Statistik II - Sitzung 5

Lena Masch

Institut für Politikwissenschaft

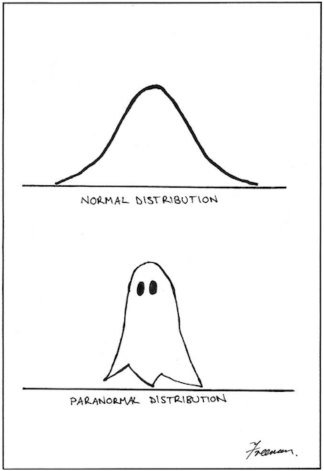
Sitzung 5

Statistik II - Sitzung 5

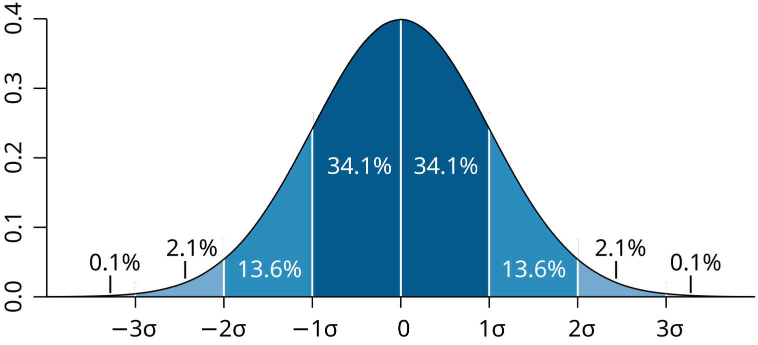
Organisatorisches

- Umfrage bitte ausfüllen! (s. Link im Learnweb)
- Chi²-Test für Zuhause folgt (nun endlich!)
- Tutorien diese Woche zum Themenbereich "Testen"

Wiederholung



Wiederholung



Wiederholung

- Fiktive Umfrage unter 100 Studierenden zur Social Media Nutzung
- Nutzen Studierende der Politikwissenschaft mehr Social Media als Andere (im Durchschnitt 3 Stunden)?

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
social_d	100	3.8072	.65241	.06524

One-Sample Test

	Test Value = 3						
			Significance			95% Confidence Interval of the Difference	
	t	df	One-Sided p	Two-Sided p	Mean Difference	Lower	Upper
social_d	12.373	99	<,001	<,001	.80720	.6778	.9367

One-Sample Effect Sizes

			95% Confidence Interval	
	Standardizer[a]	Point Estimate	Lower	Upper

Wiederholung

- Fiktive Umfrage unter 100 Studierenden zur Social Media Nutzung
- Nutzen Studierende der Politikwissenschaft mehr Social Media als Andere (im Durchschnitt 3 Stunden)?

```
> t.test(student_data$social_d, mu = 3)

        One Sample t-test

data:  student_data$social_d
t = 12.373, df = 99, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 3
95 percent confidence interval:
 3.677753 3.936657
sample estimates:
mean of x
 3.807205
```

Wiederholung

- Vergleich zweier Gruppen
- Gibt es einen Unterschied zwischen den Geschlechtern im Social Media Nutzungsverhalten?
- Was nehmen wir an?
- Wie ist vorzugehen?

Wiederholung

- Vergleich zweier Gruppen
- Gibt es einen Unterschied zwischen den Geschlechtern im Social Media Nutzungsverhalten??

T-Test

Group Statistics

	gender	N	Mean	Std. Deviation	Std. Error Mean
social_d	1.00	46	3.8474	.64910	.09571
	2.00	54	3.7730	.65934	.08972

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equa				Me
						Significance		
		F	Sig.	t	df	One-Sided p	Two-Sided p	Differ
social_d	Equal variances assumed	.477	.492	.567	98	.286	.572	
	Equal variances not assumed			.568	95.939	.286	.572	

Wiederholung

- Vergleich zweier Gruppen
- Gibt es einen Unterschied zwischen den Geschlechtern im Social Media Nutzungsverhalten??

