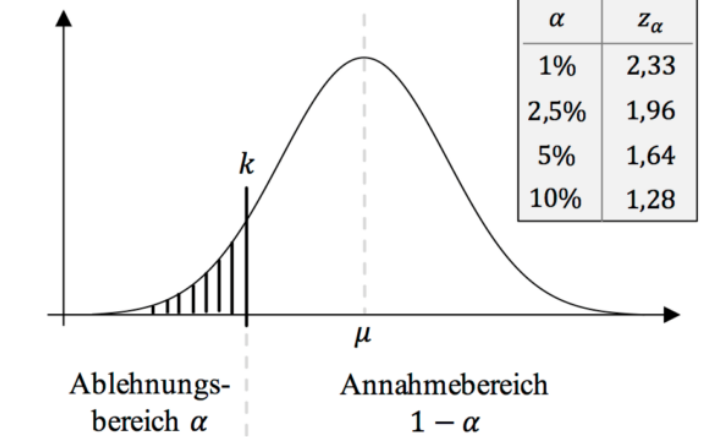
Wenn $\alpha = 5\%$ ist, verwenden wir nicht $z_{0,05} = 1{,}64$ sondern den Wert für $\alpha = 2{,}5\%$, also $z_{0,025} = 1{,}96$.

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/`

Hypothesentests - Beispiel

Beispiel (Zweiseitiger Test)

Man produziert Bleistifte, die einen Erwartungswert von 17 cm Länge haben. Man weiss, dass die Abweichungen in der Länge normalverteilt sind mit Varianz 2.25 cm. Nun macht man eine zufällige Stichprobe und misst Längen von 19.2, 17.4, 18.5, 16.5, 18.9 cm. Muss die Maschine neu kalibriert werden?

Nullhypothese H_0: $\mu_0 = 17$ cm.
Alternativhypothese H_1: $\mu_0 \neq 17$ cm.

Wir berechnen den Mittelwert der Stichprobe $\bar{x} = 18.1$ cm. Die standardisierte Prüfgrösse ist

$$z = \frac{\bar{x} - \mu_0}{\sigma}\sqrt{n} = \frac{18.1 - 17}{\sqrt{2.25}}\sqrt{5} = 1.64$$

Hypothesentests - Beispiel

Wir legen das Signifikanzniveau auf $\alpha = 0.01$ fest, d.h. die Wahrscheinlichkeit, dass wir uns fälschlicherweise für die Alternative entscheiden, ist 1%. Wir müssen nun das (1-$\alpha/2$)=0.995 Quantil der Standardverteilung bestimmen (=2.5758). Wenn somit $|z| > 2.5758$, dann muss die Maschine neu kalibriert werden und weil unser z darunter liegt, kann man schliessen, dass die Produktion in Ordnung ist.

Das Beispiel stammt aus Fahrmeir 10.1.3. Wenn man sich fragt, warum nicht die t-Verteilung verwendet wird: Wenn die Standardabweichung σ der Verteilung bekannt ist und nicht nur eine Standardabweichung S der Stichprobe geschätzt wird, dann wird in der Testgrösse z σ anstelle von S eingesetzt. Dann besitzt die Testgrösse die Normalverteilung $N(0, 1)$, sodass die t-Verteilung nicht benötigt wird, obwohl hier die Stichprobe sehr klein ist.

Hypothesentests: Erwartungswert bei bekannter Varianz

Für einseitige und zweiseitige Hypothesentests kann man folgendermassen zusammenfassen, wann H_0 abgelehnt wird.

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

Hier bezeichnet μ_0 den Erwartungswert, $\bar{x}$ den gemessenen Mittelwert, σ_0 die (bekannte) Varianz, n die Anzahl Datenpunkte der Stichprobe und μ den Mittelwert der Grundgesamtheit.

Parametrische vs. nicht-parametrische Hypothesentests

Man kann Hypothesen in parametrische und nicht-parametrische unterscheiden. Parametrische Hypothesentests setzen bestimmte Eigenschaften der Stichprobe voraus, z.B. normalverteilte Daten. Nicht-parametrische Tests machen keine Annahme zur Verteilung.

