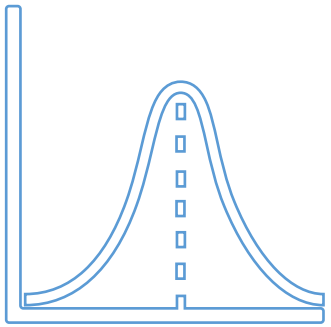
Statistik 1 – Tutorate
Sitzung 10: Mittelwertvergleiche

Marco Giesselmann, Norma De Min, Mara Moos, Lea Elina Hofer, Rémy Blum

Lernziele dieser Sitzung



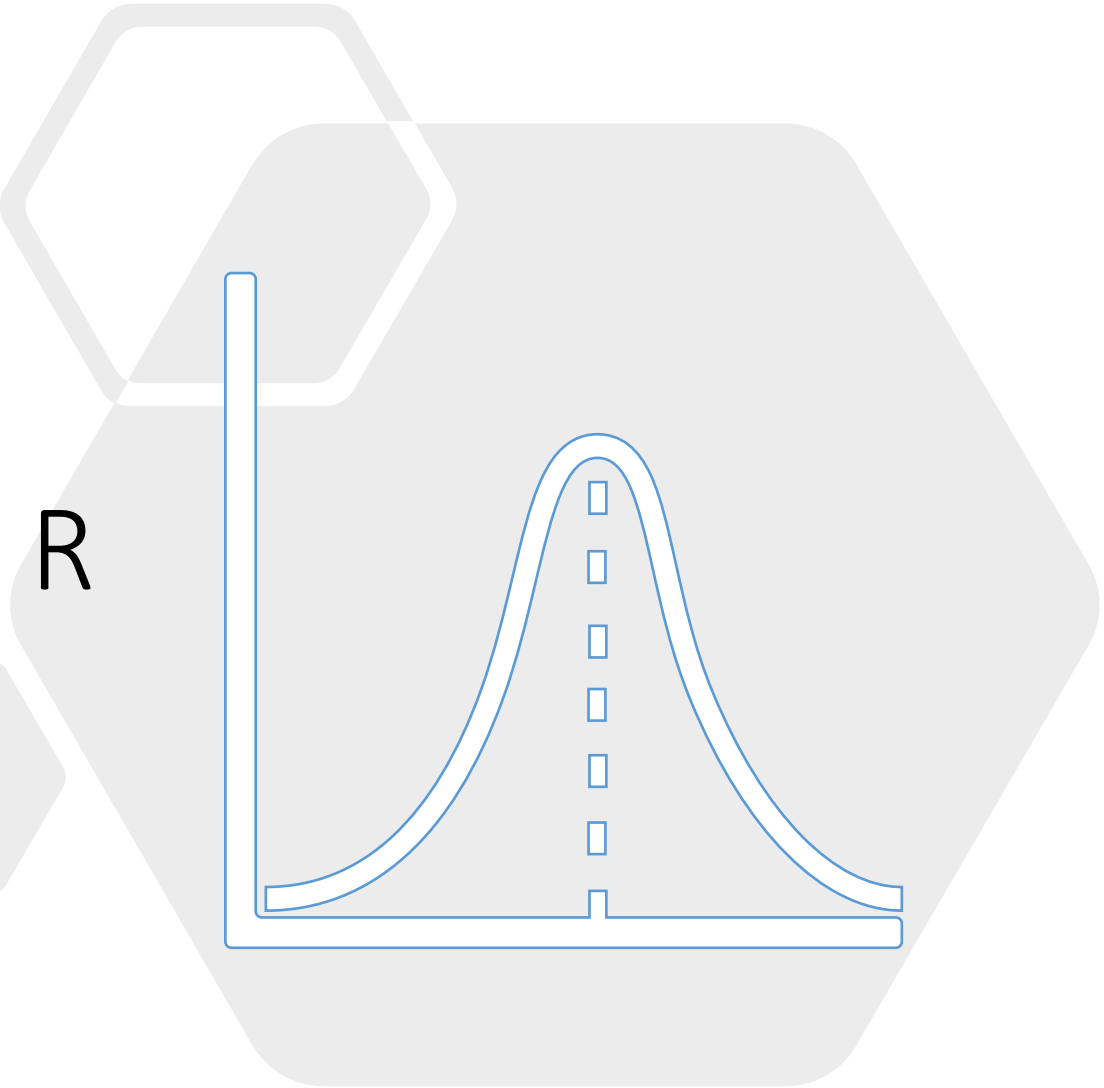
Mittelwertvergleiche in R

Recap: Mittelwertvergleich durch Gruppierung

Mittelwertvergleich durch **t.test**

(Visualisierung von Mittelwertvergleichen)

