Übung: Artificial grammar learning

Statistische Auswertung (in a nutshell)

Laura Becker

University of Freiburg

February 18, 2022

Unsere Fragestellungen

Unterscheidet sich die Anzahl von kurzen und langen Antworten wesentlich voneinander in:

1. Version `A` insgesamt:
   als Baseline, weicht die Verteilung von 50:50 (keine Präferenz) ab?

2. Version `A` vs. Version `B`:
   hat unsere Manipulation der Frequenzen auch zu einem Unterschied zwischen den zwei Versionen geführt?

3. Version `B` `freq` vs. `infreq`:
   die Hypothese ist, dass wir mehr kurze Formen in der `freq`-Kondition finden als in der `infreq`-Kondition
   (stärkere Präferenz für kurze Formen bei höherer Frequenz)

Unsere Datenstruktur

Zu welchem Datentyp gehören unsere Daten?

- numerisch
- ordinal
- kategorisch (binär)

☞ Unsere Daten sind **binär** (kategorisch), da jede Pluralform (Beobachtung) entweder `kurz` oder `lang` sein kann, d.h., wir unterscheiden **2 Kategorien**.

- um die Daten auszuwerten, zählen wir die einzelnen Antworten zusammen

☞ "Count data": Uns interessieren die **Anteile/Proportionen** von `langen` und `kurzen` Formen.

Möglicher statistischer Test

Chi-Quadrat-Test

- mit dem Chi-Quadrat-Test lässt sich einschätzen, wie sicher wir sein können, ob zwei (oder mehr) Kategorien gleichmäßig verteilt sind (in unserem Fall: N `lang` = N `kurz`)
- "The chi-square goodness of fit test is used to compare the observed distribution to an expected distribution, in a situation where we have two or more categories in a discrete data. In other words, it compares multiple observed proportions to expected probabilities. " `http://www.sthda.com/english/wiki/chi-square-goodness-of-fit-test-in-r`

☞ **goodness of fit test**

- mit dem χ-Quadrat Test lässt sich auch einschätzen, wie sicher wir sein können, ob mehrere Verteilungen, z.B. die Anteile von langen und kurzen Formen in Version A und B des Experiments sich von einander unterscheiden.
- "The chi-square test of independence is used to analyze the frequency table (i.e. contingency table) formed by two categorical variables. The chi-square test evaluates whether there is a significant association between the categories of the two variables. "
  `http://www.sthda.com/english/wiki/chi-square-test-of-independence-in-r`

☞ **test of independence**

Annahmen für den Chi-Quadrat-Test

- Unabhängigkeit der Daten: alle Zellen der Kontingenztabelle sind unabhängig voneinander

☞ Wir dürfen den Chi-Quadrat-Test eigentlich nicht anwenden, um zu prüfen, ob sich die Frequenzen von kurzen und langen Formen unterscheiden.

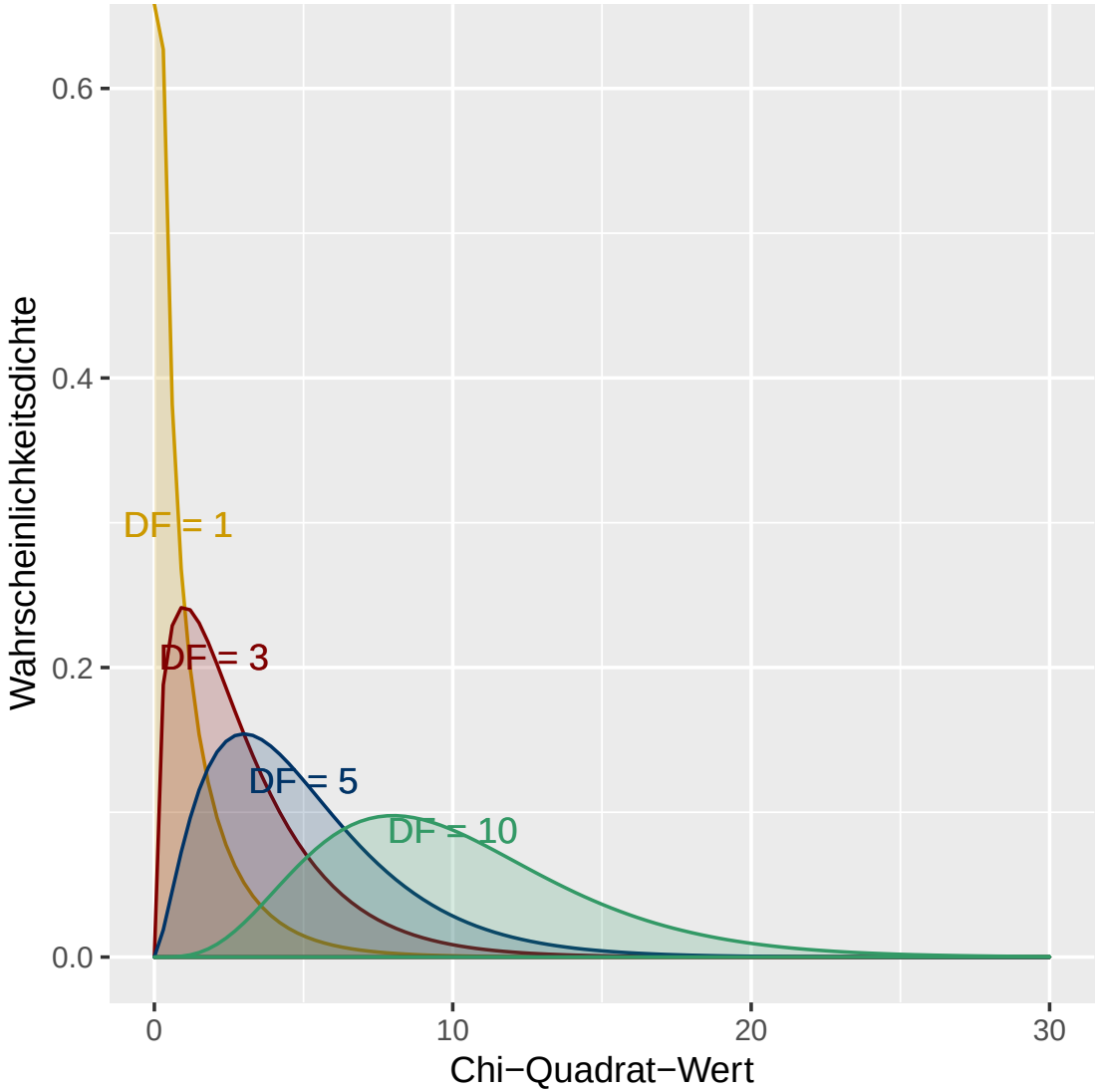
Unsere Daten als Kontingenztabelle:

	Version A	Version B	Σ_{Zeile}
kurz	132 (0.37)	161 (0.53)	293
lang	223 (0.63)	145 (0.47)	368
Σ_{Spalte}	355	306	$\Sigma_{tot} = 661$

- Wir gehen einen Chi-Quadrat-Test aber trotzdem einmal durch, um zu verstehen, wie ein solcher Test funktioniert.

Was ist ein Chi-Quadrat-Test?

- der Test heißt so, weil er auf der χ^2-Verteilung basiert



- er basiert auch auf "null hypothesis significance testing" (NHST)

Die Schritte für den Chi-Quadrat-Test

1. Aufstellen der Null-Hypothese

2. Ausrechnen des χ^2-Werts unserer beobachteten Verteilung

3. Abgleichen mit der Wahrscheinlichkeit, unter der Null-Hypothese unseren Wert zu sehen:
   - Wenn die Wahrscheinlichkeit unseres Werts über 0.05 liegt ($p(\chi^2) > 0.05$), dann können wir die Null-Hypothese nicht mit ausreichender Sicherheit verwerfen!!
   - wenn die Wahrscheinlichkeit unseres Werts unter 0.05 liegt ($p(\chi^2) < 0.05$), dann haben wir uns darauf geeinigt, dass die Wahrscheinlichkeit unseres Werts gering genug ist, um die Null-Hypothese zu verwerfen. Wir dürfen damit unsere alternative Hypothese annehmen und davon ausgehen, dass es einen statistisch signifikanten Unterschied gibt!

Ein Chi-Quadrat-Test

kurze und lange Formen in Version A und B

⚠ Der Test ist eigentlich nicht ganz zulässig für unsere Daten.

Null-Hypothese und alternative Hypothese

? Sind kurze und lange Antworten in Version A und gleich wahrscheinlich?

