Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Datenmanagement & -analyse

Statistische Inferenz und lineare Regressionen

Prof. Dr. Christoph M. Flath
Lehrstuhl für WI & BA
Julius-Maximilians-Universität Würzburg

Sommersemester 2021

Wie man zufällige Größen vergleicht

- Wann immer wir es mit Zufallsgrößen zu tun haben, müssen wir eine Stichprobe ziehen, um Rückschlüsse auf die zugrunde liegende Verteilung zu ziehen
- Die große Frage: Wie erhält man ein Beurteilungssicherheitsniveau bezüglich dieser Größen?

- Anwendungsbeispiele:
  - Wie können wir nach einer medizinischen Studie die Ergebnisse interpretieren, um die Wirksamkeit einer Behandlung zu erforschen?
  - Wie kann man die Ergebnisse von zwei aufeinanderfolgenden Vorlesungsevaluationen vergleichen?

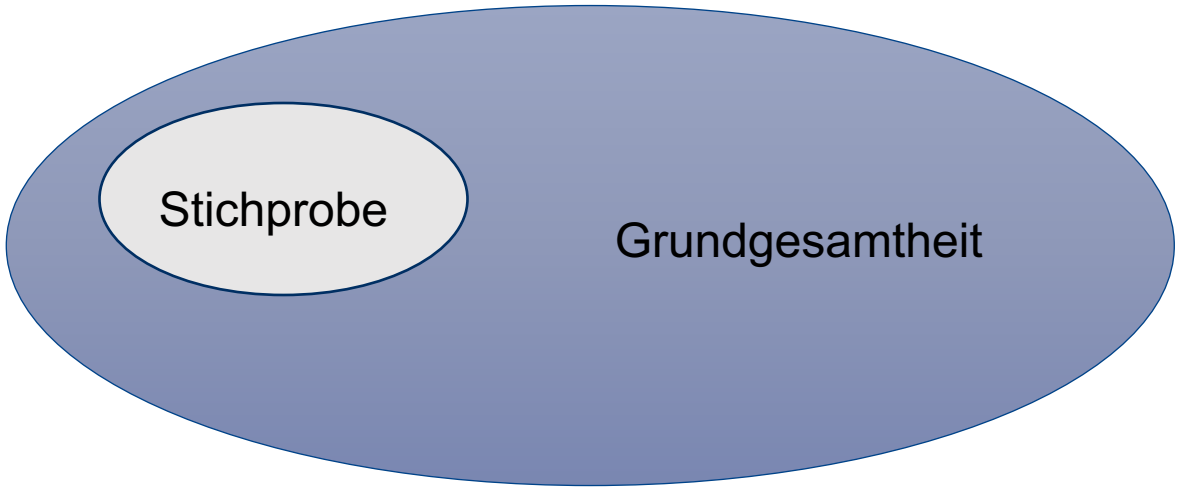
Statistische Inferenz / Hypothesentest

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Agenda

Warum Hypothesentests?

- Der erste Schritt in jeder Studie ist der Test gegen den Zufall.
- Wir können keine Schlussfolgerungen über unsere Ergebnisse ziehen, ohne sicherzustellen, dass unsere Ergebnisse nicht zufällig sind.
- Wir wissen nie genau, wie die wahre Situation ist, aber wir versuchen, die Möglichkeit eines Fehlers zu minimieren.

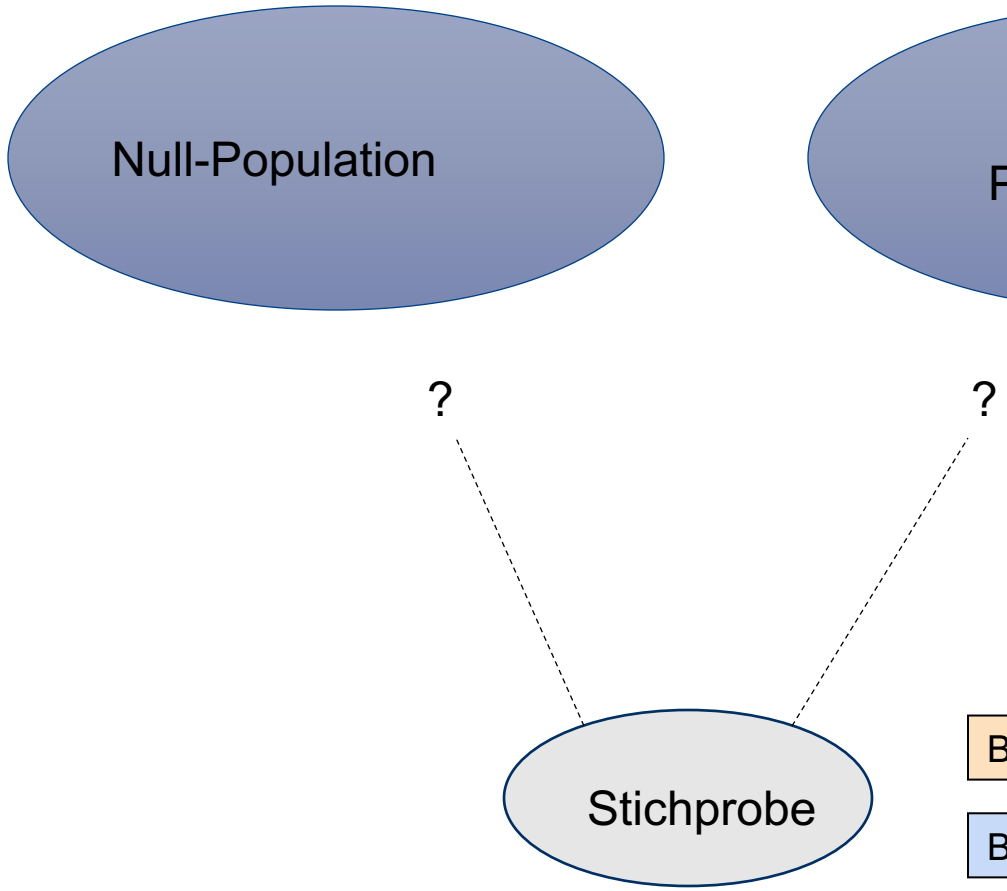


Was können wir aus einer Stichprobe über die Grundgesamtheit schließen?

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Agenda

Hypothesen sind Annahmen über den Ursprung einer Stichprobe



Eine Hypothese ist eine Behauptung (Annahme) über einen Bevölkerungsparameter:

- Bevölkerungsdurchschnitt
- Bevölkerungsanteil

Beispiel: Die mittlere monatliche Handyrechnung dieser Stadt beträgt μ = 21€

Beispiel: Der Anteil der Erwachsenen in dieser Stadt mit Mobiltelefonen ist p = 0,90

Nullhypothese und alternative Hypothese

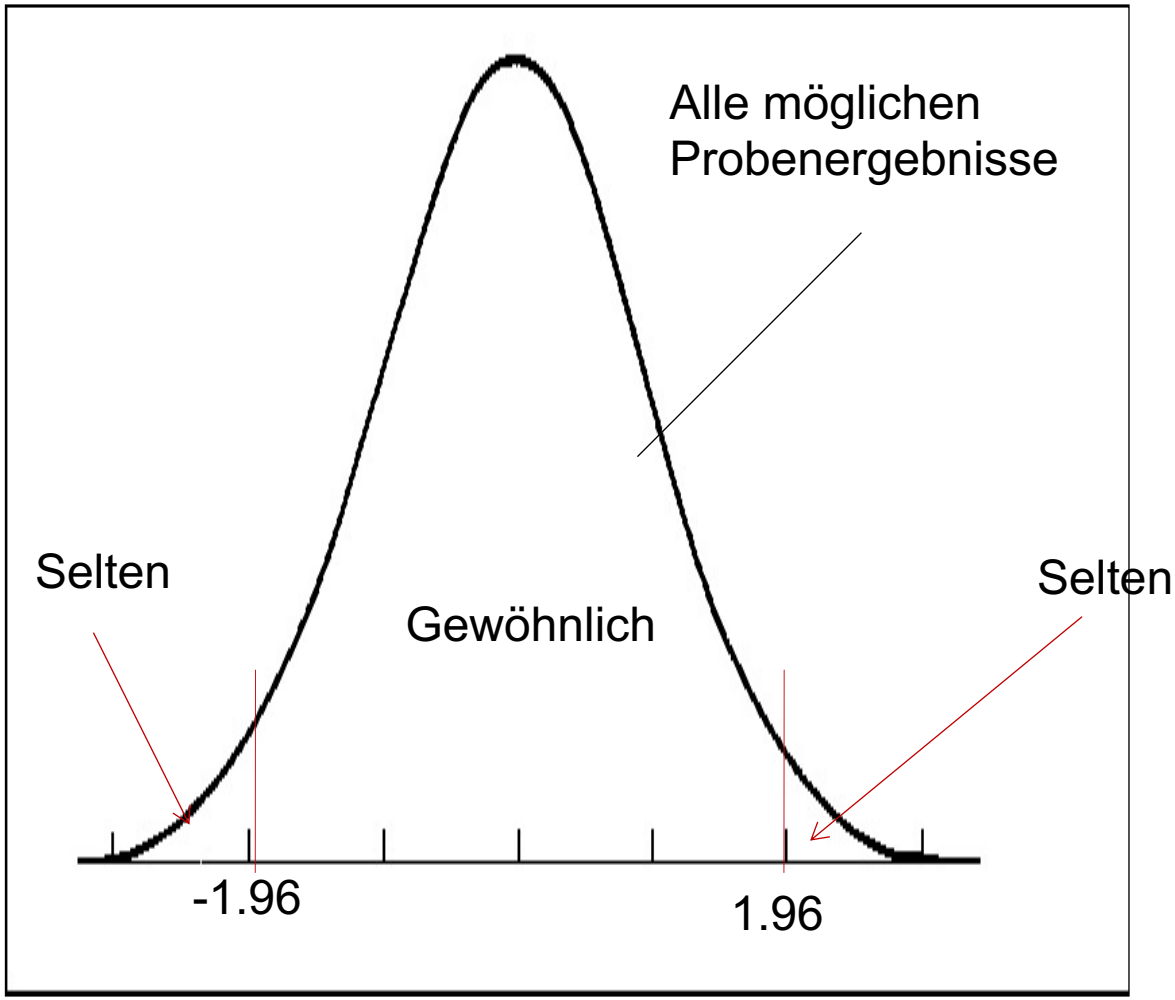
- Welche Grundgesamtheit ist die wahre Grundgesamtheit für unsere Stichprobe?
  - Wir können es nicht mit Sicherheit wissen.