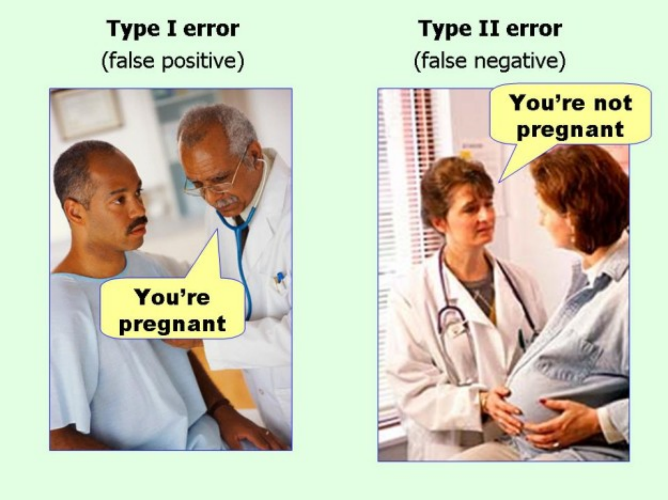
```
> t.test(student_data$social_d ~student_data$gender)

	Welch Two Sample t-test

data:  student_data$social_d by student_data$gender
t = 0.56752, df = 95.939, p-value = 0.5717
alternative hypothesis: true difference in means between group 1
95 percent confidence interval:
 -0.1859542  0.3348568
sample estimates:
mean in group 1 mean in group 2
       3.847409        3.772957
```

Probleme des statistischen Tests

- α-**Fehler oder Fehler Type I** – Ablehnung von H_0 aufgrund des Stichprobenergebnisses, obwohl H_0 in GG zutrifft
- β-**Fehler oder Fehler Type II** – Nicht-Ablehnung von H_0 aufgrund des Stichprobenergebnisses, obwohl H_0 in GG nicht zutrifft

Entscheidung aufgrund Stichprobe	H_0 in GG wahr	H_0 in GG falsch
H_0 wird **nicht abgelehnt**	Korrekter Schluss	β-Fehler (Type II)
H_0 wird **abgelehnt**	α-Fehler (Type I)	Korrekter Schluss

Probleme des statistischen Tests



https://economics.stackexchange.com/questions/27677/type-i-error-type-ii-error-pregnancy-test-analogy-is-it-legit

Logik der statistischen Signifikanz

- Aufgrund der Logik des statistischen Tests (von H_0) primäres Interesse an Vermeidung / Minimierung von α-Fehler (Type I)
- Wahrscheinlichkeit des α-Fehler - H_0 in GG ist wahr, wird aber aufgrund Stichprobenwerts abgelehnt - soll möglichst gering sein
- Annahme, dass eine Wahrscheinlichkeit des α-Fehler von unter 5% genügend Sicherheit bietet
- α-Wahrscheinlichkeit ($=$ probability, p) soll also $< 5\%$ oder $< .05$ sein
- **p**$<$**.05** $=>$ *95%-Signifikanzniveau*!

Logik der statistischen Signifikanz

- Wir lehnen also die Nullhypothese H_0 ab (und bestätigen indirekt die Alternativhypothese H_A), wenn wir sicher sein können, dass - unter der Annahme, dass es in der GG keinen Zusammenhang gibt – der über die Stichprobe erhaltene (Zusammenhangs-)Wert in weniger als 5 von 100 Stichproben vorkommt.

Logik der statistischen Signifikanz

- Wir lehnen also die Nullhypothese H_0 ab (und bestätigen indirekt die Alternativhypothese H_A), wenn wir sicher sein können, dass der über die Stichprobe erhaltene (Zusammenhangs-)Wert in weniger als 5 von 100 Stichproben unter der Annahme vorkommt, dass es in der GG keinen Zusammenhang gibt
- Weitere Signifikanzniveaus sind das $p < .01 (99\%)$ und das $p < .001 (99.9\%)$ Signifikanzniveau
  - Ablehnung von H_0, wenn der über die Stichprobe erhaltene Wert in weniger als 1 von 100 ($p < .01$) bzw. 1 von 1000 ($p < .001$) Stichproben unter Annahme von H_0 in GG vorkommt

Warum p$< .05$?

- 'The criterion of 95%confidence, or a .05-probability, forms the basis of modern statistics, and yet there is very little justification for it. How it arose is a complicated mystery to unravel.' (Fields 2013: 52)
- Verkürzt gesagt: p$< .05$ (bzw. p$< .01$, p$< .001$) hat sich einfach so eingebürgert.

Warum p$< .05$?

Auswahl des p-Levels muss reflektiert werden

Da - wie wir noch lernen werden - der Signifikanztest von einigen Faktoren abhängig ist, muss die Festlegung des Signifikanzniveaus (der Fehler-Wahrscheinlichkeit p) immer nach theoretischer Überlegung erfolgen: Welche Wahrscheinlichkeit scheint mir als ForscherIn gering genug, um ein Ergebnis als gesichert bezeichnen zu können?

