VO Statistik

Sitzung 7: Richtig vorhergesagt: Inferenzstatistik II

Dominik Duell

Universität Innsbruck

Etwas Admin

- Liste abgegebener Hausaufgaben 1 auf OLAT
- Zwischenstand Pop Quiz auf OLAT
- Hausaufgabe 2 online später diese Woche, haltet nach einer Email Ausschau

Hypothesenprüfverfahren

Das Idee hinter dem Hypothesenprüfverfahren

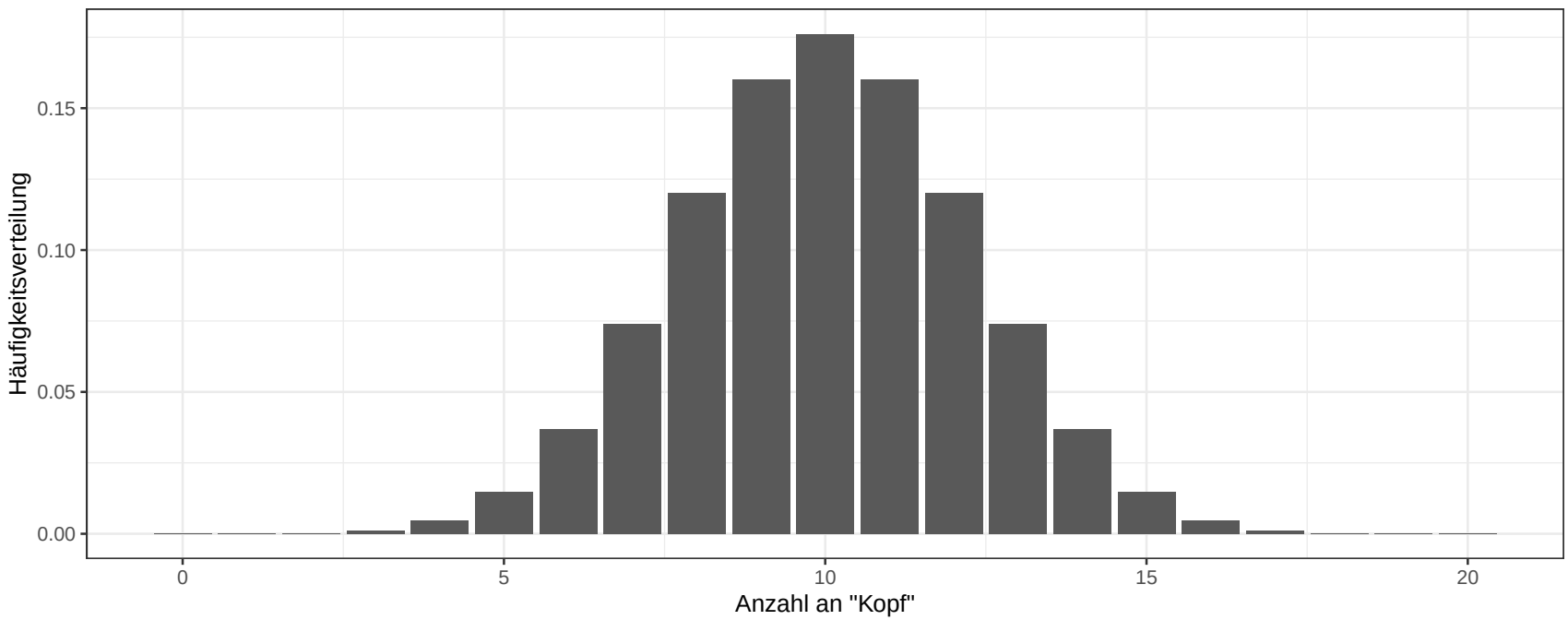
Wahrscheinlichkeit und Hypothesen prüfen

```
read.csv('../data/indicators.csv') %>%
  lm(GDPPerCapita~euJoin2004,data=.) %>%
  summary()
```

```
##
## Call:
## lm(formula = GDPPerCapita ~ euJoin2004, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8960.1 -2741.9  -474.1  2538.8 16372.9
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               11128.9      335.3   33.19   <2e-16 ***
## euJoin2004No EU Member    -7607.7      443.6  -17.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4694 on 455 degrees of freedom
##   (56 observations deleted due to missingness)
## Multiple R-squared:  0.3926, Adjusted R-squared:  0.3912
## F-statistic: 294.1 on 1 and 455 DF,  p-value: < 2.2e-16
```

Wahrscheinlichkeit und Hypothesen prüfen

```
data.frame(y=dbinom(0:20,20,.5),x=0:20) %>%
  ggplot(aes(y=y,x=x)) +
    geom_bar(stat='identity') +
    labs(y='Häufigkeitsverteilung',x='Anzahl an "Kopf"') +
    theme_bw()
```



Wahrscheinlichkeit und Hypothesen prüfen

```
dbinom(15,20,.5)
```

```
## [1] 0.01478577
```