- Wir nehmen an, dass die Nullpopulation wahr ist, und sehen dann, wie wahrscheinlich die Stichprobe unter diesen Umständen wäre → Nullhypothese H_0
  - Bezieht sich typischerweise auf den Status Quo
  - Ähnlich wie dem Rechtsgrundsatz in dubio pro reo
  - Es geht immer um einen Populationsparameter, nicht um eine Stichprobenstatistik (~~$\bar{x} = 3$~~)
  - Enthält immer = , ≤ oder ≥
  - Kann verworfen werden aber nicht für richtig bewiesen werden (falsifizierbar)
  - Beispiel: Die durchschnittliche Anzahl von Fernsehern in US-Haushalten ist drei ($H_0: \mu = 3$)

- Das Gegenteil der Nullhypothese ist die alternative Hypothese H_1
  - Hinterfragt den Status quo
  - Enthält niemals das = , ≤ oder ≥
  - Ist im Allgemeinen die Hypothese, die der Forscher zu unterstützen versucht
  - Wird als gültig angenommen wenn die Nullhypothese verworfen warden kann
  - z. B., Die durchschnittliche Anzahl von TV-Geräten in US-Haushalten ist ungleich 3 (H_1: μ ≠ 3)

Häufige vs. seltene Ergebnisse

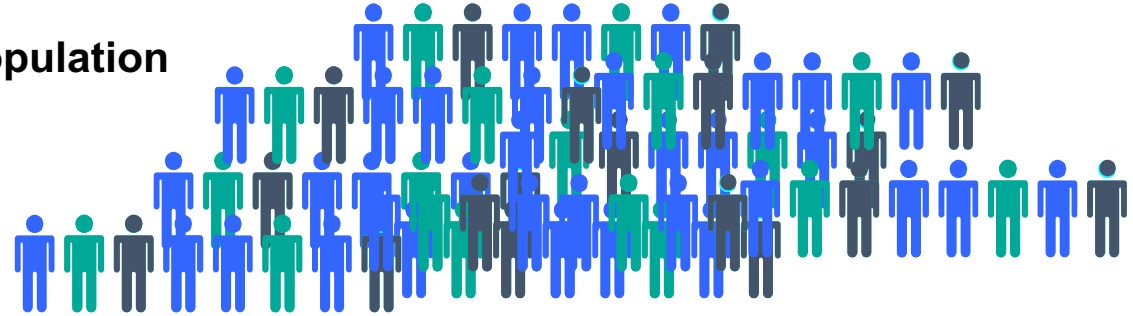
- Wir beginnen immer mit der Annahme, dass die Nullpopulation wahr ist. In diesem Fall fragen wir:
  - Wie wahrscheinlich wäre unser aktuelles Ergebnis, wenn die Nullpopulation die wahre Grundgesamtheit wäre?
- Ein Ergebnis in der Mitte der Normalkurve ist sehr wahrscheinlich (gewöhnlich).
- Ein Ergebnis in den Schwänzen der Normalkurve ist viel unwahrscheinlicher (selten).



Wir verwenden +/- 1,96 (2 Standardabweichungen) als Grenze zwischen häufig und selten.

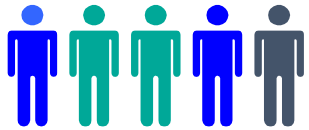
Grundsätzlicher Prozess

Population



Behauptung: Das Durchschnittsalter in Deutschland ist größer als 50 Jahre.
(Null-Hypothese: H0: $\mu = 50$)

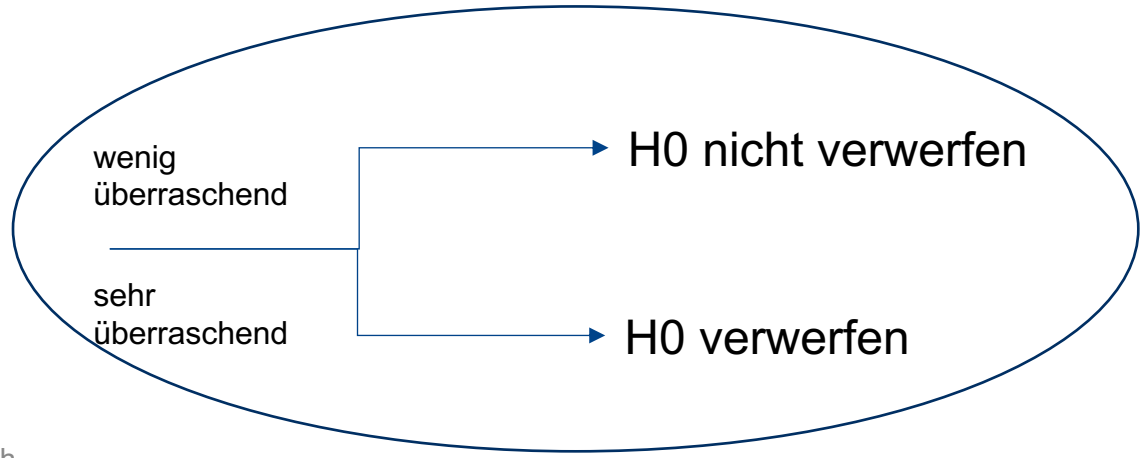
Stichprobe



Messung der Stichprobe: $\bar{x} = 20$

Große Frage

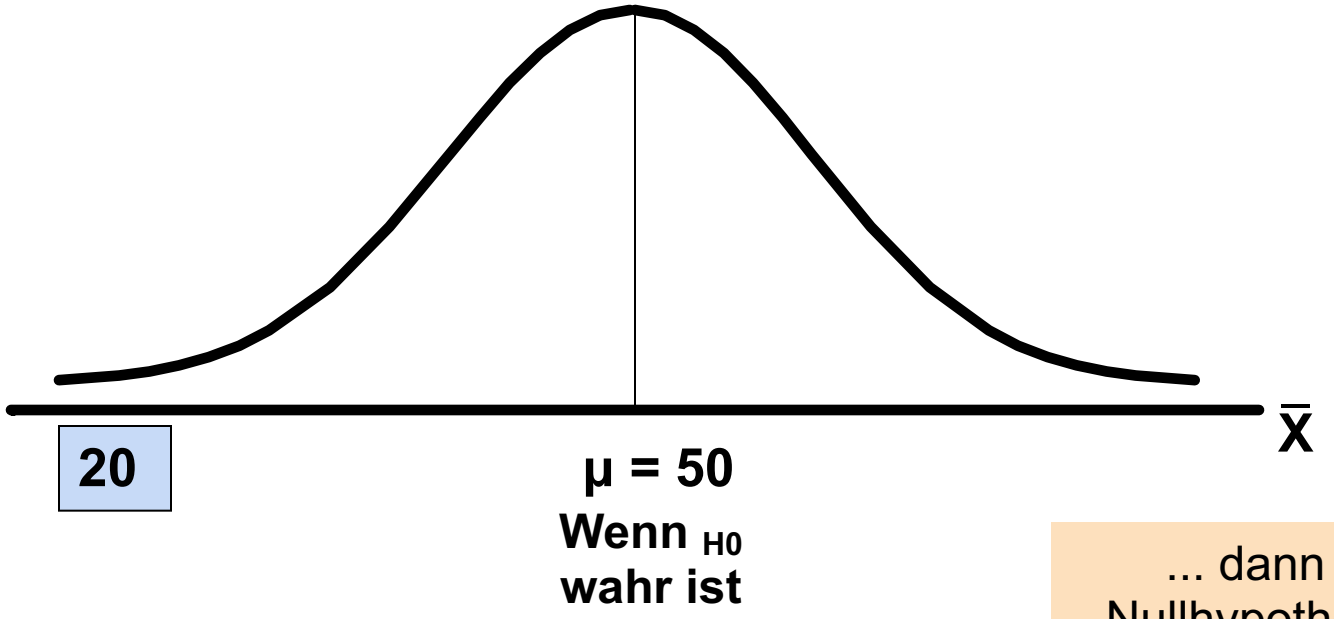
Wie wahrscheinlich ist die beobachtete Stichprobenstatistik unter der Nullhypothese



