

## Rechnerübungen Statistik mit Python / Beispiele für die Einführungsstunde

Für alle Beispiele benötigen Sie eine Datei `einf_daten.jpynb`.  
Für Beispiel 1 ist zusätzlich die Datei `einf_bl.txt` erforderlich.

Wenn Sie die Lösungswege nachvollziehen wollen:

- In der Datei `einf_beispiele_mit_lösungsweg.pdf` stehen ebenfalls die unten genannten Beispiele zusammen mit Lösungsvorschlägen.

Damit keine Missverständnisse auftreten: Die hier genannten Beispiele sind nicht die Testat-Aufgaben `sr_aufg_1` bis `sr_aufg_3`, die Sie selbständig bearbeiten sollen.

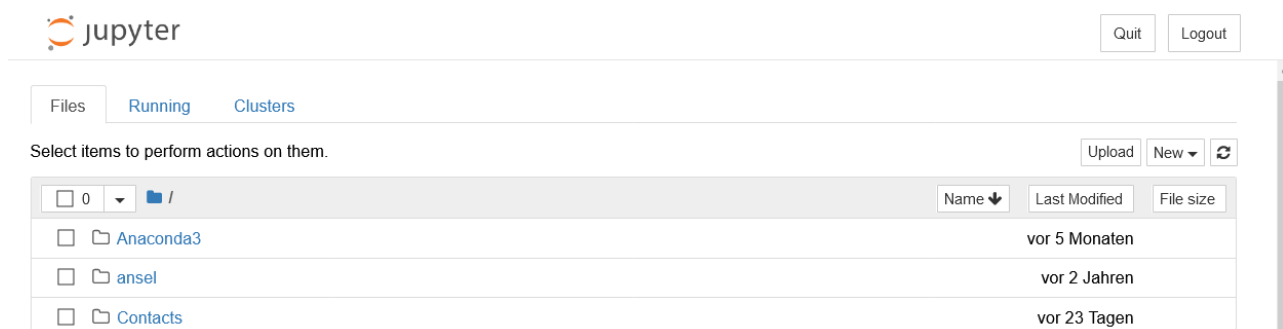
In den hier vorliegenden Beispielen werden einige wichtige Statistik- Funktionen der Software Python erläutert. Es können aber nicht alle Funktionen und Optionen angesprochen werden, die bei der Bearbeitung der Testataufgaben benötigt werden.

### 1. Installation *Python* und *Jupyter Notebooks*

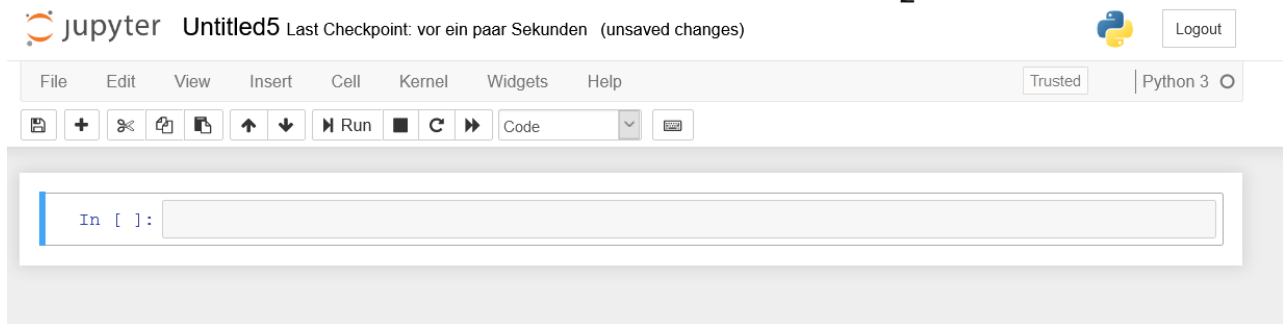
Ein Jupyter Notebook vereint ausführbaren Python-Code und seine Ausgaben und weiteren Text in einem einzigen Dokument. Zusätzlich können in einem solchen Dokument auch Graphiken, beschreibender Text, mathematische Gleichungen und andere Bestandteile vorkommen. Ein Beispiel für ein Jupyter Notebook finden Sie unter `einf_daten.ipynb`.