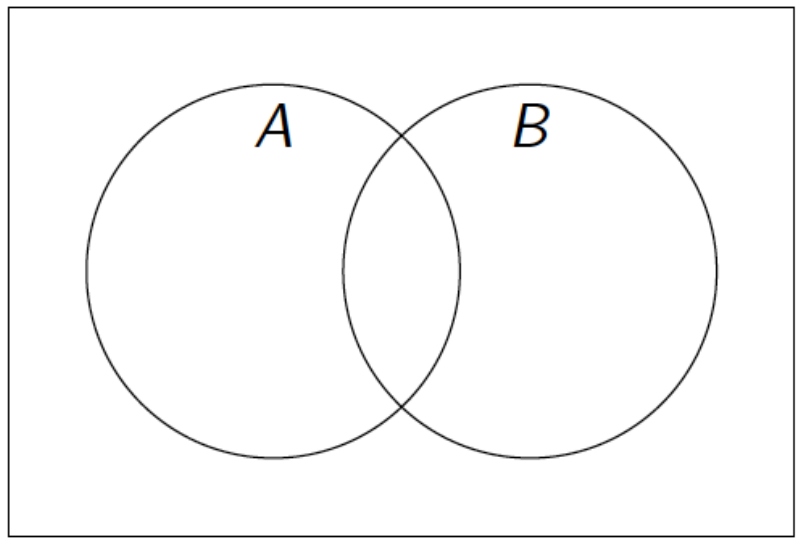
```
sum(dbinom(15:20,20,.5))
```

```
## [1] 0.02069473
```

Wahrscheinlichkeit und Hypothesen prüfen

- ein p-Wert drückt eine bedingte Wahrscheinlichkeit aus
- **nehmen wir an, dies sei die Häufigkeitsverteilung der Statistik in der Grundgesamtheit für die wir uns interessieren, was wäre dann die Wahrscheinlichkeit gegeben dieser Verteilung wenn die Statistik einen Wert annimmt, der so extreme ist wie den Wert unserer Stichprobenstatistik?**

Wahrscheinlichkeit und Hypothesen prüfen



Prüfverteilung

Prüfverteilung

- auch: Stichprobenverteilung der Statistik
- woher kriegen wir die Prüfverteilung?

Prüfverteilung

- auch: Stichprobenverteilung der Statistik
- woher kriegen wir die Prüfverteilung?
  - Kombinatorik $\Rightarrow$ exakte Häufigkeitsverteilungen

Prüfverteilung

- auch: Stichprobenverteilung der Statistik
- woher kriegen wir die Prüfverteilung?
  - Kombinatorik $\Rightarrow$ exakte Häufigkeitsverteilungen
  - theoretische Herleitung
    - aufbauend auf Gesetz der großen Zahlen
    - zentraler Grenzwertsatz
    - bedarf einer zufälligen Stichprobe

Gesetz der großen Zahlen

- für eine Zufallsvariable $X_1, \ldots, X_n$, die unabhängig und identisch verteilt ist (iid), Mittelwert μ und Standardabweichung σ.

Gesetz der großen Zahlen

- für eine Zufallsvariable $X_1, \ldots, X_n$, die unabhängig und identisch verteilt ist (iid), Mittelwert μ und Standardabweichung σ.
- wenn wir den Mittelwert der Grundgesamtheit μ mit Hilfe des Stichprobenmittelwert schätzen $\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \ldots$, dann kommen wir mit einer Stichprobe, die groß genug ist, beliebig nahe an μ heran, oder:

Gesetz der großen Zahlen

- für eine Zufallsvariable $X_1, \ldots, X_n$, die unabhängig und identisch verteilt ist (iid), Mittelwert μ und Standardabweichung σ.
- wenn wir den Mittelwert der Grundgesamtheit μ mit Hilfe des Stichprobenmittelwert schätzen $\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \ldots$, dann kommen wir mit einer Stichprobe, die groß genug ist, beliebig nahe an μ heran, oder:
- $plim(\overline{X}_n) = \mu$
  - wobei $plim()$ aussagt: $\lim_{n \to \infty} P(|\overline{X}_n - \mu|) < \epsilon) = 1$ für jedes beliebig kleine ϵ.

Gesetz der großen Zahlen

Der Durschhnitt vieler Beobachtungen der selben Größe ist genauer als jede einzelne Beobachtung alleine. Wenn die Stichprobengröße n steigt, geht die Wahrscheinlichkeit, dass $\overline{X}_n$ gleich μ ist gegen 1

Zentraler Grenzwertsatz

- für eine Zufallsvariable $X_1, \ldots, X_n$, die unabhängig und identisch verteilt ist (iid), Mittelwert μ, Standardabweichung σ und $\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$.

Zentraler Grenzwertsatz

- für eine Zufallsvariable $X_1, \ldots, X_n$, die unabhängig und identisch verteilt ist (iid), Mittelwert μ, Standardabweichung σ und $\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$.
- Dann folgt $\overline{X_n}$ **asymptotisch** (d.h. mit n konvergierend) der Normalverteilung mit Mittelwert μ und Standardabweichung σ.

Zentraler Grenzwertsatz

Wenn die Stichprobengröße n steigt, konvergiert die Häufigkeitsverteilung der Zufallsvariable X_n zur Normalverteilung mit $N(\mu, \sigma)$.

