Wiederholung: Was bisher geschah . . .

1. Erhebung von Daten im Rahmen von Studien

In der Medizin sind nur **prospektiv kontrollierte Studien mit Randomisierung** zum Nachweis der Wirksamkeit von Medikamenten zugelassen.

Alle anderen Studien können verfälscht sein durch **konfundierende Faktoren**.

Beispiel: PISA-Studie

Z.B. unterscheiden sich die einzelnen Ländern hinsichtlich der Muttersprache der Schüler mit Migrationshintergrund . . .

Seltsam: Mit solchen Studien kann man zwar begründen, dass man das Bildungssystem reformieren muss, aber kein neues Medikament zulassen . . .

2. Erhebung von Daten im Rahmen von Umfragen

Die Ergebnisse von Umfragen können systematisch verfälscht werden durch

- "untypische" Auswahl der Befragten aus der Menge aller interessierenden Personen (**sampling bias**). Am besten (aber oft nicht wirklich möglich) ist hier eine rein zufällige Auswahl der Befragten . . .

- Verweigerung der Teilnahme an der Befragung (**non-response bias**).

Beispiel: Wahlumfragen

Hier lassen sich weder sampling bias noch non-response bias völlig vermeiden.

3. Beschreibende Statistik

Datenmengen werden durch wenige Zahlen (**statistische Maßzahlen** wie z.B. (empirisches) arithmetisches Mittel bzw. (empirische) Varianz) oder Abbildungen (wie z.B. **Histogramme** oder **Boxplots**) beschrieben.

Beispiel: Sprechen Frauen mehr als Männer ?

Im Rahmen einer Studie an der Universität Arizona wurden bei 210 Studentinnen und 186 Studenten approximativ die Anzahl der gesprochenen Worte über einen Zeitraum von mehreren Tagen bestimmt. Für die empirischen arithmetischen Mittel der Anzahlen der gesprochenen Wörter pro Tag ergab sich:

Frauen: 16215
Männer: 15669

Also haben hier die betrachteten Studentinnen im Durchschnitt etwas mehr Wörter pro Tag gesprochen als die betrachteten Studenten.

4. Modellierung von Datensätzen mit Hilfe der Wahrscheinlichkeitstheorie

Frage im vorigen Beispiel:
Wie kann man die Aussage über den betrachteten Datensatz hinaus verallgemeinern?

Vorgehen in der Statistik (vereinfacht):

1. Fasse den Datensatz als Realisierung von Zufallsvariablen auf.

2. Wähle ein (von Parametern, also reellen Zahlen) abhängendes Modell für die Verteilung dieser Zufallsvariablen.

3. Passe Parameter der Verteilung dieser Zufallsvariablen an die beobachteten Daten an.

4. Beantworte die betrachtete Frage durch Betrachtung der Verteilungen der Zufallsvariablen.

Zu 1: Wir fassen die Beobachtungen (Anzahl gesprochener Wörter pro Tag und pro Versuchsperson) auf als Ergebnis eines **Zufallsexperiments:**

Dieses hat ein **unbestimmtes Ergebnis** $X(\omega) \in \mathbb{R}$, und für große Anzahlen von Wiederholungen nähert sich für jedes **Ereignis** $A \subseteq \mathbb{R}$ die relative Häufigkeit des Auftretens eines Ergebnisses, dass in der Menge A liegt, einer Zahl

$$\mathbf{P}[X \in A] \in [0, 1]$$

(sog. **Wahrscheinlichkeit von** A) an.

Z.B. $X =$ Anzahl der von einer (zufällig ausgewählten) Frau gesprochenen Wörter an einem (zufällig ausgewählten) Tag,

$\mathbf{P}[X \in [16000, 16500]] =$ Wahrscheinlichkeit, dass eine (zufällig ausgewählte) Frau an einem (zufällig ausgewählten) Tag zwischen 16000 und 16500 Wörtern spricht.

Zu 2: Die Zuordnung von Wahrscheinlichkeiten zu Mengen, also

$$A \mapsto \mathbf{P}[X \in A],$$

heißt **Verteilung** der sogenannten **Zufallsvariablen** X.

Eine Möglichkeit, solche Verteilungen festzulegen, sind **diskrete Verteilungen mit Zähldichte**. Bei diesen wird für jedes $k \in \mathbb{N}_0$ die Wahrscheinlichkeit

$$\mathbf{P}[X = k]$$

festgelegt. Anschließend bestimmen wir die Wahrscheinlichkeit, dass das unbestimmte Ergebnis in einer Menge $A \subseteq \mathbb{R}$ zu liegen kommt, als Summe der Wahrscheinlichkeiten aller natürlichen Zahlen in A:

$$\mathbf{P}[X \in A] := \sum_{k \in A \cap \mathbb{N}_0} \mathbf{P}[X = k].$$

Im obigen Beispiel können wir die folgenden diskreten Verteilungen verwenden:

1. Seien $n \in \mathbb{N}$ und $p \in [0,1]$. Bei einer **binomialverteilten ZV mit Parametern n und p** (kurz: $b(n,p)$-verteilte ZV) wird

$$\mathbf{P}[X=k] = \binom{n}{k} p^k (1-p)^{n-k} \text{ für } k \in \{0,\ldots,n\}, \mathbf{P}[X=k] = 0 \text{ für } k > n$$

gesetzt und alle weiteren Wahrscheinlichkeiten werden wie oben berechnet.