Jupyter Notebooks sind kostenlos und lassen sich am Besten über das „[Anaconda data science toolkit](#)“ herunterladen. Nach einem erfolgreichen Herunterladen des „Anaconda data science toolkits“ sollten Sie unter Ihrem Eingangsmenü unter dem Menüpunkt „Anaconda“ auch den Eintrag „Jupyter Notebook“ sehen.

Durch Anklicken öffnet sich im Browser ein Menü Ihrer Dateien:



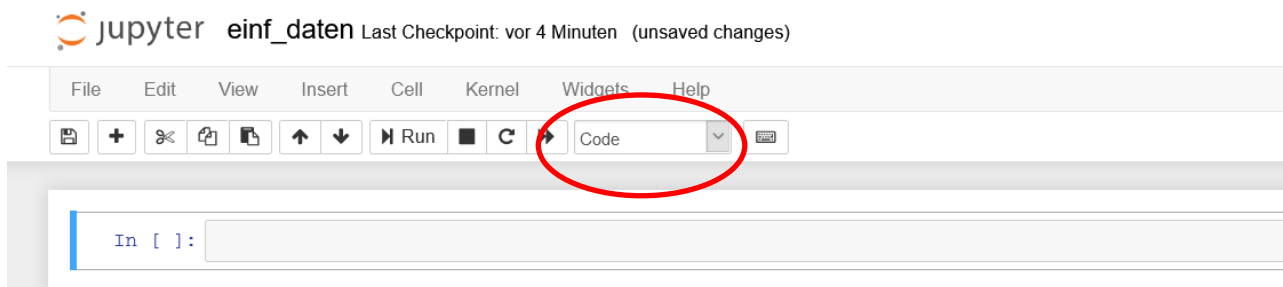
Über den Menüpunkt **New** kann ein neues Notebook mit **Python 3** aufgerufen werden. Ein neues Notebook hat zu Beginn die folgende Form:



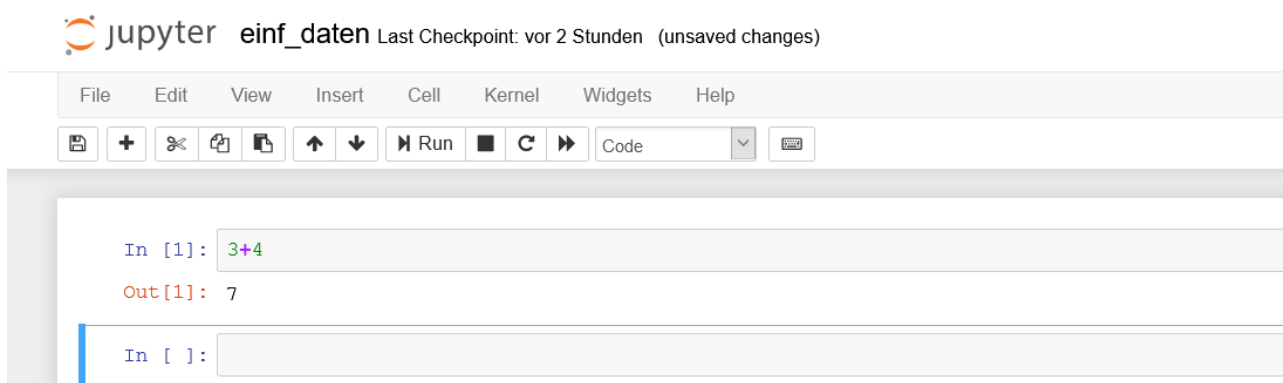
Speichern Sie ihr Jupyter Notebook unter dem Namen `einf_datens.jupyter` in einem entsprechenden Laufwerk ab.

