

# Induktive Statistik

## Induktive Statistik

Induktive Statistik, auch schließende Statistik genannt, ist ein Teilbereich der Statistik, der darauf abzielt, von einer Stichprobe auf die Grundgesamtheit zu schließen. Die induktive Statistik wird verwendet, um auf Basis von Daten einer Stichprobe Aussagen über die Population zu machen, ohne alle Daten der Population zu kennen.

## Stichprobe

Eine Stichprobe ist eine Teilmenge von Objekten, die aus der Grundgesamtheit zufällig ausgewählt werden. Stichproben werden in der Statistik verwendet, um Daten zu erheben, ohne die Grundgesamtheit untersuchen zu müssen.

Eine zufällige Stichprobe vom Umfang  $n$  ist eine Folge  $X_1, X_2, \dots, X_n$  von unabhängig, identisch verteilten Zufallsvariablen, wobei  $X_i$  die Ausprägung des Merkmals des  $i$ -ten Objekts beschreibt. Wir nennen  $X_i$  eine Stichprobenvariable.

## Schätzfunktion

Eine Funktion  $T(X_1, \dots, X_n)$  nennen wir Stichprobenfunktion und auch diese Funktion ist ebenfalls eine Zufallsvariable. Falls wir sie für die Schätzung eines Parameters  $\psi$  der Grundgesamtheit verwenden, nennen wir sie Schätzfunktion bzw. Schätzer für den Parameter  $\psi$ .

# Punktschätzer

## Punktschätzer

Aus Stichproben können ein oder auch mehrere Werte errechnet werden (z.B. Median und Varianz), die Schätzwerte für die Grundgesamtheit darstellen. Wir sprechen in diesem Fall von Punktschätzern, weil sie jeweils genau einen Wert schätzen.

## Eigenschaften von Schätzern

- ▶ Erwartungstreue:  $E(T) = \psi$ . In Worten: Der Schätzer trifft im Durchschnitt den wahren Wert  $\psi$ . Das bedeutet, dass der Schätzer im Durchschnitt den wahren Wert liefert, auch wenn er bei einzelnen Schätzungen vom wahren Wert abweicht.
- ▶ Effizienz: Je kleiner die Varianz des Schätzers, desto näher wird ein Schätzwert am wahren Wert  $\psi$  liegen. Ein effizienter Schätzer ist derjenige, der die geringstmögliche Varianz aufweist.
- ▶ Konsistenz: Mit wachsender Stichprobe werden die Abweichungen vom wahren Wert kleiner.

## Bias (Verzerrung)

Ein Schätzer, der nicht erwartungstreu ist, ist verzerrt. Die Verzerrung (auch Bias genannt) eines Schätzers ist die Differenz zwischen dem Erwartungswert des Schätzers und dem wahren Wert:  $Bias(T) = E(T) - \psi$ . Wir nennen einen Schätzer unverzerrt, falls  $Bias(T) = 0$  ist.

## Beispiele für erwartungstreue Schätzer

- ▶ Das arithmetische Mittel  $\bar{x}$  ist ein erwartungstreuer Schätzer des Erwartungswertes der Grundgesamtheit.
- ▶ Die empirische Varianz  $s^2$  ist ein erwartungstreuer Schätzer für die Varianz der Grundgesamtheit.
- ▶ Die korrigierte empirische Kovarianz  $s_{xy}$  ist ein erwartungstreuer Schätzer der Kovarianz der Grundgesamtheit.

## Erwartungstreue des arithmetischen Mittels

Seien  $X_1, \dots, X_n$  i.i.d. mit  $E(X_1) = \dots = E(X_n) = \mu$ . Das arithmetische Mittel der Zufallsvariablen berechnet sich wie folgt  $\bar{X} = \frac{X_1 + \dots + X_n}{n}$  und ist ein erwartungstreuer Schätzer des Erwartungswerts der Grundgesamtheit, da das Folgende gilt:

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{X_1 + \dots + X_n}{n}\right) \\ &= \frac{1}{n} E(X_1 + \dots + X_n) \\ &= \frac{1}{n} (E(X_1) + \dots + E(X_n)) \\ &= \frac{1}{n} \cdot n \cdot \mu \\ &= \mu. \end{aligned}$$

## Erwartungstreue der empirischen Varianz $s^2$

Seien  $X_1, \dots, X_n$  i.i.d. mit  $E(X_1) = \dots = E(X_n) = \mu$  und  $\text{Var}(X_1) = \dots = \text{Var}(X_n) = \sigma^2$ .  
 $s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$  ist ein erwartungstreuer Schätzer für die Varianz, da:

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

# Intervallschätzung



Warum könnte es Sinn machen Punktschätzer durch Intervallschätzer zu ersetzen?

