1 Grundlagen der Inferenzstatistik

1.1 Ziel der Inferenzstatistik

Lernziele

Die grundlegende Problemstellung der Inferenzstatistik ist, von Daten, die an Stichproben gewonnen wurden, auf die Gegebenheiten in der Grundgesamtheit zu schließen (daher auch *schließende Statistik*). Die Stichprobe stellt einen möglichst repräsentativen Ausschnitt der Grundgesamtheit dar. Sie spiegelt die Gegebenheiten der Grundgesamtheit aber nicht exakt wider, sondern fehlerbelastet (*Stichprobenfehler*). Mittels der Inferenzstatistik lässt sich abschätzen, wie gut auf Basis der Stichprobe auf die Grundgesamtheit geschlossen werden kann.

Mit den Methoden der deskriptiven Statistik werden empirische Daten, die an Stichproben gewonnen wurden, durch zusammenfassende Kennwerte, grafische oder tabellarische Darstellungen beschrieben. Auf diese Weise können Verteilungseigenschaften von umfangreichen Einzeldaten ökonomisch und leicht fassbar dargestellt werden.

In der Wissenschaft geht es aber meist darum, allgemeingültige Aussagen zu treffen. Das heißt, dass das eigentliche Ziel nicht darin besteht, die Stichprobe darzustellen, sondern auf Basis der Stichprobendaten Aussagen zu treffen, die über diese hinausgehen: Es geht darum, die Stichprobenergebnisse auf die Grundgesamtheit zu verallgemeinern. Man spricht daher im Gegensatz zur beschreibenden oder deskriptiven Statistik auch von der schließenden Statistik oder Inferenzstatistik, da auf Basis einer relativ kleinen Menge von Untersuchungseinheiten (Stichprobe) auf alle

potentiellen Untersuchungseinheiten (Grundgesamtheit) geschlossen werden soll.

1.2 Stichprobe und Grundgesamtheit

Definition: Grundgesamtheit und Stichprobe

Unter der *Grundgesamtheit* (auch: Population) versteht man die Gesamtmenge aller potentiellen Untersuchungseinheiten. Eine *Stichprobe* ist ein Ausschnitt aus der Grundgesamtheit. Bei der Stichprobenziehung werden nach bestimmten Methoden aus der Population Untersuchungseinheiten ausgewählt, die im Rahmen der Datenerhebung tatsächlich untersucht werden und die Population möglichst gut repräsentieren sollen.

Je nachdem über welche Grundgesamtheit man Aussagen treffen will, kann diese breiter (z. B. alle Menschen) oder enger definiert sein (z. B. alle in Deutschland lebenden Menschen oder alle Psychologiestudierenden an deutschen Hochschulen).

In aller Regel ist die Population zu groß, um sie vollständig zu untersuchen. Es wäre z. B. unrealistisch, alle in Deutschland lebenden Menschen untersuchen zu wollen (selbst wenn man nur eine Minute pro Person bräuchte, wäre man damit fast 150 Jahre beschäftigt – und zwar 24 h/Tag, 7 Tage/Woche, ohne zu schlafen oder Pausen zu machen). Stattdessen untersucht man lediglich einen Teil der Grundgesamtheit, eine sogenannte Stichprobe.

1.2.1 Stichprobenkennwerte und Populationsparameter

Stichprobendaten lassen sich mit Hilfe von Stichprobenkennwerten darstellen. So gibt das arithmetische Mittel ($\bar{x}$) den durchschnittlichen Wert einer Variablen über alle Untersuchungsteilnehmer an. In gleicher Weise gibt es auch in der Population einen Durchschnittswert (μ; ausgesprochen: »mü«). Um statistische Kennwerte, die sich auf eine Stichprobe beziehen, und statistische Parameter, die sich auf die Grundgesamtheit beziehen, schnell unterscheiden zu können, kennzeichnet man diese bei Stichproben mit lateinischen Buchstaben. Wenn es dagegen um die Beschreibung von Grundgesamtheiten geht, zieht man griechische Buchstaben heran (▶ Tab. 1.1).

