KAPITEL 1

Stichproben und Stichprobenfunktion

In diesem Kapitel werden wir auf Stichproben und Stichprobenfunktionen eingehen. Als Einstieg beginnen wir mit zwei kleinen Beispielen.

1.1. Stichproben

BEISPIEL 1.1.1. Wir betrachten ein Experiment, bei dem eine physikalische Konstante (z.B. die Lichtgeschwindigkeit) bestimmt werden soll. Da das Ergebnis des Experiments fehlerbehaftet ist, wird das Experiment mehrmals durchgeführt. Wir bezeichnen die Anzahl der Messungen mit n. Das Resultat der i-ten Messung sei mit $x_i \in \mathbb{R}$ bezeichnet. Fassen wir nun die Resultate aller Messungen zusammen, so erhalten wir eine sogenannte *Stichprobe*

$$(x_1, \ldots, x_n) \in \mathbb{R}^n.$$

Die Anzahl der Messungen (also n) nennen wir den *Stichprobenumfang*. Die Menge aller vorstellbaren Stichproben wird der *Stichprobenraum* genannt und ist in diesem Beispiel $\mathbb{R}^n$.

BEISPIEL 1.1.2. Wir betrachten eine biometrische Studie, in der ein gewisses biometrisches Merkmal, z.B. die Körpergröße, in einer bestimmten Population untersucht werden soll. Da die Population sehr groß ist, ist es nicht möglich, alle Personen in der Population zu untersuchen. Deshalb werden für die Studie n Personen, die wir mit $1, \ldots, n$ bezeichnen, aus der Population ausgewählt und gewogen. Mit $x_i \in \mathbb{R}$ wird das Gewicht von Person i bezeichnet. Das Ergebnis der Studie kann man dann in einer Stichprobe

$$(x_1, \ldots, x_n) \in \mathbb{R}^n$$

zusammenfassen. Die Auswahl der n Personen aus der Population erfolgt zufällig und kann somit als ein Zufallsexperiment betrachten werden. Die Grundmenge dieses Experiments sei mit Ω bezeichnet. Die genaue Gestalt von Ω wird im Weiteren keine Rolle spielen. Das Gewicht von Person i kann als eine Zufallsvariable $X_i : \Omega \to \mathbb{R}$ aufgefasst werden. Den Zusammenhang zwischen $(X_1, \ldots, X_n)$ und $(x_1, \ldots, x_n)$ kann man folgendermaßen beschreiben. Jede konkrete Auswahl von n Personen aus der Population entspricht einem Element (Ausgang) ω in der Grundmenge Ω. Das Gewicht der i-ten Person ist dann der Wert der Funktion X_i an der Stelle ω, also $X_i(\omega)$. Es gilt somit

$$x_1 = X_1(\omega), \ldots, x_n = X_n(\omega).$$

Man sagt auch, dass $(x_1, \ldots, x_n)$ eine *Realisierung* des Zufallsvektors $(X_1, \ldots, X_n)$ ist. Oft nennt man $(x_1, \ldots, x_n)$ die *konkrete Stichprobe* und $(X_1, \ldots, X_n)$ die *Zufallsstichprobe*. Es sei noch einmal bemerkt, dass x_i reelle Zahlen, wohingegen $X_i : \Omega \to \mathbb{R}$ Zufallsvariablen (also Funktionen auf einem Wahrscheinlichkeitsraum) sind.

Im Folgenden werden wir sehr oft annehmen, dass $X_1, \ldots, X_n : \Omega \to \mathbb{R}$ unabhängige und identisch verteilte Zufallsvariablen sind. Die Verteilungsfunktion von X_i bezeichnen wir mit

$$F(t) = \mathbb{P}[X_i \leq t], \quad t \in \mathbb{R}.$$

1.2. Stichprobenfunktionen, empirischer Mittelwert und empirische Varianz

DEFINITION 1.2.1. Eine beliebige Borel-Funktion $\varphi : \mathbb{R}^n \to \mathbb{R}^m$ heißt *Stichprobenfunktion.*

