12. Vorlesung Statistik II Modellierung wissenschaftlicher Fragestellungen

Fazit Statistik 1 und 2

- Wir haben nun die nötigen statistischen Hilfsmittel um für unterschiedliche Konfigurationen von AV und UV(s) aus Stichprobendaten Schlüsse auf Gegebenheiten in einer Population zu ziehen (= statistische Inferenz)
- Obwohl es starke Zusammenhänge zwischen den beiden Bereichen gibt, unterscheiden wir dabei
  1. die Schätzung von Parametern bzw. Populationsgrößen mit **Konfidenzintervallen**
  2. die **Testung von Nullhypothesen** über Parameter bzw. Populationsgrößen
- Je nach inhaltlicher Fragestellung sind dann bestimmte (aber nicht unbedingt alle) Parameter eines Modells für uns von besonderem Interesse.
- Je nach inhaltlicher Fragestellung kann es für die Interpretation von Parametern (bzw. deren Hypothesentests) hilfreich oder sogar notwendig sein, Variablen im Modell zu
  - **zentrieren** (um der Ausprägung 0 eine Bedeutung zu verleihen), oder sogar zu
  - **z-standardisieren** (um zusätzlich unabhängig von verwendeten Einheiten einer Variable zu sein)

Fazit Regressionsmodelle

- Mit den in Statistik 2 besprochenen Regressionsmethoden ist eine große Vielfalt an verschiedenen Forschungsfragen möglich.
- AV (Kriterium):
  - Stetig (lineare Regression) oder
  - Binär (logistische Regression)
- UVs (Prädiktoren):
  - Eine (einfache Regression) oder mehrere (multiple Regression) stetige UVs
  - Eine (einfache Regression) oder mehrere (multiple Regression) diskrete (kategoriale) UVs mit zwei oder mehr Ausprägungen
  - Interaktionen zwischen zwei oder mehr stetigen und/oder diskreten (kategorialen) Uvs
  - Beliebige Kombinationen der oben genannten Möglichkeiten
- Stetige AVs und UVs können jeweils unstandardisiert, standardisiert oder zentriert verwendet werden. Auch andere Transformationen sind möglich.

Übersicht inferenzstatistischer Modelle in Statistik 1 & 2

Modell	AV	UV
Einstichproben t-Test	stetig (Mittelwert)	-
Zweistichproben t-Test	stetig (Mittelwert)	diskret (2 Ausprägungen)
Binomialtest	diskret/binär (Häufigkeit)	-
ANOVA (einfaktoriell)	stetig (Mittelwert)	diskret (2+ Ausprägungen)
ANOVA (mehrfaktoriell)	stetig (Mittelwert)	mehrere diskrete (mit je 2+ Ausprägungen)
Lineare Regression	stetig (Mittelwert)	eine oder mehrere diskrete (mit je 2+ Auspr.) oder stetige
Logistische Regression	diskret/binär (Häufigkeit)	eine oder mehrere diskrete (mit je 2+ Auspr.) oder stetige

Disclaimer:

Die im Folgenden gezeigten Modelle sind teilweise deutlich komplexer als die Modelle der letzten Vorlesungen. Sie dienen hier **ausschließlich** dem Zweck, Ähnlichkeiten zwischen Methoden zu zeigen und Ihnen damit idealerweise das Verständnis der Modelle zu erleichtern. Es ist **ausdrücklich nicht** meine Erwartung, dass Sie alle im Folgenden gezeigten Modelle reproduzieren, erklären oder für neue, ebenso komplexe Fragestellungen anpassen können.

Nutzen Sie diese Vorlesung also, um einen übergeordneten Blick auf die Methoden von Statistik 2 zu bekommen. Für die Prüfungsfragen orientieren wir uns jedoch an den anderen Vorlesungen.

Korrelation

Korrelation vs. Regression

Auch zur in Vorlesung 3 von Statistik 1 vorgestellten Methode der Korrelation zur Beschreibung des Zusammenhangs zweier stetiger Variablen gibt es mit der Effektstärke β_z einer ELR eine Möglichkeit zur statistischen Inferenz:

Korrelation:

```
> cor(Sportdaten$Liegestuetz_prae, Sportdaten$Standweitsp_prae)
[1] 0.1790058
```

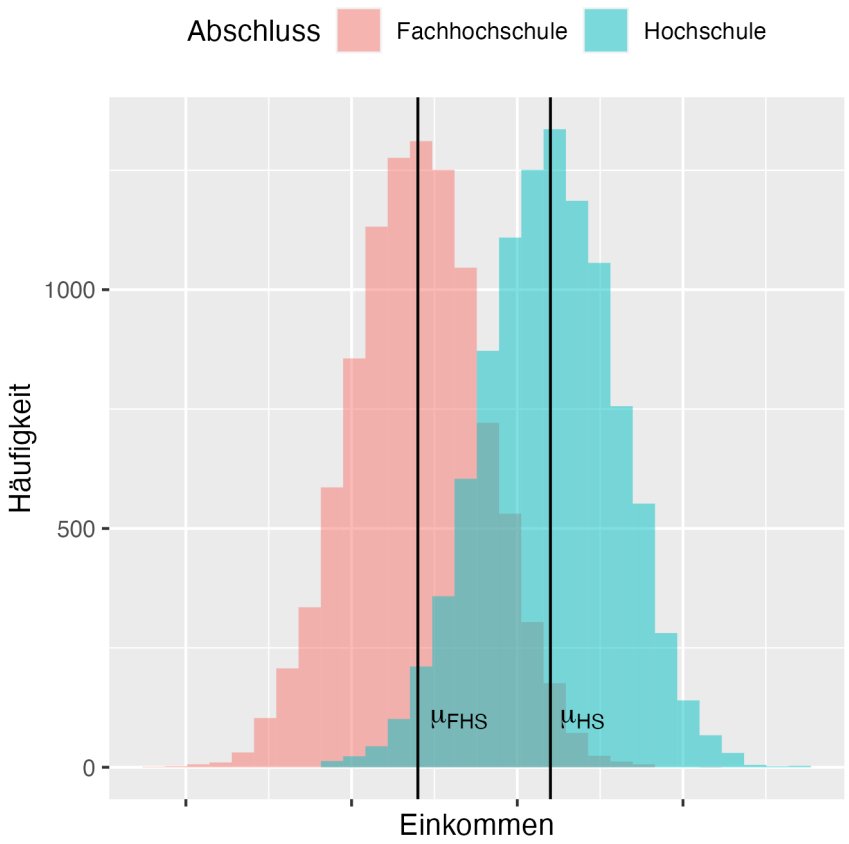
Regression mit z-Standardisierter AV und UV:

```
> r <- lm(scale(Liegestuetz_prae) ~ 0 + scale(Standweitsp_prae),
Sportdaten)
> summary(r)

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
scale(Standweitsp_prae)    0.1790     0.1796   0.997    0.327
```