Grundlage für die Ablehnung von H0

Stichprobenverteilung von $\bar{X}$

Wenn es unwahrscheinlich ist, dass wir einen Stichprobenmittelwert dieses Wertes erhalten würden ...



... dann verwerfen wir die Nullhypothese, dass μ = 50 ist.

... wenn dies in der Tat der Mittelwert der Bevölkerung...

Risikoverständnis bei der Hypothesenprüfung

Wir laufen immer Gefahr, eine falsche Schlussfolgerung zu ziehen:

1. H_0 ist wahr und der Test verwirft H_0 korrekterweise nicht
2. H_0 ist falsch und der Test verwirft H_0 korrekterweise
3. H_0 ist wahr aber Test verwirft H_0 fälschlicherweise
4. H_0 ist falsch aber der Test verwirft H_0 fälschlicherweise nicht

- Das Ergebnis 3 wird als Fehler vom Typ I bezeichnet.
- Das Ergebnis 4 wird als Fehler vom Typ II bezeichnet.
- Typischerweise sind wir am meisten über Fehler vom Typ I besorgt:
  - Unschuldig verurteilte Person
  - Unwirksame Behandlung zugelassen
  - Kranke Person gilt als gesund

Risikoverständnis bei Hypothesentests (Fortsetzung)

- Die Wahrscheinlichkeit eines Typ I Fehlers:

α = P(Ablehnung von H_0 | H_0 ist wahr)

- Die Wahrscheinlichkeit eines Typ II Fehlers:

β = P(nicht Ablehnung von H_0 | H_0 ist falsch)

- Wert von α kann gesteuert werden (wie konservativ ist unser Test?)
  - α wird typischerweise auf 0,01, 0,05, oder 0,10 gesetzt
  - Definiert den Ablehnungsbereich der Stichprobenverteilung über den/die kritischen Wert(e) des Tests
- Wert von β kann nicht im Voraus festgelegt werden und hängt vom Wert des (unbekannten) Populationsparameters ab

Ablehnungsbereich zum Signifikanzniveau α

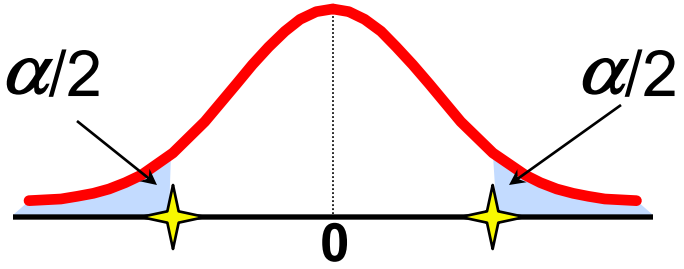
Grenzwert

Ablehnungsbereich

Zweiseitiger Test

H_0: $\mu = 3$

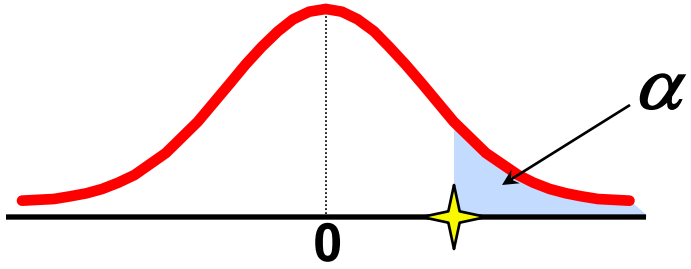
H_1: $\mu \neq 3$



Einseitiger Test

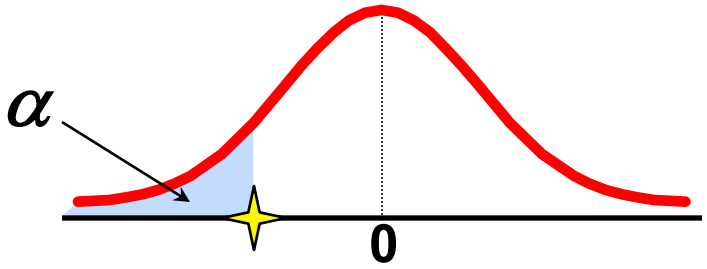
H_0: $\mu \leq 3$

H_1: $\mu > 3$



H_0: $\mu \geq 3$

H_1: $\mu < 3$



Faktoren, die den Fehler vom Typ II beeinflussen

- Ceteris paribus,
  - β steigt, wenn die Differenz zwischen hypothetischem Parameter und seinem wahren Wert gering ist
  - β steigt wenn α fällt
  - β steigt wenn σ (Varianz der Stichprobe) steigt
  - β fällt wenn Stichprobe größer wird
- Nicht vergessen – α ist festgelegt!

- Die Trennschärfe eines Tests ist die Wahrscheinlichkeit, eine Nullhypothese, die falsch ist, zurückzuweisen
  - d. h., P(Ablehnung von H_0 | H_1 ist wahr)
- Die Aussagekraft des Tests steigt mit zunehmendem Stichprobenumfang

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Schritte des Hypothesentestverfahrens

1. Identifizieren Sie den Populationsparameter und formulieren Sie die zu prüfenden Hypothesen.

2. Wählen Sie ein *Signifikanzniveau* (Risiko, eine falsche Schlussfolgerung zu ziehen).

3. Bestimmen Sie eine Entscheidungsregel, auf die Sie eine Schlussfolgerung stützen können.

---

4. Daten sammeln und Teststatistik berechnen.

5. Wenden Sie die Entscheidungsregel an und ziehen Sie eine Schlussfolgerung.

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Agenda

Arten von Ein-Stichproben-Hypothesentests

1. H_0: Parameter = konstant
   H_1: Parameter ≠ konstant

2. H_0: Parameter ≤ konstant
   H_1: Parameter > Konstante

3. H_0: Parameter ≥ konstant
   H_1: Parameter < Konstante

Der Gleichheitsteil des Hypothesenzeichens steht *immer* in der Nullhypothese. (konservativer)

Ein erstes Beispiel

- Eine Gruppe von Spielern wirft eine Münze und während 20 Würfen gibt es 15 Mal Kopf
- Ein verlierender Spieler sieht die Möglichkeit des Schummelns und möchte, dass wir einen Ein-Stichproben-Hypothesentest konstruieren, wobei p die Wahrscheinlichkeit für Kopf ist:
  - H_0: $p = 0{,}5$
  - H_1: $p > 0{,}5$
- Die Idee: Reverse Engineering des Testproblems

- In einer Simulation vieler Stichproben der gleichen Größe (20 Würfe) mit dem Nullhypothesenparameter (p=0,5), wie oft ergibt sich das realisierte Ereignis (15xKopf)?

- R-Befehl für Binomialverteilung:

```
rbinom(n, size, prob)
```

Ein erstes Beispiel

- Eine Gruppe von Spielern wirft eine Münze und während 20 Würfen gibt es 15 Mal Kopf
- Ein verlierender Spieler sieht die Möglichkeit des Schummelns und möchte, dass wir einen Ein-

Ceteris paribus,
β steigt, wenn die Differenz zwischen hypothetischem Parameter und seinem wahren Wert gering ist
β steigt wenn α fällt
β steigt wenn σ (Varianz der Stichprobe) steigt
β fällt wenn Stichprobe größer wird

- In einer Simulation vieler Stichproben der gleichen Größe (20 Würfe) mit dem Nullhypothesenparameter (p=0,5), wie oft ergibt sich das realisierte Ereignis (15xKopf)?