Ein Jupyter Notebook stellt innerhalb einer Webanwendung alle Inhalte eines Python Programms dar, dazu gehören Eingabedaten, Ausgabedaten, erklärender Text, mathematische Formeln oder Bilder. Die unterschiedlichen Eingabefelder können, wie folgt beschrieben werden:

**Code Cells:** Eine Code Cell erlaubt es Code zu schreiben und darzustellen und später auch auszuführen.



Das Resultat einer Code Cell wird nach der Ausführung über den Run Button in einer Output Cell dargestellt.



**Markdown Cells:** In einer Markdown Cell können Text, Erklärung und Erläuterungen stehen.

jupyter einf\_daten Last Checkpoint: vor 6 Stunden (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help

Run Markdown

```
In [1]: 3+4
Out[1]: 7
```

**Hierbei** handelt es sich um eine einfache *Addition*

Durch Betätigen des Run Buttons werden sie in normalen Text umgewandelt.

jupyter einf\_daten Last Checkpoint: vor 6 Stunden (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help

Run Code

```
In [1]: 3+4
Out[1]: 7
```

Hierbei handelt es sich um eine einfache *Addition*

In [ ]:

Um Überschriften in das Jupyter Notebook einzubinden können in Markdown Cells auch 1 bis 6 Hashtags # verwendet werden.

jupyter einf\_daten Last Checkpoint: vor 6 Stunden (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help

Run Markdown

```
In [1]: 3+4
Out[1]: 7
```

Hierbei handelt es sich um eine einfache *Addition*

## # Beispielaufgabe 1

jupyter einf\_daten Last Checkpoint: vor 6 Stunden (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help

Run

```
In [1]: 3+4
```

```
Out[1]: 7
```

Hierbei handelt es sich um eine einfache *Addition*

## Beispielaufgabe 1

```
In [ ]:
```

Außerdem können für mathematische Formeln auch LaTeX-Befehle verwendet werden.

jupyter einf\_daten Last Checkpoint: vor 6 Stunden (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help

Run

```
In [1]: 3+4
```

```
Out[1]: 7
```

Hierbei handelt es sich um eine einfache *Addition*

## Beispielaufgabe 1

```
$$\bar{x}=\frac{1}{n}(x_1+x_2+\dots+x_n)$$
```

```
In [ ]:
```

Durch Drücken des Run Buttons erhält man die folgende Ausgabe.

jupyter einf\_daten Last Checkpoint: vor 6 Stunden (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help

Run

```
In [1]: 3+4
```

```
Out[1]: 7
```

Hierbei handelt es sich um eine einfache *Addition*

## Beispielaufgabe 1

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \dots + x_n)$$

```
In [ ]:
```

Im Folgenden werden vier Beispielaufgaben behandelt, deren Bearbeitung ihnen auch für die eigentlichen Laboraufgaben helfen kann.

### **Beispiel 1**

In diesem Beispiel sollen die Noten und Punktzahlen einer Klausur ausgewertet werden. Die dazu erforderlichen Daten liegen noch in der Datei `einf_bl.txt` in drei Feldern (laufende Nummer, Punktzahl, Note), die durch Leerzeichen getrennt sind. Die Daten sollen in das Jupyter Notebook `einf_daten.ipynb` geladen werden und dort weiterbearbeitet werden.

Folgende Schritte sollen durchgeführt werden:


Textdatei einlesen

- a) Öffnen Sie das Notebook `einf_daten.ipynb`
- b) Lesen Sie die Daten aus der Datei `einf_bl.txt` in das Notebook ein. Verwenden Sie dazu die Bibliothek **Pandas**.