Unsere Daten als Kontingenztabelle:

	Version A	Version B	Σ_{Zeile}
kurz	132	161	293
lang	223	145	368
Σ_{Spalte}	355	306	$\Sigma_{tot} = 661$

- **Schritt 1:** Formulieren der **Null-Hypothese** H_0:
  Die Frequenz von kurzen und langen Antworten in Version A und B unterscheidet sich **nicht**.
  $N_{A.kurz/lang} = N_{B.kurz/lang}$
- Formulieren der **Alternativ-Hypothese** H_A:
  Die Frequenz von kurzen und langen Antworten in Version A und B **unterscheidet sich**.
  $N_{A.kurz/lang} \neq N_{B.kurz/lang}$

Berechnen der erwarteten Werte unter H_0

- **Schritt 3:**
  Wir wollen nun unsere erwarteten Werte unter Annahme von H_0 berechnen.
- Das machen wir für jede Zelle unserer Tabelle folgendermaßen:

$$erw_{zelle} = \frac{\Sigma_{zeile} \cdot \Sigma_{spalte}}{\Sigma_{tot}}$$

- Für unsere Verteilung bedeutet das, dass wir 4 erwartete Werte *erw* ausrechnen müssen.
  - $erw_{A:kurz}$
  - $erw_{A:lang}$
  - $erw_{B:kurz}$
  - $erw_{B:lang}$

Berechnen der erwarteten Werte

- Der erwartete Wert für kurze Formen in Version A

$$\begin{aligned} erw_{A:kurz} &= \frac{\sum_{Zeile:kurz} \cdot \sum_{Spalte:A}}{\sum_{tot}} \\ &= \frac{293 \cdot 355}{661} \\ &= 157.3601 \end{aligned}$$

- Der erwartete Wert für lange Formen in Version A

$$\begin{aligned} erw_{A:lang} &= \frac{\sum_{Zeile:lang} \cdot \sum_{Spalte:A}}{\sum_{tot}} \\ &= \frac{368 \cdot 355}{661} \\ &= 197.6399 \end{aligned}$$

Berechnen der erwarteten Werte

- Der erwartete Wert für kurze Formen in Version B

$$
\begin{aligned}
erw_{B:kurz} &= \frac{\sum_{Zeile:kurz} \cdot \sum_{Spalte:B}}{\sum_{tot}} \\
&= \frac{293 \cdot 306}{661} \\
&= 135.6399
\end{aligned}
$$

- Der erwartete Wert für lange Formen in Version B

$$
\begin{aligned}
erw_{B:lang} &= \frac{\sum_{Zeile:lang} \cdot \sum_{Spalte:B}}{\sum_{tot}} \\
&= \frac{368 \cdot 306}{661} \\
&= 170.3601
\end{aligned}
$$

Berechnen des χ^2-Werts

- Wir haben nun alle ewarteten Werte und können alles in die Formel für χ^2 einsetzen:

$$
\begin{aligned}
\chi^2 &= \sum_{1}^{n} \frac{(beobachtet_i - erwartet_i)^2}{erwartet_i} \\
&= \frac{(b_{A:kurz} - e_{A:kurz})^2}{e_{A:kurz}} + \frac{(b_{A:lang} - e_{A:lang})^2}{e_{A:lang}} + \frac{(b_{B:kurz} - e_{B:kurz})^2}{e_{B:kurz}} + \frac{(b_{B:lang} - e_{B:lang})^2}{e_{B:lang}} \\
&= \frac{(132 - 157.3601)^2}{157.3601} + \frac{(223 - 197.6399)^2}{197.6399} + \frac{(161 - 135.6399)^2}{135.6399} + \frac{(145 - 170.3601)^2}{170.3601} \\
&= 4.087025 + 3.254073 + 4.741486 + 3.775149 \\
&= 15.85773
\end{aligned}
$$

Vergleich unseres χ^2-Werts mit der χ^2-Verteilung

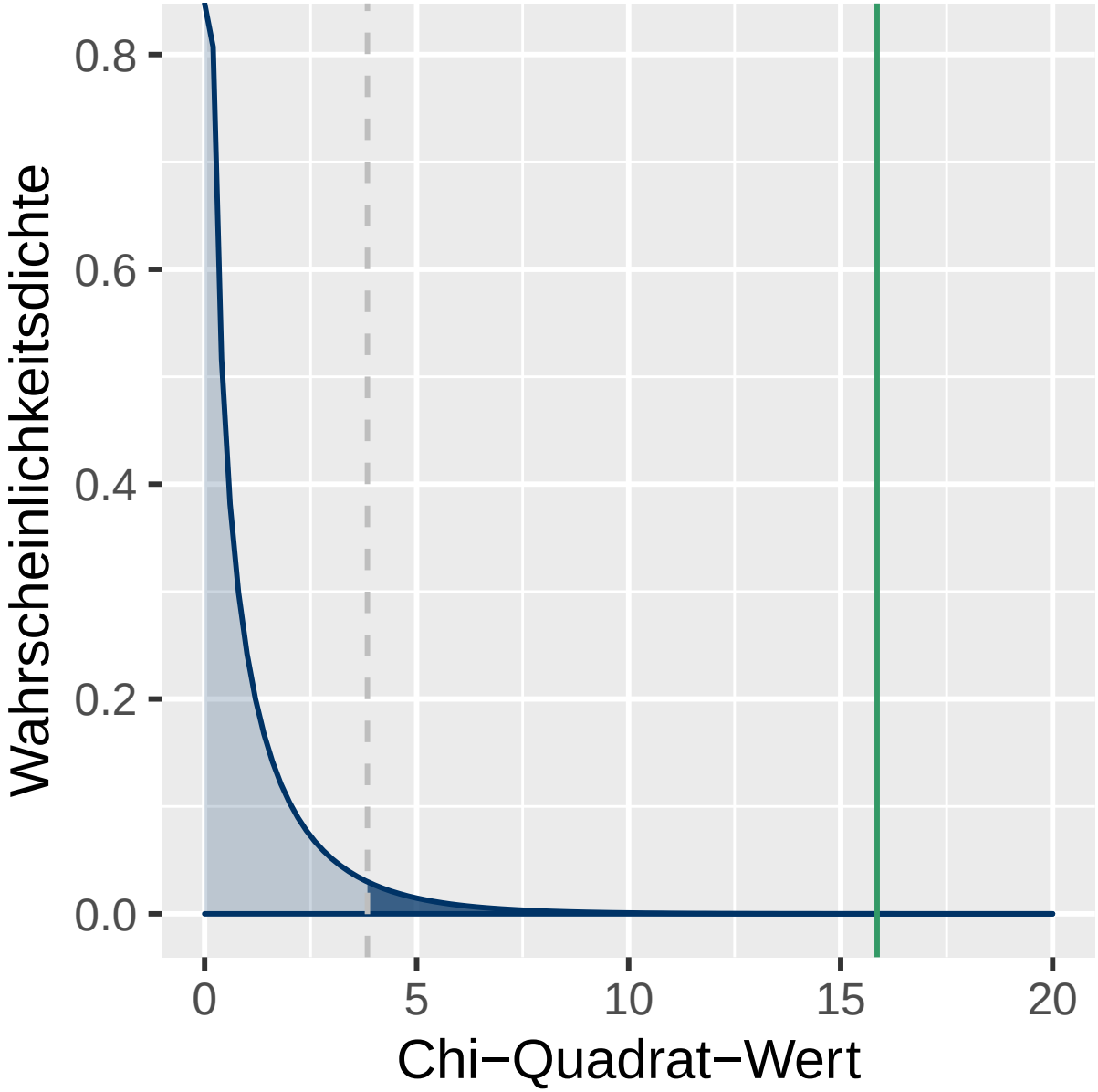
- Wir kennen jetzt den χ^2-Wert unserer Daten.
- Um einzuschätzen, wie wahrscheinlich unser Wert ist, brauchen wir die Freiheitsgrade unserer Verteilung, die wir folgendermaßen berechnen können:

$$\begin{aligned} df &= (N_{zeilen} - 1)(N_{spalten} - 1) \\ &= (2-1)(2-1) \\ &= 1 \end{aligned}$$

- Damit können wir nachsehen, wie wahrscheinlich ein Wert von 15.86 bei 1 Freiheitsgrad ist.

☞ p-Wert!

Die χ^2-Verteilung für DF=1



Kritische Werte der χ^2-Verteilung

A.4 Critical values of the chi-square distribution

	p	
df	**0.05**	**0.01**
1	3.84	6.63
2	5.99	9.21
3	7.81	11.34
4	9.49	13.28
5	11.07	15.09
6	12.59	16.81

	p	
df	**0.05**	**0.01**
25	37.65	44.31
26	38.89	45.64
27	40.11	46.96
28	41.34	48.28
29	42.56	49.59
30	43.77	50.89

- unser Wert von $\chi^2 = 15.86$ liegt für df=1 gut über dem kritischen Wert von 3.84 für p=0.05.

☞ Unser χ^2-Wert ist für unsere Art von Verteilung so hoch, dass es unter der Annahme von H_0 sehr, sehr unwahrscheinlich ist, unsere Daten bzw. Verteilung zu sehen.

☞ Wären unsere Datenpunkte alle unabhängig, dürfen wir die Null-Hypothese verwerfen und könnten annehmen, dass der Unterschied in der Anzahl an kurzen und langen Formen in Version A des Experiments signifikant wäre.

Chi-Quadrat-Test in R

- Wir bauen zunächst eine Kontingenztabelle aus unseren Werten:

```
data <- matrix(c(132, 161, 223, 145),
               nrow=2, byrow = TRUE,
               dimnames=list(c("short", "long"),c("a", "b")))
```

- Wir können das Objekt `data` aufrufen, um uns die Tabelle anzusehen:

```
> data
        a   b
short 132 161
long  223 145
```

- Wir können jetzt die Funktion `chisq.test()` mit unserer Tabelle `data` als Argument ausführen, um den Test durchzuführen:

```
chisq.test(data, correct = FALSE)
```

- R liefert uns dann folgendes Ergebnis:

```
> chisq.test(data, correct = FALSE)
        Pearson's Chi-squared test
data:  data
X-squared = 15.858, df = 1, p-value = 6.829e-05
```

McNemar's Test für paarweise abhängige Daten:

kurze und lange Formen in `freq` und `infreq` Konditionen von Version B

McNemar's Test

- Für paarweise abhängige Werte, wie z.B. wenn Partizipanten vor und nach einer Intervention getestet werden, gibt es eine spezielle Art von Chi-Quadrat-Test: den **McNemar's Test**.

- Wir können unsere Kondition von `freq` und `infreq` auch als paarweise abhängige Daten verstehen: jede Teilnehmer:in hat genau eine Form für die `infreq` und `freq` Variante jeder Pluralform produziert.