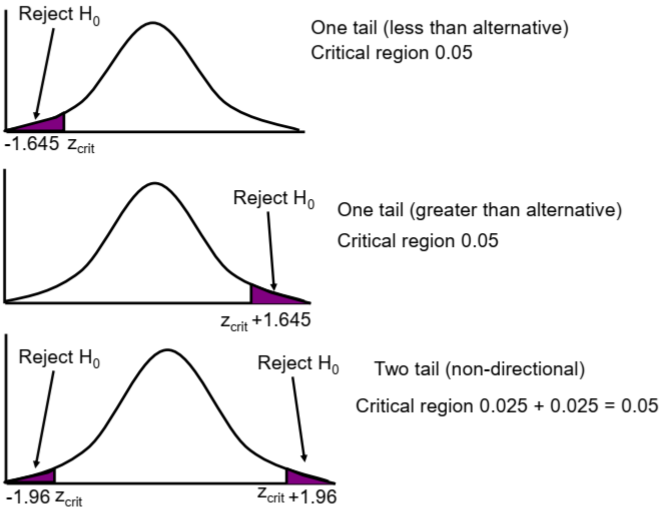
Auswahl des p-Levels VOR statistischem Test

Die Festlegung des Signifikanzniveaus für die Widerlegung der Nullhypothese erfolgt immer VOR dem statistischen Test!

Exkurs - einseitiger/zweiseitiger Test

- Ein zweiseitiger Signifikanztest wird dann vorgenommen, wenn **keine** Erwartung an die Richtung des Unterschieds zwischen zwei Gruppen vorliegt
- und die zugrundeliegende Verteilung zwei kritische Bereiche (positiv und negativ) hat, z.B. t- und z-Verteilungen
  - Dann muss der α-Fehler nach beiden Seiten abgesichert werden. Der zu erreichende Wahrscheinlichkeitswert liegt dann **nicht** bei p$< .05$, sondern bei p$< .05/2 = .025$!

Exkurs - einseitiger/zweiseitiger Test



Exkurs - einseitiger/zweiseitiger Test

- Ein zweiseitiger Signifikanztest wird dann vorgenommen, wenn **keine** Erwartung an die Richtung des Unterschieds zwischen zwei Gruppen vorliegt
  - Dann muss der α-Fehler nach beiden Seiten abgesichert werden. Der zu erreichende Wahrscheinlichkeitswert liegt dann **nicht** bei p$< .05$, sondern bei p$< .05/2 = .025$!
- Ein einseitiger Signifikanztest wird dann vorgenommen, wenn eine Erwartung an die Richtung des Unterschieds zwischen zwei Gruppen vorliegt.
  - Dann muss der α-Fehler nur nach einer Seite abgesichert werden. Der zu erreichende Wahrscheinlichkeitswert liegt dann wie angenommen bei p$< .05$!

Exkurs - einseitiger/zweiseitiger Test

Achtung: p$< .05 =>$ .025!

Für das 95%-Signifikanzniveau (p$< .05$) muss in einem zweiseitigen Test der Wahrscheinlichkeitswert p = .025 (= z-Wert: 1.96) erreicht werden. In einem einseitigen Test reicht der Wahrscheinlichkeitswert p = .05 aus (= z-Wert: 1.65).

Statistische Signifikanz

- Wir haben in den letzten Sitzungen die Signifikanz bzw. das Signifikanz-Niveau von Ergebnissen berechnet.
- Dabei war das Ziel, den α-Fehler weitestgehend auszuschließen bzw. die Wahrscheinlichkeit seines Auftretens möglichst gering zu halten.
- Wiederholung: Der α-Fehler bezeichnet den Fehler, dass wir aufgrund der Stichproben-Ergebnisse die Nullhypothese widerlegen, obwohl die Nullhypothese in der Grundgesamtheit zutrifft.