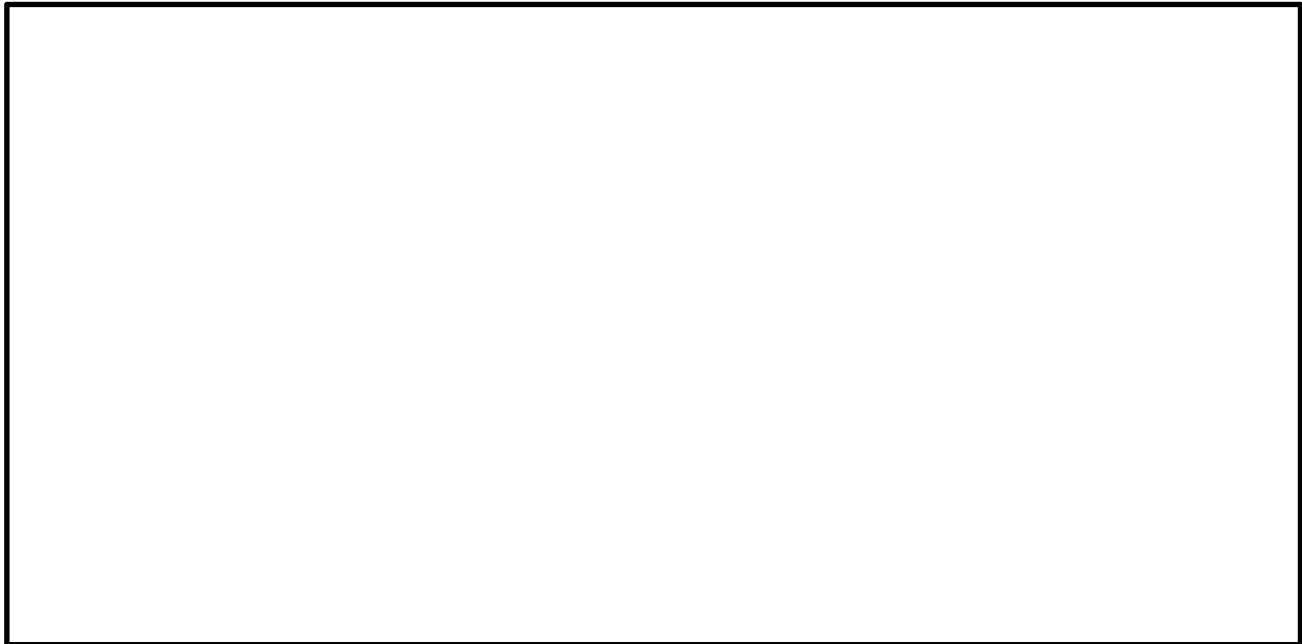
Häufigkeiten und Prozentanteile berechnen

- c) Berechnen Sie die Häufigkeiten der Noten „sehr gut“ (1,0 und 1,3), „gut“ (1,7 bis 2,3), „befriedigend“ (2,7 bis 3,3), „ausreichend“ (3,7 und 4,0) und „mangelhaft“ (4,7 und 5,0).

d) Ergänzen Sie bei c) die Summe.

