Exercise 1.a)

We are given a text of $n = 1'000'000$ words. The word *Agatha* occurs 30 times, *Christie* has 117 occurrences and the bigram *Agatha Christie* occurs 20 times.

From the given data we can directly compute the probabilities

$$\begin{aligned} P(Agatha) &= \frac{30}{1'000'000} = 3 \cdot 10^{-5}, \\ P(Christie) &= \frac{117}{1'000'000} = 0.000117. \end{aligned}$$

Following the Null-Hypothesis (H_0) of *Agatha* and *Christie* being independent, we can estimate the probability of the bigram *Agatha Christie* as

$$p_0 = P(Agatha) \cdot P(Christie) = 3.51 \cdot 10^{-9}.$$

However, directly computing the probability of the bigram from the given data yields

$$p = P(Agatha\ Christie) = \frac{20}{1'000'0000} = 2 \cdot 10^{-5}.$$

If we see this as a Bernoulli process, the mean and variance are given by

$$\begin{aligned} \bar{X} &= p = 2 \cdot 10^{-5}, \\ s^2 &= p \cdot (1 - p) \approx 1.2 \cdot 10^{-5}. \end{aligned}$$

The mean value from the Null-Hypothesis is given by

$$\mu_0 = p_0 = 3.51 \cdot 10^{-9}$$

Therefore, we get a t-value of

$$t_{obs} = \frac{\bar{X} - \mu_0}{\sqrt{\frac{s^2}{n}}} \approx 4.471$$

From the normal table, with a significance level of 1% and dof[1] $= \infty$, we can get the critical value $t_{lim} = 2.576$. The fact that $t_{obs} > t_{lim}$ means that the Null-Hypothesis is rejected. Therefore, the words *Agatha* and *Christie* are **not** independent.

[1] dof = degrees of freedom