

# VO Statistik Summer 2023

## Hausaufgabe 1

Dominik Duell

Universität Innsbruck

### Ziele der Hausaufgabe

In der Bewältigung dieser Hausaufgabe solltet ihr folgende Dinge mitnehmen:

- Verständnis dafür, wie man aus einer Forschungsfrage relevante Variablen identifiziert, die dafür notwendig sind eine empirische Antwort zu finden.
- Übung darin, diese Variablen als Datensatz zu finden und in ein Format zu bringen, das statistisch analysiert werden kann.
- Anwendung grundlegender deskriptiver Statistik, Lagemaße, Streuungsmaße und Häufigkeiten, und deren grafische Darstellung.
- Berechnung einer linearen Regression.

### Anleitung

1. Diese Hausaufgabe ist **15 Punkte wert** (15% eurer Endnote)
2. Eure Antworten müsst ihr bis zum **14 Mai, 23:59 CET**, per Email an mich übermitteln (dominik.duell@uibk.ac.at).
3. Für die Hausaufgabe könnt ihr jegliche Materialien verwenden: Lehrbücher, Skripte, das Internet, eure Kommiliton:innen, etc. Ihr könnt auch in Gruppen arbeiten.
4. Solltet ihr eure Hausaufgabe in einer Gruppe bearbeiten, können eure Antworten inhaltlich die gleichen sein (etwa die gleiche berechnete Statistik). Der Wortlaut eurer Antwort darf jedoch nicht mit der abgegebenen Hausaufgabe anderer übereinstimmen. Ihr müsst eure eigene Antworten niederschreiben.
5. Solltet ihr eine Frage haben oder wenn ihr Lernmaterialien nicht aufrufen könnt, schreit mir eine Email.
6. Bitte gebt eure Hausaufgabe als PDF ab. Ich akzeptiere keine Formate wie docx, odt, etc. Jedes Textverarbeitungsprogramm kann als PDF ausdrucken.

### Aufgabe 1 (5 Punkte) Holt euch Daten.

Diese Aufgabe hat drei Teile:

- a. Nennt eine Forschungsfrage, die einen Zusammenhang zwischen zwei Variablen beinhaltet.
- b. Definiert was die Grundgesamtheit für diese Forschungsfrage ist und was eure Stichprobe sein wird.
- c. Findet Daten, welche die beiden Variablen operationalisieren und bringt sie als zwei Spalten in ein Datenblatt. Hier ein Beispiel einer Forschungsfrage: Hängt die Energieintensität einer Volkswirtschaft vom Wohlstand dieser Volkswirtschaft ab? Um diese Frage zu Beantworten holen wir uns Emissionsdaten (Quelle: Climate Action Tracker) aus dem Jahr 2020 und Daten zum Bruttonationalprodukt von der Weltbank Das resultierende Datenblatt könnte wie folgt aussehen (Ausschnitt):

	energyIntensity	gdp
1	2.288045	6165.747
2	3.030173	48678.684
3	2.756470	11444.961
4	5.147090	46064.254
5	3.977901	3277.816
6	1.440985	93397.047
7	2.678077	23324.525
8	5.170302	26456.389
9	3.561332	16315.815
10	3.419064	59919.820
11	3.105973	14063.982
12	2.335175	40232.211

Diese Aufgabe könnt ihr in jeglichen Datenverarbeitungsprogrammen durchführen: Excel, SPSS, R, etc. Wenn ihr nicht wisst, wie ihr einen Datendownload in ein Datenverarbeitungsprogramm bekommt. Kommt in meine Sprechstunde oder schreibt mir eine Email.

## Aufgabe 2 (5 Punkte) Zeigt eure Daten

Ihr solltet nun eure Daten in einer Datentabelle haben, zwei Spalten, eine für jede Variable, und so viele Zeilen, wie ihr Beobachtungen erhoben habt.

Hier die Aufgaben:

- Definiert, mit euren eigenen Worten, was sind absolute und relative Häufigkeit?
- Generiert eine Häufigkeitstabelle.
- Generiert zwei Histogramme, eines für jede Variable.
- Beschreibt, was ihr in diesen Tabellen und Histogrammen seht.

Wie macht man das in gängigen Datenverarbeitungsprogrammen?

- Excel: Für Häufigkeitstabellen, folgt den Schritten in dieser Beschreibung [LINK](#). *Excel ist hier etwas umständlich*. Für Histogramme, ladet erst das Analysis ToolPak ([LINK](#)), klickt auf Datenanalyse → Histogramm und füllt die Boxen im Pop-Up Fenster mit euren Spaltennamen.
- SPSS: Analysieren → Deskriptive Statistik → Häufigkeiten und Diagramme → Chart builder → Histogramm
- R:

```
install.packages('summarytools', repos = "http://cran.us.r-project.org")
library(summarytools)
freq(eureDaten$eureVariable1)
freq(eureDaten$eureVariable2)
hist(eureDaten$eureVariable1)
hist(eureDaten$eureVariable2)
```

## Aufgabe 3 (5 Punkte) Findet den Zusammenhang

Ihr habt ja nun eure zwei Variablen in einem Datenblatt. Hier die Aufgaben:

- a. Von euren beiden Variablen, was ist die abhängige, was die unabhängige Variable? *Diese Frage könnte für eure Forschungsfrage irrelevant sein, dann beantwortet hier, warum?*
- b. Berechnet mit Hilfe einer statistischen Software die Varianz eurer unabhängigen Variable (für diese Stichprobe) und die Kovarianz eure abhängigen und unabhängigen Variable (für diese Stichprobe).
- c. Berechnet, durch Anwendung der Formel, den Regressionskoeffizienten für eure unabhängige Variable.
- d. Mit Hilfe eure statistischen Software, berechnet eine Regression eure abhängigen auf die unabhängige Variable. Was ist der Regressionskoeffizient, der euch von der Software ausgespuckt wird? Zeigt mir den Output der Software, der die Regressionsergebnisse zeigt.

Eine Regression lasst ihr wie folgt laufen:

- Excel: Klickt auf Datenanalyse → Regression und füllt die Boxen im Pop-Up Fenster mit euren Spaltennamen.
- SPSS: Analysieren → Regression → Linear und dann benennt eure Variablen.
- R:

```
lm(eureAbhängigeVariable~eureUnabhängigeVariable, data=eureDaten)
```