- Hierfür betrachten wir nur die Werte, die sich zwischen den beiden Bedingungen geändert haben:
  - freq-kurz & infreq-lang
    freq-lang & infreq-kurz

		`infreq`	
		kurz	lang
`freq`	kurz	57	**18**
	lang	**29**	46

Berechnen des χ^2-Werts

- Hier haben wir $29 + 18 = 47$ Beobachtungen mit nicht-kongruenten Formen (`lang-kurz`).
- Wären diese gleich verteilt, würden wir je $47/2 = 23.5$ Beobachtungen in beiden Zellen erwarten.
- Wir können jetzt wieder testen, wie wahrscheinlich unsere beobachtete Verteilung unter H_0 ist, indem wir zunächst unseren χ^2-Wert berechnen:

$$
\begin{aligned}
\chi^2 &= \sum_{1}^{n} \frac{(beobachtet_i - erwartet_i)^2}{erwartet_i} \\
&= \frac{(29 - 23.5)^2}{23.5} + \frac{(18 - 23.5)^2}{23.5} \\
&= 1.287234 + 1.287234 \\
&= 2.574468
\end{aligned}
$$

Kritische Werte der χ^2-Verteilung

A.4 Critical values of the chi-square distribution

	p	
df	0.05	0.01
1	3.84	6.63
2	5.99	9.21
3	7.81	11.34
4	9.49	13.28
5	11.07	15.09
6	12.59	16.81

	p	
df	0.05	0.01
25	37.65	44.31
26	38.89	45.64
27	40.11	46.96
28	41.34	48.28
29	42.56	49.59
30	43.77	50.89

- mit einem Freiheitsgrad von $df = 1$ liegt der unser χ^2-Wert von 2.6 unter dem Mindestwert, den wir benötigen, um unter der Null-Hypothese ausreichend sicher sein zu können, dass unsere Werte unwahrscheinlich sind.

☞ Unsere beobachteten Werte sind unter der Annahme der Null-Hypothese nicht sehr unwahrscheinlich.

☞ wir können daraus **nicht** schließen, dass sich die Werte zwischen der `freq` und `infreq` Kondition signifikant voneinander unterscheiden.

McNemar's test in R

- Wir bauen die Kontingenztabelle aus unseren Werten:

```
data2 <- matrix(c(57, 18, 29, 46),
                nrow=2, byrow = TRUE,
                dimnames=list(c("short", "long"),c("short", "long")))
```

- Wir können das Objekt `data2` aufrufen, um uns die Tabelle anzusehen:

```
> data2
      short long
short    57   18
long     29   46
```

- Wir können jetzt die Funktion `mcnemar.test()` mit unserer Tabelle `data` als Argument ausführen, um den Test durchzuführen:

```
mcnemar.test(data2, correct = FALSE)
```

- R liefert uns dann folgendes Ergebnis:

```
> mcnemar.test(data2, correct = FALSE)
        McNemar's Chi-squared test
data:  data2
McNemar's chi-squared = 2.5745, df = 1, p-value = 0.1086
```

McNemar's Test für paarweise abhängige Daten:

kurze und lange Formen in `freq` und `infreq` Konditionen von Version A

McNemar's test für Version A

- wir können auch für Version A testen, ob die Frequenz langer und kurzer Formen in Version A sich in der frequenten und infrequenten Bedingung unterscheiden

☞ eventuell können wir so auch Version A und B vergleichen

- unserer Original-Kontingenztabelle

Kondition	freq	infreq	Σ_{zeile}
kurz	75	86	161
lang	78	67	145
Σ_{spalte}	153	153	$\Sigma_{tot} = 306$

- die Kontingenztabelle, die für den McNemar's Test relevant ist:

		`infreq`	
		kurz	lang
`freq`	kurz	32	**29**
	lang	**38**	75

Berechnen des χ^2-Werts

- Hier haben wir $38 + 29 = 67$ Beobachtungen mit nicht-kongruenten Formen (`lang-kurz`).
- Wären diese gleich verteilt, würden wir je $47/2 = 33.5$ Beobachtungen in beiden Zellen erwarten.
- Wir können jetzt wieder testen, wie wahrscheinlich unsere beobachtete Verteilung unter H_0 ist, indem wir zunächst unseren χ^2-Wert berechnen:

$$
\begin{aligned}
\chi^2 &= \sum_{1}^{n} \frac{(beobachtet_i - erwartet_i)^2}{erwartet_i} \\
&= \frac{(38 - 33.5)^2}{33.5} + \frac{(29 - 33.5)^2}{33.5} \\
&= 0.6044776 + 0.6044776 \\
&= 1.208955
\end{aligned}
$$

Kritische Werte der χ^2-Verteilung

A.4 Critical values of the chi-square distribution

	p	
df	0.05	0.01
1	3.84	6.63
2	5.99	9.21
3	7.81	11.34
4	9.49	13.28
5	11.07	15.09
6	12.59	16.81

	p	
df	0.05	0.01
25	37.65	44.31
26	38.89	45.64
27	40.11	46.96
28	41.34	48.28
29	42.56	49.59
30	43.77	50.89

- Mit einem Freiheitsgrad von $df = 1$ liegt der unser χ^2-Wert von 1.2 unter dem Mindestwert, den wir benötigen, um unter der Null-Hypothese ausreichend sicher sein zu können, dass unsere Werte unwahrscheinlich sind.

☞ Unsere beobachteten Werte sind unter der Annahme der Null-Hypothese nicht sehr unwahrscheinlich.

☞ Wir können daraus **nicht** schließen, dass sich die Werte zwischen der `freq` und `infreq` Kondition signifikant voneinander unterscheiden.
(Das haben wir aber auch schon vermutet!)

McNemar's test in R

- Wir bauen die Kontingenztabelle aus unseren Werten:

```
data <- matrix(c(32, 29, 38, 75),
               nrow=2, byrow = TRUE,
               dimnames=list(c("short", "long"),c("short", "long")))
```

- Wir können das Objekt `data` aufrufen, um uns die Tabelle anzusehen:

```
> data
      short long
short    32   29
long     38   75
```

- Wir können jetzt die Funktion `mcnemar.test()` mit unserer Tabelle `data` als Argument ausführen, um den Test durchzuführen:

```
mcnemar.test(data, correct = FALSE)
```

- R liefert uns dann folgendes Ergebnis:

```
> mcnemar.test(data, correct = FALSE)
        McNemar's Chi-squared test
data:  data
McNemar's chi-squared = 1.209, df = 1, p-value = 0.2715
```

Die adäquatere Lösung, um die Verteilung kurzer und langer Formen in Version A und B zu vergleichen:

logistische Regression

1. Wozu Regression
2. Lineare Regression: eine Beispiel mit Reaktionszeiten unseres Experiments
3. (Mixed) logistische Regression
4. Zurück zum Vergleich kurzer und langer Formen in Version A und B!

Wozu Regression?

- Wir können den Zusammenhang zwischen einer **abhängigen** Variable (= **predicted**, **outcome**) und einer (oder mehrerer) **unabhängigen** Variable (= **predictor**, **input**) modellieren.
- Die Notation dafür ist:
  outcome $\sim$ input
  predicted $\sim$ predictor
  abhängige Variable $\sim$ unabhängige Variable
- Regressionsmodelle haben den Vorteil, dass sie Abhängigkeiten der Datenpunkte mit einbeziehen können:
  - wir können den Zusammenhang aller Antworten von individuellen Sprecher abbilden
  - wir können den Zusammenhang einzelner Pluralformen abbilden (falls z.B. ein Nomen besonders schwierig war und einen eigenen Effekt hatte)

☞ solche "Kontrollen" voneinander abhängiger Datenpunkte werden häufig **random effects** genannt

- im Gegensatz dazu werden die eigentlichen Prädiktoren häufig **fixed** effects genannt
- Regressionsmodelle, die beide Arten von Effekten miteinbeziehen, werden häufig **mixed effect models** genannt

Arten von Regression

Lineare vs. logistische Regression

- die einfachste Art von Regression ist **lineare Regression**

☞ dafür muss die abhängige Variable numerisch sein, wie z.B. bei Reaktionszeiten.

⚠ Count Daten sind keine numerischen Daten, sondern aggregierte kategorische Daten!

- für eine kategorische abhängigen Variable brauchen wir **logistische Regression**

☞ wie "Standard" lineare Diskussion, nur wird noch ein zusätzlicher mathematischer Schritt benötigt, damit wir unsere kategorischen Daten in einem linearen Zusammenhang abbilden können.

- Logistische Regression ist ein Typ der Familie von **generalisierten** Regressionsmodellen, die alle im Prinzip wie lineare Regression funktionieren, aber noch einen Extra-Schritt benötigen.

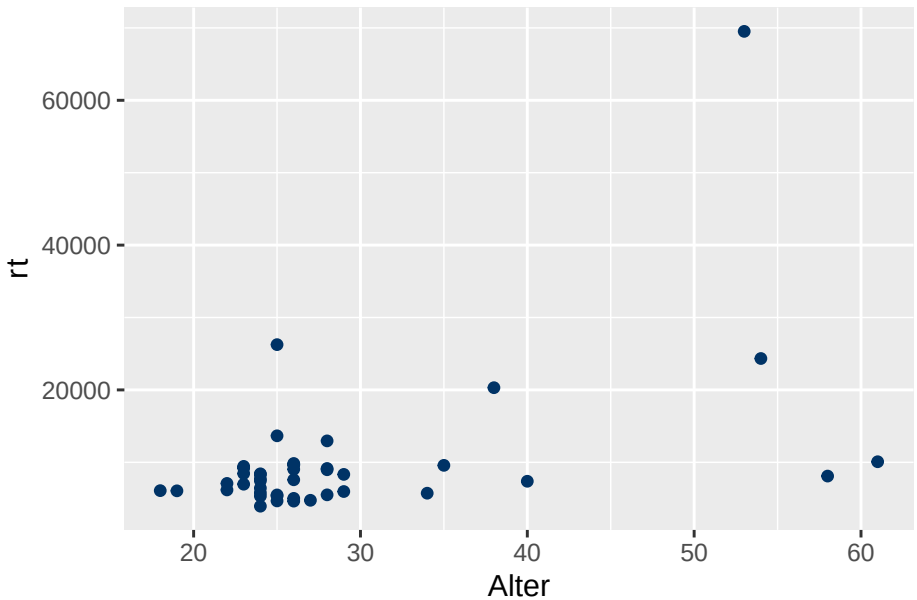
Mixed effect Regression

- für sowohl lineare als auch logistische Regression können wir **random effects** miteinbeziehen

☞ dann sprechen wir von linear/logistic mixed effect regression models

Lineare Regression: ein Beispiel

- Wir wollen untersuchen, ob es einen Zusammenhang zwischen Alter unserer Partizipanten und ihrer durchschnittlichen Reaktionszeit (RT) im Pluraltest gibt.
- Wir nehmen an, dass die RT älterer Partizipanten generell höher ist als die von jüngeren Partizipanten.
- Wir können diese Assoziation folgendermaßen visualisieren:



- Es scheint eine Tendenz in die erwartete Richtung zu geben!

