5.1 Testaufgabe:

Nach der Lektüre eines neuen Romans der Autorin X haben Sie den Eindruck, dass die Autorin in diesem Werk mehr wörtliche Rede einsetzt, als in ihren früheren Romanen. Nehmen wir an, Sie haben eine Software, mit der Sie zählen können, wie viele Wörter im Buch wörtliche Rede sind und wie viele nicht.
Formulieren Sie Nullhypothese und Gegenhypothese für eine Operationalisierung Ihrer Annahme.

- Gegenhypothese H_1 (Annahme): Die Autorin setzt mehr wörtliche Rede in ihrem neuen Roman ein, als in ihren früheren Romanen => erster Schritt der Operationalisierung der Annahme geht davon aus, dass: Die Wahrscheinlichkeit für das Auftreten von mehr wörtlicher Rede im neuen Roman ist höher als im alten Roman:
  $\text{WörtlicheRede}_{neu} \neq \text{WörtlicheRede}_{alt}$
- Nullhypothese H_0: Trifft die Annahme nicht zu, gibt es nur eine einzige Alternative und das hieße es bliebe alles beim Alten, dass auch im neuen Roman genauso viel wörtliche Rede vorkommt, wie im alten Roman:
  $\text{WörtlicheRede}_{neu} = \text{WörtlicheRede}_{alt}$

5.2 Übungsaufgabe:

Kommen wir zurück zum *Genderbias*-Problem. Wenn wir einen *Bias* annehmen, wäre das Gegenteil unserer Annahme, dass verbliebene Ungleichgewichte zufällig zu Stande kommen. Unsere Nullhypothese
$H_0 : p_{weibliche\ Figur} = p_{männliche\ Figur}$
wird also exakt durch das Beispiel mit der geworfenen Münze abgebildet: Jedes Mal, wenn der Autor eine neue Figur einführt, ist die Wahrscheinlichkeit für jede der beiden Kategorien gleich hoch.

(A) Wir gehen von einem Drama mit 5 Figuren aus. Schreiben Sie alle potentiellen Kombinationen auf, die auftreten können, wenn Sie nacheinander für jede dieser Figuren die Genderzuordnung per Zufall (=Münzwurf) bestimmen. In wie vielen dieser Kombinationen kommt eine Kategorie (männlich oder weiblich) nur 1 mal oder weniger vor?

[m m m m m], [w m m m m], [m w m m m], [m m w m m], [m m m w m], [m m m m w], [w w w w w],

[m w w w w], [w m w w w], [w w m w w], [w w w m w], [w w w w m]

$\text{Möglichkeiten}_{gesamt}$: $2^5 = 32$

Möglichkeiten für (m oder w) <= 1: 12

(B) Wie hoch ist die Wahrscheinlichkeit für ein Genderungleichgewicht von ***mindestens*** *4:1 (ergo 4:1 oder 5:0) wenn wir davon ausgehen (*H0*), dass beide Zuordnungen eigentlich gleich wahrscheinlich sind?*

12/32 = 37,5 %

(C) Sie können die geworfene Münze auch in Python mit einem Zufallsgenerator simulieren. Laden Sie das Paket random *und nutzen Sie die Funktion* random.sample()*, um gemäß* H0 *eine hypothetische Genderverteilung für ein Drama mit 20 Figuren zu erzeugen.*

s. Anhang!

(D) Bauen Sie eine for*-Schleife, um 1000 Genderverteilungen über 20 Figuren zu erzeugen. Berechnen sie für jedes dieser "Dramen" das Genderverhältnis, am besten als Prozentwert, und speichern Sie alle 1000 Ergebnisse in einer Liste, oder besser noch in einem Numpy-Array oder einer Pandas-Series. Können Sie anhand der Quantile dieser Datenreihe sagen, wie viele ihrer Durchläufe ein Ungleichgewicht von* ***mindestens*** *15:5 haben?*

ab dem Quantil 45.0 gibt es ein Ungleichgewicht von 5:15

5.3 Testaufgabe:

(A) Ermitteln Sie mit dem Binomialtest, ob es unter den Figuren in Ihrem Drama einen statistisch signifikanten Genderbias gibt.

id	label	gender	role	importance	per_mes_sps
1	Ghost	male	other	primary	0
2	Hamlet	male	protagonist	primary	358
3	Gertrude	female	other	secondary	69
4	Claudius	male	antagonist	primary	102
5	Ophelia	female	lover	primary	58
6	Laertes	male	antagonist	primary	62
7	Polonius	male	other	secondary	86
8	Reynaldo	male	other	minor	13
9	Horatio	male	other	secondary	109
10	Voltemand	male	other	minor	1
11	Cornelius	male	other	minor	1
12	Rosencrantz	male	other	minor	48
13	Guildenstern	male	other	minor	29
14	Osric	male	other	minor	25
16	A Lord	male	other	minor	3
17	Francisco	male	other	minor	8
18	Bernardo	male	other	minor	19
19	Marcellus	male	other	minor	37
20	Fortinbras	male	other	minor	6
21	Captain in Fortinbras Army	male	other	minor	7
26	Gravedigger = First Clown	male	other	minor	34
27	Gravedigger's companion = 2nd Clown	male	other	minor	12
28	Doctor of Divinity = Priest	male	other	minor	2
nz	Gentlemen	male	other	minor	1
nz	Ambassodors to Denmark from Englan	male	other	minor	1
nz	Players who take Proglogue	male	other	minor	1
nz	Two Messengers	male	other	minor	1
nz	Sailors	male	other	minor	2
nz	Atenndants, Lords, Guards, Musicians,	male	other	minor	0