Mittelwertvergleiche in R



1 Recap: Mittelwertvergleich durch Gruppierung

Wir analysieren den Zusammenhang der Variablen **Erwerbstätigkeit neben dem Studium** (*Ja; Nein*) und **Lebenszufriedenheit** (*0 - sehr geringe Lebenszufriedenheit; 100 - sehr hohe Lebenszufriedenheit*) in der Statistik I-Kursumfrage.

Frage:
Welche **Perspektive auf den Zusammenhang** scheint euch hier sinnvoll? (Kausalrichtung? Tendenz?)

1 Recap: Mittelwertvergleich durch Gruppierung

Was ist vor der Analyse zu tun?

Vorbereitungen:

- **Faktorisierung der unabhängigen, kategorialen Variable *sidejob* (nicht notwendig, aber sinnvoll)**
- **Bereinigung der abhängigen Variable *lezufr* (hier notwendig wegen falsch codiertem Missing)**

	sidejob sind sie neben dem studium erwerbstätig	**lezufr** Lebenszufriedenheit derzeit
1	2	78
2	2	81
3	2	60
4	1	30
5	1	16
6	1	73
7	1	-99
8	2	80
9	2	69
10	2	70

```
kursdata_anon$lezufr[kursdata_anon$lezufr == -99] <- NA

kursdata_anon$sidejob <- as_factor(kursdata_anon$sidejob)
```

1 Recap: Mittelwertvergleich durch Gruppierung

Zur Vorbereitung des Mittelwertvergleichs: Aufspaltung des Datensatzes nach Kategorien der UV

```
library(dplyr)
work <- filter(kursdata_anon, sidejob == "ja")
nowork <- filter(kursdata_anon, sidejob == "nein")
```

	sidejob sind sie neben dem studium erwerbstätig	lezufr Lebenszufriedenheit derzeit
1	ja	78
2	ja	81
3	ja	60
4	nein	30
5	nein	16
6	nein	73
7	nein	*NA*
8	ja	80
9	ja	69
10	ja	70

	sidejob sind sie neben dem studium erwerbstätig	lezufr Lebenszufriedenheit derzeit
1	ja	78
2	ja	81
3	ja	60
4	ja	80
5	ja	69
6	ja	70

	sidejob sind sie neben dem studium erwerbstätig	lezufr Lebenszufriedenheit derzeit
1	nein	30
2	nein	16
3	nein	73
4	nein	*NA*

1 Recap: Mittelwertvergleich durch Gruppierung

Ermittlung der Mittelwerte in den kategorienspezifischen Datensätzen bzw. Wertelisten

```
> mean(work$lezufr, na.rm = TRUE)
[1] 69.01923
> mean(nowork$lezufr, na.rm = TRUE)
[1] 63.33333
```

Ergebnis?
Die durchschnittliche Lebenszufriedenheit arbeitender Studierender liegt etwa **5,7 Skalenpunkte** (*Mittelwertdifferenz*) oberhalb der von nicht nicht-arbeitenden Studierenden.

```
> sd(kursdata_anon$lezufr, na.rm = TRUE)
[1] 21.046
```

Standardisierte Mittelwertdifferenz (Cohen‘s d):
Der durchschnittliche Zufriedenheitsüberhang arbeitender Studierender beträgt etwa ein Viertel (Cohen‘s d=0,27) der Standardabweichung der Lebenszufriedenheit.

1 Recap: Mittelwertvergleich durch Gruppierung

Ermittlung der Mittelwerte in den kategorienspezifischen Datensätzen bzw. Wertelisten

```
> mean(work$lezufr, na.rm = TRUE)
[1] 69.01923
> mean(nowork$lezufr, na.rm = TRUE)
[1] 63.33333
```

Ergebnis?