Zentraler Grenzwertsatz

Satz gilt auch für die Standarnormalverteilung

$$Z_n = \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma}$$

wobei, wenn die Stichprobengröße n steigt, konvergiert die Häufigkeitsverteilung der Zufallsvariable Z_n zur Standardnormalverteilung mit $N(1, 0)$.

Wichtige Prüfverteilungen: Normalverteilung

- Normalverteilung

Wichtige Prüfverteilungen: Normalverteilung

- Normalverteilung
- Standardnormalverteilung

Wichtige Prüfverteilungen: Normalverteilung

- Normalverteilung
- Standardnormalverteilung

$\rightarrow$ Sitzungsnotizen

Wichtige Prüfverteilungen: Normalverteilung

Weitere Gründe warum die Normalverteilung super ist:

- Ihr könnt die Verteilng skalieren: wenn $X \sim N(\mu, \sigma^2)$, dann $a + bX \sim N(a + b\mu, b^2\sigma^2)$

Wichtige Prüfverteilungen: Normalverteilung

Weitere Gründe warum die Normalverteilung super ist:

- Ihr könnt die Verteilng skalieren: wenn $X \sim N(\mu, \sigma^2)$, dann $a + bX \sim N(a + b\mu, b^2\sigma^2)$
- Ihr könnt Zufallsvariablen kombinieren: wenn X und Y unabhängig sind, d.h. $X \sim N(\mu_1, \sigma_1^2)$ und $Y \sim N(\mu_2, \sigma_2^2)$, dann $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

Wichtige Prüfverteilungen: Normalverteilung

Weitere Gründe warum die Normalverteilung super ist:

- Ihr könnt die Verteilng skalieren: wenn $X \sim N(\mu, \sigma^2)$, dann $a + bX \sim N(a + b\mu, b^2\sigma^2)$
- Ihr könnt Zufallsvariablen kombinieren: wenn X und Y unabhängig sind, d.h. $X \sim N(\mu_1, \sigma_1^2)$ und $Y \sim N(\mu_2, \sigma_2^2)$, dann $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$
- Ihr könnt mit quadrierten Zufallsvariablen arbeiten: Wenn $X_1, \ldots, X_n$ eine identisch und unabhängig verteile Zufallsvariable ist, dann folgt, dass $\sum_{i=1}^{n} X_i^2$ auch normalverteilt ist!

Wichtige Prüfverteilungen: Normalverteilung

Weitere Gründe warum die Normalverteilung super ist:

- Ihr könnt die Verteilng skalieren: wenn $X \sim N(\mu, \sigma^2)$, dann $a + bX \sim N(a + b\mu, b^2\sigma^2)$
- Ihr könnt Zufallsvariablen kombinieren: wenn X und Y unabhängig sind, d.h. $X \sim N(\mu_1, \sigma_1^2)$ und $Y \sim N(\mu_2, \sigma_2^2)$, dann $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$
- Ihr könnt mit quadrierten Zufallsvariablen arbeiten: Wenn $X_1, \ldots, X_n$ eine identisch und unabhängig verteile Zufallsvariable ist, dann folgt, dass $\sum_{i=1}^{n} X_i^2$ auch normalverteilt ist!
  - Das selbe gilt für Verhältnisse, Proportionen, etc von Zufallsvariablen.

Wichtige Prüfverteilungen: Normalverteilung

Weitere Gründe warum die Normalverteilung super ist:

Für eine $X \sim N(\mu, \sigma^2)$
* mit einer Wahrscheinlichkeit von 68%, liegt X zwischen $\mu - \sigma$ und $\mu + \sigma$
* mit einer Wahrscheinlichkeit von 95%, liegt X zwischen $\mu - 2\sigma$ und $\mu + \sigma$
* mit einer Wahrscheinlichkeit von 99%, liegt X zwischen $\mu - 3\sigma$ und $\mu + \sigma$

Wichtige Prüfverteilungen: t-Verteilung

$$t = \sqrt{n}\frac{\overline{x} - \mu}{s} = \frac{\overline{x} - \mu}{s_{\overline{x}}}$$

wobei μ der Mittelwert der Grundgesamheit, $\overline{x}$ der Mittelwert der Stichprobe, s die Standardabweichung der Stichprobe und n die Stichprobengröße ist.