Tab. 1.1: Abkürzungen für Stichprobenkennwerte und Populationsparameter

	Stichprobenkennwert	Populationsparameter
Arithmetischer Mittelwert	$\bar{x}$	μ
Varianz	s^2	σ^2
Korrelation	r	ρ
Wahrscheinlichkeit	p	π

Wenn nun das Ziel der Inferenzstatistik darin besteht, mittels Stichprobendaten auf die Grundgesamtheit zu schließen, so liegt es nahe, die Stichprobenkennwerte zu nutzen, um auf die Populationsparameter zu schließen. Und genau dies tut man auch. Je besser die Stichprobe die zugrundeliegende Population abbildet, desto exakter lassen sich über die Stichprobenkennwerte die Populationsparameter abschätzen. Aber die Stichprobendaten sind nie ein exaktes Abbild der Population. Man muss immer damit rechnen, dass sie mehr oder weniger fehlerbehaftet sind. Daraus ergibt sich auch, dass Stichprobenkennwerte die jeweiligen Populationsparameter nicht exakt abbilden, sondern mehr oder weniger stark davon abweichen können. Sie können das leicht selbst ausprobieren:

Experiment zur Ungenauigkeit von Stichprobendaten

Das arithmetische Mittel der Population aller Würfelwürfe beträgt μ = 3,5 (da die theoretische Verteilung von Würfelwürfen bekannt ist, ist auch der Populationsparameter bekannt).

Wenn Sie nun anhand von Stichprobendaten μ abschätzen wollen, können Sie z. B. 10-mal würfeln. Sie haben damit eine Stichprobe von $n = 10$ Würfelwürfen generiert. Bestimmen Sie den arithmetischen Mittelwert der 10 gewürfelten Augenzahlen. Vermutlich werden Sie feststellen, dass dieser Stichprobenmittelwert nicht exakt $\bar{x} = 3{,}5$ ist. Vermutlich werden Sie aber auch feststellen, dass er in der Nähe von 3,5 liegt. Führen Sie das Experiment noch einmal durch. Wieder wird der Mittelwert mehr oder weniger nahe an 3,5 liegen, aber vermutlich nicht exakt 3,5 betragen. Sie können dieses Experiment beliebig häufig wiederholen und werden feststellen, dass sich die Mittelwerte der einzelnen Stichprobenziehungen jeweils mehr oder weniger stark vom Populationsparameter $\mu = 3{,}5$ und auch voneinander unterscheiden. Sie werden aber auch feststellen, dass die Ergebnisse in der Regel recht nahe bei 3,5 liegen und größere Abweichungen selten sind.

1.3 Stichprobenkennwerteverteilung

Wenn man das Würfelexperiment mit 10 Würfelwürfen und anschließender Bestimmung des arithmetischen Mittelwertes sehr häufig wiederholt, resultiert eine Verteilung der Mittelwerte, die sich um den Erwartungswert von $\mu = 3{,}5$ verteilen wird (► Abb. 1.1).

Man könnte das Beispiel auch mit anderen Kennwerten wiederholen (z. B. könnte man die Standardabweichungen bestimmen etc.). Eines würde immer gleichbleiben: Die resultierenden Kennwerte schwanken mehr oder weniger stark um den »wahren« Wert in der Population, den