Eine stetige AV, eine diskrete UV mit 2 Ausprägungen

t-Test (2 Stichproben, unabh.) vs. Regressionsmodell



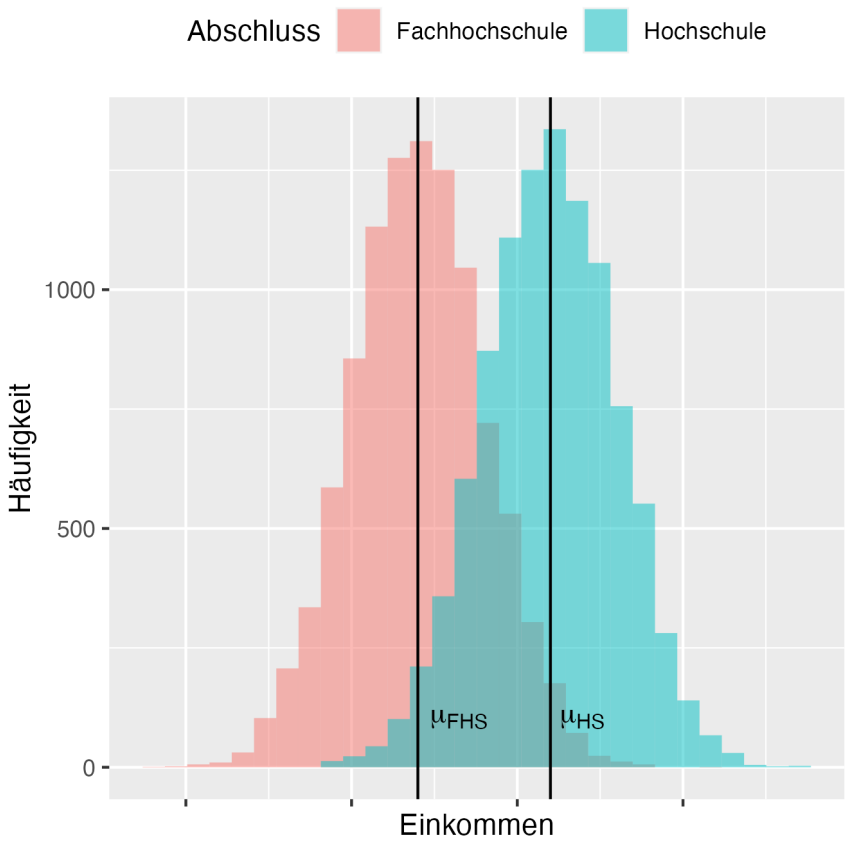
- Fragestellung in Anlehnung an Aufgabe 6/7 des Übungsblattes *Regressionsmodelle mit diskreten Prädiktoren*:

  *Sie interessieren sich dafür, **ob** sich das durchschnittliche Einkommen von Personen mit Hochschulabschluss und Personen mit Fachhochschulabschluss unterscheidet:*

- Hypothesen:

$$H_0: \mu_{FHS} - \mu_{HS} = 0$$
$$H_1: \mu_{FHS} - \mu_{HS} \neq 0$$

t-Test (2 Stichproben, unabh.) vs. Regressionsmodell



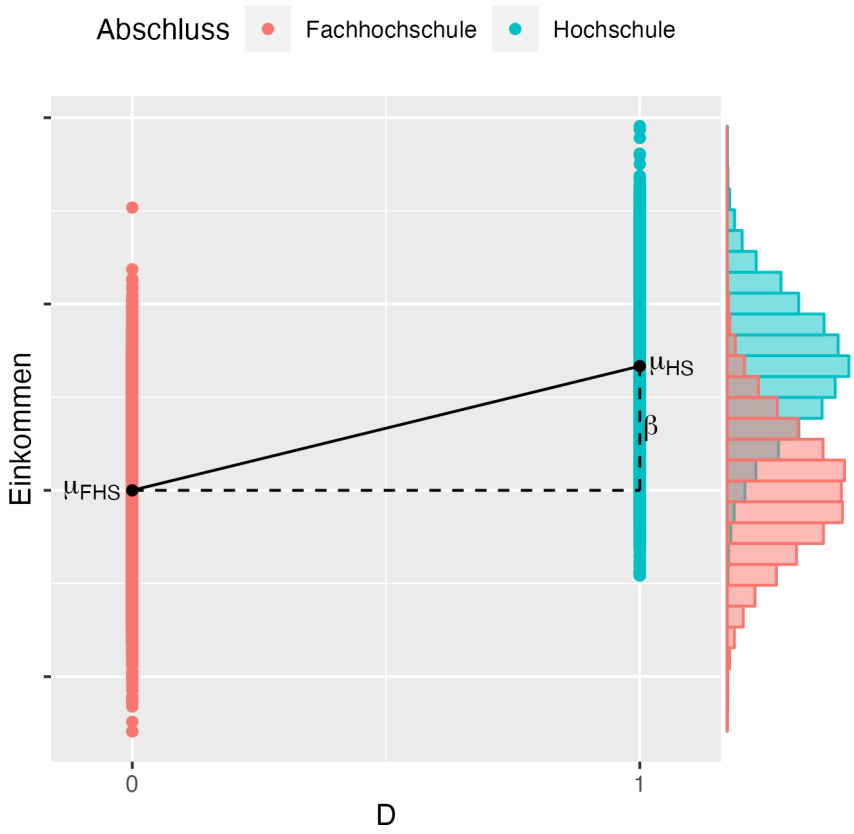
Testen der Hypothesen mit einem T-Test:

```
> t.test(Einkommen ~ Abschluss, einkommen, var.equal = TRUE, paired = FALSE)

        Two Sample t-test

data:  Einkommen by Abschluss
t = -10.623, df = 298, p-value < 2.2e-16
alternative hypothesis: true difference in means between group Fachhochschule and group Hochschule is not equal to 0
```

t-Test (2 Stichproben, unabh.) vs. Regressionsmodell



- Modellierung der Fragestellung mit einem Dummy-Regressionsmodell

$$Y_i = \alpha + \beta \cdot D_i + \epsilon_i$$

- Hypothesen:

$$H_0: \beta = 0$$
$$H_1: \beta \neq 0$$

- *Hinweis*:

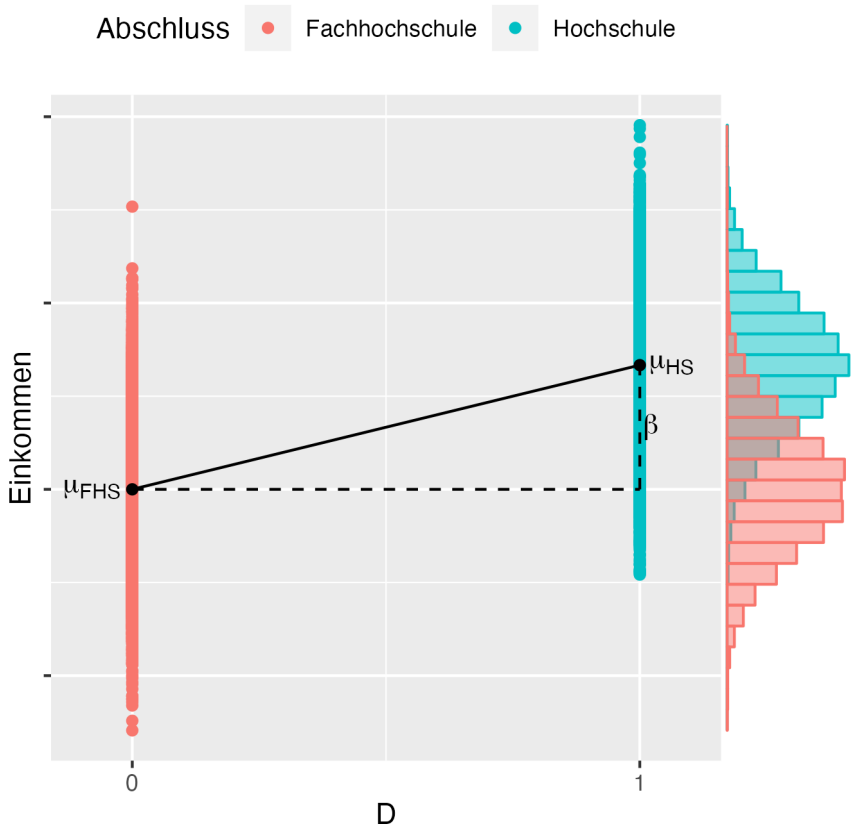
$$\beta \mathrel{\widehat{=}} \mu_{HS} - \mu_{FHS}$$

während im t-Test

$$\mu_{FHS} - \mu_{HS}$$

getestet wird.