Betrachten wir eine Stichprobe  $X_1, \dots, X_n$  aus einer  $N(\mu, 1)$  verteilten Population.

- ▶ Wir kennen  $\mu$  nicht, können diesen Parameter aber mit  $\bar{X}$  schätzen.
- ▶ Es gilt  $P(\bar{X} = \mu) = 0$ .
- ▶ Beispiel eines Intervallschätzers ist  $[\bar{X} - 1; \bar{X} + 1]$ .
- ▶ Wir verlieren an Genauigkeit der Schätzung.
- ▶ Andererseits erhalten wir mit positiver Wahrscheinlichkeit den korrekten Parameterwert, da  $P(\mu \in [\bar{X} - 1; \bar{X} + 1]) > 0$ .

## Intervallschätzung

- ▶ Die Intervallschätzung verfolgt das Ziel, einen unbekannten Parameter zu schätzen.
- ▶ Sie liefert uns ein Intervall, das den wahren Wert mit einer bestimmten Wahrscheinlichkeit enthält.
- ▶ Einen wahren Wert genau zu treffen ist schwierig und daher wird eine Bandbreite angegeben.

## Gebräuchlichste Verfahren: Konfidenzintervall KI

- ▶ Ein KI ist ein Intervall, das den wahren Parameter mit einer bestimmten Wahrscheinlichkeit enthält.
- ▶ Die Grenzen des Intervalls berechnen sich aus den Werten der Stichprobe.
- ▶ Formal kann das Konfidenzintervall wie folgt dargestellt werden:  
 $[T_u(X_1, \dots, X_n); T_o(X_1, \dots, X_n)]$ .
- ▶ Beispiel: Die Varianz liegt mit 90 % Wahrscheinlichkeit zwischen [50, 60].
- ▶ Beispiele:  
[https://shinyapps.wiwi.hu-berlin.de/mmstat\\_en/confidence\\_mean/](https://shinyapps.wiwi.hu-berlin.de/mmstat_en/confidence_mean/).

## Konfidenzniveau und Irrtumsniveau

- ▶ Das **Konfidenzniveau** gibt an, mit welcher Wahrscheinlichkeit der zu schätzende Parameter im Intervall enthalten ist.

Zum Beispiel für das Konfidenzniveau  $1 - \alpha = 0.95$ :

$$P(\psi \in [T_u(X_1, \dots, X_n); T_o(X_1, \dots, X_n)]) = 0.95.$$

- ▶ Die **Irrtumswahrscheinlichkeit** gibt die Wahrscheinlichkeit an, dass der zu schätzende Parameter nicht im Intervall liegt.

Zum Beispiel für die Irrtumswahrscheinlichkeit  $\alpha = 0.05$ :

$$P(\psi \notin [T_u(X_1, \dots, X_n); T_o(X_1, \dots, X_n)]) = 0.05.$$

## Zweiseitige Konfidenzintervalle zum Konfidenzniveau $1 - \alpha$ eines normalverteilten Merkmals

**Anmerkung:**  $n$  beschreibt die Stichprobengröße,  $\bar{x}$  das arithmetische Mittel und  $s$  die empirische Standardabweichung.

- KI für den Erwartungswert  $\mu$ , falls  $\sigma^2$  der Population bekannt ist:

$$\left[ \bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}; \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right]$$

**Anmerkung:**  $z_{1-\frac{\alpha}{2}}$  schneidet die oberen  $\frac{\alpha}{2}$  der Fläche unter der Normalverteilung ab und ist das  $(1 - \frac{\alpha}{2})$ -Quantil der Standardnormalverteilung.