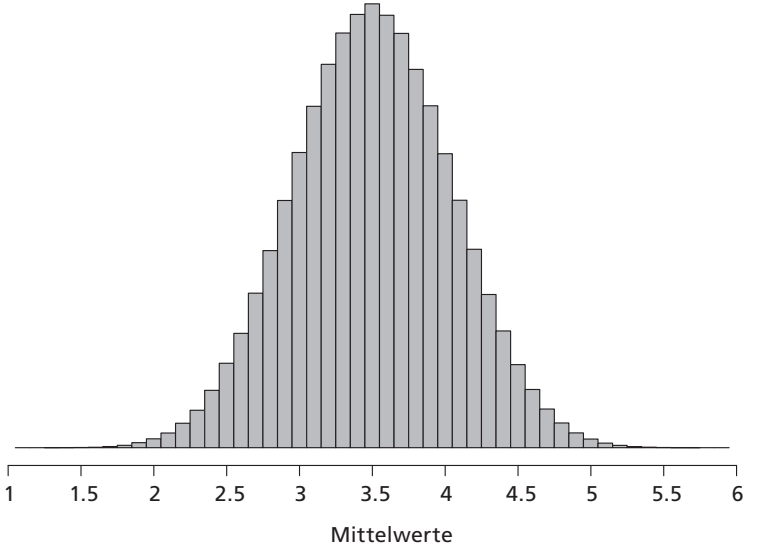

Abb. 1.1: Verteilung der Mittelwerte bei n =10 Würfelwürfen

Populationsparameter. Die resultierende Verteilung bezeichnen wir als Stichprobenkennwerteverteilung (oft auch verkürzt: Stichprobenverteilung; englisch: sampling distribution). Je größer die Schwankungen sind, desto ungenauer ist die Schätzung des Populationsparameters durch den Stichprobenkennwert und desto vorsichtiger müssen wir sein, wenn wir auf Basis der Stichprobendaten Aussagen über die Grundgesamtheit treffen wollen.

Es wäre hilfreich, wenn wir angeben könnten, wie ungenau unsere Aussagen über die Population sind, d. h., wie groß die Schwankungen der Stichprobenkennwerteverteilung sind. Ein Kennwert, der bereits aus der deskriptiven Statistik bekannt ist, um die Schwankungen der Einzelwerte auszudrücken, ist die Standardabweichung. In äquivalenter Weise können die Schwankungen der Stichprobenkennwerte durch die Standardabweichung der Stichprobenkennwerteverteilung quantifiziert werden. Da es sich um einen besonderen Fall handelt, hat die Standardabweichung der Stichprobenkennwerteverteilung einen besonderen Namen: Sie wird als Standardfehler bezeichnet.

Definition: Standardfehler

Der Standardfehler (englisch: standard error [SE]) ist die Standardabweichung der Kennwerteverteilung von gleichgroßen Zufallsstichproben einer Grundgesamtheit.

1.3.1 Standardfehler des Mittelwerts

Ein in der empirischen Forschung sehr häufiger Anwendungsfall betrifft Aussagen über den Mittelwert einer Variablen (Populationsparameter μ). Der Stichprobenmittelwert ist ein erwartungstreuer Schätzer des Populationsparameters μ (▶ Kap. 2) μ. Wie im Beispiel mit den 10 Würfelwürfen gesehen, verschätzt der Stichprobenmittelwert den Populationsparameter mehr oder weniger stark und es ergeben sich Schwankungen, so dass sich als Stichprobenkennwerteverteilung die Verteilung der Mittelwerte ergibt. Diese Variabilität der Mittelwerte wird durch den Standardfehler des Mittelwerts ($\sigma_{\bar{x}}$) ausgedrückt. Dieser bestimmt sich auf einfache Weise als:

$$\sigma_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

Wie groß der Standardfehler des Mittelwertes ist, hängt also lediglich von zwei Faktoren ab: der Varianz des Merkmals in der Population (σ^2) sowie der Größe der Stichprobe *n:*

- Je stärker das Merkmal in der Population variiert, desto eher kommt es auch beim Stichprobenmittelwert zu stärkeren Abweichungen vom Populationsparameter μ und vice versa. Lassen Sie uns als Extrembeispiel davon ausgehen, dass ein Merkmal in der Population überhaupt nicht variiert, dann würde auch ein Standardfehler von $\sigma_{\bar{x}} = 0$ resultieren, d. h., auch die Stichprobenmittelwerte wären bei jeder Stichprobe identisch (und würden exakt μ widerspiegeln).
- Je größer die Stichprobe ist, an der man den Stichprobenmittelwert ermittelt hat, desto kleiner ist der Standardfehler und vice versa. Im

theoretischen Fall einer unendlich großen Stichprobe geht der Standardfehler daher gegen 0. Realistischer gedacht: Je größer die Stichprobe ist, desto genauer sind die Aussagen, die man aus den Stichprobendaten in Bezug auf die Populationsverhältnisse treffen kann. Große Stichproben sind also in der Inferenzstatistik hilfreich, da sie bessere Schätzungen ermöglichen.