- R-Befehl für Binomialverteilung:

```
rbinom(n, size, prob)
```

Beispiel 2: Formulierung eines Ein-Stichproben-Tests

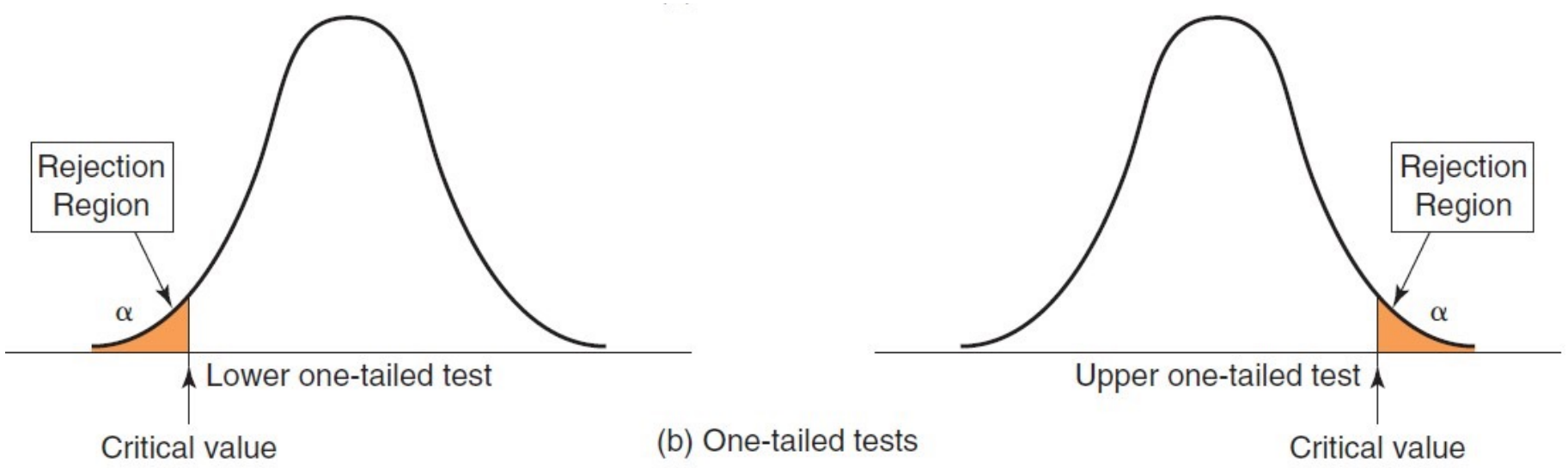
- Eine Softwarefirm geht davon aus, dass die durchschnittliche Reaktionszeit für technische Supportanfragen weniger als 25 Minuten ist
- Die Statistik der Beispieldaten ist wie folgt:
  - $n = 44$
  - Stichprobenmittelwert: 21.91
  - Standardabweichung der Probe: 19,49
- Stellen Sie die Hypothesen auf:
  - H_0 : mittlere Reaktionszeit ≥ 25
  - H_1 : mittlere Reaktionszeit < 25

20	28	19	5
12	13	47	33
15	2	24	29
11	25	19	2
22	25	17	25
6	48	13	61
39	12	8	15
19	118	33	11
12	27	21	2
13	11	2	31
13	21	15	20

Ziehen von Schlussfolgerungen bei einseitigen Tests

H_0: Parameter ≥ Konstante
H_1: Parameter < Konstante

H_0: Parameter ≤ Konstante
H_1: Parameter > Konstante



(b) One-tailed tests

Berechnen der Teststatistik

- Ein-Stichproben-Test auf einen Mittelwert, σ unbekannt $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$

- Im Beispiel ergaben die Beispieldaten für 44 Kunden eine mittlere Antwortzeit von 21,91 Minuten und eine Stichprobenstandardabweichung von 19,49 Minuten.

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{21.91 - 25}{19.49/\sqrt{44}} = \frac{-3.09}{2.938} = -1.05$$

- t = -1,05 zeigt an, dass der Stichprobenmittelwert von 21,91 um 1,05 Standardfehler unter dem hypothetischen Mittelwert von 25 Minuten liegt

Beispiel: Den kritischen Wert finden und eine Schlussfolgerung ziehen

- Unter der Annahme der t-Verteilung für die Test-Statistik ergibt sich der kritische Wert als das 0.05-Quantil der t-Verteilung

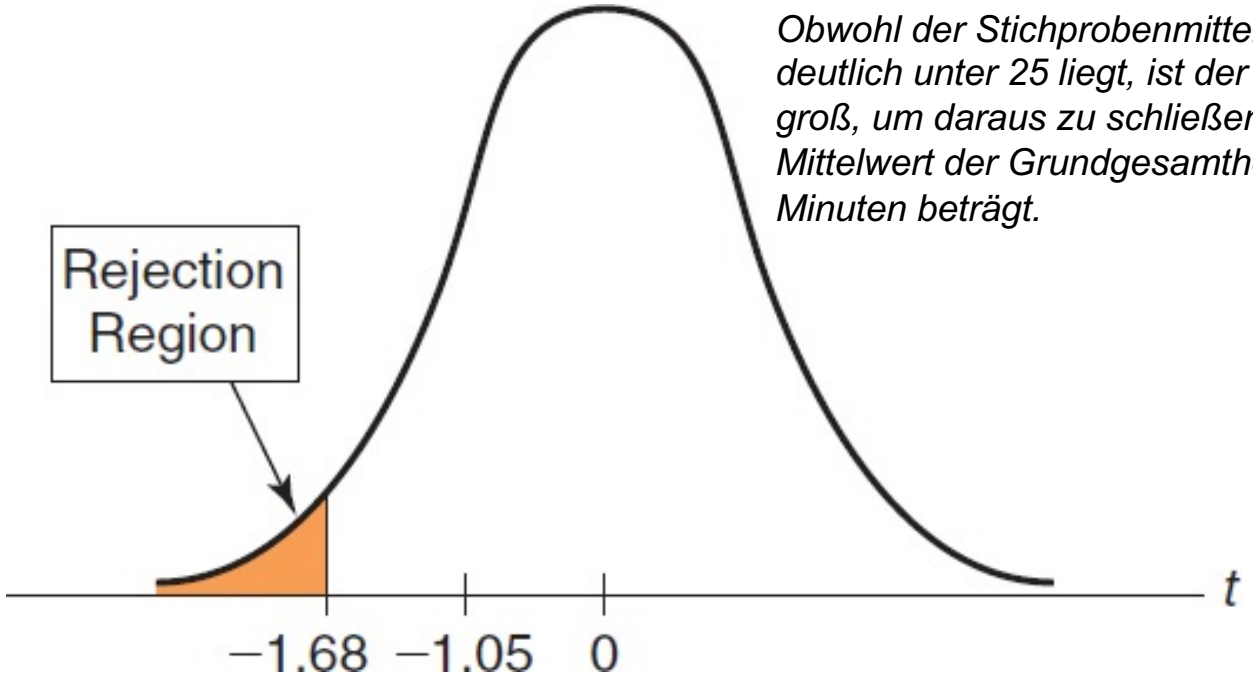
```
=qt(alpha,df)
=pt(0.05,43)
= -1.681071
```

Appendix B Statistical Tables

Degrees of Freedom	Upper Tail Areas					
	.25	.10	.05	.025	.01	.005
31	0.6825	1.3095	1.6955	2.0395	2.4528	2.7440
32	0.6822	1.3086	1.6939	2.0369	2.4487	2.7385
33	0.6820	1.3077	1.6924	2.0345	2.4448	2.7333
34	0.6818	1.3070	1.6909	2.0322	2.4411	2.7284
35	0.6816	1.3062	1.6896	2.0301	2.4377	2.7238
36	0.6814	1.3055	1.6883	2.0281	2.4345	2.7195
37	0.6812	1.3049	1.6871	2.0262	2.4314	2.7154
38	0.6810	1.3042	1.6860	2.0244	2.4286	2.7116
39	0.6808	1.3036	1.6849	2.0227	2.4258	2.7079
40	0.6807	1.3031	1.6839	2.0211	2.4233	2.7045
41	0.6805	1.3025	1.6829	2.0195	2.4208	2.7012
42	0.6804	1.3020	1.6820	2.0181	2.4185	2.6981
43	0.6802	1.3016	1.6811	2.0167	2.4163	2.6951
44						

Beispiel: Finden des kritischen Werts und Ziehen einer Schlussfolgerung (Forts.)