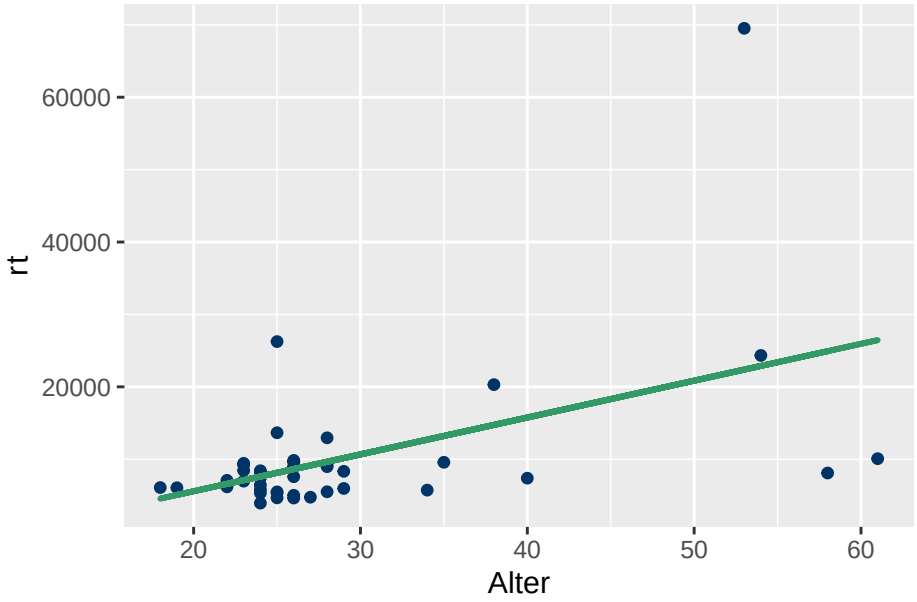
  $rt \sim age$

Lineare Regression als Modell

- Wir wollen ein Modell bauen, dass die durchschnittliche RT_i für jedes denkbare $Alter_i$ vorhersagt
- Das Modell soll außerdem einschätzen können, wie sicher wir bei der vorhergesagten RT sind.

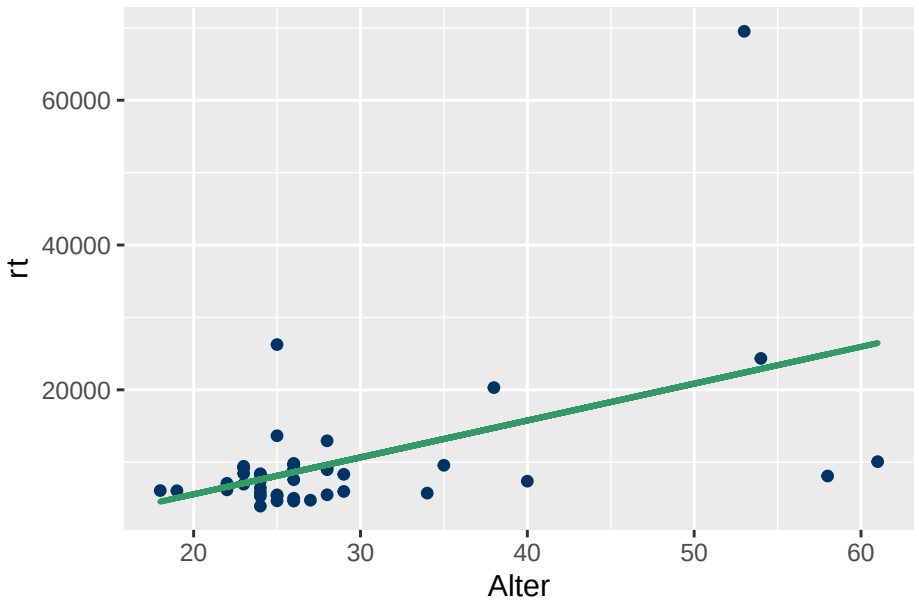
$$RT_i = Modell(Alter_i) + Fehler_i$$

- eine Möglichkeit, so ein Modell zu bauen, ist, eine Gerade durch die Verteilung zu legen
- Dabei soll die Gerade möglichst so liegen, dass der Abstand aller beobachteten Punkte zur Geraden insgesamt so klein wie möglich ist.



Eigenschaften der Regressionsgeraden

- Wie können wir jetzt unsere Regressionsgerade benutzen, um z.B. vorherzusagen, welche RT eine Partizipantin im Alter von 45 Jahren hat?
- Wir müssen dafür die relevanten Eigenschaften unserer Gerade kennen!

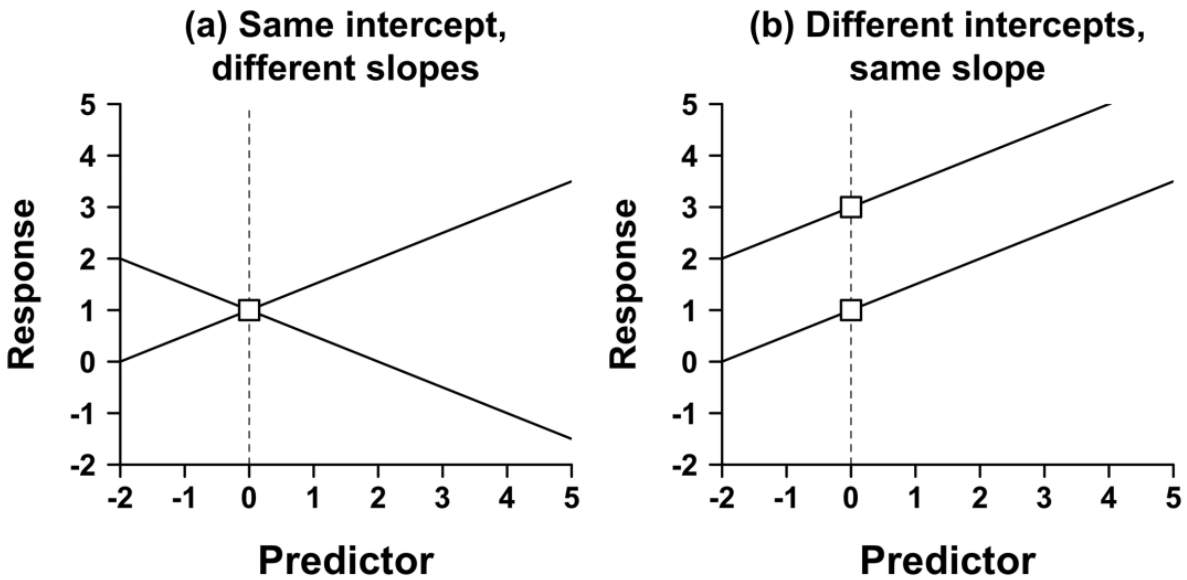


Eigenschaften der Regressionsgeraden

- unsere Werte für x (`Alter`)
- unsere Werte für y (`rt`)
- die **Steigung** β der Geraden:

$$\beta = \frac{x2 - x1}{y2 - y1}$$

- die vertikale Verschiebung β_0, also wo die Gerade die 0 auf der x-Achse schneidet (**intercept**)



Lineare Regression als Modell

- Damit können wir unsere Regressionsgerade mathematisch vollständig beschreiben.
- Allerdings wollen wir die Gerade so beschreiben, dass wir y (unsere abhängige Variable) von den restlichen Werten vorhersagen.
- Wir können das mit einer einfachen linearen Gleichung tun:

$$y = \beta_0 + \beta \cdot x$$

- für unser Beispiel ist $\beta_0 = -4602.2$ und $\beta = 509.1$
- Wenn wir jetzt mit unserem Regressionsmodell vorhersagen wollen, welche RT eine Partizipantin von 45 Jahren hat, können wir die Werte einfach in die Gleichung einsetzen:

$$\begin{aligned} RT &= \beta_0 + \beta \cdot Alter \\ &= -4602.2 + 509.1 \cdot 45 \\ &= 18307.3 \end{aligned}$$

☞ Mithilfe unserer beobachteten Daten können wir also ein Modell bauen und vorhersagen, dass eine neue Partizipantin von 45 eine vermutlich eine durchschnittliche RT in der Nähe von 18307 ms hat.

Model fit

❓ Wie gut ist unser Modell insgesamt?

- Wir haben jetzt den Modellteil von der ursprünglichen Gleichung berechnet.

$$RT_i = Modell(Alter_i) + Fehler_i$$

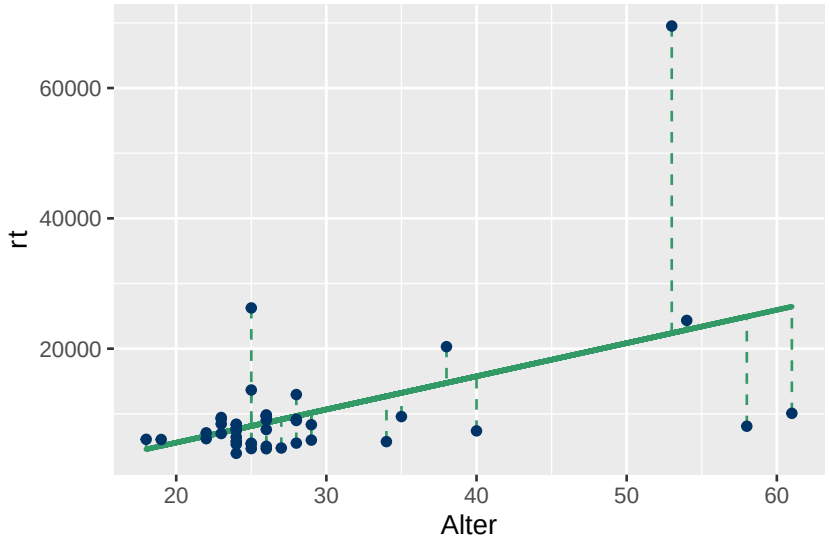
❓ Und was ist mit dem Fehlerterm?