t-Test (2 Stichproben, unabh.) vs. Regressionsmodell



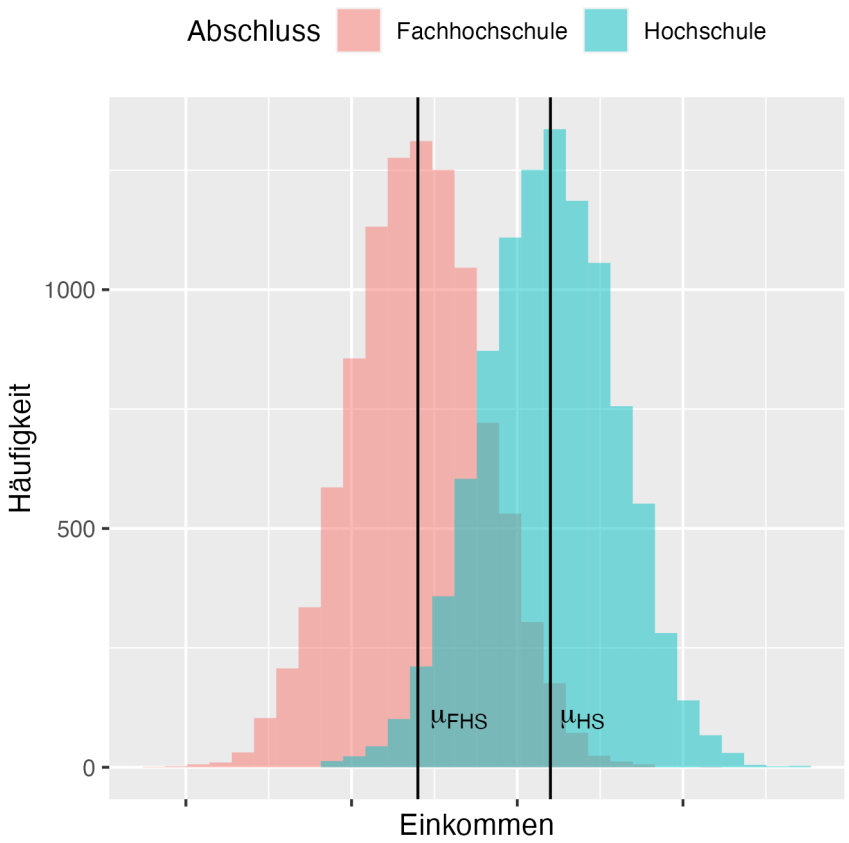
Testung der Hypothesen mit einem Dummy-Regressionsmodell

```
> r <- lm(Einkommen ~ Abschluss, einkommen)
> summary(r)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    35837       1788   20.04   <2e-16
AbschlussHS    26861       2529   10.62   <2e-16

Residual standard error: 21900 on 298 degrees of
freedom
```

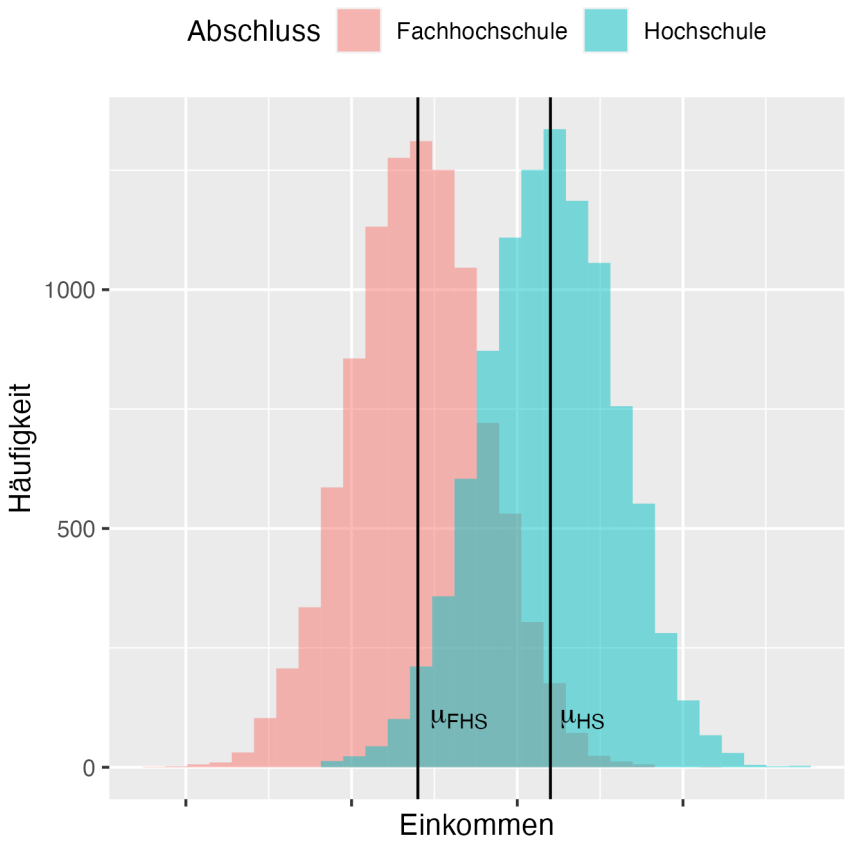
KI für $\mu_1 - \mu_2$ vs. Regressionsmodell



- Fragestellung in Anlehnung an Aufgabe 6/7 des Übungsblattes *Regressionsmodelle mit diskreten Prädiktoren*:

  *Sie interessieren sich dafür, **wie stark** sich das durchschnittliche Einkommen von Personen mit Hochschulabschluss und Personen mit Fachhochschulabschluss unterscheidet.*

KI für $\mu_1 - \mu_2$ vs. Regressionsmodell



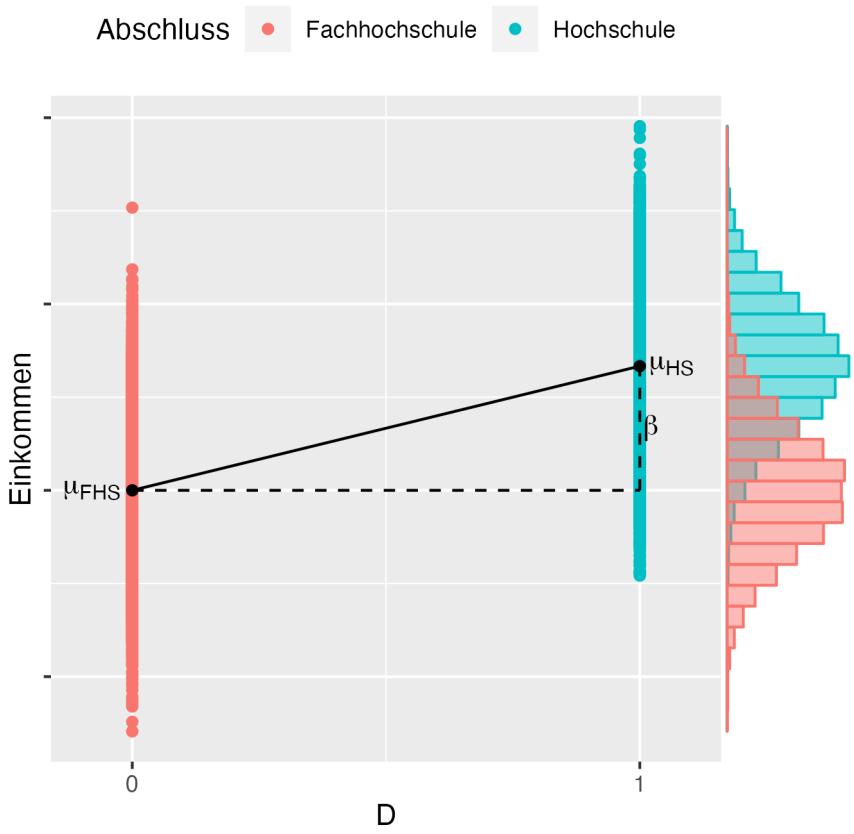
KI für $\mu_{FHS} - \mu_{HS}$ mit den Methoden aus Statistik 1:

```
> t.test(Einkommen ~ Abschluss, einkommen, var.equal = TRUE, paired = FALSE)

        Two Sample t-test

95 percent confidence interval:
 -31837.19 -21884.59
```