Die durchschnittliche Lebenszufriede
(*Mittelwertdifferenz*) oberhalb der v

```
> sd(kursdat
[1] 21.046
```

Cohen´s d effect size	Interpretation	Differences in SD
d=.0–.19	Trivial effect	<1/5 from a SD
d=.20	Small effect	1/5 from a SD
d=.50	Medium effect	1/2 from a SD
d=.80 or higher	Large effect	8/10 from a SD

Interpretation of the results on d according to Cohen's (1992) recommendations

Standardisierte Mittelwertdifferenz (Cohen‘s d):

Der durchschnittliche Zufriedenheitsüberhang arbeitender Studierender beträgt etwa ein Viertel (Cohen‘s d=0,27) der Standardabweichung der Lebenszufriedenheit.

2 Inferenzstatistik zur Mittelwertdifferenz per *t.test*

t-test mit R: Direkte Umsetzung

```
t.test(nowork$lezufr, work$lezufr, var.equal = TRUE)
```

Vergleiche die Mittelwerte der zuvor nach UV-Kategorien differenzierten Wertlisten
(Differenzierung der Werteliste erfolgte vorher, ausserhalb des Befehls)

Basisversion des t-tests
(mit Annahme, das Varianz in beiden Gruppen näherungsweise identisch ist)

Differenziere die Werteliste der Variable *vor* der Tilde nach Kategorien der Variable *hinter* der Tilde. Vergleiche dann die beiden Mittelwerte der Teillisten
(Differenzierung der Werteliste erfolgt innerhalb des Befehls)

```
t.test(formula = lezufr ~ sidejob, var.equal = TRUE, data = kursdata_anon)
```

t-test mit R: Standardumsetzung

2 Inferenzstatistik zur Mittelwertdifferenz per *t.test*

```
t.test(formula = lezufr ~ sidejob, var.equal = TRUE, data = kursdata_anon)
```

```
        Two Sample t-test

data:  lezufr by sidejob
t = -1.0363, df = 71, p-value = 0.3036
alternative hypothesis: true difference in means between group nein
and group ja is not equal to 0
95 percent confidence interval:
 -16.62656   5.25476
sample estimates:
mean in group nein   mean in group ja
          63.33333           69.01923
```

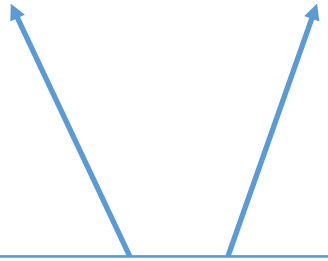
Beschreibt die Elemente des Outputs zum t.test-Kommando (ohne die konkreten Werte bereits zu interpretieren)

2 Inferenzstatistik zur Mittelwertdifferenz per *t.test*

```
t.test(formula = lezufr ~ sidejob, var.equal = TRUE, data = kursdata_anon)
```

```
        Two Sample t-test

data:  lezufr by sidejob
t = -1.0363, df = 71, p-value = 0.3036
alternative hypothesis: true difference in means between group nein
and group ja is not equal to 0
95 percent confidence interval:
 -16.62656   5.25476
sample estimates:
mean in group nein   mean in group ja
          63.33333            69.01923
```



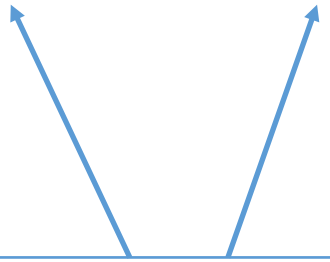
Die **Mittelwerte** entsprechen den manuell von uns berechneten Werten. Die von R abgeleitete Mittelwertdifferenz (Differenz zwischen erster und zweiter Gruppe) beträgt **-5,7**: *Nichterwerbstätige Studierende (hier Gruppe 1) sind im Mittel 5.7 Skalenpunkte weniger zufrieden als erwerbstätige Studierende*

2 Inferenzstatistik zur Mittelwertdifferenz per *t.test*

```
t.test(formula = lezufr ~ sidejob, var.equal = TRUE, data = kursdata_anon)
```

```
        Two Sample t-test

data:  lezufr by sidejob
t = -1.0363, df = 71, p-value = 0.3036
alternative hypothesis: true difference in means between group nein
and group ja is not equal to 0
95 percent confidence interval:
 -16.62656   5.25476
sample estimates:
mean in group nein    mean in group ja
          63.33333            69.01923
```



Wofür stehen in der Logik von R also ***negative Mittelwertdifferenzen*** in diesem empirischen Fall?