Statistische Signifikanz

- Wie gehen wir aber im Forschungs-Alltag mit der Signifikanz um?
- Leider ist die Interpretation der Signifikanz schwieriger oder: weniger eindeutig als angenommen – und wird oft fehlerhaft interpretiert
  - Siehe *Haller und Krauss 2002*
- Statistische Signifikanz besagt nur, wie wahrscheinlich das Auftreten eines Stichprobenwertes unter angenommener Gültigkeit der Nullhypothese in GG ist

Statistische Signifikanz

- Daher lässt die Signifikanz eines Stichprobenwertes auf dem 95%-Niveau streng genommen nur folgende Aussagen zu:
  - Wenn die Nullhypothese in der GG zutrifft, taucht der gefundene oder ein noch extremerer Stichproben-Wert in weniger als 5% aller aus der GG gezogenen Stichproben auf ODER
  - Da der gefundene oder ein noch extremerer Wert nur in 5 von 100 Stichproben aus einer GG, in der die Nullhypothese zutrifft, auftauchen würde, können wir die Annahme, dass die Nullhypothese in der GG zutrifft, als unwahrscheinlich ablehnen

Statistische Signifikanz

- Was wir NICHT aufgrund der Signifikanz sagen können:
  - Dieser Wert spiegelt mit 95% Wahrscheinlichkeit den Wert der Grundgesamtheit wider
- Für die Anwendung in der täglichen Arbeit:
  - Da der Wert A auf dem 95%-Niveau signifikant ist, können wir die Nullhypothese ablehnen
  - Der Wert A ist auf dem 99%-Niveau signifikant. Wir akzeptieren daher vorläufig die Alternativhypothese / Die Alternativhypothese kann vorläufig bestätigt werden

Ursachen der statistischen Signifikanz

- Welche Faktoren beeinflussen das Ergebnis des statistischen Tests (und damit die Entscheidung über signifikant oder nicht)?
  - Klar: Die **Größe des Effektes** (des Unterschiedes zwischen Gruppen bzw. die Stärke des Effekts/Zusammenhangs)
  - Auch klar: Das gewählte **Wahrscheinlichkeitsniveau** (p$< .05$, $< .01$, $< .001$)
  - die **Stichprobengröße**
  - das spezifische Test- oder Analyseverfahren
  - die Anzahl der zu vergleichenden Gruppen
  - die Entscheidung ob ein- oder zweiseitig

Varianzanalyse (ANOVA)

- Die Varianzanalyse nutzt die Varianz (Streuung), um Hypothen zu testen
- zurückzuführen auf R.A. Fisher
- Varianzanalyse, analysis of variance (ANOVA), F-Test
- Vergleich von mehr als zwei Gruppen
- einfache Varianzanalyse (one-way ANOVA) und für Messwiederholungen
- oftmals in der Psychologie angewandt (Experimente)

Varianzanalyse (ANOVA)

- in der Psychologie klassischerweise ein Vergleich als Kontroll- und Experimentalgruppen
- die unabhängige Variable ist dabei ein Faktor (kategorielle Variable)
- es ist nicht möglich, mehrere t-Tests durchzuführen, um Gruppen zu vergleichen (alpha-Fehler)

Was ist die Varianzanalyse (ANOVA)? F-Test?

- Der F-Test wird verwendet, um zu prüfen, ob mehrere Gruppen signifikant unterschiedliche Mittelwerte haben.
- F-Tests verwendet werden, um zu analysieren, ob verschiedene Gruppen signifikante Unterschiede in bestimmten Merkmalen aufweisen.
- Der Test basiert auf dem Vergleich der Streuungen (Varianzen) innerhalb und zwischen Gruppen.
- Die Alternativhypothese ist ungerichtet. Sie geht davon aus, dass ein oder mehrere Gruppen einen Einfluss auf die abhängige Variable haben
- Die Nullhypothese geht davon aus, dass die beobachteten Werte der Gruppen sich nicht unterscheiden

Die F-Statistik

Die F-Statistik wird wie folgt berechnet:

$$F = \frac{\text{Zwischen-Gruppen-Varianz}}{\text{Innerhalb-Gruppen-Varianz}} = \frac{\frac{\text{QS}_{\text{zwischen}}}{k-1}}{\frac{\text{QS}_{\text{innerhalb}}}{N-k}} \tag{1}$$

- **QS**: Quadratsumme
- **k**: Anzahl der Gruppen
- **N**: Gesamtanzahl der Beobachtungen

Ein hoher F-Wert deutet darauf hin, dass die Unterschiede zwischen den Gruppen signifikant sind.

