1 Grundfragen der Statistik

Die Statistik befasst sich mit der Frage, wie man aus gegebenen Beobachtungen die zugrunde liegende Wahrscheinlichkeit oder andere Kenngrössen bestimmt bzw. abschätzt.

1.1 Typische Problemstellungen

Bei der Auswertung von Daten unterscheidet zwischen:

- der beschreibenden Statistik
- der schliessenden Statistik

Beschreibende Statistik
Wiedergabe vorhandener Daten durch Auswahl, Zusammenfassung und geeignete Darstellung.
→ Für mehr Informationen: siehe Hübner Kap. 1.5

Schliessende Statistik
Auswertung von gezielt erhobenen Stichproben, die über Eigenschaften einer grösseren Gesamtheit Aufschluss geben sollen.

Die schliessende Statistik kennt 3 Fragestellungen:

1. Punktschätzung: charakteristischen Wert möglichst genau ermitteln
2. Intervallschätzung: unter und obere Schranken ermitteln
3. Testen von Hypothesen: vermutetes Ergebnis bestätigen

1.2 Punktschätzung

Aufgrund einer vorliegenden Stichprobe soll eine Kenngrösse g durch eine Schätzfunktion $h(X_1,...,X_n)$ geschätzt werden.

Schätzfunktion
Gibt aus einer Stichprobe den entsprechenden Wert g zurück.
Eine Schätzfunktion ist erwartungstreu wenn ihr Erwartungswert dem zu schätzendem Parameter entspricht.

Beispiel für die Schätzung von Erwartungswerten:

$$g = h(X_1,...,X_n) = \frac{(X_1 + ... + X_n)}{n}$$

$$g = h(X_1,...,X_n) = \frac{\max(X_1,...,X_n) + \min(X_1,...,X_n)}{2}$$

Varianzschätzung
Zur Schätzung von $VarX = E(X_1 - EX_1)\text{^}2$ liegt es nahe den Schätzer $\frac{1}{n} * \sum_{i=1}^{n} (X_i - EX_i)^2$ zu benutzen. In den meisten Fällen wird aber bei der Schätzung von VarX auch EX unbekannt sein. EX wird durch dessen Schätzwert ($\overline{X_n}$) ersetzt. Wir erhalten die so genannte Stichprobenvarianz:

Über GEsammtheit: $$\overline{V} = \frac{1}{n} * \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 = \frac{1}{n}\left[\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2\right]$$

Für Stichproben: $$\overline{V} = \frac{1}{n-1} * \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2\right]$$

Maximum-Likelihood-Prinzip
Gegeben sind Werte einer Stichprobe sowie die allgemeine (parametrisierte) Verteilungsfunktion. Mit der Maximum-Likelihood-Prinzip werden die Parameter der Verteilungsfunktion so bestimmt, dass die Werte aus der Stichprobe am wahrscheinlichsten sind.

1.3 Intervallschätzung

Nach einer Punktschätzung weiss man nicht wie weit der Schätzwert vom unbekannten wahren Wert entfernt sein könnte. Deshalb sollte man in der Praxis Intervallschätzungen bevorzugen.

Das 100- prozentige Intervall gibt es nicht. Deshalb gibt man eine Sicherheitsanforderung vor (z. B. 1-α =95%). Dies bedeutet, dass ein geschätztes Intervall den wahren Wert hier mit der Wahrscheinlichkeit von 95% enthalten soll. In der Literatur spricht man meist von Schätzungen „zum Niveau α“ (Beispiel: zum Niveau 5%).

Unter der Voraussetzung einer $N(a,\sigma^2)$ Verteilung ergibt sich zum Niveau α (also mit der Sicherheit 1-α) die Intervallschätzung:

$$\left[\overline{X_n} - \mu_{1-\frac{\alpha}{2}} * \sigma/\sqrt{n} \;;\; \overline{X_n} + \mu_{1-\frac{\alpha}{2}} * \sigma/\sqrt{n}\right]$$

$\overline{X_n}$ = geschätzter Erwartungswert
σ = (geschätzte) Streuung
α = Niveau
μ = Quantil der Standardnormalverteilung (siehe Hübner S. 196)
n = Anzahl der Beobachtungen

Chi-Quadrat-Verteilung

Die durch die Chi-Quadrat-Verteilung χ^2 verteilen Zufallsvariablen können für Signifikanztests und zur Berechnung von Konfidenzintervallen eingesetzt werden. Die Chi-Quadrat-Verteilung entsteht durch die Summierung von n quadrierten standardnormalverteilten Zufallsvariablen (sog. z-Variablen). Es gibt also viele verschiedene Chi-Quadratverteilungen χ_n^2, die sich durch die Zahl der aufsummierten z-Variablen unterscheiden.

$$\chi_n^2 = X_1^2 + X_2^2 + ... + X_n^2 = \sum_{i=1}^{n} X_i^2$$

X_i = Stochastisch unabhängige Zufallsvariablen einer N(0,1)
n = sog. Freiheitsgrad

Student-Verteilung

Die Student-Verteilung setzt auf der Chi-Quadrat-Verteilung auf.

$$T = \frac{X}{\sqrt{Y/n}}$$

X = Zufallsvariable der N(0,1)
Y = Zufallsvariable der χ_n^2
n = Freiheitsgrad in Y bzw. χ_n^2

Auf die Dichtefunktionen dieser beiden Verteilungen wird hier verzichtet.