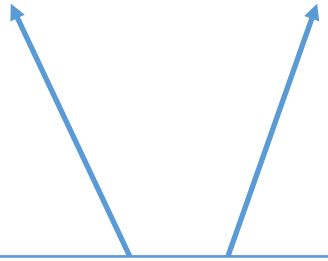
Die **Mittelwerte** entsprechen den manuell von uns berechneten Werten. Die von R abgeleitete Mittelwertdifferenz (Differenz zwischen erster und zweiter Gruppe) beträgt **-5,7**: *Nichterwerbstätige Studierende (hier Gruppe 1) sind im Mittel 5.7 Skalenpunkte weniger zufrieden als erwerbstätige Studierende*

2 Inferenzstatistik zur Mittelwertdifferenz per *t.test*

```
t.test(formula = lezufr ~ sidejob, var.equal = TRUE, data = kursdata_anon)

        Two Sample t-test

data:  lezufr by sidejob
t = -1.0363, df = 71, p-value = 0.3036
alternative hypothesis: true difference in means between group nein
and group ja is not equal to 0
95 percent confidence interval:
 -16.62656   5.25476
sample estimates:
mean in group nein    mean in group ja
          63.33333            69.01923
```



Wofür stehen in der Logik von R also ***negative Mittelwertdifferenzen*** in diesem empirischen Fall?

→ Negative Mittelwertdifferenzen zeigen einen **höheren Mittelwert der zweiten Gruppe** im Vergleich zur ersten Gruppe an

→ ….verweisen also in diesem Fall auf einen mittleren Zufriedenheitsunterschied **zugunsten der *erwerbstätigen* Studierenden**

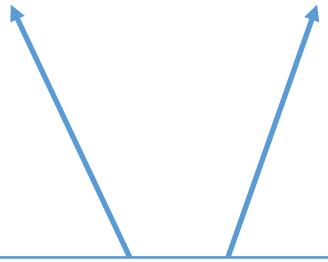
Die **Mittelwerte** entsprechen den manuell von uns berechneten Werten. Die von R abgeleitete Mittelwertdifferenz (Differenz zwischen erster und zweiter Gruppe) beträgt **-5,7**: *Nichterwerbstätige Studierende (hier Gruppe 1) sind im Mittel 5.7 Skalenpunkte weniger zufrieden als erwerbstätige Studierende*

2 Inferenzstatistik zur Mittelwertdifferenz per *t.test*

```
t.test(formula = lezufr ~ sidejob, var.equal = TRUE, data = kursdata_anon)

        Two Sample t-test

data:  lezufr by sidejob
t = -1.0363, df = 71, p-value = 0.3036
alternative hypothesis: true difference in means between group nein
and group ja is not equal to 0
95 percent confidence interval:
 -16.62656   5.25476
sample estimates:
mean in group nein    mean in group ja
          63.33333            69.01923
```



Wofür stehen in der Logik von R also ***positive Mittelwertdifferenzen*** in diesem empirischen Fall?

→ Positive Mittelwertdifferenzen zeigen einen **tieferen Mittelwert der zweiten Gruppe** im Vergleich zur ersten Gruppe an

→ ….verweisen also in diesem Fall auf einen mittleren Zufriedenheitsunterschied **zugunsten der *nicht-erwerbstätigen* Studierenden**

Die **Mittelwerte** entsprechen den manuell von uns berechneten Werten. Die von R abgeleitete Mittelwertdifferenz (Differenz zwischen erster und zweiter Gruppe) beträgt **-5,7**: *Nichterwerbstätige Studierende (hier Gruppe 1) sind im Mittel 5.7 Skalenpunkte weniger zufrieden als erwerbstätige Studierende*

2 Inferenzstatistik zur Mittelwertdifferenz per ***t.test***

```
t.test(formula = lezufr ~ sidejob, var.equal = TRUE, data = kursdata_anon)

        Two Sample t-test

data:  lezufr by sidejob
t = -1.0363, df = 71, p-value = 0.3036
alternative hypothesis: true difference in means between group
and group ja is not equal to 0
95 percent confidence interval:
 -16.62656   5.25476
sample estimates:
mean in group nein    mean in group ja
          63.33333            69.01923
```

Sowohl ein wahrer Zufriedenheitsunterschied zugunsten der Erwerbstätigen als auch zugunsten der Nicht-erwerbstätigen liegt ausweislich des ermittelten Konfidenzintervalls im Bereich des Möglichen!

