Kapitel 2

Beschreibende Statistik

In den bisherigen Kapiteln waren Wahrscheinlichkeitsverteilungen explizit vorgegeben, und wir haben Kenngrößen dieser Verteilungen durch explizite Rechnung bestimmt. Jetzt betrachten wir die Situation von experimenteller Seite: Endlich viele Messungen bei einem Experiment ergeben eine sogenannte *empirische* Wahrscheinlichkeitsverteilung. Daraus möchte man dann — so gut wie möglich — auf die zugrundeliegende Wahrscheinlichkeitsverteilung schließen.

2.1 Stichprobe

In einem Experiment führen wir n Messungen durch. Man spricht auch von einer Stichprobe vom Umfang n. Die Werte $x_1, \ldots, x_n$ betrachten wir als Realisierungen einer Zufallsvariable X. Im allgemeinen werden Werte in der Stichprobe mehrfach auftreten. Um auf die Zufallsvariable X schließen zu können, ist es wichtig, dass die Messungen "unabhängig" voneinander sind. Zur Verdeutlichung besprechen wir ein Beispiel im Detail.

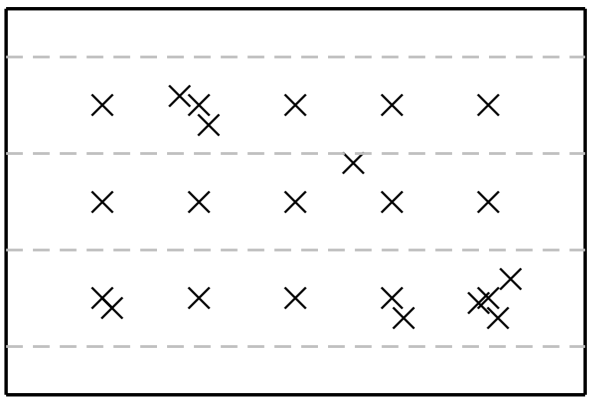

Abbildung 2.1: ein Bohnenfeld

Beispiel. Wir betrachten ein Bohnenfeld; unsere ZV X soll dabei der Anzahl der Bohnen pro Schote entsprechen. Messungen bei eng benachbarten Pflanzen sind i.a. nicht unabhängig, denn solche Pflanzen konkurrieren um Nährstoffe. Also beeinflussen sie sich gegenseitig! Ein möglicher Ausweg besteht darin, durch Würfeln die Pflanzen zufällig auszuwählen, deren Anzahl Bohnen pro Schote untersucht werden soll. In der Praxis ist Unabhängigkeit nicht immer leicht zu garantieren. Es gibt aber leider kein Patentrezept, um praktische Probleme statistisch handhabbar zu machen!

2.2 Skalenniveaus

Merkmale haben unterschiedliche mathematische Qualitäten. Man unterscheidet bei Messungen drei Skalen:

(i) **Nominalskala**: Merkmal qualitativer Art, ohne Rangordnung (z.B. Farbe, Geschlecht)

(ii) **Ordinalskala**: Rangordnung: Man kann für je zwei Werte sagen, welcher der größere ist, aber man kann nicht sagen, um wieviel größer. (z.B. Bachelor $<$ Master $<$ Promotion)

(iii) **Intervallskala**: quantitatives Merkmal: Man kann Differenzen bilden, und gleiche Differenzen bedeuten gleiche Abstände (Temperatur, Konzentration). Damit haben *Mittelwert* und *Varianz* eine Bedeutung! Diese Skala bildet die Mindestvoraussetzung für viele statistische Tests.

Gelegentlich spricht man auch von *Proportionalskalen*, dies sind Intervallskalen, bei denen auch die Bedeutung "doppelt so groß" inhaltlich sinnvoll ist (Körpergewicht, Fahrtdauer). *Absolutskalen* sind Intervallskalen, bei denen der Nullpunkt und die Einheiten festliegen (Studierende einer Veranstaltung). Siehe auch das entsprechende Übungsblatt.

2.3 Darstellung von Daten für diskrete Merkmale

Zur Darstellung von Daten für ein ordinal- oder intervallskaliertes Merkmal mit *diskreten* Ausprägungen geht man wie folgt vor.

Urliste :	$x_1, x_2, \ldots, x_n$ (ungeordnet)
Häufigkeitstabelle :	gleiche Werte zusammenfassen,
	der 'Größe nach' ordnen und Häufigkeit
	in Urliste notieren
	$\xi_1 < \xi_2 < \ldots < \xi_r,\ r \leq n$
absolute Häufigkeit :	$n_1,\ n_2, \ldots, n_r,\ \sum_j n_j = n$
relative Häufigkeit :	$h_j := \frac{n_j}{n}, \sum_j h_j = 1$

Für große Stichproben sollten die relativen Häufigkeiten die zugrundeliegenden Wahrscheinlichkeiten approximieren, wir erwarten eine Eigenschaft wie etwa

$$\text{“} \lim_{n\to\infty} h_j \;=\; \mathbb{P}(X = \xi_j) \text{”}.$$

Wir wollen hier nicht erklären, in welchem Sinn die obige Aussage mathematisch zu verstehen ist, sondern es bei einem anschaulichen Verständnis belassen. Wer sich genauer informieren möchte, schlägt unter dem Stichwort “Gesetz der großen Zahl” in den empfohlenen Lehrbüchern nach.