KI für $\mu_1 - \mu_2$ vs. Regressionsmodell



- Modellierung der Fragestellung mit einem Dummy-Regressionsmodell
- Modellgleichung:

$$Y_i = \alpha + \beta \cdot D_i + \epsilon_i$$

- KI für β:

```
> confint(r, par = "AbschlussHochschule")
                        2.5 %    97.5 %
AbschlussHochschule 21884.59 31837.19
```

Eine stetige AV, eine diskrete, mehrstufige UV

Einfaktorielle ANOVA

- Aufgabe in Anlehnung an Übungsblatt 4, Aufgabe 1

  a) Sie wollen untersuchen, ob die Farbe des Klausurpapiers (rot, grün oder blau) einen Einfluss auf die (stetige) Prüfungsleistung von Student:innen hat. (Omnibustest)
  b) Sie haben die Hypothese, dass die Prüfungsleistung bei Klausuren mit rotem Klausurpapier schlechter ist als bei Klausuren mit grünem Klausurpapier.

- Modellgleichung der ANOVA:

$$Y_{ij} = \mu_j + \epsilon_{ij}, \qquad \text{mit j} = \text{rot, grün, blau}$$

- Hypothesen:

  a) $H_0: \mu_{rot} = \mu_{grün} = \mu_{blau}$; $H_1: \mu_j \neq \mu_k$ für mindestens ein Paar j, k

  b) $H_0: \mu_{rot} \geq \mu_{grün}$; $H_1: \mu_{rot} < \mu_{grün}$

Einfaktorielle ANOVA

Hypothesen:

a) $H_0: \mu_{rot} = \mu_{grün} = \mu_{blau}$; $H_1: \mu_j \neq \mu_k$ für mindestens ein Paar j, k

```
> anova <- aov(Pruefungsleistung ~ Farbe, farbe)
> summary(anova)
             Df Sum Sq Mean Sq F value   Pr(>F)
Farbe         2   2933  1466.4    11.6 2.11e-05 ***
Residuals   147  18589   126.5
```

b) $H_0: \mu_{rot} \geq \mu_{grün}$; $H_1: \mu_{rot} < \mu_{grün}$

```
> hypB <- glht(anova, linfct = mcp(Farbe = "rot - gruen >= 0"))
> summary(hypB)

       Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: User-defined Contrasts
Linear Hypotheses:
                  Estimate Std. Error t value Pr(<t)
rot - gruen >= 0    9.380      2.249   4.171      1
```

Regression mit zwei Dummyvariablen

- Modellgleichung der Regression (Farbe = rot ist die Referenzkategorie):

$$Y_i = \alpha + \beta_1 D_1 + \beta_2 D_2 + \epsilon_i$$

$$D_1 = \begin{cases} 1, \text{falls Farbe} = grün \\ 0, \text{sonst.} \end{cases}$$

$$D_2 = \begin{cases} 1, \text{falls Farbe} = blau \\ 0, \text{sonst.} \end{cases}$$

- $\mu_{rot} = E(Y_i | D_1 = 0, D_2 = 0) = \alpha$

  $\mu_{grün} = E(Y_i | D_1 = 1, D_2 = 0) = \alpha + \beta_1$

  $\mu_{blau} = E(Y_i | D_1 = 0, D_2 = 1) = \alpha + \beta_2$

➢ $\mu_{grün} - \mu_{rot} = \alpha + \beta_1 - \alpha = \beta_1$

➢ Hypothesen:

a) $H_0: \beta_j = 0 \; für \; alle \; j; \;\; H_1: \beta_j \neq 0$ für mindestens ein j

b) $H_0: \beta_1 \leq 0; \;\; H_1: \beta_1 > 0$

Regression mit zwei Dummyvariablen

a) $H_0: \beta_j = 0$ *für alle* j; $H_1: \beta_j \neq 0$ für mindestens ein j

```
> r <- lm(Pruefungsleistung ~ Farbe, farbe)
> summary(r)

Residual standard error: 11.25 on 147 degrees of freedom
Multiple R-squared:  0.1363,    Adjusted R-squared:  0.1245
F-statistic:  11.6 on 2 and 147 DF,  p-value: 2.107e-05
```

b) $H_0: \beta_1 \leq 0$; $H_1: \beta_1 > 0$

```
> summary(glht(r, linfct = "Farbegruen <= 0"))

      Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = Pruefungsleistung ~ Farbe, data = farbe)

Linear Hypotheses:
                 Estimate Std. Error t value Pr(>t)
Farbegruen <= 0   -9.380      2.249  -4.171      1
(Adjusted p values reported -- single-step method)
```

