CAS in Statistical Data Science 2020–2021 am IMSV der Universität Bern
Modul: Lineare Modelle

Übungsserie 1

Datum: 26. Februar 2021
Thema: Einweg-ANOVA
Dauer: 1 Lektion

Aufgabe 1

Ziel dieser Aufgabe ist es, den Zusammenhang einer kategoriellen und einer numerischen Variable zu beschreiben. Dazu betrachten wir den Datensatz `Caffeine.txt`. Er enthält das Resultat einer Studie, die den Einfluss von Koffein auf die Tippgeschwindigkeit untersuchen sollte. Dabei wurde 30 männlichen Collegestudenten das Schreibmaschinenschreiben beigebracht. Danach wurden sie zufällig in drei gleich grosse Gruppen aufgeteilt und erhielten verschiedene Dosen Koffein (0, 100 und 200 mg). Zwei Stunden später musste jeder Proband tippen und die Anzahl Anschläge pro Minute wurde aufgezeichnet.

(a) Wie gross sind die Mittelwerte von `fingertaps` pro Gruppe? Beschreiben Sie damit den Zusammenhang zwischen `fingertaps` und Koffeindosis.

(b) Erstellen Sie pro Gruppe einen Boxplot für `fingertaps`. Können Sie damit die Beurteilung aus Teil a) bestätigen?

(c) Führen Sie nun eine entsprechende ANOVA durch. Wie hoch ist das Bestimmtheitsmass? Was können Sie damit kurz über die Varianz von 'fingertabs' sagen?

(d) Bestimmen Sie die Varianz der Residuen und jene von `fingertaps`. Berechnen Sie daraus das R^2 von Hand.

Aufgabe 2

Fassen wir nun die Daten aus Aufgabe 1 als Zufallsstichprobe aus der Population aller Collegestudenten auf.

(a) In welchen Bereichen liegen die tatsächlichen Gruppenmittelwerte μ_i, $i = 1, 2, 3$, mit je etwa 95% Sicherheit?

(b) Formulieren Sie Arbeits- und Nullhypothesen des F-Tests hinsichtlich der μ_i.

(c) Formulieren Sie die Hypothesen nun auch noch hinsichlich des wirklichen Bestimmtheitsmasses θ.

(d) Können Sie die Nullhypothese(n) auf dem 5%-Niveau verwerfen? Was bedeutet dies für den Zusammenhang zwischen `fingertaps` und Koffeindosis?

(e) Bestimmen sie TSS und RSS. Berechnen Sie damit die F-Teststatistik. Wie findet man daraus den p-Wert?

(f) Trauen Sie dem p-Wert des F-Tests? Beantworten Sie diese Frage mithilfe deskriptiver Überlegungen.

(g) Eine wichtige Voraussetzung des F-Tests ist die gleiche Streuung zwischen den Gruppen. Angenommen, ein Test der Nullhypothese "Alle Varianzen sind gleich" (z. B. Levenes Test) ergibt einen p-Wert von 0.3. Weshalb hilft uns das nicht zur Beurteilung der Voraussetzung?

(h) Was können Sie anhand des p-Werts aus d) über die untere 95%-Konfidenzschranke für θ sagen?

(i) Prüfen Sie nun auf zwei Arten "post-hoc", zwischen welchen der drei Gruppen wirklich ein mittlerer Unterschied besteht: Einerseits ohne Korrektur für multiples Testen (5%-Niveau pro Test), andererseits mit Bonferroni-Holm-Korrektur (insgesamt 5%-Niveau). Überlegen Sie, wie Sie die p-Werte nach der Bonferroni-Holm-Korrektur aus den p-Werten ohne Korrektur erhalten.

Aufgabe 3

Hier wollen wir sehen, dass die F-Teststatistik der Einweg-ANOVA als Funktion des R^2 geschrieben werden kann. Dabei verwenden wir die Notation aus dem Vorlesungsskript.

(a) Zeigen Sie, dass

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} \quad \text{und} \quad 1 - R^2 = \frac{\text{RSS}}{\text{TSS}}.$$

(b) Verwenden Sie die Formeln aus (a), um zu zeigen, dass die F-Teststatistik T folgende alternative Formel hat:

$$T = \frac{R^2}{1 - R^2} \frac{n - p}{p - 1}.$$

(c) Überlegen Sie anhand dieser Formel, in welchen Situationen besonders viel gegen die Nullhypothese von keinem Zusammenhang spricht.

(d) Verifizieren Sie die Formel aus (b) mit dem Datensatz aus Aufgabe 1.