- t = -1,05 fällt nicht in den Ablehnungsbereich
- H_0 kann nicht verworfen werden



Obwohl der Stichprobenmittelwert von 21,91 deutlich unter 25 liegt, ist der Stichprobenfehler zu groß, um daraus zu schließen, dass der wahre Mittelwert der Grundgesamtheit weniger als 25 Minuten beträgt.

p-Werte

- Ein alternativer Ansatz für Schritt 3 eines jeden Hypothesentests (Aufstellen einer Entscheidungsregel) verwendet den p-Wert anstelle des kritischen Werts
- Der p-Wert ist das *beobachtete Signifikanzniveau*

- Die p-Wert-Entscheidungsregel ist dann:

Verwerfen Sie H_0, wenn der *p-Wert* < α

t-test in R

```
> t.test(x,alternative = "less", mu = 25)

        One Sample t-test

data:  x
t = -1.0522, df = 43, p-value = 0.1493
alternative hypothesis: true mean is less than 25
95 percent confidence interval:
    -Inf 26.8475
sample estimates:
mean of x
 21.90909
```

Im Beispiel ist der p-Wert 0,1493

Verwerfen Sie H_0 nicht, weil die
p-Wert ist nicht kleiner als α
0,1498 ist nicht kleiner als 0,05

t-test in R

```
> t.test(x,alternative = "less", mu = 25)

        One Sample t-test

data:  x
t = -1.0522, df = 43, p-value = 0.1493
alternative hypothesis: true mean is less than 25
95 percent confidence interval:
    -Inf 26.8475
sample estimates:
mean of x
 21.90909
```

Ceteris paribus,

β steigt, wenn die Differenz zwischen hypothetischem Parameter und seinem wahren Wert gering ist

β steigt wenn α fällt

β steigt wenn σ (Varianz der Stichprobe) steigt

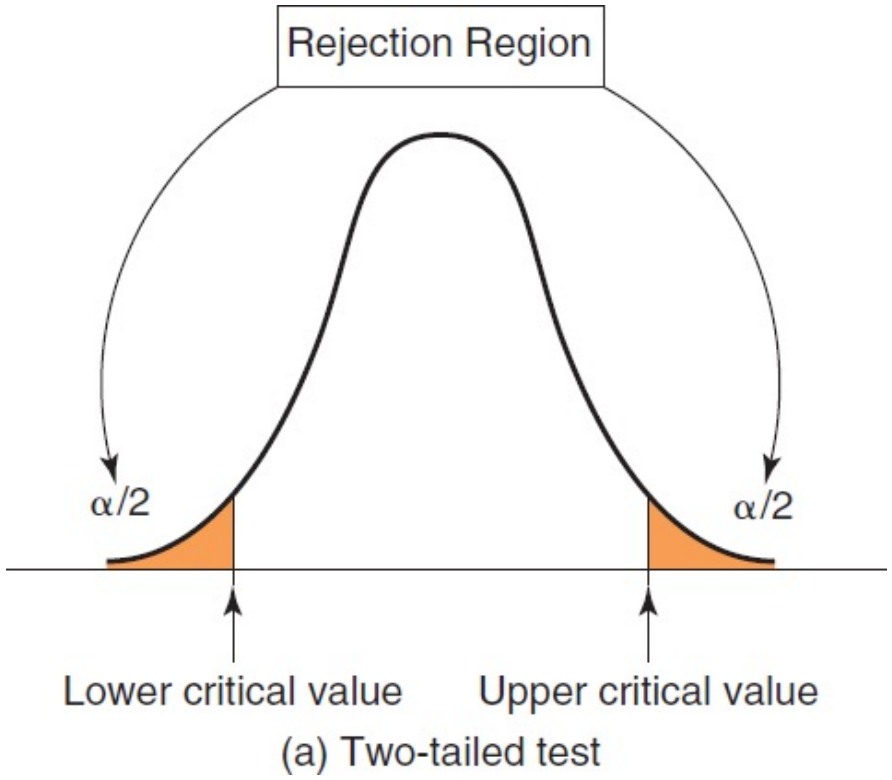
β fällt wenn Stichprobe größer wird

Beispiel: Durchführen eines zweiseitigen Hypothesentests für den Mittelwert

- Testen Sie zu einem Signifikanzniveau von 5 %, ob das Durchschnittsalter der Befragten gleich 35 ist

Age	24	26	28	33	45	49	29	37	37	38	38	39	39	40	42	42	43	44	44	45	46	46	48	48	24	26	28	32	34	37	39	46	49	50

H_0: Parameter = konstant
H_1: Parameter ≠ konstant



Nicht schwierig

```
> t.test(x2, mu = 35)

        One Sample t-test

data:  x2
t = 2.7283, df = 33, p-value = 0.01012
alternative hypothesis: true mean is not equal to 35
95 percent confidence interval:
 35.93485 41.41809
sample estimates:
mean of x
 38.67647
```

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Agenda

Vergleich von Populationen

- Sehr oft führen wir statistische Tests durch, um auf die Existenz von Unterschieden zwischen zwei verschiedenen Gruppen (z. B. Behandlungsgruppen) zu schließen
- Wenn Beobachtungen über Populationen hinweg voneinander abhängen, wendet man einen "gepaarten" Stichprobentest an, indem man das entsprechende Argument anpassen

Stellen Sie einen Hypothesentest auf, um festzustellen, ob die mittlere Lieferzeit für den Anbieter alum (μ1) größer ist als die mittlere Lieferzeit für den Anbieter durrable (μ_2).

H_0: $\mu_1 \leq \mu_2$

```
alum=c(5,7,6,9,7,9,6,7)
durrable =
c(5,5,5,5,6,5,5,5,3,4,6,5,5)
```

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Es wird monoton ☺

```
> t.test(alum, durrable, alternative = "greater")

        Welch Two Sample t-test

data:  alum and durrable
t = 3.828, df = 9.5306, p-value = 0.001818
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 1.088604      Inf
sample estimates:
mean of x mean of y
 7.000000  4.923077
```

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Agenda

Regressionsanalyse

- Idee: Vorhersage einer wirtschaftlichen Größe (= abhängige Variable) aufgrund bekannter und messbarer Einflussfaktoren (= unabhängige Variablen)

- Beispiel
- Abhängige Variable: Umsatz mit Badeanzügen
- Unabhängige Variablen (u.a.)
  - Verkaufspreis
  - Preise der Mitbewerber
  - Temperatur
  - Marketing-Ausgaben