Wofür stehen in der Logik von R also ***positive Mittelwertdifferenzen*** in diesem empirischen Fall?

→ Positive Mittelwertdifferenzen zeigen einen **tieferen Mittelwert der zweiten Gruppe** im Vergleich zur ersten Gruppe an

→ ….verweisen also in diesem Fall auf einen mittleren Zufriedenheitsunterschied **zugunsten der *nicht-erwerbstätigen* Studierenden**

Die **Mittelwerte** entsprechen den manuell von uns berechneten Werten. Die von R abgeleitete Mittelwertdifferenz (Differenz zwischen erster und zweiter Gruppe) beträgt **-5,7**: *Nichterwerbstätige Studierende (hier Gruppe 1) sind im Mittel 5.7 Skalenpunkte weniger zufrieden als erwerbstätige Studierende*

2 Inferenzstatistik zur Mittelwertdifferenz per *t.test*

```
t.test(formula = lezufr ~ sidejob, var.equal = TRUE, data = kursdata_anon)

        Two Sample t-test

data:  lezufr by sidejob
t = -1.0363, df = 71, p-value = 0.3036
alternative hypothesis: true difference in means between group ne
and group ja is not equal to 0
95 percent confidence interval:
 -16.62656   5.25476
sample estimates:
mean in group nein    mean in group ja
          63.33333            69.01923
```

Konkret: Mit 95% Sicherheit liegt der wahre mittlere Unterschied der Lebenszufriedenheit zwischen 16.6 Punkten zugunsten der Erwerbstätigen und 5.2 Punkten zugunsten der Nicht-Erwerbstätigen.

Wofür stehen in der Logik von R also ***positive Mittelwertdifferenzen*** in diesem empirischen Fall?

→ Positive Mittelwertdifferenzen zeigen einen **tieferen Mittelwert der zweiten Gruppe** im Vergleich zur ersten Gruppe an

→ ….verweisen also in diesem Fall auf einen mittleren Zufriedenheitsunterschied **zugunsten der *nicht-erwerbstätigen* Studierenden**

Die **Mittelwerte** entsprechen den manuell von uns berechneten Werten. Die von R abgeleitete Mittelwertdifferenz (Differenz zwischen erster und zweiter Gruppe) beträgt **-5,7**: *Nichterwerbstätige Studierende (hier Gruppe 1) sind im Mittel 5.7 Skalenpunkte weniger zufrieden als erwerbstätige Studierende*

2 Inferenzstatistik zur Mittelwertdifferenz per ***t.test***

```
t.test(formula = lezufr ~ sidejob, var.equal = TRUE, data = kursdata_anon)

        Two Sample t-test

data:  lezufr by sidejob
t = -1.0363, df = 71, p-value = 0.3036
alternative hypothesis: true difference in means between group nein
and group ja is not equal to 0
95 percent confidence interval:
 -16.62656   5.25476
sample estimates:
mean in group nein    mean in group ja
          63.33333            69.01923
```

Oder: Mit 95% Sicherheit sind die Erwerbstätigen in der Population zwischen 16.6 Skalenpunkte zufriedener und 5.3 Skalenpunkten unzufriedener als die Nicht-Erwerbstätigen.

Wofür stehen in der Logik von R also ***positive Mittelwertdifferenzen*** in diesem empirischen Fall?

→ Positive Mittelwertdifferenzen zeigen einen **tieferen Mittelwert der zweiten Gruppe** im Vergleich zur ersten Gruppe an

→ ….verweisen also in diesem Fall auf einen mittleren Zufriedenheitsunterschied **zugunsten der *nicht-erwerbstätigen* Studierenden**

Die **Mittelwerte** entsprechen den manuell von uns berechneten Werten. Die von R abgeleitete Mittelwertdifferenz (Differenz zwischen erster und zweiter Gruppe) beträgt **-5,7**: *Nichterwerbstätige Studierende (hier Gruppe 1) sind im Mittel 5.7 Skalenpunkte weniger zufrieden als erwerbstätige Studierende*