Wichtige Prüfverteilungen: t-Verteilung

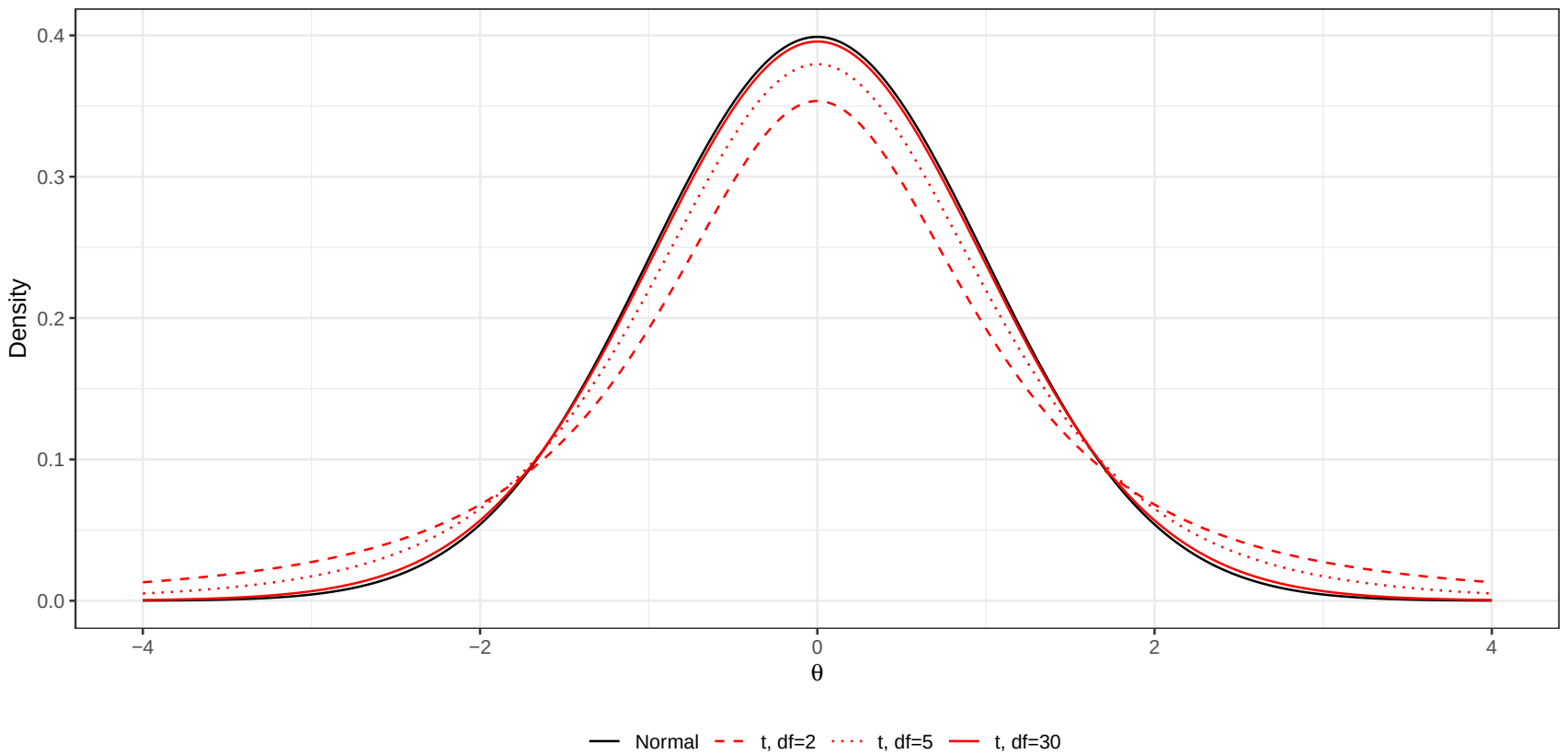
```
out <- data.frame(theta = seq(-4,4,.001)) %>%
  mutate(
    x1=dnorm(theta,0,1),
    x2=dt(theta,2),
    x3=dt(theta,5),
    x4=dt(theta,30)) %>%
  pivot_longer(cols=x1:x4) %>%
  mutate(name=factor(recode(name,
    'x1'='Normal','x2'='t, df=2','x3'='t, df=5','x4'='t, df=30'),
    levels=c('Normal','t, df=2','t, df=5','t, df=30')))
```

Wichtige Prüfverteilungen: t-Verteilung

```
plot <- out %>% ggplot(aes(y=value,x=theta,color=name,linetype=name)) +
    geom_line() +
    scale_color_manual(values=c('black','red','red','red')) +
    scale_linetype_manual(values=c(1,2,3,1)) +
    labs(y='Density',x=expression(theta)) +
    theme_bw() +
    theme(legend.position='bottom',legend.title=element_blank())
```

Wichtige Prüfverteilungen: t-Verteilung

```
plot
```



Wichtige Prüfverteilungen: χ^2-Verteilung

Gegeben n unabhängige und identisch standardnormalverteilte Zufallsvariablen $Z_1, \ldots, Z_n$, dann heißt die Verteilung der Summe der quadrierten Zufallsvariablen

$$X = \sum_{i=1}^{n} Z_i \sim \chi^2(df)$$

wobei df sind die Freiheitsgrade (eine Funktion von n)

Wichtige Prüfverteilungen: χ^2-Verteilung

Gegeben n unabhängige und identisch standardnormalverteilte Zufallsvariablen $Z_1, \ldots, Z_n$, dann heißt die Verteilung der Summe der quadrierten Zufallsvariablen

$$X = \sum_{i=1}^{n} Z_i \sim \chi^2(df)$$

wobei df sind die Freiheitsgrade (eine Funktion von n)

$\rightarrow$ Beispiel? Summe der quadrierten Fehler in der Regressionsanalyse.