DEFINITION 1.2.2. Bezeichne mit $X = (X_1, \ldots, X_n) : \Omega \to \mathbb{R}^n$ eine Zufallsstichprobe. Dann heißt die zusammengesetzte Funktion $\varphi \circ X : \Omega \to \mathbb{R}^m$ eine *Statistik*:

$$\varphi \circ X : \Omega \to \mathbb{R}^n \to \mathbb{R}^m, \quad \omega \mapsto (X_1(\omega), \ldots, X_n(\omega)) \mapsto \varphi(X_1(\omega), \ldots, X_n(\omega)).$$

Im Folgenden werden wir zwei wichtige Beispiele von Stichprobenfunktionen, den empirischen Mittelwert und die empirische Varianz, betrachten. Es sei $(x_1, \ldots, x_n) \in \mathbb{R}^n$ eine Stichprobe.

DEFINITION 1.2.3. Der *empirische Mittelwert* (auch das *Stichprobenmittel* oder das *arithmetische Mittel* genannt) ist definiert durch

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Analog benutzen wir auch die Notation

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Dabei ist $\bar{x}_n$ eine Stichprobenfunktion und $\bar{X}_n$ eine Statistik. Im Weiteren werden wir meistens keinen Unterschied zwischen diesen Begriffen machen.

SATZ 1.2.4. *Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit $\mu = \mathbb{E}X_i$ und $\sigma^2 = \operatorname{Var} X_i$. Dann gilt*

$$\mathbb{E}\bar{X}_n = \mu \text{ und } \operatorname{Var} \bar{X}_n = \frac{\sigma^2}{n}.$$

BEWEIS. Indem wir die Linearität des Erwartungswertes benutzen, erhalten wir

$$\mathbb{E}\bar{X}_n = \mathbb{E}\left[\frac{X_1 + \ldots + X_n}{n}\right] = \frac{1}{n} \cdot \mathbb{E}[X_1 + \ldots + X_n] = \frac{1}{n} \cdot n\mathbb{E}[X_1] = \mathbb{E}[X_1] = \mu.$$

Indem wir die Additivität der Varianz (bei unabhängigen Zufallsvariablen) benutzen, erhalten wir

$$\operatorname{Var} \bar{X}_n = \operatorname{Var}\left(\frac{X_1 + \ldots + X_n}{n}\right) = \frac{1}{n^2} \operatorname{Var}(X_1 + \ldots + X_n) = \frac{1}{n^2} \cdot n \operatorname{Var}(X_1) = \frac{\sigma^2}{n}.$$

□

BEMERKUNG 1.2.5. In der Statistik nimmt man an, dass die Stichprobe $(x_1, \ldots, x_n)$ bekannt ist und fragt dann, wie anhand dieser Stichprobe verschiedene Kenngrößen der Zufallsvariablen X_i (etwa der Erwartungswert, die Varianz, die Verteilungsfunktion) "geschätzt" werden können. Zum Beispiel bietet sich der empirische Mittelwert $\bar{x}_n$ (oder $\bar{X}_n$) als ein natürlicher Schätzer für den theoretischen Erwartungswert $\mu = \mathbb{E}X_i$. Der obige Satz zeigt, dass durch

eine solche Schätzung kein systematischer Fehler entsteht, in dem Sinne, dass der Erwartungswert des Schätzers $\bar{X}_n$ mit dem zu schätzenden Parameter μ übereinstimmt: $\mathbb{E}\bar{X}_n = \mu$. Man sagt, dass $\bar{X}_n$ ein *erwartungstreuer Schätzer* für μ ist.

DEFINITION 1.2.6. Die *empirische Varianz* oder die *Stichprobenvarianz* ist definiert durch

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2.$$

Analog benutzen wir auch die Notation

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Die Rolle des Faktors $\frac{1}{n-1}$ (anstelle von $\frac{1}{n}$) wird im folgenden Satz klar.