IQ-Werte sind so normiert, dass sie in der Normalbevölkerung einen Mittelwert von $\bar{x} = 100$ und eine Streuung von $\sigma = 15$ haben.[1] Bei einer Stichprobenerhebung mit $n = 50$ würde sich demnach eine Stichprobenkennwerteverteilung des Mittelwerts mit einem Standardfehler von $\sigma_{\bar{x}} = 15/\sqrt{50} = 2{,}12$ ergeben.

Ein Problem bei der praktischen Bestimmung des Standardfehlers liegt darin, dass die dafür benötigte Populationsvarianz σ^2 in der Regel gar nicht bekannt ist. Dementsprechend muss die Varianz und in der Folge der Standardfehler aus den Stichprobendaten geschätzt werden. Um zu verdeutlichen, dass es sich hier um einen geschätzten Standardfehler handelt, bezeichnet man ihn mit dem lateinischen Buchstaben. Die Bestimmung erfolgt auf der Basis der Stichprobenvarianz $[s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)]$. Die so bestimmte Stichprobenvarianz ist wiederum ein erwartungstreuer Schätzer der Populationsvarianz, so dass sich für den Standardfehler des Mittelwerts ergibt:

$$s_{\bar{x}} = \sqrt{\frac{s^2}{n}}$$

1 Kleine Anekdote nebenbei: David Wechsler, der Erfinder des modernen IQs, wählte übrigens eine Streuung von 15, weil er wollte, dass zwischen einem IQ von 90 und 110 genau 50 % der Menschen liegen, die man dann als durchschnittlich intelligent bezeichnen könnte.

Datenbeispiel

In einer Untersuchung mit einer Stichprobe der Größe $n = 30$ ergibt sich eine Varianz von $s^2 = 100$. Somit ergibt sich für den geschätzten Standardfehler: $s_{\bar{x}} = \sqrt{s^2/n} = \sqrt{100/30} = 1{,}83$

1.3.2 Besondere Stichprobenkonstellationen

Die dargestellte Bestimmung des Standardfehlers bezieht sich auf den »Normalfall«, dass eine Zufallsstichprobe aus einer (theoretisch unendlich) großen Grundgesamtheit gezogen wurde. Es gibt Stichprobenkonstellationen, bei denen die Repräsentativität erhöht ist, sei es, weil die Grundgesamtheit im Vergleich zur Stichprobengröße relativ klein ist (finite Grundgesamtheit) oder bestimmte Strategien bei der Stichprobenziehung angewendet wurden, um die Repräsentativität zu verbessern, z. B. die Ziehung einer geschichteten Stichprobe. Auf diese beiden Sonderfälle soll im Folgenden eingegangen werden, da der Standardfehler in diesen Konstellationen gegenüber einer »normalen« Zufallsstichprobe geringer ausfällt.

Finite Grundgesamtheiten

Wenn die Grundgesamtheit in ihrer Größe (N) beschränkt ist, so ergibt sich die Situation, dass eine Stichprobenerhebung mit der Größe n nahe an eine Vollerhebung kommen kann. Das hat zur Folge, dass Stichproben, die von der Population extrem abweichen, unwahrscheinlicher werden. Dies kann bei der Bestimmung des Standardfehlers Berücksichtigung finden, indem eine sogenannte Endlichkeitskorrektur vorgenommen wird:

$$s_{\bar{x}_f} = \sqrt{\frac{s^2}{n} \cdot \frac{N-n}{N-1}}$$

Beispiel

Wenn im vorherigen Beispiel ($n = 30$; $s^2 = 100$) die Stichprobenziehung aus einer relativ kleinen Gesamtpopulation (z. B. $N = 425$) gezogen worden wäre, so könnte dies bei der Bestimmung des Standardfehlers berücksichtigt werden:

$$s_{\bar{x}_f} = \sqrt{\frac{s^2}{n} \cdot \frac{N-n}{N-1}} = \sqrt{\frac{100}{30} \cdot \frac{425-30}{424}} = 1{,}76$$

Man sieht, dass der Standardfehler sichtbar niedriger ausfällt. Die Endlichkeitskorrektur fällt aber nur ins Gewicht, wenn die Population relativ klein ist bzw. die Stichprobengröße der Gesamtpopulation relativ nahekommt. Als Faustregel lässt sich sagen, dass man bei $N/n > 100$ darauf verzichten kann und stattdessen den unkorrigierten Standardfehler bestimmt.