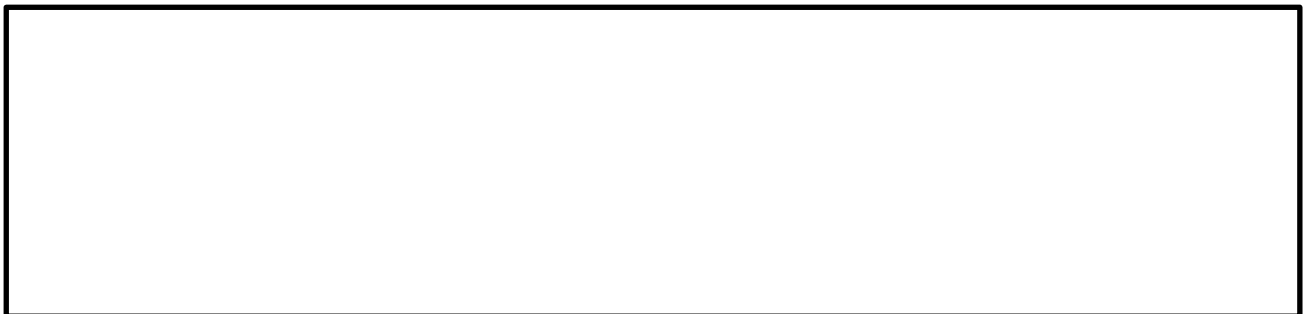


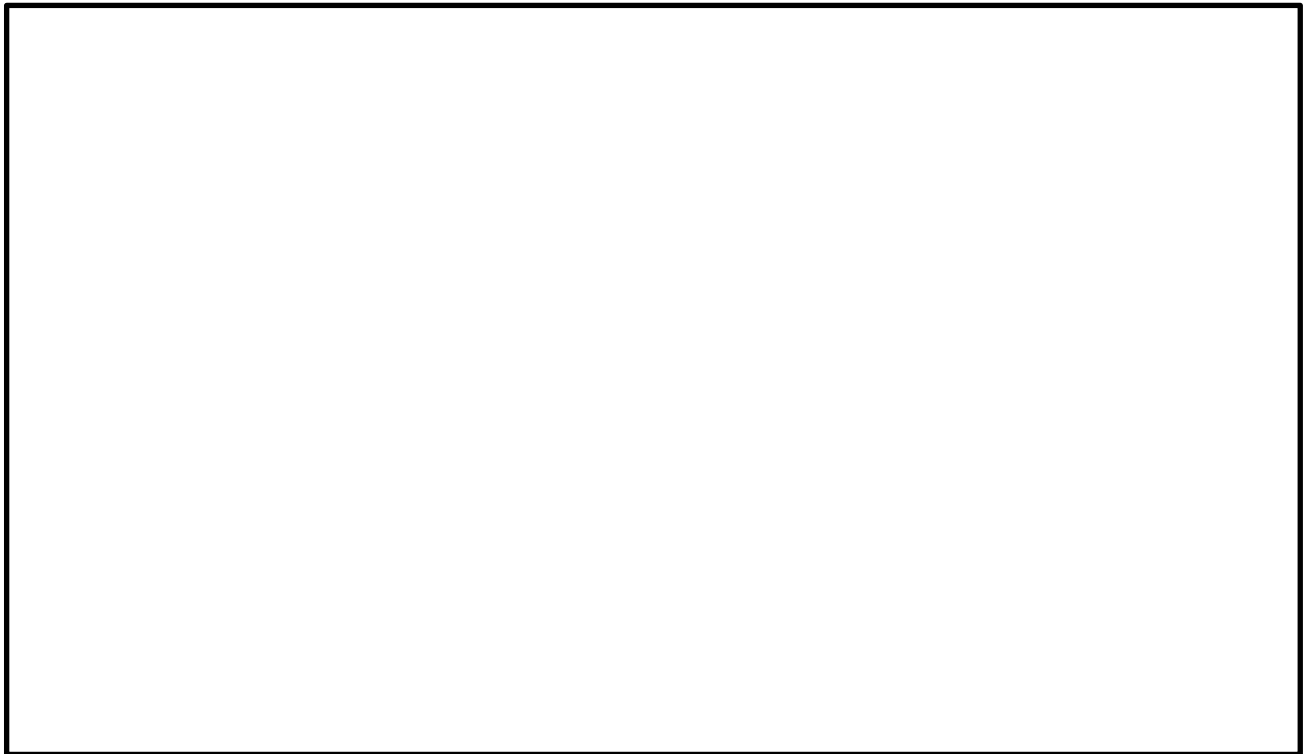
- e) Berechnen Sie, welche prozentualen Anteile auf die Notenstufen „sehr gut“ bis „mangelhaft“ entfallen. Geben Sie die Prozentzahlen als ganze Zahlen (ohne Nachkommastellen) an.



- f) Säulendiagramm erstellen

Stellen Sie die Häufigkeiten aus c) mit einem Säulendiagramm dar. Geben Sie dem Diagramm einen passenden Titel und beschriften Sie die Achsen.





g) Führen Sie in dem Diagramm die folgenden Umformatierungen durch:

- g1) Die Farbe der Säulen soll dunkelblau sein.
- g2) Jede Säule soll mit der zugehörigen Häufigkeit (wie oft gab es diese Note?) beschriftet sein.