2 Inferenzstatistik zur Mittelwertdifferenz per *t.test*

Die Wahrscheinlichkeit, dass eine Population ohne Mittelwertdifferenz ein Stichprobenergebnis erzeugt, welches mindestens eine so hohe Mittelwertdifferenz wie unsere Stichprobe aufweist, liegt bei etwa 30,4%

```
t.test(formula = lezufr ~ sidejob, var.equal = TRUE, data = kursdata_anon)

        Two Sample t-test

data:  lezufr by sidejob
t = -1.0363, df = 71, p-value = 0.3036
alternative hypothesis: true difference in means between group nein
and group ja is not equal to 0
95 percent confidence interval:
 -16.62656   5.25476
sample estimates:
mean in group nein    mean in group ja
          63.33333            69.01923
```

Oder: Mit 95% Sicherheit sind die Erwerbstätigen in der Population zwischen 16.6 Skalenpunkte zufriedener und 5.3 Skalenpunkten unzufriedener als die Nicht-Erwerbstätigen.

Wofür stehen in der Logik von R also ***positive Mittelwertdifferenzen*** in diesem empirischen Fall?

→ Positive Mittelwertdifferenzen zeigen einen **tieferen Mittelwert der zweiten Gruppe** im Vergleich zur ersten Gruppe an

→ ….verweisen also in diesem Fall auf einen mittleren Zufriedenheitsunterschied **zugunsten der *nicht-erwerbstätigen* Studierenden**

Die **Mittelwerte** entsprechen den manuell von uns berechneten Werten. Die von R abgeleitete Mittelwertdifferenz (Differenz zwischen erster und zweiter Gruppe) beträgt **-5,7**: *Nichterwerbstätige Studierende (hier Gruppe 1) sind im Mittel 5.7 Skalenpunkte weniger zufrieden als erwerbstätige Studierende*

2 Inferenzstatistik zur Mittelwertdifferenz per *t.test*

…die Nullhypothese, dass arbeitende und nicht-arbeitende Soziologiestudierende sich in der Population nicht in der mittleren Lebenszufriedenheit unterscheiden, kann also wegen p>0.05 nicht abgelehnt werden

```
t.test(formula = lezufr ~ sidejob, var.equal = TRUE, data = kursdata_anon)

        Two Sample t-test

data:  lezufr by sidejob
t = -1.0363, df = 71, p-value = 0.3036
alternative hypothesis: true difference in means between group nein
and group ja is not equal to 0
95 percent confidence interval:
 -16.62656   5.25476
sample estimates:
mean in group nein    mean in group ja
          63.33333            69.01923
```

Oder: Mit 95% Sicherheit sind die Erwerbstätigen in der Population zwischen 16.6 Skalenpunkte zufriedener und 5.3 Skalenpunkten unzufriedener als die Nicht-Erwerbstätigen.

Wofür stehen in der Logik von R also ***positive Mittelwertdifferenzen*** in diesem empirischen Fall?

→ Positive Mittelwertdifferenzen zeigen einen **tieferen Mittelwert der zweiten Gruppe** im Vergleich zur ersten Gruppe an

→ ….verweisen also in diesem Fall auf einen mittleren Zufriedenheitsunterschied **zugunsten der *nicht-erwerbstätigen* Studierenden**

Die **Mittelwerte** entsprechen den manuell von uns berechneten Werten. Die von R abgeleitete Mittelwertdifferenz (Differenz zwischen erster und zweiter Gruppe) beträgt **-5,7**: *Nichterwerbstätige Studierende (hier Gruppe 1) sind im Mittel 5.7 Skalenpunkte weniger zufrieden als erwerbstätige Studierende*

2 Vergleich: Unterschiedliche Varianten des *t.test*-Kommandos

```
t.test(formula = lezufr ~ sidejob, var.equal = TRUE, data = kursdata_anon)

data:  lezufr by sidejob
t = -1.0363, df = 71, p-value = 0.3036
alternative hypothesis: true difference in means between group nein
and group ja is not equal to 0
95 percent confidence interval:
 -16.62656   5.25476
sample estimates:
mean in group nein   mean in group ja
          63.33333           69.01923
```

Unterschiede im Ergebnisoutput zwischen den beiden R-basierten *t.test* - Umsetzungsvarianten?