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Agenda

1 Hypothesentests

2 Lineare Regression

2.1 Einfache Lineare Regression

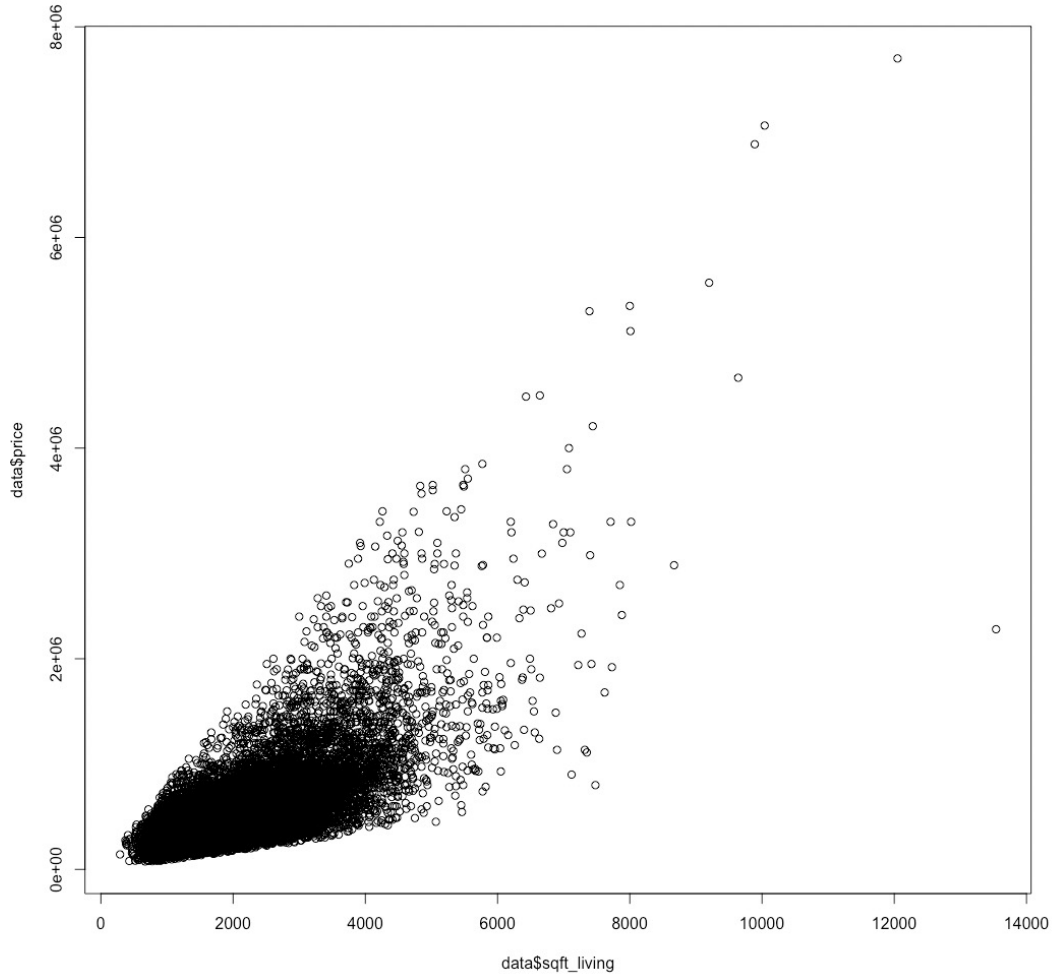
2.2 Multiple Lineare Regression

2.3 Dummy-Variablen

2.4 Variablen-Selektion

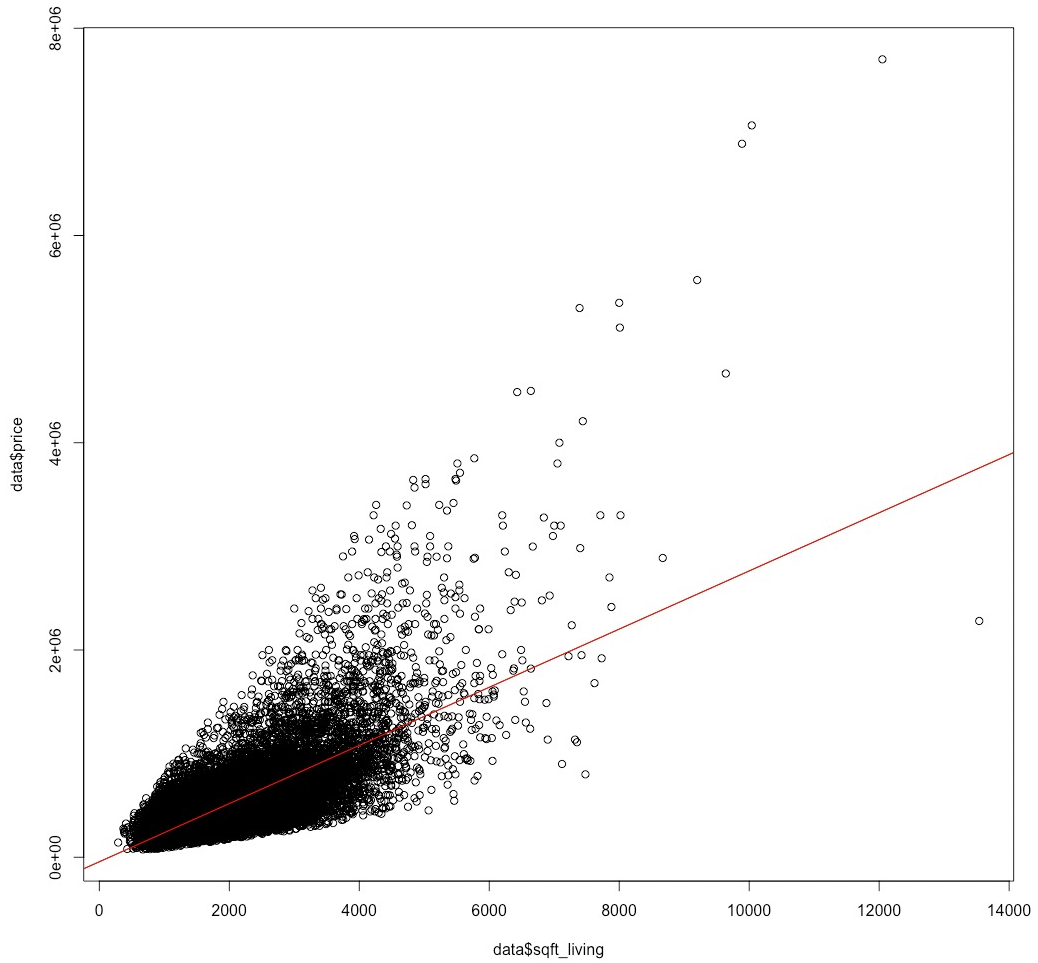
Beispiel für eine einfache lineare Regression

- Beispiel: Vorhersage des Verkaufspreises eines Hauses (abhängige Variable y) basierend auf der Grundstücksgröße (unabhängige Variable x)
  - Verkaufspreis und Größe von 21613 kürzlich verkauften Häusern sind verfügbar
- Identifizieren Sie die erklärendste (in Bezug auf die Beispieldaten) Funktion der Form
- Die anfängliche Annahme eines linearen Zusammenhangs kann mit Hilfe eines Streudiagramms überprüft werden



Bestimmen der "bestpassenden" Regressionsfunktion für die Daten

- Die Regressionskoeffizienten und müssen optimal bestimmt werden
- Der mittlere quadratische Fehler (MSE) zwischen den Daten und den Schätzungen ist ein gutes Maß für die Anpassungsgüte
- Eine einfache lineare Regression versucht, den MSE zu minimieren



Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Das zugrunde liegende Regressionsmodell

```
> reg = lm(price ~ sqft_living, data = data)
> summary(reg)

Call:
lm(formula = price ~ sqft_living, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1476062  -147486   -24043   106182  4362067

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -43580.743   4402.690  -9.899   <2e-16 ***
sqft_living    280.624      1.936 144.920   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 261500 on 21611 degrees of freedom
Multiple R-squared:  0.4929,    Adjusted R-squared:  0.4928
F-statistic: 2.1e+04 on 1 and 21611 DF,  p-value: < 2.2e-16
```

Das zugrunde liegende Regressionsmodell

```
> reg = lm(price ~ sqft_living, data = data)
> summary(reg)

Call:
lm(formula = price ~ sqft_living, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1476062  -147486   -24043   106182  4362067

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -43580.743   4402.690  -9.899   <2e-16 ***
sqft_living    280.624      1.936 144.920   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 261500 on 21611 degrees of freedom
Multiple R-squared:  0.4929,    Adjusted R-squared:  0.4928
F-statistic: 2.1e+04 on 1 and 21611 DF,  p-value: < 2.2e-16
```

Achsenabschnitt

Steigung

Koeffizient signifikant unterschiedlich von Null? (t-Test)

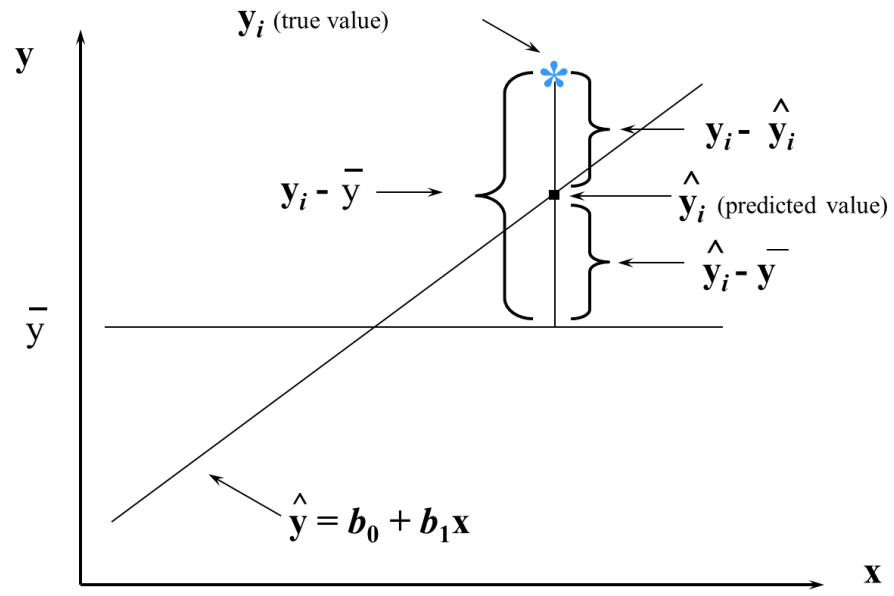
Anpassungsgüte

Interpretieren des Modells

- Preis = -43,581+ 0,2806 * Größe
- Interpretation der Steigung
  - Im Durchschnitt führt eine Vergrößerung des Hauses um 1.000 Quadratmeter zu einem Anstieg des Verkaufspreises um $280,6
- Interpretation des Abschnitts
  - Ein Haus der Größe 0 wird im Durchschnitt für -$43.581 verkauft
  - *Keine direkte praktische Auswirkung, da es keine Häuser der Größe 0 gibt!*

- Das R^2-Maß quantifiziert die Anpassungsgüte (Goodness-of-Fit) des Regressionsmodells
  - Bereich: $0 < R2 < 1$
  - Kann als Prozentsatz interpretiert werden: 0,493 → 49,3 % der Verkaufspreisvariation von Häusern wird durch die Hausgröße erklärt
- Formal ist R^2 der Anteil der Gesamtvariation, der durch das Modell erklärt wird