Eine stetige AV, keine UV (nur eine Stichprobe)

KI für μ vs. Regressionsmodell

Siehe Statistik 1, Übungsblatt "Intervallschätzung II", Aufgabe 6a:

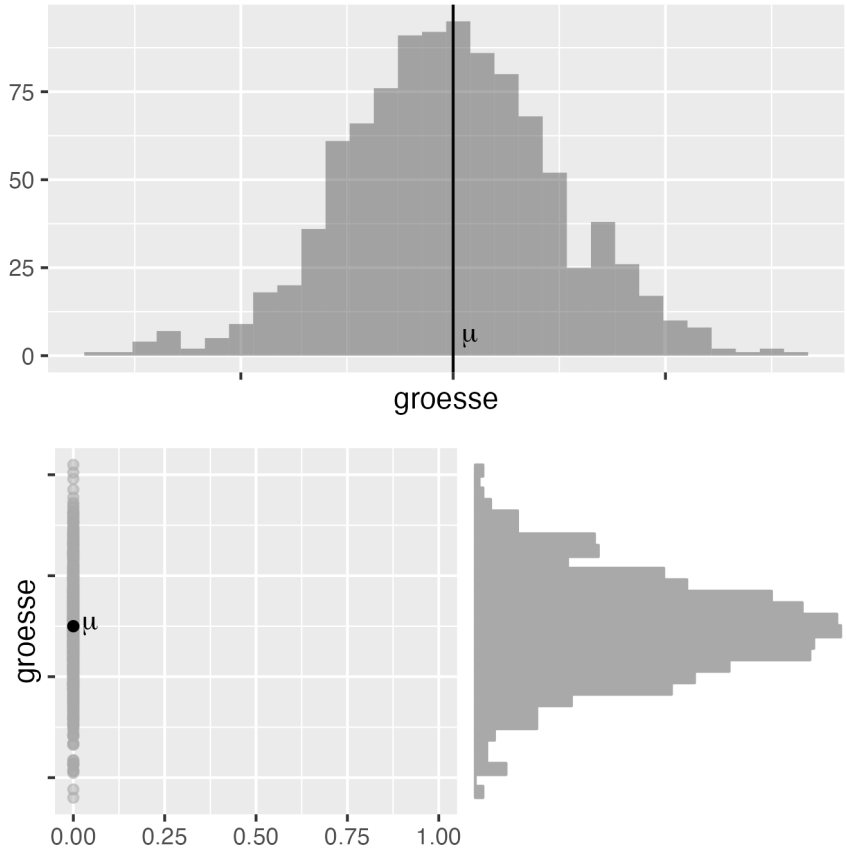
- *Der Datensatz enthält die Körpergröße in cm und die Augenfarbe von n = 1000 Personen, die zufällig aus einer Population gezogen wurden (einfache Zufallsstichprobe). Sie können davon ausgehen, dass das Histogramm der Körpergröße in der Population durch die Wahrscheinlichkeitsdichtefunktion einer Normalverteilung approximiert werden kann.*
  Berechnen Sie ein 0.95-Konfidenzintervall für die durchschnittliche Körpergröße in der Population.

➢
```
t.test(daten$groesse, conf.level = 0.95)
```

```
    One Sample t-test

data:  daten$groesse
t = 538.18, df = 999, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 169.1151 170.3529
```

KI für μ vs. Regressionsmodell



- Im Einstichprobenfall haben wir gar keine UV und wir haben gelernt, dass selbst die ELR eine UV braucht.
- Es ist aber tatsächlich möglich, ein Regressionsmodell ohne UV und nur mit einem Achsenabschnitt zu betrachten (sog. Intercept-only Modell):

```
> r <- lm(groesse ~ 1, Daten)
> confint(r)
                2.5 %    97.5 %
(Intercept) 169.1151 170.3529
```

Eine diskrete, binäre AV, keine UV (nur eine Stichprobe)

Binomialtest und KI für π vs. Regressionsmodell

Siehe Statistik 1, Übungsblatt "Intervallschätzung II", Aufgabe 6b:

- *Der Datensatz enthält die Körpergröße in cm und die Augenfarbe von n = 1000 Personen, die zufällig aus einer Population gezogen wurden (einfache Zufallsstichprobe). Sie können davon ausgehen, dass das Histogramm der Körpergröße in der Population durch die Wahrscheinlichkeitsdichtefunktion einer Normalverteilung approximiert werden kann.*
  Berechnen Sie ein 0.95-Konfidenzintervall für die relative Häufigkeit der Augenfarbe Braun in der Population. Interpretieren Sie dieses.

```
> table(Daten$augenfarbe)
 blau braun gruen
  310   354   336
> DescTools::BinomCI(354, 1000, method = 'wald', conf.level =
0.95)
       est    lwr.ci    upr.ci
[1,] 0.354 0.3243608 0.3836392
```

Binomialtest und KI für π vs. Regressionsmodell

- Wir können nun auch in der logistischen Regression die UV gänzlich weg lassen und nur einen einzelnen Parameter α schätzen.