Wichtige Prüfverteilungen: χ^2-Verteilung

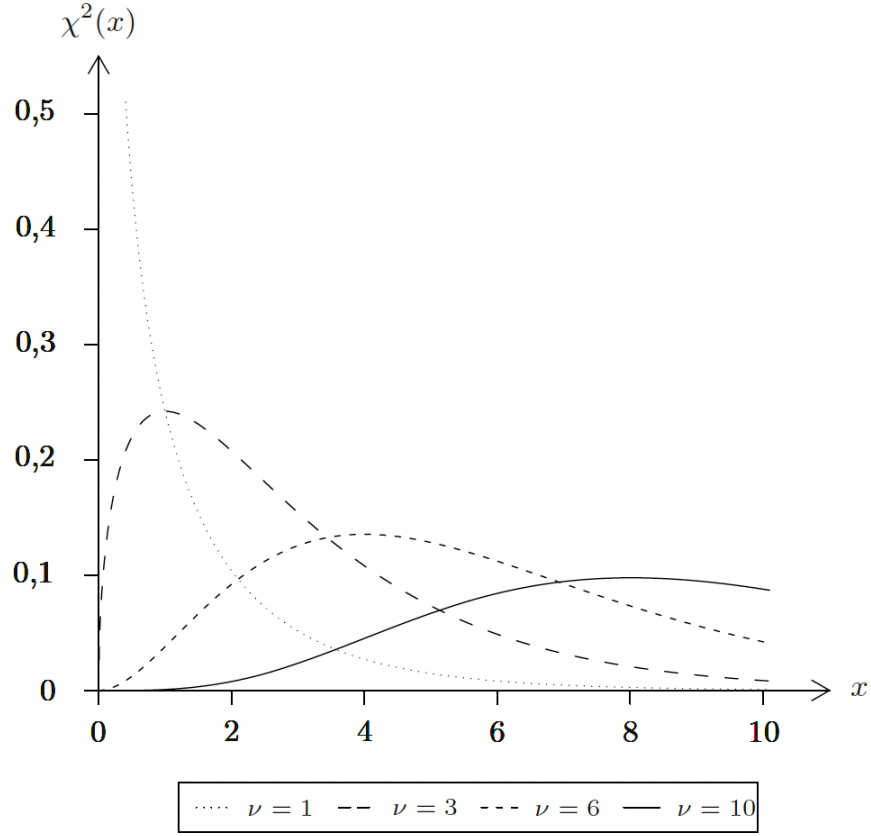


Figure 1: Source: Sibbertsen/Lehne, S.301

Wichtige Prüfverteilungen: F-Verteilung

Gegeben zwei unabhängige und identisch standardnormalverteilte Zufallsvariablen X_1 und X_2 mit Freiheitsgraden df_1 und df_2, dann heißt die Verteilung der Zufallsvariable

$$F = \frac{\frac{X_1}{df_1}}{\frac{X_2}{df_2}} \sim F(df_1, df_2)$$

Wichtige Prüfverteilungen

```
read.csv('../data/indicators.csv') %>%
  lm(GDPPerCapita~euJoin2004,data=.) %>%
  summary()
```

```
##
## Call:
## lm(formula = GDPPerCapita ~ euJoin2004, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8960.1 -2741.9  -474.1  2538.8 16372.9
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               11128.9      335.3   33.19   <2e-16 ***
## euJoin2004No EU Member    -7607.7      443.6  -17.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4694 on 455 degrees of freedom
##   (56 observations deleted due to missingness)
## Multiple R-squared:  0.3926, Adjusted R-squared:  0.3912
## F-statistic: 294.1 on 1 and 455 DF,  p-value: < 2.2e-16
```

Zusammenfassung Hypothesenprüfverfahren

Hypothesenprüfverfahren

1. Definiere eine aussagekräftige, testbare Hypothese.

Hypothesenprüfverfahren

1. Definiere eine aussagekräftige, testbare Hypothese.
2. Finde eine angemessene, valide Prüfstatistik.

Hypothesenprüfverfahren

1. Definiere eine aussagekräftige, testbare Hypothese.
2. Finde eine angemessene, valide Prüfstatistik.
3. Leite die Verteilung der Prüfstatistik ab – exakte, theoretisch (, basierend auf Simulation) $\rightarrow$ das ist die Verteilung der Statistik in der Grundgesamtheit gegeben das die Null-Hypothese wahr ist.

Hypothesenprüfverfahren

1. Definiere eine aussagekräftige, testbare Hypothese.
2. Finde eine angemessene, valide Prüfstatistik.
3. Leite die Verteilung der Prüfstatistik ab – exakte, theoretisch (, basierend auf Simulation) $\rightarrow$ das ist die Verteilung der Statistik in der Grundgesamtheit gegeben das die Null-Hypothese wahr ist.
4. Identifiziere basierend auf der Prüfverteilung den Wert der Statistik der den Ablehnungsbereich der Hypothese abgrenzt (**Kritischer Wert**) und den p-Wert, und treffe eine Entscheidung ob die Null-Hypothese abzulehnen ist.