Intervallschätzung für die Varianz

Die bisherige Intervallschätzung heisst Intervallschätzung zum Mittelwert. Im Folgenden wird eine Intervallschätzung für die Varianz betrachtet. Die Varianz einer Stichprobe entspricht nicht der Varianz über die Gesamtheit. Mit einem Konfidenzintervall für die Varianz wird eine obere und eine untere Grenze für die wahre Varianz bestimmt. Je grösser die Sicherheitsanforderung (1-α), dessto grösser wird das Intervall.

$$(n-1)\overline{V_n} / \overline{\mu} \leq \sigma^2 \leq (n-1)\overline{V_n} / \underline{\mu}$$

$\overline{V_n}$ = (geschätzte) Varianz
α = Niveau
μ = Quantile der Chi Quadrat Verteilung ($\underline{\mu} = \chi^2_{n-1,a/2}$, $\overline{\mu} = \chi^2_{n-1,1-a/2}$)
→ Für mehr Informationen: siehe Hübner S. 197

n = Anzahl der Beobachtungen

1.4 Testen von Hypothesen

Dieses Kapitel hilft bei der Entscheidung von Ja-Nein-Fragen aufgrund einer Stichprobe. Die Frage kann z. B. lauten: „Ist das neue Produktionsverfahren im Mittel besser als das alte?“.

Der Bereich der durch eine Stichprobe in Frage gestellt wird, bezeichnet man als Hypothese H (ja, neues Verfahren besser), den anderen Bereich als Alternative H^c (nein, altes Verfahren besser).

Wenn ein Entscheidungskriterium formuliert wurde, das jeder Stichprobe eine Entscheidung für oder gegen H zuordnet, unterscheidet man zwei Arten von Fehl-Entscheidungen.

Fehler 1. Art	Wenn eine Hypothese zutrifft, aber trotzdem die Alternative verwendet wird
Fehler 2. Art	Wenn die Hypothese nicht zutrifft, aber trotzdem die Hypothese verwendet wird.

Ein kritischer Bereich K, bestimmt für welche Werte die Hypothese abgelehnt wird. Das wesentliche Problem für das Testen von Hypothesen besteht in der Suche nach einem geeigneten kritischen Bereich K.

Für eine Hypothese $H = \{N(a, \sigma_0^2), a \le a_0\}$ erhält man einen Test zum Niveau α durch den kritischen Bereich

$$K = \{\overline{X} \ge c_n\} \qquad \text{mit} \qquad c_n = a_0 + \mu_{1-\alpha}\sigma_0 / \sqrt{n}$$

1.5 Vergleiche mehrerer Stichproben

Es besteht die Möglichkeit, Vergleiche zwischen Stichproben aus unterschiedlichen Quellen zu ziehen.
Hier unterscheidet man zwischen 3 verschiedenen Fällen:

1. ***Verbundene Stichproben:***
   Daten kommen von einem Objekt oder einer Person.

   Statt X_i gibt es zwei nicht notwendig stochastisch unabhängige Zufallsvariablen $X_i^{(1)}$ und $X_i^{(2)}$. Sie sollten jedoch unabhängig identisch verteilt sein. Man betrachtet in diesem Falle die Differenz $X_i = X_i^{(1)} - X_i^{(2)}$, sofern die X_i als $N(a, \sigma^2)$-verteilt angenommen werden kann.

   Hier benützt man die in 1.4 besprochenen Verfahren.

2. ***2 Unverbundene Stichproben*** $\sigma_1 \cong \sigma_2$**:**
Streuung σ der Daten zweier Objekte stimmen näherungsweise überein.

Zwei Messreihen $X^{(1)}$ und $X^{(2)}$ mit $N\left(a_j, \sigma_j^2\right)$ - Verteilung sollen auf Gleichheit der Mittelwerte getestet werden.

a. $\overline{V}^* = \frac{\left[\left(n_1 - 1\right) * \overline{V}_{n_1}^{(1)} + \left(n_2 - 1\right) * \overline{V}_{n_2}^{(2)}\right]}{\left(n_1 + n_2 - 2\right)}$

b. Test-Statistik: $T\left(X\right) = \frac{\left(\overline{X}_{n_1}^{(1)} - \overline{X}_{n_2}^{(2)}\right)}{\left(\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} * \sqrt{\overline{V}^*}\right)}$

→ Für mehr Informationen: siehe Hübner S. 185

3. n **Unverbundene Stichproben** $\sigma_1 \cong \ldots \cong \sigma_n$**:**
Streuung σ der Daten mehrerer Objekte stimmen näherungsweise überein.