- Wir können den Fehlerterm als Maß verstehen, wie stark die Regressionsgerade von unseren tatsächlichen Werten abweicht.

- Wir können den Fehlerterm auch als Maß dafür verstehen, wie gut unser Modell insgesamt unsere Daten repräsentiert.

Model fit: Residuals und Sum of Squares

- Wir haben die Regressionsgerade so gewählt, dass insgesamt der Abstand zwischen unseren Beobachtungen und den Punkten auf der Geraden minimal ist.



- damit negative und positive Residuals sich nicht gegenseitig auslöschen, können wir alle Residuals quadrieren und dann addieren

☞ **Residual Sum of Squares**

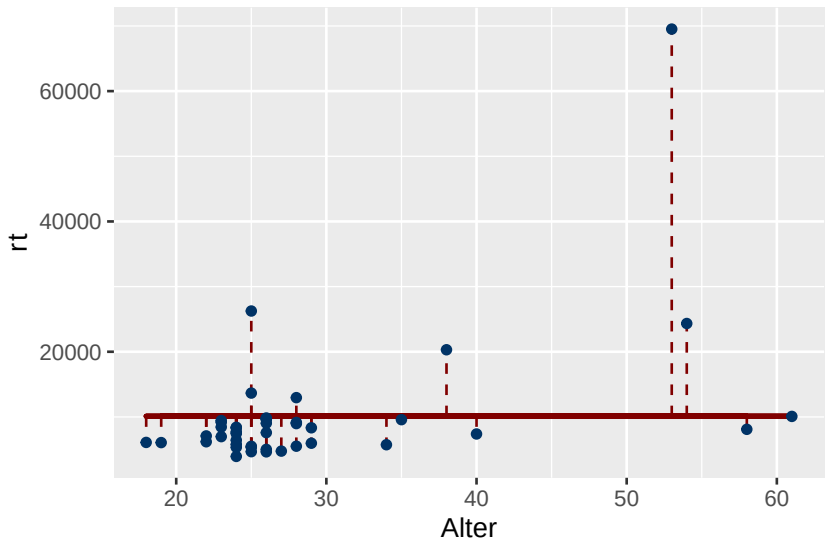
$$SS_{res} = res_1^2 + res_2^2 + res_3^2 + \ldots + res_i^2$$

- für unsere Daten und unser Modell heißt das:

$$\begin{aligned} SS_{res} &= -503^2 + 3306^2 + 1341^2 + \ldots + (-3460)^2 \\ &= 3467285131 \end{aligned}$$

Model fit

- Wir kennen nun den Sum of Squares-Wert unserer Residuals $SS_{res} = 3467285131$.
- Um einzuschätzen, ob dieser Wert hoch oder niedrig ist, können wir ihn mit dem Wert für ein Modell unter der Null-Hypothese vergleichen.
- Bei einem Modell unter der Null-Hypothese gäbe es keine Assoziation zwischen `Alter` und `rt`.



- Unter dieser Annahme wäre das beste Modell, für alle Werte von `Alter` den Durchschnittswert der beobachteten RTs anzunehmen.
- Der Durchschnittswert liegt bei 10103.2.
- Auch hier können wir die Residual Sum of Squares für unser Null-Modell SS_0 ausrechnen.

Model fit: Sum of Squares SS_0

- Die Gesamt-Abweichung unserer beobachteten Werte von unserem Null-Modell (residual Sum of Squares SS_0) ist:

$$\begin{aligned} SS_0 &= res_1^2 + res_2^2 + res_3^2 + \ldots + res_i^2 \\ &= -953^2 + 2856^2 + (-1655)^2 + \ldots + (-2658)^2 \\ &= 4541029957 \end{aligned}$$

Model fit: R-squared

- Wir haben die Residual Sum of Squares für unsere Regressionsgerade SS_{res} und für ein Null-Modell SS_0 berechnet.

☞ Wir können damit vergleichen, wieviel besser unser Modell die beobachteten Daten repräsentiert als das Null-Modell.

- Das können wir folgendermaßen tun:
  - wir bilden die Differenz von SS_0 und SS_{res}
  - zum Normalisieren des Werts (damit er zwischen 0 und 1 liegt), teilen wir diese Differenz noch durch SS_0

☞ **R-squared-Wert**

$$
\begin{aligned}
R^2 &= \frac{SS_0 - SS_{mod}}{SS_0} \\
&= \frac{4541029957 - 3467285131}{4541029957} \\
&= \frac{1073744826}{4541029957} \\
&= 0.236454
\end{aligned}
$$

Model fit: R-squared

- Unser R-squared-Wert $R^2 = 0.236454$ gibt an, welchen Teil der Variation in unseren beobachteten RTs wir mit unserem Regressionsmodell erfassen.

☞ unser Modell erfasst etwa 24% der Variation in den RTs, was absolut gesehen nicht besonders viel ist, aber doch beachtlich ist, wenn wir bedenken, dass das Regressionsmodell nur das Alter der Partizipanten und sonst nichts "kennt".

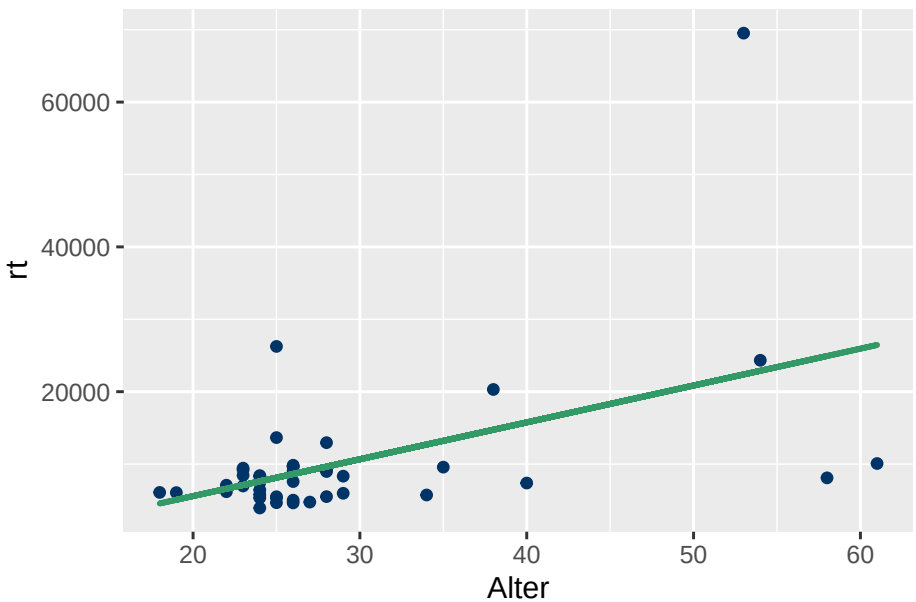
? Es bleibt aber noch die Frage, wie sicher wir sein können, dass es eine Assoziation zwischen Alter und RTs gibt!

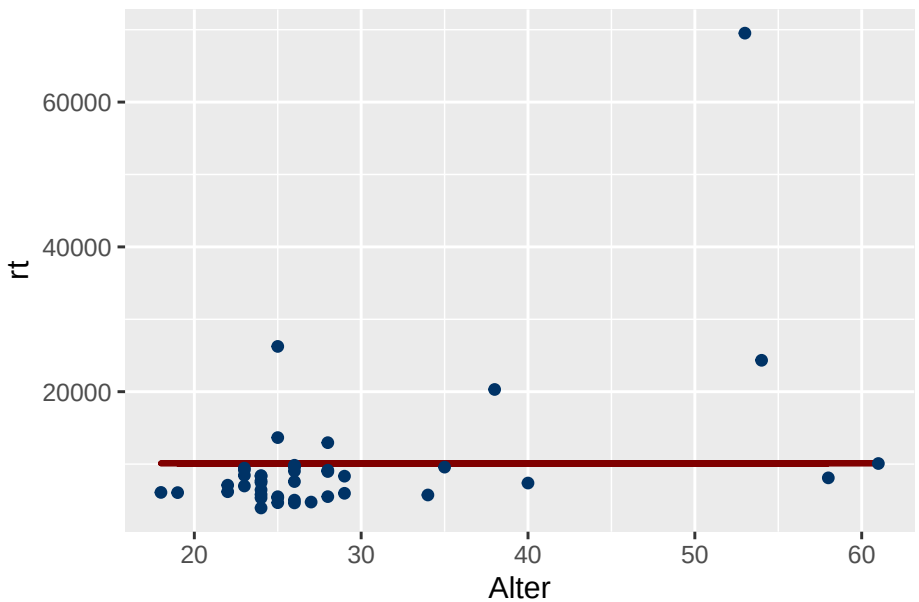
☞ Wie wahrscheinlich sind unsere RTs unter der Annahme, dass es keinen Zusammenhang gibt (H_0)?

Das können wir mit einem t-test für unseren Koeffizienten β (Steigung der Gerade) ermitteln!