Zerlegung des Vorhersagefehlers



$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{TSS}} = \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{RSS}} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{ESS}}$$

TSS	=	RSS	+	ESS
Total deviations		Unexplained (residual) deviations		Explained deviations

Verwendung des Regressionsmodells

- Beispiel
  - Prognostizierter Verkaufspreis für ein 3.100 Quadratmeter großes Haus

    X = 3.10
  - Y = -43.581 + 280.6 * 3.1 = 295
    - Der erwartete Verkaufspreis beträgt $826,279

- Wichtige Hinweise
  - Die Werte der unabhängigen Variablen müssen zugänglich sein
  - Das Modell ist nur für die vom Datensatz abgedeckten X-Bereiche gültig
    - Vorhersagen für X-Werte außerhalb der bekannten Daten sollten hinterfragt werden

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Agenda

1 Hypothesentests

2 Lineare Regression

2.1 Einfache Lineare Regression

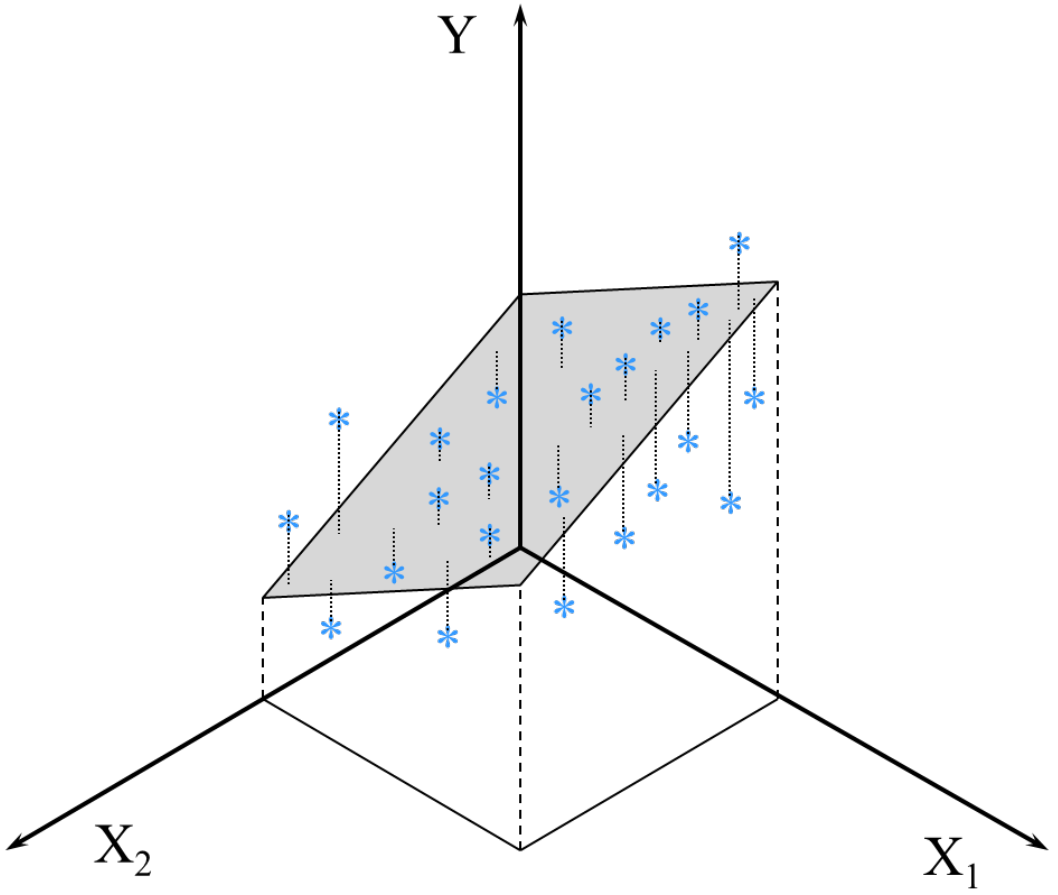
2.2 Multiple Lineare Regression

2.3 Dummy-Variablen

2.4 Variablen-Selektion

Quantitative Vorhersage mit multiplen linearen Regressionsmodellen

Situation: *p* unabhängige Variablen



mit

=Konstante (y-Achsenabschnitt)

=Steigung entlang der Dimension i

Beispiel von zuvor mit Grundstücksgröße

```
> reg2 = lm(price ~ sqft_living + sqft_lot, data = data)
> summary(reg2)

Call:
lm(formula = price ~ sqft_living + sqft_lot, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1417234  -147122   -23174   106305  4343197

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.390e+04  4.399e+03  -9.981  < 2e-16 ***
sqft_living  2.829e+02  1.964e+00 144.030  < 2e-16 ***
sqft_lot    -2.893e-01  4.355e-02  -6.644 3.13e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 261200 on 21610 degrees of freedom
Multiple R-squared:  0.4939,    Adjusted R-squared:  0.4938
F-statistic: 1.054e+04 on 2 and 21610 DF,  p-value: < 2.2e-16
```

Interpretation der Ergebnisse

- Nicht signifikante Variablen könnten prinzipiell aus dem Modell entfernt werden
  - Dies spiegelt sich im "t-Test" wider
  - P-Wert sollte < 0,05 sein ($0 \notin \mathrm{CI}$)

- "Beispiel "Hausverkaufspreise
  - *Sowohl Landfläche als auch Wohnungsgröße* sind signifikant, haben aber unterschiedliche Vorzeichen
    - Nicht überraschend, da beide Variablen ein ähnliches Konzept ausdrücken
    - Die Variablen sind höchstwahrscheinlich hoch korreliert

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Agenda

Kategoriale unabhängige Variablen

- In den bisherigen Beispielen waren die unabhängigen Variablen kontinuierliche numerische Variablen
- Oft sind wir an der Wirkung von kategorischen Variablen (z. B. Ethnie, Geschlecht) interessiert
- Frage: Wie können wir kategoriale Variablen in die Regression einbeziehen?

- Option 1: Analysieren Sie jede Untergruppe separat
  - Erzeugt unterschiedliche Steigung, konstant für jede Gruppe
  - Unpraktisch da Daten nicht gemeinsam genutzt werden können
- Option 2: Dummy-Variablen
  - "Dummy" = eine binäre Variable, die kodiert ist, um das Vorhandensein oder Fehlen von etwas anzuzeigen
  - Abwesenheit kodiert als Null, Anwesenheit kodiert als 1.
  - Durch die Einführung von Dummy-Variablen (künstlich definierte Variablen) können wir ein Modell umwandeln, um kategoriale unabhängige Variablen zu berücksichtigen

In unserem Datensatz gibt es ein schönes Beispiel

In unserem Datensatz gibt es ein schönes Beispiel

Dummy-Variable:
waterfront

!!!

```
> reg2 = lm(price ~ sqft_living + waterfront, data = data)
> summary(reg2)

Call:
lm(formula = price ~ sqft_living + waterfront, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1376782  -142867   -21360   107201  4449253

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -32957.851   4242.971  -7.768 8.35e-15 ***
sqft_living    272.507      1.873 145.499  < 2e-16 ***
waterfront  829983.104  19882.279  41.745  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 251500 on 21610 degrees of freedom
Multiple R-squared:  0.5307,    Adjusted R-squared:  0.5307
F-statistic: 1.222e+04 on 2 and 21610 DF,  p-value: < 2.2e-16
```

Wenn es nicht für Uferlage reicht

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Ergibt das Sinn? Was sollten wir noch tun?

```
Call:
lm(formula = price ~ sqft_living + waterfront + view, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1487434  -138469   -18344   104874  4391875

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -16322.391   4186.474  -3.899 9.69e-05 ***
sqft_living    256.797      1.902 135.044  < 2e-16 ***
waterfront  574358.138  21133.213  27.178  < 2e-16 ***
view         76680.750   2475.421  30.977  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 246100 on 21609 degrees of freedom
Multiple R-squared:  0.5507,    Adjusted R-squared:  0.5506
F-statistic:  8827 on 3 and 21609 DF,  p-value: < 2.2e-16
```