Beispiel. Wir besprechen das Bohnen-Beispiel 2.1. Es entspreche X der (diskreten) Anzahl von Samen pro Schote. Urliste: $1, 2, 1, 4, 5, 3, \ldots$ ($n = 50$ Werte)

ξ_j	1	2	3	4	5	6	
n_j	4	7	9	15	11	4	$\sum$ = 50
h_j	0.08	0.14	0.18	0.3	0.22	0.08	$\sum$ = 1

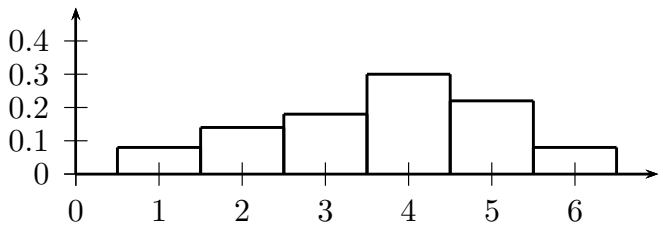

Abbildung 2.2: Säulendiagramm

Bemerkung. Bei kontinuierlichen ZVn oder bei einer großen Zahl verschiedener Messwerte ist eine Klassenbildung notwendig, d.h. eine Zusammenfassung von Messwerten in (halboffene) Intervalle. Dann heißt das Säulendiagramm meist **Histogramm**. Meistens wählt man die Intervalle gleich lang (äquidistant).

2.4 Kenngrößen einer Stichprobe

Wir besprechen zwei Kenngrößen einer intervallskalierten Stichprobe, ihren (empirischen) Mittelwert und ihre empirische Varianz.

Mittelwert $\overline{x}$

$$\overline{x} := \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{j=1}^{r} n_j\,\xi_j.$$

Der Mittelwert $\overline{x}$ ist ein Näherungswert (oder Schätzwert) für den Erwartungswert $\underline{m} = \mu$ der zugrundeliegenden Zufallsvariable. Wir erwarten (in einem später noch zu präzisierenden Sinn)

$$\text{“}\lim_{n\to\infty} \overline{x} \longrightarrow \mu\text{”},$$

wobei mit μ der Erwartungswert der zugrundeliegenden Zufallsvariablen gemeint ist, was dem $\underline{m}$ aus dem ersten Kapitel entspricht.

Empirische Varianz s^2

$$\begin{aligned} s^2 &:= \frac{1}{n-1}\sum_{i=1}^{n}\big(x_i - \overline{x}\big)^2 \\ &= \frac{1}{n-1}\left(\left(\sum_{i=1}^{n} x_i^2\right) - n\overline{x}^2\right) \\ &= \frac{1}{n-1}\sum_{j=1}^{r} n_j\big(\xi_j - \overline{x}\big)^2 \end{aligned}$$

Die empirische Varianz s^2 ist ein Näherungswert für die Varianz σ^2 der zugrundeliegenden Zufallsvariable. Wir erwarten (in einem später noch zu präzisierenden Sinn)

$$\text{“}\lim_{n\to\infty} s^2 \longrightarrow \sigma^2\text{”}.$$

Die Größe $s = \sqrt{s^2}$ heißt **empirische Standardabweichung**.

Warum taucht bei der empirischen Varianz der Vorfaktor $1/(n-1)$ anstelle von $1/n$ auf? Der Schätzwert $\overline{x}$ für den Erwartungswert μ wurde schon aus der Stichprobe ermittelt. Dies reduziert den Stichprobenumfang zur Varianzbestimmung um 1. Tatsächlich kann man leicht zeigen (siehe dazu die Übungen), dass der Ausdruck

$\sum(x_i - \overline{x})^2$ im Mittel kleiner ist als $\sum(x_i - \mu)^2$, und dass mit der angegebenen Normierung s^2 im Mittel gleich der Varianz σ^2 ist. (Man sagt: Die Größe s^2 ist ein erwartungstreuer Schätzer für die Varianz σ^2. Dies wird im Kapitel über schließende Statistik noch genauer thematisiert.)

Beispiel. Für das obige Bohnenbeispiel ergibt sich der folgende empirische Mittelwert.

$$\overline{x} \;=\; \frac{1}{50}(3+4+4+\ldots) \;=\; \frac{1}{50}(4\cdot 1 + 7\cdot 2 + 9\cdot 3 + \ldots) \;\approx\; 3.68.$$

Als empirische Varianz erhalten wir

$$s^2 \;=\; \frac{1}{49}\left(4\cdot(1-3.68)^2 + 7\cdot(2-3.68)^2 + \ldots + 4\cdot(6-3.68)^2\right) \;\approx\; 1.94.$$

Dies ergibt für die empirische Standardabweichung den Wert $s \approx 1.39$.