Signifikanz unseres Effekts mit einem t-Test

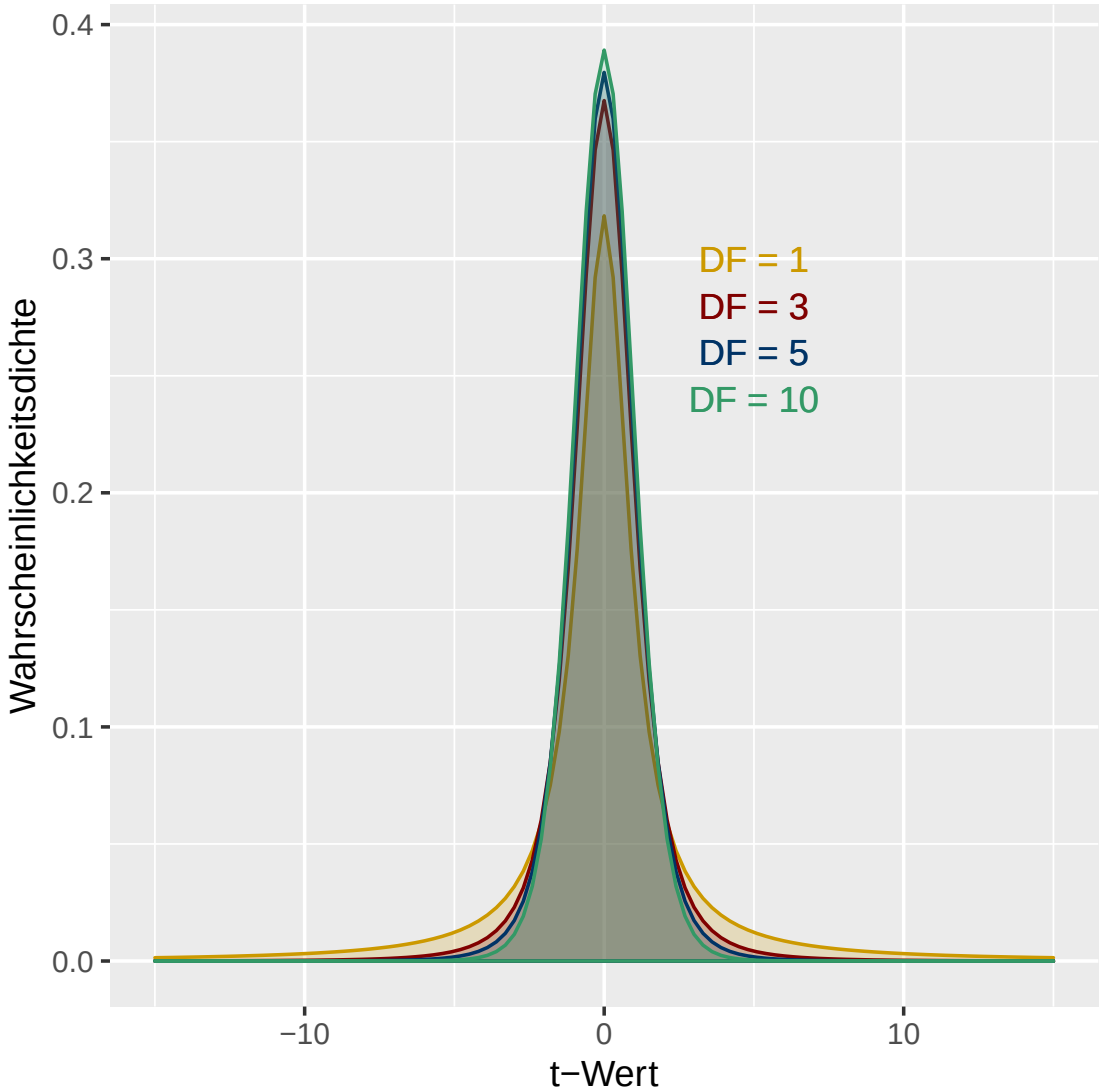
- Wir können mit einem t-Test ausrechnen, wie wahrscheinlich unser erwarteter Koeffizient β (links) unter der Null-Hypothese ist.
- die Null-Hypothese ist in diesem Fall, dass keine Assoziation zwischen `Alter` und `rt` besteht, und dass $\beta = 0$ ist (rechts).





t-Test

- ein t-Test basiert auf der Testgröße t, die wie χ^2 eine bestimmte Wahrscheinlichkeitsverteilung für jeden Freiheitsgrad hat.



t-Test

- ein t-Wert, wie auch der Chi-Quadrat-Wert, quantifiziert folgendes Verhältnis

$$Testgroesse\,(z.B.\ t, \chi^2, F...) = \frac{Varianz\ erklaert}{Varianz\ nicht\ erklaert} = \frac{Effekt}{Fehler}$$

- Effekt: hier Differenz zwischen $\beta = 509.1$ unseres Modells und $\beta = 0$ des Null-Modells
- Fehler:
  - Angenommen, wir wiederholen unser Experiment mehrere Male:
  - Wie stark weichen die β-Werte von vergleichbaren Daten von unserem Wert ab?

  ☞ Wir müssen unser Experiment nicht wiederholen, sondern können diese Varianz mathematisch approximieren:
  Standard Error (SE)
  - Wenn SE klein ist, dann können wir davon ausgehen, dass wir ganz ähnliche β-Werte finden würden, wenn wir unser Experiment wiederholen würden. Unser Wert ist dann repräsentativ für die Population.
  - Wenn SE groß ist, dann können wir nicht so sicher sein, dass unser β-Wert repräsentativ und damit robust ist und uns viel über den Zusammenhang von `Alter` und `rt` im Allgemeinen aussagt.

t-Test

- wir können unseren t-Wert folgendermaßen berechnen:

$$t = \frac{\beta_{Modell} - \beta_{Null-Modell}}{SE_{\beta_{Modell}}}$$

$$t = \frac{\beta_{Modell}}{SE_{\beta_{Modell}}}$$

- uns fehlt also noch der Standard Error *SE*:

$$SE = \frac{variability\ of\ the\ data}{sample\ size\ and\ sample\ distribution}$$

Standard Error

- Der Standard Error SE ist ein approximiertes Maß dafür, wie stark unsere Reaktionszeiten in verschiedenen Wiederholungen des Experiments variieren würden.
- Die Gleichung für den SE ist:

$$SE = \frac{\sqrt{\frac{SS_{res}}{N-2}}}{\sqrt{SS_x}}$$

- Um den SE unseres Regressionskoeffizienten β berechnen zu können, brauchen wir:
  - Residual Sum of Squares $SS_{res} = 3467285131$
  - Die quadrierte Summe der Differenz unserer x-Werte (`Alter`) und ihres Durchschnitts (SS_x)

  ☞ müssen wir noch berechnen

  - unsere Anzahl an Beobachtungen $N = 43$

Sum of Squares der x-Werte SS_x

- Wir müssen noch ausrechnen, wie groß die Sum of Squares der Abweichungen unserer x-Werte von ihrem Durchschnittswert ist.
- Der Durchschnittswert für `Alter` unserer Studie beträgt $\bar{x} = 28.88372$ Jahre.
- Wir können die Sum of Squares der x-Werte folgendermaßen berechnen:

$$\begin{aligned} SS_x &= (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \ldots + (x_{43} - \bar{x})^2 \\ &= 0.78096268 + 34.61817198 + 15.08328826 + \ldots + 23.85073012 \\ &= 4142.419 \end{aligned}$$

Standard Error

- Jetzt können wir alles einsetzen:
  - $SS_{res} = 3467285131$
  - $N = 43$
  - $SS_x = 4142.419$

$$SE = \frac{\sqrt{\frac{SS_{res}}{N-2}}}{\sqrt{SS_x}} = \frac{\sqrt{\frac{3467285131}{43-2}}}{\sqrt{4142.419}} = \frac{\sqrt{84567930}}{\sqrt{4142.419}} = \frac{9196.082}{64.36163} = 142.8814$$

☞ Unser Standard Error beträgt 142.8814 ms.

Zurück zum t-Wert und t-Test

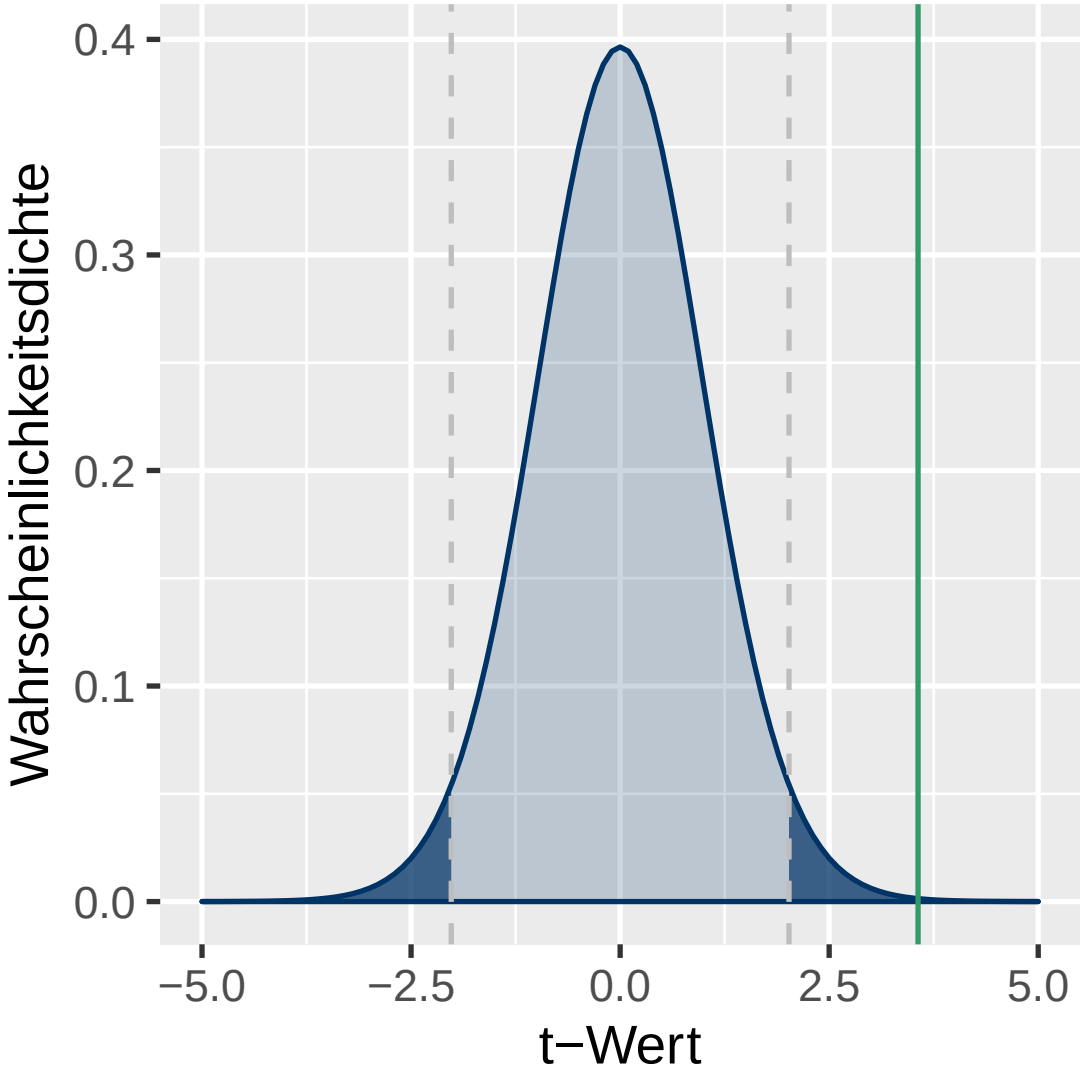
- Wir haben jetzt alle relevanten Größen für die Berechnung des t-Werts und können sie in die Formel einsetzen:

$$\begin{aligned} t &= \frac{\beta_{modell}}{SE} \\ &= \frac{509.1}{142.8814} \\ &= 3.563095 \end{aligned}$$

- Wir müssen diesen nur noch auf der Wahrscheinlichkeitsverteilung der t-Werte für unsere Freiheitsgrade einordnen:
- Wenn unser t-Wert weniger wahrscheinlich als 5% ist ($p<0.05$), dann ist unser β-Koeffizient signifikant: wir haben eine signifikante Assoziation zwischen `Alter` und `rt`
- die Freiheitsgrade können wir in diesem Fall wie folgt berechnen (wie auch schon in der Formel für *SE*):

$$\begin{aligned} DF &= N_{Beobachtungen} - N_{Praediktoren} - 1 \\ &43 - 1 - 1 \\ &41 \end{aligned}$$

Die t-Verteilung für DF=41



- grau-gestrichelte Linie: die Grenz-t-Werte, ab denen $p<0.05$ ist
- grüne Linie: unser t-Wert von 3.563095

Zurück zum t-Test für unseren β-Koeffizienten

- Wir haben jetzt unseren t-Wert von 3.563095.
- Wir können wieder den kritischen t-Wert für p=0.05 suchen:
- $t \geq 2.019541$ für $p = 0.05$ bei 41 Freiheitsgraden
- Mit unserem t-Wert von rund 3.6 ist der Effekt damit **statistisch signifikant**, d.h. wir können sicher genug sein, dass unsere Daten unter der Null-Hypothese sehr unwahrscheinlich sind.
- Damit können wir also annehmen, dass es eine Assoziation zwischen dem Alter der Partizipanten und ihren Reaktionszeiten gibt.

Lineare Regression in R

- Viel einfacher können wir eine Regression mit statistischer Software durchführen, die uns u.a. die Koeffizienten, ihre Signifikanz, und den R-Squared-Wert gibt.
- Um Regression in R durchzuführen, brauchen wir ein Dataframe (`df`) mit den Prädiktoren und der vorherzusagenden Variable als Spalten:

```
  Alter        rt
1    28  9149.869
2    28 12959.656
3    23  8448.188
4    25  4666.062
5    24  6437.938
6    29  8332.650
...
```

- Das Regressionsmodell können wir dann mit der `lm()` Funktion bauen:

```
m_rt <- lm(rt ~ Alter, data = df_rt)
```

Lineare Regression in R

- Wir können uns dann mit der `summary()` Funktion einen Überblick unseres Regressionsmodells anzeigen lassen:

```
> summary(m_rt)

Residuals:
   Min     1Q Median     3Q    Max
-16829  -3536   -411   1123  47139

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4602.2     4358.7  -1.056 0.297214
age            509.1      142.9   3.563 0.000946 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9196 on 41 degrees of freedom
Multiple R-squared:  0.2365,    Adjusted R-squared:  0.2178
F-statistic:  12.7 on 1 and 41 DF,  p-value: 0.0009462
```

- Wir sehen eine Zusammenfassung über die Verteilung unserer Residuals.
- Wir bekommen Details zu den Koeffizienten β_0 (Intercept) und β (Steigung): Für beide wird uns der Standard Error gegeben, sowie der daraus berechnete t-Wert und der dazugehörige p-Wert mit seiner Signifikanz (für unsere 41 Freiheitsgrade).
- Wir bekommen dann noch den R-squared-Wert und den F-Wert, der ebenso quantifiziert, wie gut das Modell insgesamt unsere Daten erfasst.

❷ Vergleich von Version A und B: logistische Regression

Logistische Regression

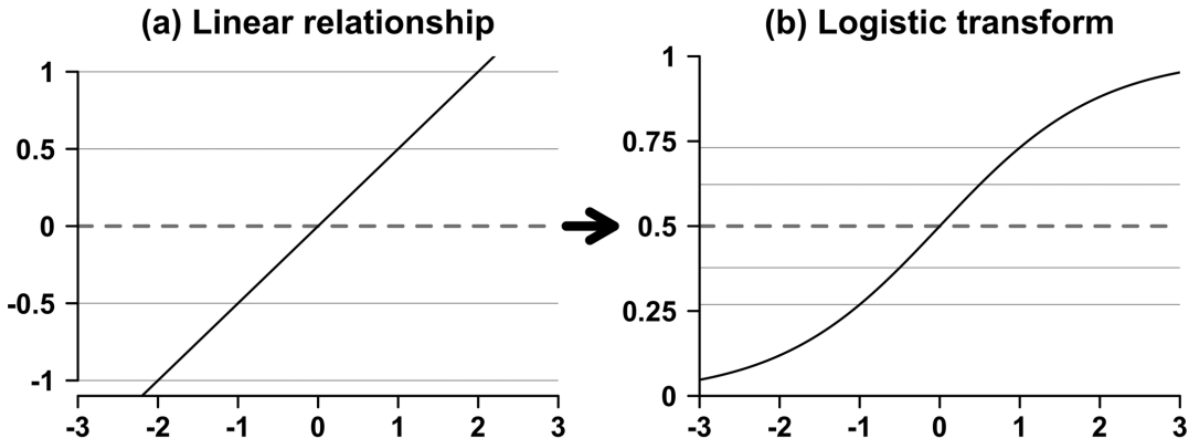
- Wir können jetzt auf unsere `langen` und `kurzen` Formen zurückkommen:
- Wir wollen eigentlich keinen numerischen Wert für unser y berechnen, sondern z.B. die Wahrscheinlichkeit p von `kurzen` Formen in Abhängigkeit der Version (`A` vs. `B`).
- Wahrscheinlichkeit sind Werte zwischen 0 und 1.
- Das heißt, wir wollen unsere Gleichung $y = \beta_0 + \beta \cdot x$ so verändern, dass y keine Werte zwischen $-\infty$ und $+\infty$ einnehmen kann, sondern nur zwischen 0 und 1.
- Das können wir mit der logistischen Funktion machen, die wir auf unsere ursprüngliche Gleichung legen.
- Unsere Gleichung sieht nun so aus:

$$y = logistic(\beta_0 + \beta \cdot x)$$

Effekt der logistischen Funktion

- Die logistische Funktion wandelt eine lineare Funktion um, sodass deren y-Werte immer zwischen [0,1] bleiben.

☞ So könnten wir gut Wahrscheinlichkeiten abbilden, die sich genau in diesem Intervall bewegen.



⚠ diese Art von Verteilung ist vor allem für große und kleine Werte nahe an den Randbereichen visuell schwer zu interpretieren

☞ statt Wahrscheinlichkeiten vorherzusagen, gehen wir noch zwei weitere Transformationsschritte (odds und dann log odds), um am Ende wieder eine “lesbare” Gerade zu bekommen

Odds (Chancen)

Wahrscheinlichkeiten drücken wir im Alltag häufig auch als **Odds** (Chancen) aus.

- Angenommen, die Wahrscheinlichkeit für Partizipanten, eine kurze Form zu produzieren, ist 0.5 oder 50%.
- Wir könnten dann auch sagen:

  Die **Chancen stehen 1:1** für Partizipanten, eine kurze Form zu produzieren.
- Wir können die Odds aus den Wahrscheinlichkeiten mit folgender Formel berechnen:

$$odds = \frac{p}{1-p} = \frac{0.5}{1-0.5} = \frac{0.5}{0.5} = 1$$

  (die Wahrscheinlichkeit, dass Ereignis stattfindet, geteilt durch die Wahrscheinlichkeit, dass das Ereignis nicht stattfindet)

Log Odds

- Odds liegen im Intervall $-\infty$ bis $+\infty$, aber nicht auf einer Geraden!
- Wir können aber wieder eine andere Funktion über die Odds legen, damit sie auf einer Geraden liegen:

☞ **logarithmische** Funktion: **log odds**

$$log(odds) = log(\frac{p}{1-p})$$

Wahrscheinlichkeit	Odds	Log odds
0.1	0.11 zu 1	-2.20
0.2	0.25 zu 1	-1.39
0.3	0.43 zu 1	-0.85
0.4	0.67 zu 1	-0.41
0.5	1 zu 1	0.00
0.6	1.5 zu 1	0.41
0.7	2.33 zu 1	0.85
0.8	4 zu 1	1.39
0.9	9 zu 1	2.20

Logistische Regression: Zurück zu unseren Daten

- Wenn wir der Einfachheit halber nur produzierte Formen betrachten, die entweder `kurz` oder `lang` sind, haben wir folgende Verteilung:

	A	A freq	A infreq	B	B freq	B infreq
N kurz	132	61	71	161	75	86
prop kurz	**0.37**	0.34	0.4	**0.53**	0.49	0.56

- Bevor wir zum Modell kommen, können wir uns die Proportionen ansehen:
  - Es ist in beiden Versionen nicht sehr wahrscheinlich, dass wir einen signifikanten Unterschied zwischen der `freq` und `infreq` Kondition sehen werden.
  - Eventuell gibt es aber einen leichten Unterschied zwischen Version A und B in der Proportion von kurzen Formen.
  - ☞ Der Einfachheit halber werden wir nur auf einen **Unterschied zwischen Version A und B** testen.