Hypothesenprüfverfahren

Definitionen:

- Kritischer Wert: kleinster Wert der Statistik, gegeben des gewählten Levels an statistischer Signifikanz des Prüfverfahrens, an dem die Null-Hypothese abgelehnt wird.
- Type 1-Fehler: Ablehnen der Null-Hypothese auch wenn die Null-Hypothese wahr ist
  - Level statistischer Signifikanz des Prüfverfahrens α ist die Wahrscheinlichkeit eines Type 1-Fehlers of a type 1 error
- Type 2-Fehler: Versäumnis die Null-Hypothese abzulehnen obwohl diese falsch ist
  - β ist die Wahrscheinlichkeit des Type 2-Fehlers
- **statistical power** eines Prüfverfahrens: $1 - \beta$

Hypothesenprüfverfahren

→ Sitzungsnotizen

Ein Wort der Warnung bezüglich p-Werten

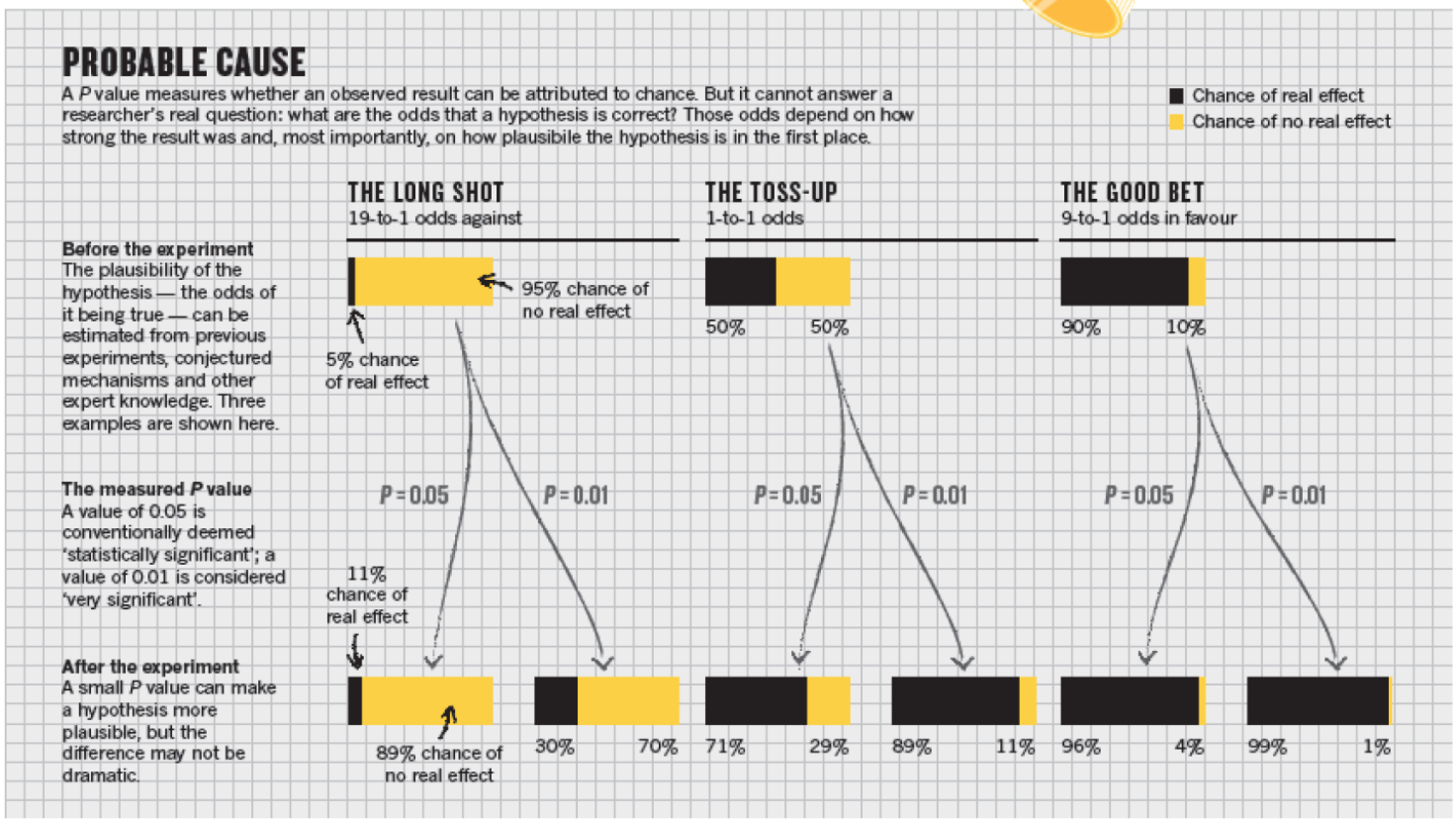


Figure 2: Source: @nuzzoValuesGoldStandard2014, p.2

Ein Wort der Warnung bezüglich p-Werten

Nochmal zu den Sitzungsnotizen:

→ Sitzungsnotizen

Vorhersagen

Regression: Beschreiben, Inferenz oder Vorhersagen?

- Regressionsanalyse gibt uns eine Beschreibung des Zusammenhangs zwischen einer abhängigen und einer oder mehrerer unabhängigen Variablen
- Die Methode der kleinsten quadrierten Fehler gibt uns Regressionskoeffizienten, welche die best-passende Linie durch die Datenwolke zieht

Regression: Beschreiben, Inferenz oder Vorhersagen?