Geschichtete Stichproben

Wenn man die Grundgesamtheit in verschiedene Teilmengen aufteilt (z. B. die Bewohner der BRD nach den Bundesländern oder Studierende nach den Studienfächern etc.) und nun in den jeweiligen Schichten getrennt Stichproben zieht, deren Größe den Anteil in der Grundgesamtheit widerspiegelt, dann spricht man von einer *geschichteten Stichprobe.* Der Vorteil einer geschichteten Stichprobe gegenüber einer einfachen Zufallsstichprobe ist, dass die Repräsentativität der Stichprobe steigt, weil sichergestellt ist, dass die Stichprobe in Bezug auf die Schichtungsvariable exakt der Population entspricht. Kennwerte einer sinnvoll geschichteten Stichprobe liefern deswegen bessere Schätzwerte als ungeschichtete Stichproben und dies kann man bei der Bestimmung des Standardfehlers berücksichtigen. Die entsprechende Formel lautet:

$$s_{\bar{x}_g} = \sqrt{\frac{s^2 - s^2_{\bar{x}(m)}}{n}}$$

Dabei wird die Variabilität des erfassten Merkmals, die die Fehleranfälligkeit bei der Stichprobenziehung bedingt (s^2), um den Anteil reduziert, der sich über die Unterschiedlichkeit in den Stufen der Schichtungsvariablen ergeben würde ($s^2_{\bar{x}_{(m)}}$, weil diese ja korrekt (und fehlerfrei) repräsentiert ist). Bei $s^2_{\bar{x}_{(m)}}$ handelt es sich um die Varianz der Mittelwerte über die verschiedenen Schichten:

$$s^2_{\bar{x}(m)} = \frac{1}{n} \cdot \left[\sum_{m=1}^{k} n_m \cdot (\bar{x}_m - \bar{x})^2 \right]$$

Beispiel

Um die Zufriedenheit mit dem öffentlichen Nahverkehr in der Bevölkerung abzuschätzen, wurde eine Stichprobe von $N = 622$ in Deutschland lebenden Personen befragt. Auf einer Skala von 1 (sehr unzufrieden) bis 7 (sehr zufrieden) ergab sich ein Mittelwert von $\bar{x} = 3{,}76$ ($s = 1{,}56$). Bei der Stichprobenziehung wurde nach Stadt- und Landbevölkerung unterschieden und eine geschichtete Stichprobe gezogen:

$n_{\text{Stadt}} = 325$; $\bar{x}_{Stadt} = 4{,}21$
$n_{\text{Land}} = 97$; $\bar{x}_{Land} = 2{,}24$

Als Standardfehler des (Gesamt-)Mittelwerts würde sich nun nicht der einfache Standardfehler ergeben (das wäre: $s_{\bar{x}} = 1{,}56/\sqrt{422} = 0{,}08$). Stattdessen kann, und sollte, die Stratifizierung berücksichtigt werden. Die Unterschiedlichkeit in der Zufriedenheit, die auf die Gruppenunterschiede zwischen Stadt- und Landbevölkerung zurückgeht, beträgt:

$$s^2_{\bar{x}(m)} = \frac{1}{422} \cdot \left[325 \cdot (4{,}21 - 3{,}76)^2 + 97 \cdot (2{,}24 - 3{,}76)^2 \right] = 0{,}69$$

Die verbleibende Fehlervarianz beträgt daher nur noch:

$$s_{\bar{x}_g} = \sqrt{\frac{s^2 - s^2_{\bar{x}(m)}}{n}} = \sqrt{\frac{1{,}56^2 - 0{,}69}{422}} = 0{,}06$$