- Diesen bzw. sein KI müssen wir dann aber noch wegen $P(Y_i = 1) = \frac{e^{\alpha}}{1+e^{\alpha}}$ in Wahrscheinlichkeiten umrechnen:

```
> r <- glm(braun ~ 1, Daten,  family = 'binomial')
> confint(r)
     2.5 %     97.5 %
-0.7319761 -0.4726674
> ki <- confint(r)
> exp(ki) / (1 + exp(ki))
    2.5 %    97.5 %
0.3247612 0.3839851
```

Erweiterung der Modelle durch die Möglichkeiten der Regression

Mögliche Fragestellungen I

- Frage: Gibt es Unterschiede im Zusammenhang zwischen der Erfahrung des Therapeuten/der Therapeutin (in Jahren) und der stetigen Depressionsschwere nach der Therapie **je nach Therapieform** (KVT, TT oder ET)?
- Modellgleichung:

$$Y_i = \alpha + \beta_1 \cdot X_i + \beta_2 \cdot D_{KVT,i} + \beta_3 \cdot D_{TT,i} + \beta_4 \cdot \left(X_i \cdot D_{KVT,i}\right) + \beta_5 \left(X_i \cdot D_{TT,i}\right) + \varepsilon_i$$

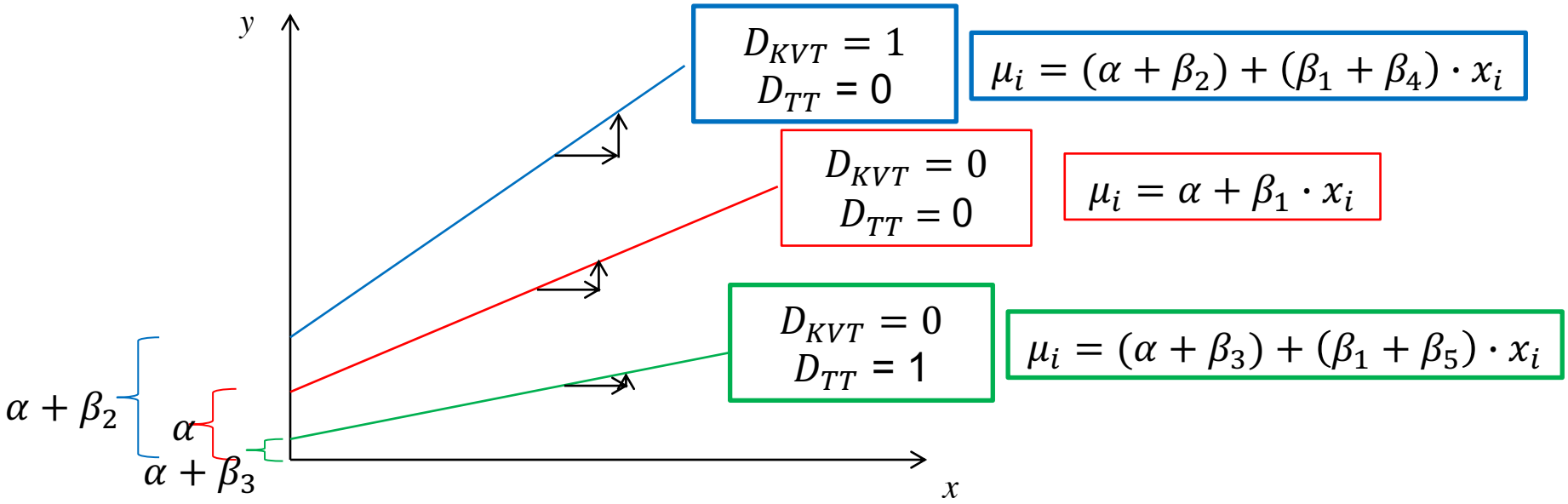
mit ET als Referenzkategorie und zwei Dummy-Variablen, die kennzeichnen ob eine Person i mit KVT, bzw. TT therapiert wurde.

- Hypothesen:

$H_0: \beta_4 = 0$ und $\beta_5 = 0$ und $\beta_5 - \beta_4 = 0$

$H_1: \beta_4 \neq 0$ oder $\beta_5 \neq 0$ oder $\beta_5 - \beta_4 \neq 0$

Mögliche Fragestellungen I



- Der Modellparameter β_4 quantifiziert den Unterschied in der Steigung zwischen KVT und ET, während sich β_5 auf den Unterschied in der Steigung zwischen TT und ET bezieht.
- β_2, bzw. β_3 quantifiziert den therapiebezogenen Unterschied in der durchschnittlichen Depressionsschwere von Personen, die von einem Therapeuten/ einer Therapeutin ohne Berufserfahrung therapiert wurden

Mögliche Fragestellungen II

- Frage: Ändert sich das Einkommen mit steigender Intelligenz positiver für Menschen mit Hochschulabschluss und ohne Migrationshintergrund als für Menschen mit Migrationshintergrund und ohne Hochschulabschluss?
- Modellgleichung (es wird komplex):

$$Y_i = \alpha + \beta_1 \cdot X_i + \beta_2 D_{1,i} + \beta_3 D_{2,i} + \beta_4 \left(D_{1,i} \cdot D_{2,i}\right) + \beta_5 \left(X_i \cdot D_{1,i}\right) + \beta_6 \left(X_i \cdot D_{2,i}\right) + \beta_7 \left(X_i \cdot D_{1,i} \cdot D_{2,i}\right) + \varepsilon_i$$

  mit zwei Dummy-Variablen: D_1 kennzeichnet ob eine Person i einen Hochschulabschluss besitzt (Referenzkategorie: kein Hochschulabschluss) und D_2 gibt an, ob eine Person einen Migrationshintergrund besitzt (Referenzkat.: kein Migrationshintergrund).
- Hypothesen:

$$H_0: \beta_5 \leq \beta_6$$
$$H_1: \beta_5 > \beta_6$$

Mögliche Fragestellungen II

Modellgleichung (es wird komplex):

$$Y_i = \alpha + \beta_1 \cdot X_i + \beta_2 D_{1,i} + \beta_3 D_{2,i} + \beta_4 \left(D_{1,i} \cdot D_{2,i}\right) + \beta_5 \left(X_i \cdot D_{1,i}\right) + \beta_6 \left(X_i \cdot D_{2,i}\right) + \beta_7 \left(X_i \cdot D_{1,i} \cdot D_{2,i}\right) + \varepsilon_i$$