Kennzahlen berechnen

- h) Berechnen Sie Mittelwert, empirische Varianz und empirische Standardabweichung der Punktzahlen und geben Sie sie mit 4 Nachkommastellen an.

--



- i) Berechnen Sie Median und Spannweite der Punktzahlen, ohne die Punktzahlenliste zu sortieren. (Spannweite = größter Datenwert minus kleinster Datenwert.)

- j) Speichern Sie die geänderte Datei einf\_daten.xls in Ihr persönliches Verzeichnis ab.

## Beispiel 2

Die Daten, die diesem Beispiel zugrunde liegen, sind Angaben über die Weltproduktion von Mais (Körnermais) in Millionen Tonnen. Sie stehen in der Datei **Maisproduktion.txt** in Moodle im Ordner **Einführung** (Quelle der Daten: Deutsches Maiskomitee; Stand: Oktober 2005.).

Streudiagramm zeichnen

- a) Erstellen Sie ein Streudiagramm der Maisdaten. Legen Sie dabei die Jahreszahlen auf die x-Achse und die Maisproduktion auf die y-Achse. Geben Sie dem Diagramm einen passenden Titel und beschriften Sie Achsen.



Regressionsgerade einzeichnen

- b) Zeichnen Sie in Ihr Diagramm aus a) die lineare Regressionsgerade ein. Geben Sie die Gleichung der linearen Regressionsgerade und das zugehörige Bestimmtheitsmaß  $R^2$  an

Korrelationskoeffizienten berechnen

- c) Berechnen Sie den empirischen Korrelationskoeffizienten  $r$  zwischen Jahr und produzierter Maismenge. Geben Sie  $r$  mit 4 Nachkommastellen an.

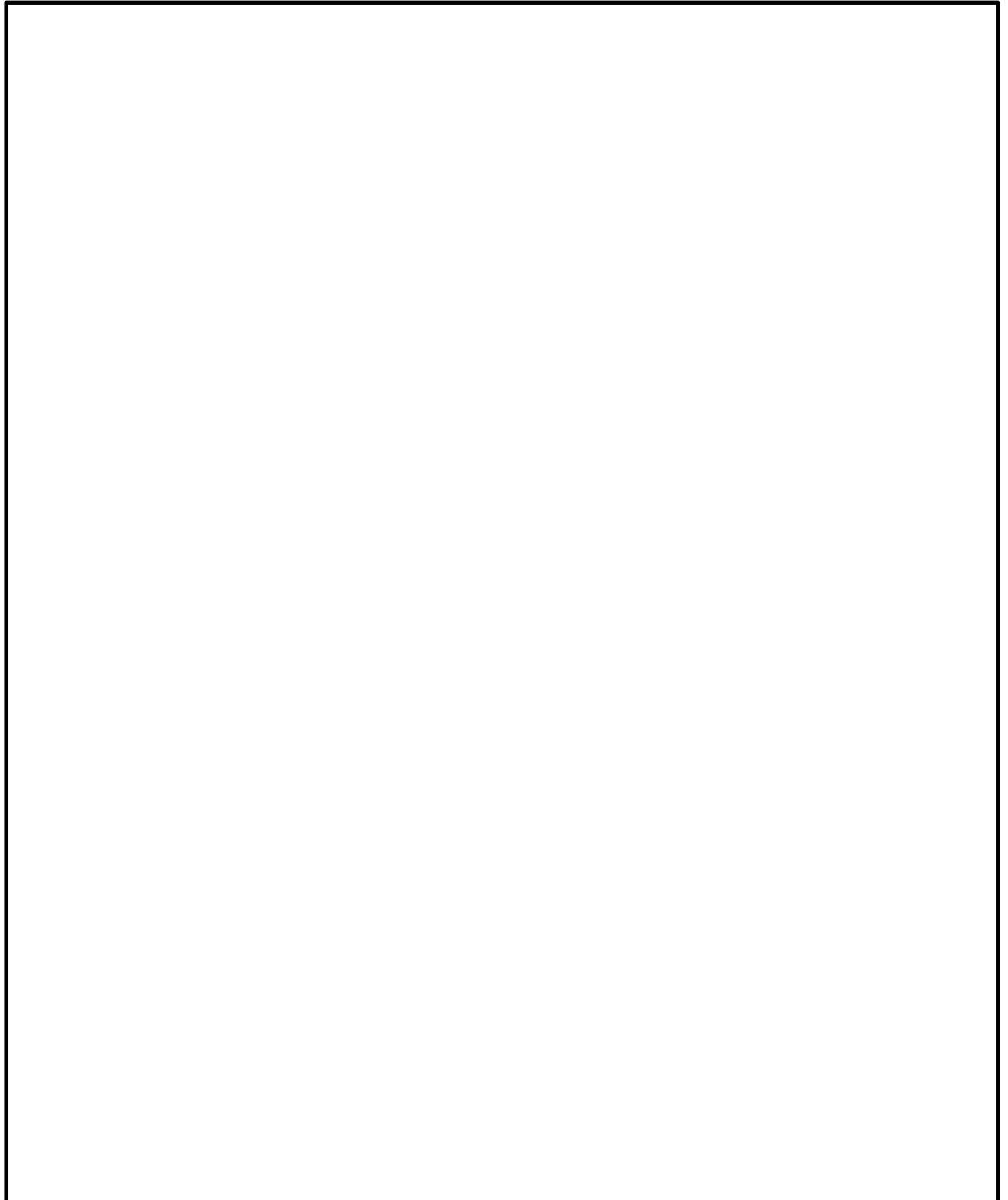
- d) Speichern Sie die geänderte Datei `einf_daten.ipynb` in Ihr persönliches Verzeichnis ab.