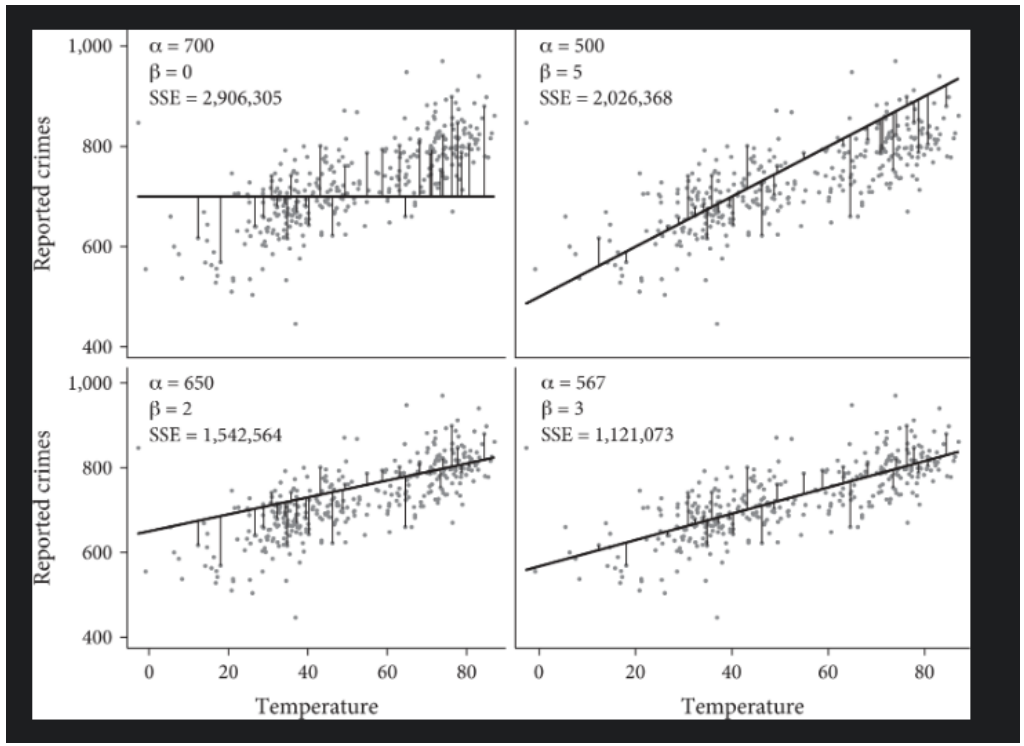


Figure 3: Source: de Mesquita/Fowler, S.99

Regression: Beschreiben, Inferenz oder Vorhersagen?

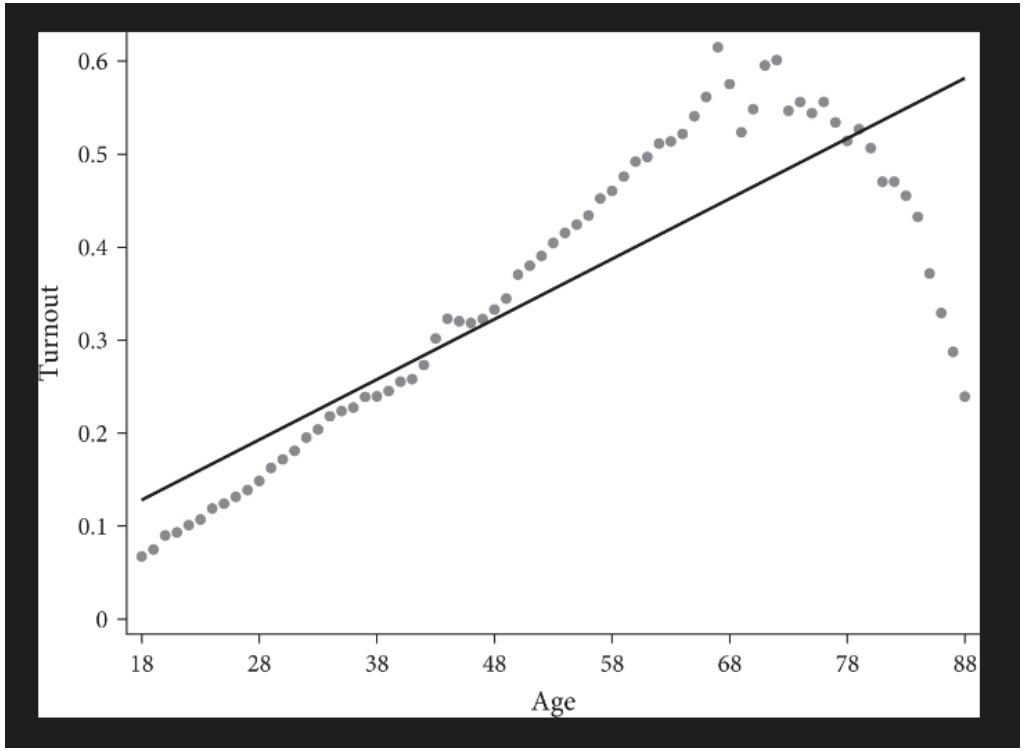
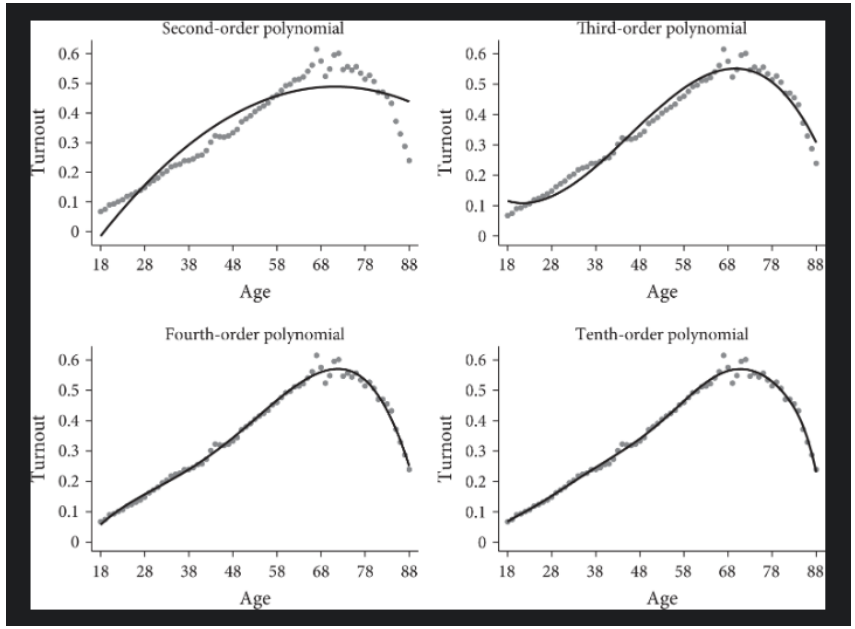