SATZ 1.2.7. *Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit $\mathbb{E}X_i = \mu$ und $\operatorname{Var} X_i = \sigma^2$. Dann gilt*

$$\mathbb{E}[S_n^2] = \sigma^2.$$

BEWEIS. Zuerst beweisen wir die Formel

$$S_n^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}_n^2\right).$$

Das geht folgendermaßen:

$$\begin{aligned}
S_n^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i^2 - 2X_i\bar{X}_n + \bar{X}_n^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} 2X_i\bar{X}_n + n\bar{X}_n^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - 2\bar{X}_n n\bar{X}_n + n\bar{X}_n^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}_n^2\right).
\end{aligned}$$

Nun ergibt sich

$$\begin{aligned}
\mathbb{E}[S_n^2] &= \mathbb{E}\left[\frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}_n^2\right)\right] \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}\mathbb{E}[X_i^2] - n\mathbb{E}[\bar{X}_n^2]\right) \\
&= \frac{1}{n-1}\left(n(\sigma^2+\mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right) \\
&= \sigma^2.
\end{aligned}$$

Dabei haben wir verwendet, dass

$$\mathbb{E}[X_i^2] = \operatorname{Var} X_i + (\mathbb{E}X_i)^2 = \sigma^2 + \mu^2$$

und (mit Satz 1.2.4)

$$\mathbb{E}[\bar{X}_n^2] = \operatorname{Var} \bar{X}_n + (\mathbb{E}\bar{X}_n)^2 = \frac{\sigma^2}{n} + \mu^2.$$

□

BEMERKUNG 1.2.8. Die empirische Varianz s_n^2 (bzw. S_n^2) ist ein natürlicher Schätzer für die theoretische Varianz $\sigma^2 = \operatorname{Var} X_i$. Der obige Satz besagt, dass S_n^2 ein erwartungstreuer Schätzer für σ^2 ist im Sinne, dass der Erwartungswert des Schätzers mit dem zu schätzenden Parameter σ^2 übereinstimmt: $\mathbb{E}S_n^2 = \sigma^2$.

BEMERKUNG 1.2.9. An Stelle von S_n^2 kann auch folgende Stichprobenfunktion betrachtet werden

$$\tilde{S}_n^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Der Unterschied zwischen S_n^2 und $\tilde{S}_n^2$ ist also nur der Vorfaktor $\frac{1}{n-1}$ bzw. $\frac{1}{n}$. Allerdings ist $\tilde{S}_n^2$ *kein* erwartungstreuer Schätzer für σ^2, denn

$$\mathbb{E}[\tilde{S}_n^2] = \mathbb{E}\left[\frac{n-1}{n}S_n^2\right] = \frac{n-1}{n} \cdot \mathbb{E}[S_n^2] = \frac{n-1}{n} \cdot \sigma^2 < \sigma^2.$$

Somit wird die Varianz σ^2 "unterschätzt". Schätzt man σ^2 durch $\tilde{S}_n^2$, so entsteht ein systematischer Fehler von $-\frac{1}{n}\sigma^2$.

BEMERKUNG 1.2.10. Die *empirische Standardabweichung* ist definiert durch

$$s_n = \sqrt{s_n^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}.$$

BEMERKUNG 1.2.11. Das Stichprobenmittel $\bar{x}_n$ ist ein Lageparameter (beschreibt die Lage der Stichprobe). Die Stichprobenvarianz s_n^2 (bzw. die empirische Standardabweichung s_n) ist ein Streuungsparameter (beschreibt die Ausdehnung der Stichprobe).

BEMERKUNG 1.2.12. Das Stichprobenmittel ist kein robuster Parameter, d.h. es wird stark von Ausreißern beeinflusst. Dies zeigt folgendes Beispiel: Betrachte zuerst die Stichprobe $(1, 2, 2, 2, 1, 1, 1, 2)$. Somit ist $\bar{x}_n = 1,5$. Ändert man nur den letzten Wert der Stichprobe in 20 um, also $(1, 2, 2, 2, 1, 1, 1, 20)$, dann gilt $\bar{x}_n = 3,75$. Wir konnten also den Wert des Stichprobenmittels stark verändern, indem wir nur ein einziges Element aus der Stichprobe verändert haben. Die Stichprobenvarianz ist ebenfalls nicht robust. Im weiteren werden wir robuste Lage- und Streuungsparameter einführen, d.h. solche Parameter, die sich bei einer Änderung (und zwar sogar bei einer sehr starken Änderung) von nur wenigen Elementen aus der Stichprobe nicht sehr stark verändern.