Andere Regressionskurven ausprobieren

- e) Ändern Sie den Typ der Regressionskurve von linear in quadratisch.

- f) Das Bestimmtheitsmaß  $R^2$  gibt an, wie gut die Regressionskurve die Punktwolke beschreibt (0 = gar nicht, 1 = alle Datenpunkte liegen auf der Regressionskurve). Bei quadratischer Regression ist  $R^2$  größer als bei linearer Regression. Warum ist bei diesem Datensatz trotzdem eine lineare Regression sinnvoller als eine quadratische?

- g) Probieren Sie außerdem eine Regression mit einem Polynom sechsten Grades. Was stellen Sie hier fest?



### Beispiel 3

In diesem Beispiel lernen Sie einige Statistik-Funktionen kennen, mit denen man Berechnungen der wichtigsten **diskreten Wahrscheinlichkeitsverteilungen** (hypergeometrische Verteilung, Binomialverteilung, Poissonverteilung) durchführen kann. Genaueres über diese Verteilungen erfahren Sie in der Vorlesung.

#### Hypergeometrische Verteilung

- a) Sie erhalten eine Lieferung von 50 elektronischen Bauteilen. Daraus entnehmen Sie eine Stichprobe von 20 Bauteilen und testen diese 20 Bauteile auf Funktionsfähigkeit. Die Zufallsvariable  $X$  gebe die Anzahl der defekten Bauteile unter den 20 Bauteilen der Stichprobe an. Angenommen, in der Lieferung sind 5 defekte elektronische Bauteile. Unter diesen Annahmen folgt  $X$  einer so genannten hypergeometrischen Verteilung  $X \sim H(20; 50; 5)$ . Berechnen Sie hierfür

a1) die Wahrscheinlichkeit, dass in Ihrer Stichprobe kein defektes Bauteil ist;

a2) die Wahrscheinlichkeit, dass in Ihrer Stichprobe genau 1 defektes Bauteil ist;

a3) die Wahrscheinlichkeit, dass in Ihrer Stichprobe genau 2 defekte Bauteile sind;

a4) die Wahrscheinlichkeit, dass in Ihrer Stichprobe genau 3 defekte Bauteile sind;

a5) die Wahrscheinlichkeit, dass in Ihrer Stichprobe höchstens 3 defekte Bauteile sind.