Logistische Regression in R

- Damit wir ein Regressionsmodell bilden können, brauchen wir wieder ein Dataframe mit den relevanten Variablen als Spalten:
- Wir können die `pl_produktion` so kodieren, dass wir eine `1` für kurze Formen und eine `0` für lange Formen vergeben.

```
  Version pl_produktion
1       A             1
2       A             1
3       A             0
4       B             0
5       B             1
6       B             0
...
```

- In R können wir ein logistisches Regressionsmodell mit der `glm()` Funktion ("generalized linear model") bauen.

```
m_pl <- glm(pl_produktion ~ Version, data = df, family = 'binomial')
```

⚠ (Das ist noch nicht unser finales Modell, weil wir auch hier die Annahme von unabhängigen Datenpunkten verletzen würden!)

Logistische Regression in R

- Wir können uns die Eigenschaften des Modells wieder mit der `summary()` Funktion anzeigen lassen.

```
> summary(m_pl)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2222  -0.9643  -0.9643   1.1333   1.4066

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.5244     0.1098  -4.775 1.80e-06 ***
VersionB      0.6290     0.1586   3.965 7.34e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 907.81  on 660  degrees of freedom
Residual deviance: 891.91  on 659  degrees of freedom
AIC: 895.91

Number of Fisher Scoring iterations: 4
```

- Wir bekommen ähnliche Informationen über das Modell wie bei linearer Regression.
- Bei logistischer Regression wird allerdings die Signifikanz der Koeffizienten über eine andere Testgröße ermittelt (z-Wert statt t-Wert).

Estimates der Koeffizienten

- Unsere Estimates für die Koeffizienten sind
  - -0.5244 für β_0 oder das Intercept (=Version A)
  - 0.6290 für die β oder die Steigung von Version A zu Version B

? Was heißt das??

☞ Beide Estimates sind in log odds.

- Wenn wir den Wert für das Intercept in eine Wahrscheinlichkeit umwandeln, erhalten wir die Wahrscheinlichkeit, die das Modell für kurze Antworten in Version A vorhersagt.
- Wir können die Exponentialfunktion benutzen, um aus Log Odds wieder Odds zu machen.

$$prob(kurz) = \frac{exp(log.odds)}{1 + exp(log.odds)} = \frac{exp(-0.5244)}{1 + (-0.5244)} = 0.3718239$$

- Für **Version A** sagen wir also eine **Wahrscheinlichkeit** von etwa **0.37** für **kurze** Formen voraus.

Estimates der Koeffizienten

- Für Version B gibt uns das Model ein Estimate für β, die Steigung von Version A zu B.
- ☞ Eine Veränderung von Version A zu Version B führt zu einer Veränderung der log odds von 0.6290.
- Um den absoluten Estimate für Version B zu ermitteln, müssen wir beide Estimates addieren: $-0.5244 + 0.6290 = 0.1046$
- Diesen Wert von 0.1046 können wir wieder in eine Wahrscheinlichkeit umwandeln.

$$prob = \frac{exp(log.odds)}{1 + exp(log.odds)} = \frac{exp(0.1046)}{1 + (0.1046)} = 0.5261262$$

- Für **Version B** sagen wir eine **Wahrscheinlichkeit** von etwa **0.53** für **kurze** Formen voraus.

- wir sehen außerdem, dass beide Estimates einen statistisch signifikanten z-Wert haben.
- Unter der Annahme, dass wir unabhängige Datenpunkte haben, würde unser Modell also vorhersagen, dass wir einen signifikanten Unterschied zwischen Version A und B in der Proportion der kurzen Antworten haben.

Logistische mixed-effects Regression

⚠ Wir haben allerdings keine unabhängigen Datenpunkte, sondern Partizipanten geben mehr als eine Antwort, und auch unsere Nomen wurden wiederholt getestet.

- Für die Effekte / Abhängigkeiten zwischen Datenpunkten einzelner Partizipanten und Nomen können wir in Regressionsmodellen kontrollieren.

  (Anders als bei einem Chi-Quadrat-Test, weshalb wir ja ein Regressionsmodell bilden wollen.)

- Wir können `Partizipant` und `Nomen` als sogenante **random effects** in das Modell einbringen. Das Modell weiß damit, dass die Datenpunkte, die von denselben Partizipanten oder Items kommen, nicht unabhängig voneinander sind:

☞ Das Modell berechnet individuelle Intercepts (β_0) für die Variablen, die wir als Random Effects eingeben. Damit kann es erfassen, dass jede Partizipantin eine andere "Baseline" für die Präferenz von kurzen und langen Formen haben kann.

(Man kann auch die Steigung β individuell variieren lassen; das machen wir hier der Einfachheit halber nicht.)

Logistische mixed-effects Regression in R

- Damit wir unser Regressionsmodell bauen können, brauchen wir wieder ein Dataframe mit den relevanten Variablen als Spalten:

```
  Version pl_produktion Partizipant  Nomen
1       A             1      0j38la    bee
2       A             1      0j38la  horse
3       A             0      0j38la    owl
4       B             0      81e52w    bee
5       B             1      81e52w  horse
6       B             0      81e52w    owl
...
```

- In R können wir ein mixed logistisches Regressionsmodell mit der `glmer()` Funktion bauen.

```
m_pl <- glmer(pl_produktion ~ version + (1|Partizipant) + (1|Nomen),
              data = df, family = "binomial")
```

Logistische mixed-effects Regression in R

```
> summary(m_pl)

     AIC      BIC   logLik deviance df.resid
   774.4    792.4   -383.2    766.4      657

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.0127 -0.6727 -0.2722  0.7457  2.8186

Random effects:
 Groups      Name        Variance Std.Dev.
 Partizipant (Intercept) 2.39850  1.549
 Nomen       (Intercept) 0.08763  0.296
Number of obs: 661, groups:  Partizipant, 43; Nomen, 8

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.7703     0.3665  -2.102   0.0356 *
VersionB      0.8554     0.5176   1.653   0.0984 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Logistische mixed-effects Regression in R

- Für ein mixed-effect Modell bekommen wir neben den Details für die Fixed Effects auch Details zu den Random Effects:
- Die Varianz und Standardabweichung geben an, wie stark β_0 über die verschiedenen Partizipanten und Nomen hinweg variiert.

- Wir sehen auch, wenn wir diese Variation durch individuelle Partizipanten und Nomen miteinbeziehen, dass unser β Koeffizient (der Unterschied zwischen den Proportionen von kurzen Formen in Version A und B) nicht mehr statistisch signifikant ist.
- Das heißt, wir können nicht mehr sicher genug sein, dass wir unsere Daten unter der Annahme der Null-Hypothese nicht sehen können.
- (Wir können aber auch nicht ausschließen, dass es einen zugrundeliegenden Unterschied gibt! Denn das testen wir mit dieser Methode nicht direkt!)

Logistische mixed-effects Regression in R

- Das Modell macht auch wieder Vorhersagen zu den Proportionen von kurzen Formen, die wir uns (mit Vorsicht) ansehen können.
- Die Version A ist im Intercept enthalten: Hier wird ein Koeffizient als -0.7703 vorhergesagt.
- Der Koeffizient ist in log odds. Wir können ihn wieder mit der folgenden Formel in eine Wahrscheinlichkeit umwandeln:

$$prob = \frac{exp(log.odds)}{1 + exp(log.odds)} = \frac{exp(-0.7703)}{1 + (-0.7703)} = 0.3164142$$

☞ Für **Version A** sagen wir also eine **Wahrscheinlichkeit** für **kurze** Formen von etwa **0.32** voraus.

- Für Version B sagt das Modell die log odds von $-0.7703 + 0.8554 = 0.0851$ voraus.
- Diesen Wert von 0.0851 können wir zum besseren Verständnis wieder in eine Wahrscheinlichkeit umwandeln.

$$prob = \frac{exp(log.odds)}{1 + exp(log.odds)} = \frac{exp(0.0851)}{1 + (0.0851)} = 0.5675638$$

- Für **Version B** sagt das Modell also voraus, dass **kurze** Formen eine **Wahrscheinlichkeit** von etwa **0.57** haben.

Logistische mixed-effects Regression in R

- Da dieser Unterschied zwischen Version A und B unter der Annahme von H_0 nicht unwahrscheinlich genug ist (p=0.0984), müssen wir bei diesen vorausgesagten Unterschied mit Vorsicht interpretieren.
- Da p=0.0984 allerdings auch nicht sehr groß ist, ist es nicht undenkbar, dass der Unterschied mit mehr Daten von mehr Partizipanten signifikant werden könnte.
- Wir können also trotzdem festhalten, dass wir möglicherweise eine **stärkere** Präferenz **gegen kurze** und **für lange** Formen in **Version A** als in **Version B** haben.
- Die Ergebnisse des McNemar's Tests haben gezeigt, dass wir in **beiden Versionen keinen signifikanten** Unterschied in der Präferenz kurzer und langer Formen in der **frequenten** und **infrequenten** Kondition finden.