Quadratsummen (QS)

- **Zwischen-Gruppen-Quadratsumme ($\mathbf{QS_{zwischen}}$)**: Variation zwischen den Mittelwerten der Gruppen.

$$\text{QS}_{\text{zwischen}} = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2$$

- **Innerhalb-Gruppen-Quadratsumme ($\mathbf{QS_{innerhalb}}$)**: Variation innerhalb jeder Gruppe.

$$\text{QS}_{\text{innerhalb}} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$$

- **Gesamt-Quadratsumme ($\mathbf{QS_{gesamt}}$)**:

$$\text{QS}_{\text{gesamt}} = \sum_{i=1}^{N} (X_i - \bar{X})^2$$

Exkurs: Kleines Beispiel F-Statistik

Betrachten wir drei Gruppen mit folgenden Datenpunkten:

Gruppe 1: $5, 6, 7$ Gruppe 2: $8, 9, 10$ Gruppe 3: $10, 11, 12$

- **Gesamtmittelwert** ($\bar{X}$): 8.5
- Berechnung der Mittelwerte jeder Gruppe:

$$\begin{aligned}\bar{X}_1 &= 6\\ \bar{X}_2 &= 9\\ \bar{X}_3 &= 11\end{aligned}$$

- Berechnung der Quadratsummen:

$$\begin{aligned}\text{QS}_{\text{zwischen}} &= 3 \times \left((6-8.5)^2 + (9-8.5)^2 + (11-8.5)^2\right)\\ &= 40.5\\ \text{QS}_{\text{innerhalb}} &= (5-6)^2 + (6-6)^2 + (7-6)^2 +\\ &\quad (8-9)^2 + (9-9)^2 + (10-9)^2 +\\ &\quad (10-11)^2 + (11-11)^2 + (12-11)^2\\ &= 6\end{aligned}$$

Exkurs: Kleines Beispiel F-Statistik

$$\text{Mittlere QS}_{\text{zwischen}} = \frac{\text{QS}_{\text{zwischen}}}{k-1} = \frac{40.5}{2} = 20.25$$

$$\text{Mittlere QS}_{\text{innerhalb}} = \frac{\text{QS}_{\text{innerhalb}}}{N-k} = \frac{6}{6} = 1$$

$$F = \frac{\text{Mittlere QS}_{\text{zwischen}}}{\text{Mittlere QS}_{\text{innerhalb}}} = \frac{20.25}{1} = 20.25$$

Ein hoher F-Wert (hier 20.25) könnte darauf hindeuten, dass die Gruppenmittelwerte signifikant unterschiedlich sind.

Interpretation des Ergebnisses

- Wenn der berechnete F-Wert größer ist als der kritische F-Wert aus der F-Tabelle, lehnen wir die Nullhypothese ab.
- Die Nullhypothese besagt, dass alle Gruppenmittelwerte gleich sind.
- Hier: Ein hoher F-Wert deutet darauf hin, dass es signifikante Unterschiede zwischen den Gruppen gibt.
- Hohe Werte mit kritischen Werten oder ausgegeben Signifikanz-Niveau abgleichen

Zusammenfassung

- Der F-Test hilft uns, zu analysieren, ob es signifikante Unterschiede zwischen Gruppen gibt.
- Er basiert auf dem Vergleich der Varianz zwischen und innerhalb von Gruppen.
- Ein signifikanter F-Wert weist darauf hin, dass die Gruppenmittelwerte wahrscheinlich unterschiedlich sind.
- Weitere Überlegungen:
    - Dieses Beispiel nutzt gleiche Fallzahlen pro Gruppe, dies wird oft als Voraussetzung für die ANOVA gesehen (equal sample size)
    - Wichtig ist jedoch, dass die Varianz zwischen den Gruppen gleich ist (Test auf Varianzgleichheit) oder Anwendung von robusten Tests wie Welch's ANOVA
    - soll analysiert werden, zwischen welchen Gruppen signifikante Unterschiede vorliegen, müssen Posthoc-Tests durchgeführt werden
    - Posthoc-Tests korrigieren den alpha-Fehler für das multiple Testen, damit die Irrtumswahrschenlichkeit insgesamt auf dem gewählten Niveau liegt (und nicht größer ausfällt), z.B. Bonferroni

Beispiel

- Harry Potter und politische Einstellungen?

Politics

Harry Potter and the Deathly Donald

Diana C. Mutz, *University of Pennsylvania*

ABSTRACT Few empirical studies suggest that fictional stories can influence political opinions. Nonetheless, in this study I demonstrate the relevance of Harry Potter consumption to oppositional attitudes toward Donald Trump and his worldview. Using multivariate observational models and panel data from 2014 to 2016, results suggest that the lessons of the Harry Potter series have influenced levels of opposition to punitive policies and support for tolerance of groups considered outside the American mainstream. Further, they predict public reactions to Donald Trump above and beyond their influence on policies consistent with his views.