- Zusammenhang Einkommen (Y_i) mit Intelligenz (X_i) für Personen ohne Hochschulabschluss ($D_1 = 0$) und ohne Migrationshintergrund ($D_2 = 0$):

$$Y_i = \alpha + \beta_1 \cdot X_i + \varepsilon_i$$

- Zusammenhang Einkommen (Y_i) mit Intelligenz (X_i) für Personen mit Hochschulabschluss ($D_1 = 1$) und ohne Migrationshintergrund ($D_2 = 0$):

$$Y_i = \alpha + \beta_2 + (\beta_1 + \beta_5) \cdot X_i + \varepsilon_i$$

- Zusammenhang Einkommen (Y_i) mit Intelligenz (X_i) für Personen ohne Hochschulabschluss ($D_1 = 0$) und mit Migrationshintergrund ($D_2 = 1$):

$$Y_i = \alpha + \beta_3 + (\beta_1 + \beta_6) \cdot X_i + \varepsilon_i$$

- Zusammenhang Einkommen (Y_i) mit Intelligenz (X_i) für Personen mit Hochschulabschluss ($D_1 = 1$) und mit Migrationshintergrund ($D_2 = 1$):

$$Y_i = (\alpha + \beta_2 + \beta_3 + \beta_4) + (\beta_1 + \beta_5 + \beta_6 + \beta_7) \cdot X_i + \varepsilon_i$$

Mögliche Fragestellungen III

- Frage: Hat das Einkommen einen Einfluss auf den Zusammenhang zwischen der (stetigen) Arbeitszufriedenheit und der Wahrscheinlichkeit zu kündigen?
- Modellgleichung ($Y_i = 1$ bedeutet Person i hat gekündigt):

$$P(Y_i = 1|x_{i1}, x_{i2}) = \frac{e^{\alpha+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3(x_{i1}x_{i2})}}{1 + e^{\alpha+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3(x_{i1}x_{i2})}}$$

- X_{i1} ist die Arbeitszufriedenheit einer zufällig gezogenen Person i und X_{i2} gibt das Einkommen einer zufällig gezogenen Person i an. In VL 10 (Folie 59ff.) haben wir Interaktionen zwischen stetigen Prädiktoren eingeführt (inklusive Interpretation als Moderatorvariable). Dies können wir jetzt auf die logistische Regression übertragen (Einkommen wird als Moderator betrachtet).
- Hypothesen:

$$H_0: \beta_3 = 0$$
$$H_1: \beta_3 \neq 0$$

Mögliche Regressionsmodelle

- Die eingeführten Regressionsmodelle (sowohl im Rahmen der linearen als auch im Rahmen der logistischen Regression) lassen sich beliebig um Prädiktoren und Interaktionsterme erweitern.
- Beispiel: Lineare Regression mit zwei stetigen Variablen und einer Dummy-Variable sowie den beiden Interaktionstermen zwischen der Dummy-Variable und den jeweiligen stetigen Prädiktorvariable:

$$Y_i = \alpha + \beta_1 \cdot X_{i1} + \beta_2 \cdot X_{i2} + \beta_3 \cdot D_i + \beta_4 \cdot (X_{i1} \cdot D_i) + \beta_5 (X_{i2} \cdot D_i) + \varepsilon_i$$

- Oder ergänzt um die Interaktion der stetigen Variablen und der Dreifachinteraktion mit der Dummy-Variable:

$$Y_i = \alpha + \beta_1 \cdot X_{i1} + \beta_2 \cdot X_{i2} + \beta_3 \cdot D_i + \beta_4 \cdot (X_{i1} \cdot D_i) + \beta_5 (X_{i2} \cdot D_i) + \beta_6 \cdot (X_{i1} \cdot X_{i2}) + \beta_7 \cdot (X_{i1} \cdot X_{i2} \cdot D_i) + \varepsilon_i$$

- Oder das entsprechende Beispiel mit binärer AV in der logistischen Regression:

$$P(Y_i = 1 | x_{i1}, x_{i2}, d_i) = \frac{e^{\alpha + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \beta_3 \cdot d_i + \beta_4 \cdot (x_{i1} \cdot d_i) + \beta_5 (x_{i2} \cdot d_i)}}{1 + e^{\alpha + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \beta_3 \cdot d_i + \beta_4 \cdot (x_{i1} \cdot d_i) + \beta_5 (x_{i2} \cdot d_i)}}$$

Mögliche Regressionsmodelle

- Die eingeführten Regressionsmodelle (sowohl im Rahmen der linearen als auch im Rahmen der logistischen Regression) lassen sich beliebig um Prädiktoren und Interaktionsterme erweitern.
- Theoretisch lassen sich auch transformierte Prädiktorvariablen in der Regressionsanalyse ergänzen. Zum Beispiel könnte eine quadrierte Variable im Modell genutzt werden um einen möglichen quadratischen Zusammenhang abzubilden:

$$Y_i = \alpha + \beta_1 \cdot X_i + \beta_2 \cdot X_i^2 + \varepsilon_i$$

- Die Beispiele auf den letzten Folien zeigen, dass lineare Modelle sehr flexibel erweitert werden können, dann allerdings auch relativ schnell schwer zu interpretieren sind (insbesondere, wenn viele Interaktionen im Modell sind).
- Deshalb sollten theoretische Überlegung bei der Modellbildung immer berücksichtigt werden und Prädiktoren nicht wahllos hinzugefügt werden (siehe auch VL 9, ab Folie 44ff.)