### Binomialverteilung

b) Bei der Massenproduktion bestimmter elektronischer Kleinteile entsteht eine Ausschussquote von 10 %. Sie entnehmen der laufenden Produktion eine Stichprobe vom Umfang 20. Man kann davon ausgehen, dass hierbei verschiedene Stichprobenteile unabhängig voneinander defekt sind. Die Zufallsvariable  $X$  gebe die Anzahl der defekten Kleinteile unter diesen 20 Teilen an. Unter den genannten Annahmen folgt  $X$  einer so genannten Binomialverteilung  $X \sim B(20; 0.1)$ . Berechnen Sie hierfür

b1) die Wahrscheinlichkeit, dass in Ihrer Stichprobe genau 3 defekte Kleinteile sind;

b2) die Wahrscheinlichkeit, dass in Ihrer Stichprobe höchstens 3 defekte Kleinteile sind;

### Poissonverteilung

c) Bei der Produktion einer bestimmten Textilart entstehen zufallsbedingt Gewebefehler. Im Mittel sind es 2 Gewebefehler auf  $1 \text{ m}^2$ . Sie entnehmen zufällig ein Textilstück von  $1 \text{ m}^2$  und zählen, wie viele Gewebefehler auf diesem Stück sind. Die Zufallsvariable  $X$  gebe die Anzahl festgestellter Gewebefehler an. Unter den genannten Annahmen folgt  $X$  einer so



genannten Poissonverteilung  $X \sim Po(2)$ , dabei ist  $\lambda = 2$  der Erwartungswert von  $X$  (mittlere, d. h. erwartete Anzahl von Fehlern). Berechnen Sie hierfür

c1) die Wahrscheinlichkeit, dass auf einem Textilstück genau 3 Gewebefehler sind;

c2) die Wahrscheinlichkeit, dass auf Ihrem Textilstück höchstens 3 Gewebefehler sind.

#### Beispiel 4

In diesem Beispiel lernen Sie einige Statistik-Funktionen kennen, mit denen man Berechnungen bei der wichtigsten stetigen Wahrscheinlichkeitsverteilung, nämlich der Normalverteilung, durchführen kann. Genauer über die Normalverteilungen erfahren Sie später in der Vorlesung.

Eine Maschine füllt Zucker in Packungen. Die Füllmenge variiert zufällig. Die Zufallsvariable  $X$  gebe die Füllmenge [in g] einer zufällig ausgewählten Zuckerpackung an. Wir gehen in diesem Beispiel davon aus dass die Zufallsvariable  $X$  einer Normalverteilung  $X \sim N(1000;9)$  folgt. In diesem Beispiel ist also der Erwartungswert der Füllmenge  $\mu=1.000$  [g], Varianz der Füllmenge  $\sigma^2=9$  [g<sup>2</sup>] und Standardabweichung der Füllmenge  $\sigma=3$  [g].

a) Berechnen Sie die Wahrscheinlichkeit, dass die Füllmenge einer zufällig ausgewählten Zuckerpackung bei höchstens 994 g liegt.

b) Berechnen Sie das 1-%-Quantil der Normalverteilung . Das ist diejenige Füllmenge, die von einer zufällig ausgewählten Zuckerpackung nur mit einer Wahrscheinlichkeit von 0,01 unterschritten wird.

c) Mit welcher Funktion können die Quantile der so genannten t-Verteilung berechnen kann? (Die Quantile der t-Verteilung werden in Kapitel 5 der Vorlesung genauer erläutert.)

d) Zeichnen Sie die Dichtefunktion der Zufallsvariablen  $X$  im Intervall  $[990, 1010]$ .