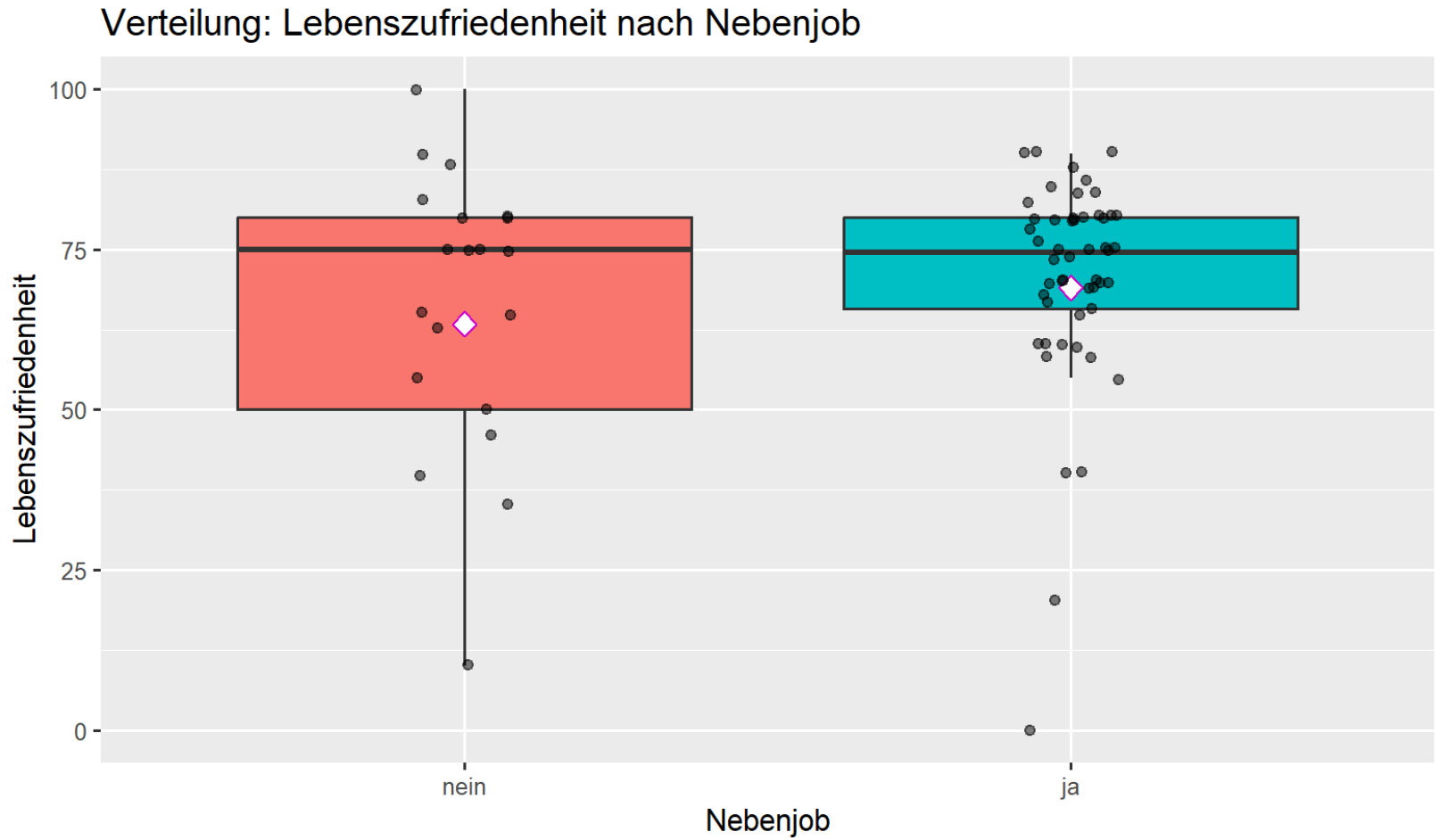
- Ergebnisdarstellung weitgehend identisch
- Kategorienspezifische Mittelwerte sind in der Standardvariante (oben) mit Kategoriennamen gelabelt, in der direkten Variante (unten) dagegen abstrakt mit x / y (Positionen korrespondieren dabei zur Reihenfolge unter «data: … and …»)

```
t.test(nowork$lezufr, work$lezufr, var.equal = TRUE)

data:  nowork$lezufr and work$lezufr
t = -1.0363, df = 71, p-value = 0.3036
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -16.62656   5.25476
sample estimates:
mean of x mean of y
 63.33333  69.01923
```

3 Visualisierung von Mittelwertvergleichen

Siehe Homepage…



Quelle: Kursbefragung Statistik I (n = 73)