Nun könnte man meinen, es sei am sichersten, wenn man immer nicht-parametrische Hypothesentests verwendet. Während nicht-parametrische Tests den Vorteil haben, dass die Verteilung nicht bekannt sein muss und dass sie für kleine Stichproben anwendbar sind, haben sie den Nachteil, dass ihre Aussagekräftigkeit immer tiefer (weniger genau und präzis) ist als bei einem vergleichbaren parametrischen Test.

Parametrische vs. nicht-parametrische Hypothesentests

Für parametrische Tests verweisen wir auf Kapitel 10.1, wo der exakte und approximative Binomialtest und der Gauss-Test erklärt werden. Weitere gebräuchliche Tests sind:

	parametrisch	nicht-parametrisch
Daten	normalverteilt	nicht normalverteilt
Vergleich von 2 unabhängigen Stichproben	t-Test	Mann-Whitney-U Test
Vergleich von 2 abhängigen Stichproben	gepaarter t-Test	Wilcoxon Paarvergleichstest (=Signed-Rank Test)
Vergleich von >2 unabhängigen Stichproben	einfaktorielle Varianzanalyse (ANOVA)	Kruskal-Wallis-Test
Vergleich von >2 abhängigen Stichproben	Varianzanalyse mit Messwiederholung	Friedman-Test
Korrelation zwischen 2 Stichproben	Pearson-Korrelation	Spearman-Korrelation

Quelle der Tabelle:
`https://statistik-und-beratung.de/2012/09/parametrisch-oder-nichtparametrisch-das-ist-hier-die-frage/`

Anwendung Hypothesentests 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 Hypothesentests auf lineare Regression

Eine Ablehnung von H_0 zugunsten von H_1 bedeutet, dass es sinnvoll ist, X mittels $\alpha + \beta x$ zur Erklärung von Y einzusetzen. Man widerlegt somit das Modell $y = \alpha$.

Als Teststatistiken verwendet man die standardisierten Schätzer. Aufgrund ihrer Verteilungseigenschaften erhält man Entscheidungsvorschriften, die analog zu denen des t-Tests im Ein-Stichproben-Fall sind:

$$t = \frac{\beta - 0}{\sigma_\beta}$$

wobei hier t die Teststatistik ist, 0 wegen der Aussage in der Nullhypothese erscheint und σ_β die Standardabweichung der Steigung ist.

Anwendung Hypothesentests auf lineare Regression

Dieser Parameter t wird meistens auch von Statistikprogrammen ausgegeben, in der Grafik unter "T-value" und "Pr (probability)" und zwar für beide Fitparameter.

```
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:

Hide

```
model = lm(note ~ stunden)
summary(model)
```

```
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
```

Abbildung: Beispiel eines R-Outputs für das Beispiel mit der Lerndauer vs. Note

Pr (0.122) ist hier über der normalerweise definierten Signifikanz (0.05) und deswegen wird die Nullhypothese nicht abgelehnt, d.h. in diesem Fall kann es sein, dass die Note nicht von der Lerndauer abhängt.

Anwendung Hypothesentests auf lineare Regression

Ein weiteres Beispiel zeigt den umgekehrten Fall: Prob. level ist hier weit unter der normalerweise definierten Signifikanz (0.05) und deswegen wird die Nullhypothese verworfen, d.h. in diesem Fall sind beide Koeffizienten, α und β relevant.

Parameter	Estimate	Standard Error	T Value	(P-Wert)	Prob. Level
(Intercept)	$\widehat{\alpha} = 1.54492$	$\text{se}^{(\alpha)} = 0.06639$	$T^{(\alpha)} = 23.3$		< 2e-16
temp	$\widehat{\beta} = -0.01611$	$\text{se}^{(\beta)} = 0.00374$	$T^{(\beta)} = -4.3$		1e-04

Residual standard error: $= \widehat{\sigma} = 0.211$ on $n - 2 = 41$ degrees of freedom
R-squared $= 0.503 = r^2_{XY}$
F-statistic: 40.6 on 1 and 40 DF, p-value: 1.47e-07

Abbildung: Beispiel eines Outputs nach einem Fit in R. Quelle: ETH Skript Statistik, W. Stahel

List der Änderungen

- 11.7.21: Folie 11 präzisiert. Hypothesentest/lin Regr. ausgebaut.
- 19.11.20: Bsp S.14: H_0 und H_1 explizit aufgeschrieben. S.12: mathematisch besser: $H_0 \leq$ statt $=$. Ergänzung S.9. Variablen erklärt auf S.16.