Man kann zeigen:

Hier liegt die Vorstellung zugrunde, dass die (zufällig ausgewählte) Frau am Tag genau n Möglichkeiten hat, ein einzelnes Wort zu sprechen, und jedes dieser Wörter unbeeinflusst voneinander mit Wahrscheinlichkeit p spricht.

2. Sei $\lambda > 0$. Bei einer **Poisson-verteilten ZV mit Parameter λ** (kurz: $\pi(\lambda)$-verteilte ZV) wird

$$\mathbf{P}[X = k] = \frac{\lambda^k}{k!} \cdot e^{-\lambda} \quad \text{für } k \in \mathbb{N}_0$$

gesetzt und alle weiteren Wahrscheinlichkeiten werden wie oben berechnet.

Man kann zeigen:

Dieses Modell kann als Approximation einer $b(n, p)$-Verteilung für n groß und p klein aufgefasst werden, sofern man $\lambda = n \cdot p$ setzt.

Bei einer stetig verteilten Zufallsvariablen mit Dichte wählen wir eine sogenannte Dichte $f : \mathbb{R} \to \mathbb{R}$, also eine Funktion $f : \mathbb{R} \to \mathbb{R}$ mit

$$\mathbf{f(x) \geq 0} \quad \text{für alle } \mathbf{x \in \mathbb{R}} \quad \text{und} \quad \int_{-\infty}^{\infty} \mathbf{f(x)\,dx = 1},$$

und bestimmen die Wahrscheinlichkeit, dass das unbestimmte Ergebnis in einer Menge $A \subseteq \mathbb{R}$ zu liegen kommt, als Flächeninhalt zwischen der Dichte und der x-Achse im Bereich der Menge A:

$$\mathbf{P[X \in A]} := \int_{\mathbf{A}} \mathbf{f(x)\,dx}.$$

Wählt man für f die Funktion

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (x \in \mathbb{R}),$$

so spricht man von einer sogenannten **Normalverteilung** mit Parametern μ und σ^2 ist.

Wir werden später sehen:
Summen von Zufallsvariablen der gleichen Art, die sich gegenseitig nicht beeinflussen, können durch Normalverteilungen approximiert werden, daher bietet sich im obigen Beispiel auch die Normalverteilung als Modell an.

Um Punkt 4 bearbeiten zu können (Punkt 3 folgt später) wollen wir beschreiben, wie groß der **Wert** ist, der sich bei wiederholter Durchführung des Zufallsexperiments für große Anzahl von Wiederholungen **im Mittel** approximativ ergibt.

Im Beipiel oben:
Wieviele Wörter sprechen Frauen im Durchschnitt am Tag, wenn wir immer wieder einzelne Frauen und einzelne Tage zufällig auswählen ?

Definition: Sei X eine diskrete Zufallsvariable, die mit Wahrscheinlichkeit Eins nur einen der Werte $x_1, x_2, \ldots, x_K \in \mathbb{R}$ bzw. $x_1, x_2, \cdots \in \mathbb{R}$ annimmt. Dann heißt

$$\mathbf{E}X = \sum_{k=1}^{K} x_k \cdot \mathbf{P}[X = x_k] \quad \text{bzw.} \quad \mathbf{E}X = \sum_{k=1}^{\infty} x_k \cdot \mathbf{P}[X = x_k]$$

der **Erwartungswert** von X.

Beispiel: Für eine $b(n,p)$-verteilte Zufallsvariable gilt

$$\begin{aligned}
\mathbf{E}X &= \sum_{k=0}^{n} k \cdot \binom{n}{k} p^k (1-p)^{n-k} \\
&= n \cdot p \cdot \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{(n-1)-(k-1)} \\
&= n \cdot p \cdot (p + (1-p))^{n-1} = n \cdot p,
\end{aligned}$$

da $(a+b)^m = \sum_{k=0}^{m} \binom{m}{k} a^k b^{m-k}$.

Folgerung: Sind die Anzahl der gesprochenen Wörter bei Frauen $b(n_f, p_f)$- und bei Männern $b(n_M, p_M)$-verteilt, so sprechen Frauen mehr als Männer, falls gilt:

$$n_f \cdot p_f > n_M \cdot p_M.$$

Beispiel: Für eine $\pi(\lambda)$-verteilte Zufallsvariable gilt

$$\begin{aligned}
\mathbf{E}X &= \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!} \cdot e^{-\lambda} \\
&= \lambda \cdot \left(\sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \right) \cdot e^{-\lambda} \\
&= \lambda \cdot e^{\lambda} \cdot e^{-\lambda} = \lambda,
\end{aligned}$$

da

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

Folgerung: Sind die Anzahl der gesprochenen Wörter bei Frauen $\pi(\lambda_F)$- und bei Männern $\pi(\lambda_M)$-verteilt, so sprechen Frauen mehr als Männer, falls gilt:

$$\lambda_F > \lambda_M.$$