Zunächst einmal ist vor der Berechnung über Python eine Festlegung der Personenzahl nötig. Laut Personentabelle sind nicht nur Einzelpersonen im Drama vertreten sondern auch Ansammlungen von Personen (Lord, Guards, Spieler, Gesandte etc). Diese Ansammlungen werden in die Statistik nicht miteinberechnet, da sie einfach nicht zählbar sind. Der Fokus liegt also nur auf zählbaren Personen mit eindeutig erkennbarem Geschlecht (s. Screenshot -> rot markiert ist unzählbar und fällt somit weg). Wir haben im Hamlet damit also 23 zählbare Personen, darunter 21 Männer und 2 Frauen. Bezogen auf den Genderbias sollte zunächst eine Annahme formuliert werden. Diesbezüglich hätte ich ohne Hamlet gelesen zu haben, geschätzt, dass speziell in Shakespeares Werken aufgrund der damaligen Zeit gerade einmal 5% weiblich der Akteure weiblich sind. Der Test in Python ergibt dann wie folgt (Screenshot darunter bzw. Datei *aufgabe05_52_53.pdf*):

```
binom = stats.binom_test(2, n=23, p=0.05, alternative='greater')
print(binom)
```

```
0.32057955555520995
```

(B) Ermitteln Sie mit dem Binomialtest, ob es unter den Autorinnen und Autoren, zu denen Sie Lebensdaten gesammelt haben, einen statistisch signifikanten Genderbias gibt.

Insgesamt sind es 20 Autoren (17 männlich, 3 weiblich)

	Autor	Geburtsjahr	Todesjahr	Geschlecht
2	Adabert Stifter	1805	1868	m
3	Charles Dickens	1812	1870	m
4	Conrad Ferdinand Meyer	1825	1898	m
5	Émile Zola	1840	1902	m
6	Friedrich Hebbel	1813	1863	m
7	Gottfried Keller	1819	1890	m
8	Gustav Freytag	1816	1895	m
9	Gustave Flaubert	1821	1880	m
10	Henrik Ibsen	1828	1906	m
11	Herman Melville	1819	1891	m
12	Lew Nikolajewitsch Tolstoi	1828	1910	m
13	Mark Twain	1835	1910	m
14	Octave Mirbeau	1848	1917	m
15	Theodor Fontane	1819	1898	m
16	Theodor Storm	1817	1888	m
17	Wilhelm Busch	1832	1908	m
18	Wilhelm Raabe	1831	1910	m
19	Božena Němcová	1820	1862	w
20	Emily Brontë	1818	1848	w
21	Marie von Ebner-Eschenbach	1830	1916	w

Unter der Annahme, dass berühmte Autorinnen damals eher selten vorkamen, wird die These „weniger 10% der Autorinnen waren berühmt“ formuliert. Dies ergibt einen Wert von:

```
binom = stats.binom_test(3, n=20, p=0.1, alternative='greater')
print(binom)
```

```
0.32307319481053404
```

(C) Wie viele Autoren muss Ihr Datensatz mindestens haben, damit der P-Wert überhaupt ein Signifikanzniveau von 5% erreichen kann? Stellen Sie sich vor, Sie werfen eine Münze und bekommen immer das gleiche Ergebnis. ab dem wievielten Wurf ist die Beobachtung statistisch signifikant?

s. Datei *aufgabe05_52_53.pdf*

5.4 Übungsaufgabe:

Wir haben nun eine Prozedur kennen gelernt, um Hypothesen zu prüfen.

1. Annahme operationalisieren.

2. Nullhypothese finden.

3. Mit Hilfe eines Signifikanztests entscheiden, ob die Nullhypothese verworfen werden sollte oder die ursprüngliche Annahme.

Nehmen wir mal an, sie haben in allen drei Schritten alles richtig gemacht und sind zu einer Entscheidung (H_0 oder H_1) gelangt. Die Ergebnisse Ihrer Studie sind klar und eindeutig.

(A) Kann es trotzdem sein, dass Sie sich irren, auch wenn ihre Studie handwerklich einwandfrei ist? Welche Fehler können immer noch passieren?

Messfehler, nicht alle nötigen Daten konnten gesammelt werden, Daten fehlerhaft/beschädigt, Ausreißer

(B) Gibt es Möglichkeiten, das Risiko der unter (A) genannten Fehler zu reduzieren?

mehr Daten sammeln, Daten genau prüfen (Quellen prüfen, Messwerte auf Abweichungen prüfen)

(C) Können wir die Wahrscheinlichkeit solcher Fehler irgendwie abschätzen?

Nach dem Erkennen dokumentieren, wie oft bestimmte Fehler auftreten und dies dann in zukünftigen Experimenten/Tests berücksichtigen, thematische Verteilung schätzen und mögliche Ausreißer vorher berücksichtigen