- KI für den Erwartungswert  $\mu$ , falls  $\sigma^2$  der Population unbekannt ist:

$$\left[ \bar{x} - t_{n-1;1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}; \bar{x} + t_{n-1;1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \right]$$

**Anmerkung:**  $t_{n-1;1-\frac{\alpha}{2}}$  ist das  $(1 - \frac{\alpha}{2})$ -Quantil der t-Verteilung mit  $n - 1$  Freiheitsgraden.

- KI für die Varianz  $\sigma^2$ :

$$\left[ \frac{(n-1)s^2}{\chi^2_{n-1; \frac{\alpha}{2}}}; \frac{(n-1)s^2}{\chi^2_{n-1; 1-\frac{\alpha}{2}}} \right]$$

**Anmerkung:**  $\chi^2_{n-1}$  beschreibt die Chi-Quadrat-Verteilung mit  $n-1$  Freiheitsgraden.

## Freiheitsgrad

Ein Freiheitsgrad gibt die Anzahl frei wählbarer Werte für einen Parameter an. Die Anzahl der Freiheitsgrade nimmt mit zunehmender Stichprobengröße zu und fällt mit der Anzahl geschätzter Parameter.

Zum Beispiel haben drei Maschinen folgendes Gewicht: 110 kg, 120 kg und 130 kg. Der Mittelwert ist 120 kg. Basierend auf dem Mittelwert könnten die Maschinen auch folgende Gewichte aufweisen: 100 kg, 120 kg, 140 kg. Dabei sind nur die ersten beiden Gewichte der Maschinen frei wählbar und das dritte Gewicht ergibt sich aus den ersteren und dem Mittelwert. In diesem Beispiel ergeben sich somit für die Verteilung zwei Freiheitsgrade.

## Einseitige Konfidenzintervalle zum Konfidenzniveau $1 - \alpha$ eines normalverteilten Merkmals

- ▶ KI für den Erwartungswert  $\mu$ , falls  $\sigma^2$  der Population bekannt ist:

$$\left( -\infty; \bar{x} + z_{1-\alpha} \frac{\sigma}{\sqrt{n}} \right] \text{ oder } \left[ \bar{x} - z_{1-\alpha} \frac{\sigma}{\sqrt{n}}; +\infty \right)$$

- ▶ KI für den Erwartungswert  $\mu$ , falls  $\sigma^2$  der Population unbekannt ist:

$$\left( -\infty; \bar{x} + t_{n-1;1-\alpha} \frac{s}{\sqrt{n}} \right] \text{ oder } \left[ \bar{x} - t_{n-1;1-\alpha} \frac{s}{\sqrt{n}}; +\infty \right)$$

- ▶ KI für die Varianz  $\sigma^2$ :

$$\left( 0; \frac{(n-1)s^2}{\chi^2_{n-1;\alpha}} \right] \text{ oder } \left[ \frac{(n-1)s^2}{\chi^2_{n-1;1-\alpha}}; +\infty \right)$$

**Beispiel 1:** Wir betrachten das Merkmal Alter einer Stichprobe von 100 Personen aus Österreich. Der Mittelwert dieser Stichprobe liegt bei 43 Jahren und die Standardabweichung ist 10 Jahre. Zum Konfidenzniveau 0.95 beträgt das zweiseitige Konfidenzintervall:

►  $T_u(X_1, \dots, X_{100}) = 43 - 1.96 \frac{10}{\sqrt{100}} = 41.04$

►  $T_o(X_1, \dots, X_{100}) = 43 + 1.96 \frac{10}{\sqrt{100}} = 44.96$

Somit können wir schlussfolgern, dass das Durchschnittsalter in Österreich mit 95 % Wahrscheinlichkeit zwischen 41 und 45 Jahren liegt.

**Anmerkung 1:** Die Standardnormalverteilungstabelle ist unter <https://de.wikipedia.org/wiki/Standardnormalverteilungstabelle> zu finden.

**Anmerkung 2:** Die  $t$ -Verteilung ähnelt einer Normalverteilung und mit steigender Anzahl an Freiheitsgraden können wir die  $t$ -Verteilung mithilfe der Normalverteilung approximieren. Eine Faustregel besagt, dass ab 30 Freiheitsgraden die  $t$ -Verteilung durch die Normalverteilung approximiert werden kann.