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Agenda

```
lm(formula = price ~ ., data = select(data, -date, -id, -long,
    -lat, -zipcode))

Residuals:
     Min       1Q   Median       3Q      Max
-2543337  -105875    -8850    85737  4039159

Coefficients: (1 not defined because of singularities)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.097e+06  1.648e+05  30.934  < 2e-16 ***
bedrooms      -2.796e+04  1.977e+03 -14.143  < 2e-16 ***
bathrooms0.5  -1.114e+05  1.224e+05  -0.910 0.362744
bathrooms0.75 -7.213e+03  7.020e+04  -0.103 0.918163
bathrooms1    -3.178e+04  6.603e+04  -0.481 0.630323
bathrooms1.25 -2.245e+04  9.535e+04  -0.235 0.813847
bathrooms1.5  -4.820e+04  6.620e+04  -0.728 0.466603
bathrooms1.75 -5.262e+04  6.611e+04  -0.796 0.426076
bathrooms2    -4.349e+04  6.617e+04  -0.657 0.511060
bathrooms2.25 -4.387e+04  6.620e+04  -0.663 0.507492
bathrooms2.5  -6.535e+04  6.613e+04  -0.988 0.323073
bathrooms2.75 -5.019e+04  6.639e+04  -0.756 0.449697
bathrooms3    -1.364e+04  6.656e+04  -0.205 0.837590
bathrooms3.25  6.663e+04  6.676e+04   0.998 0.318216
bathrooms3.5   3.384e+04  6.670e+04   0.507 0.611895
bathrooms3.75  1.738e+05  6.837e+04   2.542 0.011014 *
bathrooms4     1.654e+05  6.866e+04   2.410 0.015979 *
bathrooms4.25  2.778e+05  7.042e+04   3.945 8.00e-05 ***
bathrooms4.5   1.838e+05  6.964e+04   2.640 0.008306 **
bathrooms4.75  5.594e+05  7.939e+04   7.046 1.89e-12 ***
bathrooms5     4.485e+05  8.042e+04   5.577 2.48e-08 ***
bathrooms5     4.485e+05  8.042e+04   5.577 2.48e-08 ***
bathrooms5.25  5.537e+05  8.791e+04   6.299 3.05e-10 ***
bathrooms5.5   9.021e+05  9.368e+04   9.630  < 2e-16 ***
bathrooms5.75  8.718e+05  1.235e+05   7.061 1.70e-12 ***
bathrooms6     1.285e+06  1.078e+05  11.920  < 2e-16 ***
bathrooms6.25  1.098e+06  1.611e+05   6.820 9.37e-12 ***
bathrooms6.5   2.452e+05  1.610e+05   1.523 0.127655
bathrooms6.75  4.395e+05  1.616e+05   2.720 0.006538 **
bathrooms7.5  -1.886e+04  2.172e+05  -0.087 0.930826
bathrooms7.75  4.581e+06  2.178e+05  21.029  < 2e-16 ***
bathrooms8     2.122e+06  1.632e+05  13.007  < 2e-16 ***
sqft_living    1.448e+02  4.547e+00  31.838  < 2e-16 ***
sqft_lot       8.296e-04  4.907e-02   0.017 0.986510
floors1.5      1.383e+04  5.587e+03   2.475 0.013323 *
floors2        2.283e+04  4.674e+03   4.884 1.05e-06 ***
floors2.5      1.452e+05  1.693e+04   8.577  < 2e-16 ***
floors3        1.351e+05  9.610e+03  14.059  < 2e-16 ***
floors3.5      2.944e+05  7.362e+04   3.999 6.39e-05 ***
waterfront     5.784e+05  1.788e+04  32.350  < 2e-16 ***
view           3.649e+04  2.181e+03  16.731  < 2e-16 ***
condition      2.640e+04  2.409e+03  10.958  < 2e-16 ***
grade          1.206e+05  2.176e+03  55.442  < 2e-16 ***
sqft_above    -1.760e+01  4.560e+00  -3.859 0.000114 ***
sqft_basement         NA         NA      NA       NA
yr_built      -2.961e+03  7.636e+01 -38.781  < 2e-16 ***
yr_renovated   2.038e+01  3.765e+00   5.411 6.33e-08 ***
sqft_living15  5.170e+01  3.526e+00  14.663  < 2e-16 ***
sqft_lot15    -6.168e-01  7.506e-02  -8.218  < 2e-16 ***
```

```
lm(formula = price ~ ., data = select(data, -date, -id, -long,
    -lat, -zipcode))

Residuals:
     Min       1Q   Median       3Q      Max
-2543337  -105875    -8850    85737  4039159

Coefficients: (1 not defined because of singularities)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.097e+06  1.648e+05  30.934  < 2e-16 ***
bedrooms      -2.796e+04  1.977e+03 -14.143  < 2e-16 ***
bathrooms0.5  -1.114e+05  1.224e+05  -0.910 0.362744
bathrooms0.75 -7.213e+03  7.020e+04  -0.103 0.918163
bathrooms1    -3.178e+04  6.603e+04  -0.481 0.630323
bathrooms1.25 -2.245e+04  9.535e+04  -0.235 0.813847
bathrooms1.5  -4.820e+04  6.620e+04  -0.728 0.466603
bathrooms1.75 -5.262e+04  6.611e+04  -0.796 0.426076
bathrooms2    -4.349e+04  6.617e+04  -0.657 0.511060
bathrooms2.25 -4.387e+04  6.620e+04  -0.663 0.507492
bathrooms2.5  -6.535e+04  6.613e+04  -0.988 0.323073
bathrooms2.75 -5.019e+04  6.639e+04  -0.756 0.449697
bathrooms3    -1.364e+04  6.656e+04  -0.205 0.837590
bathrooms3.25  6.663e+04  6.676e+04   0.998 0.318216
bathrooms3.5   3.384e+04  6.670e+04   0.507 0.611895
bathrooms3.75  1.738e+05  6.837e+04   2.542 0.011014 *
bathrooms4     1.654e+05  6.866e+04   2.410 0.015979 *
bathrooms4.25  2.778e+05  7.042e+04   3.945 8.00e-05 ***
bathrooms4.5   1.838e+05  6.964e+04   2.640 0.008306 **
bathrooms4.75  5.594e+05  7.939e+04   7.046 1.89e-12 ***
bathrooms5     4.485e+05  8.042e+04   5.577 2.48e-08 ***
bathrooms5.25  5.537e+05  8.791e+04   6.299 3.05e-10 ***
bathrooms5.5   9.021e+05  9.368e+04   9.630  < 2e-16 ***
bathrooms5.75  8.718e+05  1.235e+05   7.061 1.70e-12 ***
bathrooms6     1.285e+06  1.078e+05  11.920  < 2e-16 ***
bathrooms6.25  1.098e+06  1.611e+05   6.820 9.37e-12 ***
bathrooms6.5   2.452e+05  1.610e+05   1.523 0.127655
bathrooms6.75  4.395e+05  1.616e+05   2.720 0.006538 **
bathrooms7.5   1.8
bathrooms7.75  4.5
bathrooms8     2.1
sqft_living    1.4
sqft_lot       8.2
floors1.5      1.3
floors2        2.2
floors2.5      1.4
floors3        1.3
floors3.5      2.9
waterfront     5.7
view           3.649e+04  2.181e+03  16.731  < 2e-16 ***
condition      2.640e+04  2.409e+03  10.958  < 2e-16 ***
grade          1.206e+05  2.176e+03  55.442  < 2e-16 ***
sqft_above    -1.760e+01  4.560e+00  -3.859 0.000114 ***
sqft_basement         NA         NA      NA       NA
yr_built      -2.961e+03  7.636e+01 -38.781  < 2e-16 ***
yr_renovated   2.038e+01  3.765e+00   5.411 6.33e-08 ***
sqft_living15  5.170e+01  3.526e+00  14.663  < 2e-16 ***
sqft_lot15    -6.168e-01  7.506e-02  -8.218  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 206400 on 21566 degrees of freedom
Multiple R-squared:  0.6847,    Adjusted R-squared:  0.684
F-statistic:  1018 on 46 and 21566 DF,  p-value: < 2.2e-16
```

Warum nicht alle Variablen nehmen?

- Rechenintensiv
- Effekte verschiedener Variablen konkurrieren / überlappen
- Evtl. zu wenig Daten um die Effekte auseinander zu halten

- Der direkteste Ansatz wird als "All Subsets"- oder "Best Subsets"-Regression bezeichnet: Wir berechnen die kleinste quadratische Anpassung für alle möglichen Teilmengen und wählen dann zwischen ihnen basierend auf einem Kriterium, das den Trainingsfehler mit der Modellgröße ausgleicht.
- Allerdings können wir oft nicht alle möglichen Modelle untersuchen, da ihre Anzahl exponentiell mit der Anzahl der unabhängigen Variablen wächst
- Alternativen: Rückwärts- oder Vorwärtsauswahl

Offene Punkte

- Feature-Engineering
  - Zusammengesetzte Variablen
  - Geocoding
  - …
- Diagnostik
- Variablentransformationen
- Dichotome Zielvariablen
- …

Bisher alles retroperspektiv, Generalisierbarkeit unklar!

➔ Nächste Woche Übergang zur Prädiktion