Figure 4: Source: de Mesquita/Fowler, S.104

Regression: Beschreiben, Inferenz oder Vorhersagen?



What's this?

$$\hat{t} = \beta_0 + \beta_1 age + \beta_2 age^2 + \ldots + \beta_9 age^{10}$$

$\rightarrow$ Brilliant, aber wir haben hier das Problem von **Overfitting**

Regression: Vorhersagen

- Regressionsanalyse zur Vorhersage möchte keine Inferenz betreiben, sondern die Daten so gut wie möglich beschreiben:

Regression: Vorhersagen

- Regressionsanalyse zur Vorhersage möchte keine Inferenz betreiben, sondern die Daten so gut wie möglich beschreiben:
  - was ist das beste Modell (welche Variablen in welcher funktionaler Form), um die Daten zu beschreiben
  - Kriterium: Mittlere quadrierte Fehler

Regression: Vorhersagen

- Regressionsanalyse zur Vorhersage möchte keine Inferenz betreiben, sondern die Daten so gut wie möglich beschreiben:
  - was ist das beste Modell (welche Variablen in welcher funktionaler Form), um die Daten zu beschreiben
  - Kriterium: Mittlere quadrierte Fehler

$$\hat{y} = \hat{f}(x) + \epsilon$$

Regression: Vorhersagen

- Regressionsanalyse zur Vorhersage möchte keine Inferenz betreiben, sondern die Daten so gut wie möglich beschreiben:
  - was ist das beste Modell (welche Variablen in welcher funktionaler Form), um die Daten zu beschreiben
  - Kriterium: Mittlere quadrierte Fehler

$$\hat{y} = \hat{f}(x) + \epsilon$$

Regression: Vorhersagen

- Vorhersagen heißt:
  1. finde $\hat{f}(x)$ (könnte das einfache Regressionsmodell $\beta_0 + \beta_1 x_1$ etc sein) mit Daten die wir schon haben ("train data")
  2. Ob es ein gutes Model ist, sagt uns dann der Mittlere quadrierte Fehler, wenn wir es auf neue Daten anwenden ("test data")

Regression: Vorhersagen

- Vorhersagen heißt:
  1. finde $\hat{f}(x)$ (könnte das einfache Regressionsmodell $\beta_0 + \beta_1 x_1$ etc sein) mit Daten die wir schon haben ("train data")
  2. Ob es ein gutes Model ist, sagt uns dann der Mittlere quadrierte Fehler, wenn wir es auf neue Daten anwenden ("test data")

$\rightarrow$ Gängigerweise machen das oft Computer: Machine Learning

Regression: Vorhersagen

- Vorhersagen heißt:
  1. finde $\hat{f}(x)$ (könnte das einfache Regressionsmodell $\beta_0 + \beta_1 x_1$ etc sein) mit Daten die wir schon haben ("train data")
  2. Ob es ein gutes Model ist, sagt uns dann der Mittlere quadrierte Fehler, wenn wir es auf neue Daten anwenden ("test data")

$\rightarrow$ Gängigerweise machen das oft Computer: Machine Learning

$\rightarrow$ Immer ein trade-off zwischen Varianz und Bias

$\leftrightarrow$ Warum ist das nicht Inferenz im Sinne wie oben?

Was sollt ihr aus der heutigen Sitzung mitnehmen?

- versteht den Grundgedanken hinter Hypothesenprüfverfahren
- wisst woher die Prüfverteilungen kommen können
- wisst wie ihr das Hypothesenprüfverfahren anwenden könnt
- verstehet, dass das Hypothesenprüfverfahren und wie wir p-Werte berechnen und für Inferenz nutzen problematisch sind