Die Messreihen $X^{(1)}$ und $X^{(n)}$ mit $N\left(a_j, \sigma^2\right)$ - Verteilung sollen auf Gleichheit der Mittelwerte getestet werden.

a. Mittlere Stichproben Varianz: $\overline{V}_{Ind} = \frac{1}{n-k} \sum_{j=1}^{k} \left(n_j - 1\right) * \overline{V}_{n_j}^{(j)}$

b. Gesamt Stichprobenmittel: $\overline{X}_n = \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_j} X_i^{(j)}$

c. Gesamt Stichprobenvarianz: $\overline{V}_n = \frac{1}{n-1} \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(X_i^{(j)} - \overline{X}_n\right)^2$

d. Varianz zwischen den Stichproben: $\overline{V}_{ZW} = \frac{1}{k-1} \sum_{j=1}^{k} n_j * \left(X_{n_j}^{(j)} - \overline{X}_n\right)^2$

e. Test-Statistik: $T\left(X\right) = \frac{\overline{V}_{ZW}}{\overline{V}_{Ind}}$

→ Für mehr Informationen: siehe Hübner S. 186

1.6 Chi-Quadrat-Anpassungstest

Der Chi-Quadrat-Test beruht auf dem Vergleich zwischen den aus einer Stichprobe vom Umfang n gewonnen Häufigkeiten $H_1,\ldots, H_k$ und der theoretisch erwarteten Verteilung der Verteilungsfunktion F, die als wahr angenommen wird. Die Aussage des Tests hat eine Sicherheit von 1-α.

$$Z_k = \sum_{i=1}^{k} (H_i \underbrace{- np_i}_{\text{Theoretische Häufigkeit nach } p_i})^2 / (np_i)$$

H_i = Häufigkeit eines Zustandes in der Stichprobe
p_i = gegebene Z-Dichten (Wahrscheinlichkeit in F)
n = Anzahl Elemente in der Stichprobe
k = Anzahl Zustände in Stichprobe

Diese Zufallsvariable wird mit dem (1-α)-Quantil der Chi(k-1)-Quadrat-Verteilung, $\chi^2_{k-1,1-\alpha}$, verglichen. Ist dabei $Z_k > \chi^2_{k-1,1-\alpha}$, dann gilt die Abweichung von der gegebenen Z-Dichte als erwiesen. Andernfalls ist zwar nicht gesichert, dass die Z-Dichte vorliegt, aber die Schwankungen liegen in einem akzeptablen Bereich.

1.7 Test auf Unabhängigkeit

Bei diesem Test wird eine mögliche Abhängigkeit bzw. Unabhängigkeit zweier Merkmale festgestellt. Man kann herausfinden, ob zwei scheinbar zusammenhängende Begebenheiten stochastisch abhängig oder unabhängig sind.

Beispiel:
Die Antworten (Y_i, Z_i) von n Befragten wurden in:

Y_i Jährliches Einkommen à r Klassen

Z_i Zeit seit dem letzten Arztbesuch à s Klassen

... eingeteilt

Die Antworten werden in eine Matrix $H_{j,k}$ (Spalte: Y_i, Reihe: Z_i) eingefüllt.

1. Man berechnet als erstes die Randsummen $H_j^Y = \sum_{k=1}^{s} H_{j,k}$ und $H_k^Z = \sum_{j=1}^{r} H_{j,k}$
2. **Nur Kontrolle:** Die Summe sollte n ergeben: $n = \sum_{j=1}^{r} H_j^Y + \sum_{k=1}^{s} H_k^Z$
3. Berechnung der Wahrscheinlichkeiten bzw des Schätzwertes:
   $p_{j,k} = \frac{H_j^Y}{n} * \frac{H_k^Z}{n}$
4. **Nur Kontrolle:** Die Summe sollte 1 ergeben: $1 = \sum_{j=1}^{r} \sum_{k=1}^{s} p_{j,k}$
5. Berechnung der Testgrösse Z: $Z = \sum_{j=1}^{r} \sum_{k=1}^{s} \frac{(H_{j,k} - n * p_{j,k})^2}{n * p_{j,k}}$
6. Ist Z Grösser als das geforderte Quantil der Chi-Quadrat-Verteilung ($\chi^2_{(r-1)*(s-1)}$) wird die Hypothese abgelehnt, andernfalls angenommen.s