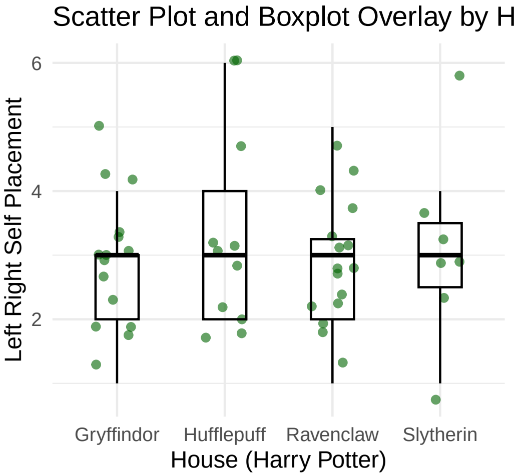
Varianzanalyse

- Links-Rechts-Selbsteinstufung und Harry Potter Haus Präferenz
- Mittelwerte, Standardabweichung und Varianzanalyse

```
>
> tapply(ds_filtered$PE01_01, ds_filtered$hp, mean, na.rm =
Gryffindor Hufflepuff  Ravenclaw  Slytherin
  2.866667   3.363636   2.875000   3.142857
> tapply(ds_filtered$PE01_01, ds_filtered$hp, sd, na.rm =TR
Gryffindor Hufflepuff  Ravenclaw  Slytherin
 0.9904304  1.5666989  1.0246951  1.5735916
> summary(aov(PE01_01 ~ hp, data = ds_filtered))
            Df Sum Sq Mean Sq F value Pr(>F)
hp           3   2.09  0.6979   0.456  0.714
Residuals   45  68.89  1.5308
```

Varianzanalyse

- Links-Rechts-Selbsteinstufung und Harry Potter Haus Präferenz



Varianzanalyse

- Links-Rechts-Selbsteinstufung und Harry Potter Haus Präferenz
- Beispiel einer Power-Analyse für eine große Effekstärke

```
> pwr::pwr.anova.test(k = 4, f = 0.35, power = 0.8, sig.level = 0.05)

     Balanced one-way analysis of variance power calculation

              k = 4
              n = 23.25056
              f = 0.35
      sig.level = 0.05
          power = 0.8

NOTE: n is number in each group
```

Varianzanalyse

- Links-Rechts-Selbsteinstufung und Harry Potter Haus Präferenz
- Beispiel einer Power-Analyse für eine mittlere Effekstärke

```
>
> pwr::pwr.anova.test(k = 4, f = 0.15, power = 0.8, sig.level = 0.05)

     Balanced one-way analysis of variance power calculation

              k = 4
              n = 122.1209
              f = 0.15
      sig.level = 0.05
          power = 0.8

NOTE: n is number in each group
```

Varianzanalyse

- Links-Rechts-Selbsteinstufung und Harry Potter Haus Präferenz
- Beispiel einer Welch's ANOVA (ohne Annahme der Varianzgleichheit)

```
>
> oneway.test(PE01_01 ~ hp, data = ds_filtered)

	One-way analysis of means (not assuming equal variances)

data:  PE01_01 and hp
F = 0.33082, num df = 3.000, denom df = 18.549, p-value =
0.8031

>
```

Varianzanalyse

- Links-Rechts-Selbsteinstufung und Harry Potter Haus Präferenz
- Beispiel eines Posthoc-Tests

```
> TukeyHSD(aov(PE01_01 ~ hp, data = ds_filtered))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = PE01_01 ~ hp, data = ds_filtered)

$hp
                              diff        lwr       upr     p adj
Hufflepuff-Gryffindor  0.496969697 -0.8132394 1.8071788 0.7433548
Ravenclaw-Gryffindor   0.008333333 -1.1779028 1.1945694 0.9999976
Slytherin-Gryffindor   0.276190476 -1.2346291 1.7870100 0.9614486
Ravenclaw-Hufflepuff  -0.488636364 -1.7814069 0.8041342 0.7453908
Slytherin-Hufflepuff  -0.220779221 -1.8166099 1.3750514 0.9826112
Slytherin-Ravenclaw    0.267857143 -1.2278646 1.7635789 0.9636223

>
```

Ausblick

- nächste Woche: Zusammenhang des Testens und der Regression
- nächste Einheit: multivariate Regression
- Tipps: Tutorienbesuch und zusätzliche Literatur auf Learnweb beachten