**Beispiel 2a:** Eine Stichprobe aus 10 Studierenden des Studienfachs Mechatronik hat an einem IQ-Test teilgenommen. Das Ergebnis war: 105, 100, 123, 90, 100, 128, 105, 109, 104, 115. Wir sind am Konfidenzintervall zum Konfidenzniveau  $1 - \alpha = 90\%$  für den Erwartungswert der IQ-Werte für die Population der Studierenden des Studienfachs Mechatronik interessiert.

**Lösung:**

- ▶ Mittelwert der Stichprobe:  $\bar{x} = 107.9$ .
- ▶ Geschätzte Populationsvarianz:  $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = 128.99$ .
- ▶  $t$ -Wert für  $n - 1 = 9$  Freiheitsgrade und  $1 - \alpha = 0.9$ :  $t_{n-1; 1-\frac{\alpha}{2}} = t_{9; 0.95} = 1.833$ .
- ▶  $T_u(X_1, \dots, X_{10}) = 107.9 - 1.833 \cdot \frac{\sqrt{128.99}}{\sqrt{10}} = 101.32$ .
- ▶  $T_o(X_1, \dots, X_{10}) = 107.9 + 1.833 \cdot \frac{\sqrt{128.99}}{\sqrt{10}} = 114.48$ .

Demnach kann mit 90 % Sicherheit gesagt werden, dass der wahre IQ-Wert zwischen 101.32 und 114.48 liegt.

**Anmerkung:** Die Tabelle einiger  $t$ -Quantile finden wir unter [https://de.wikipedia.org/wiki/Studentsche\\_t-Verteilung](https://de.wikipedia.org/wiki/Studentsche_t-Verteilung).

**Beispiel 2b:** Wir sind am einseitigen Konfidenzintervall  $[a, +\infty)$  zum Konfidenzniveau 90 % interessiert.

**Lösung:**

- ▶ Mittelwert der Stichprobe:  $\bar{x} = 107.9$ .
- ▶ Geschätzte Populationsvarianz:  $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = 128.99$ .
- ▶  $t$ -Wert für  $n - 1 = 9$  Freiheitsgrade und  $1 - \alpha = 0.9$ :  $t_{n-1;1-\alpha} = t_{9;0.9} = 1.383$ .
- ▶  $T_u(X_1, \dots, X_{10}) = 107.9 - 1.383 \cdot \frac{\sqrt{128.99}}{\sqrt{10}} = 102.93$ .

Demnach kann mit 90 % Sicherheit gesagt werden, dass der wahre IQ-Wert zwischen 102.93 und  $+\infty$  liegt.

**Beispiel 3:** Uns liegt eine zufällige Stichprobe vom Umfang  $n = 20$  Haushalten vor, die ein mittleres Haushaltsnettoeinkommen von  $\bar{x} = 2800$  Euro haben. Als Punktschätzer für die unbekannte Varianz  $\sigma^2$  erhalten wir aus der Stichprobe  $s^2 = 500$ . Wir sind am zweiseitigen Konfidenzintervall für die Varianz  $\sigma^2$  zum Konfidenzniveau 95 % interessiert.

**Lösung:**

- ▶ Anzahl an Freiheitsgraden:  $n - 1 = 20 - 1 = 19$ .
- ▶  $\chi^2_{n-1; 1 - \frac{\alpha}{2}} = \chi^2_{19; 0.975} = 8.907$ .
- ▶  $\chi^2_{n-1; \frac{\alpha}{2}} = \chi^2_{19; 0.025} = 32.852$ .
- ▶ Untere Schranke:  $\frac{19 \cdot 500}{32.852} = 289.18$ .
- ▶ Obere Schranke:  $\frac{19 \cdot 500}{8.907} = 1066.58$ .

Demnach kann mit 95 % Sicherheit gesagt werden, dass die Varianz der Population zwischen 289.18 und 1066.58 liegt.

**Anmerkung:** Die Tabelle einiger Quantile der Chi-Quadrat-Verteilung finden wir unter <https://datatab.de/tutorial/tabelle-chi-quadrat> sowie unter [https://www.uibk.ac.at/econometrics/einf/tab\\_chisq.pdf](https://www.uibk.ac.at/econometrics/einf/tab_chisq.pdf).