Statistische Konzepte und die Berechnung von Konfidenzintervallen

1 Normalverteilung, Standardnormalverteilung, t-Student-Verteilung und Chi²-Verteilung

1.1 Normalverteilung

Die **Normalverteilung**, auch als Gauß-Verteilung bezeichnet, ist eine der grundlegendsten Wahrscheinlichkeitsverteilungen in der Statistik. Sie beschreibt Zufallsvariablen, die symmetrisch um ihren Mittelwert verteilt sind. Die Form dieser Verteilung ist charakteristisch glockenförmig und wird durch zwei Parameter bestimmt:

- Der *Mittelwert* (μ), der den Schwerpunkt der Verteilung angibt.
- Die *Standardabweichung* (σ), die die Breite der Glockenkurve beschreibt, also wie stark die Werte um den Mittelwert streuen.

Die Dichtefunktion der Normalverteilung ist gegeben durch:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Ein zentrales Merkmal der Normalverteilung ist die sogenannte *68-95-99,7-Regel*. Diese besagt, dass etwa 68% der Daten innerhalb einer Standardabweichung um den Mittelwert liegen, 95% innerhalb von zwei Standardabweichungen und 99,7% innerhalb von drei Standardabweichungen.

In der Praxis wird die Normalverteilung zur Modellierung vieler realer Phänomene verwendet, wie z.B. Fehlerverteilungen, Körpergrößen oder IQ-Werte, da viele natürliche Prozesse einem normalverteilten Muster folgen.

1.2 Standardnormalverteilung

Die **Standardnormalverteilung** ist ein Spezialfall der Normalverteilung, bei der der Mittelwert $\mu = 0$ und die Standardabweichung $\sigma = 1$ ist. Die Dichtefunktion der Standardnormalverteilung lautet:

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

Die Standardnormalverteilung ist ein nützliches Werkzeug in der Statistik, da sie es ermöglicht, beliebige Normalverteilungen durch eine Transformation auf diese Standardform zu bringen. Dies wird durch die Berechnung des *z-Scores* erreicht:

$$z = \frac{X - \mu}{\sigma}$$

Der z-Wert gibt an, wie viele Standardabweichungen ein Wert X vom Mittelwert entfernt ist. Diese Transformation erleichtert die Berechnung von Wahrscheinlichkeiten und statistischen Tests, da die Eigenschaften der Standardnormalverteilung bekannt sind und in Tabellen dokumentiert werden.

1.3 t-Student-Verteilung

Die **t-Student-Verteilung** wird verwendet, wenn die Stichprobengröße klein ist ($n < 30$) und die Varianz der Grundgesamtheit unbekannt ist. Sie ähnelt der Normalverteilung, hat jedoch schwerere“ Ränder. Dies bedeutet, dass extremere Werte wahrscheinlicher sind als bei der Normalverteilung.

Die t-Student-Verteilung hat einen Parameter, die Freiheitsgrade *df*, der die Form der Verteilung bestimmt. Je größer die Freiheitsgrade, desto mehr nähert sich die t-Verteilung der Normalverteilung an. Bei einer großen Stichprobengröße (ca. $df > 30$) ist der Unterschied zur Normalverteilung vernachlässigbar.

In der Statistik wird die t-Verteilung häufig für Hypothesentests verwendet, insbesondere wenn es um kleine Stichproben geht. Sie spielt auch bei der Berechnung von Konfidenzintervallen eine wichtige Rolle, wenn die Varianz der Grundgesamtheit nicht bekannt ist.

1.4 Chi²-Verteilung

Die **Chi²-Verteilung** entsteht als Summe der quadrierten z-Werte aus einer Standardnormalverteilung. Sie wird häufig in der Statistik verwendet, um Varianzen zu schätzen und in Hypothesentests wie dem *Chi²-Anpassungstest* oder dem *Chi²-Test zur Unabhängigkeit*.

Die Chi²-Verteilung hat einen Parameter, die Freiheitsgrade *df*, der die Form der Verteilung bestimmt. Sie ist asymmetrisch und nur für positive Werte definiert. Für $df \geq 30$ nähert sich die Chi²-Verteilung zunehmend der Normalverteilung an.

Ein häufiges Anwendungsgebiet der Chi²-Verteilung ist die Analyse von Kontingenztafeln, bei denen überprüft wird, ob zwei kategorische Variablen unabhängig voneinander sind.

2 Die Trapezregel zur numerischen Integration

2.1 Erklärung der Trapezregel

Die **Trapezregel** ist eine Methode der numerischen Integration, die darauf basiert, die Fläche unter einer Kurve durch Trapeze zu approximieren. Wenn eine Funktion $f(x)$ in einem Intervall $[a, b]$ integriert werden soll, wird das Intervall in kleinere Teilintervalle unterteilt und die Fläche unter der Kurve wird durch die Summe der Flächen der Trapeze angenähert.

Die Trapezregel lautet für ein Intervall $[a, b]$:

$$\int_a^b f(x)dx \approx \frac{b-a}{2} [f(a) + f(b)]$$

Durch die Unterteilung des Intervalls in viele kleine Abschnitte (mit gleichmäßiger Schrittweite Δx) verbessert sich die Genauigkeit der Approximation.

2.2 Anwendung der Trapezregel im Code

In unserem Python-Code wird die Trapezregel verwendet, um die Fläche unter der Standardnormalverteilung zu berechnen. Diese Fläche wird benötigt, um die Konstante c zu finden, die für die Berechnung des Konfidenzintervalls wichtig ist. Die Funktion integriert die Standardnormalverteilung $f(x) = \frac{1}{\sqrt{2\pi}} e^{-0.5x^2}$, bis der Anteil der Fläche dem Konfidenzniveau γ entspricht.

Im Code wird dies folgendermaßen umgesetzt:

```
while flaeche < gesucht:
    flaeche += (ende - start) / 2 * (standardnormalverteilung(start) + standardnormal
    start += 0.001
    ende += 0.001
```

Die Schleife summiert iterativ die Flächen der Trapeze, bis der gewünschte Wert erreicht ist.

3 Bedeutung des Konfidenzniveaus Gamma

Das **Konfidenzniveau** γ gibt an, wie sicher man sein kann, dass ein berechnetes Konfidenzintervall den wahren Parameter (z.B. den Mittelwert) der Grundgesamtheit enthält. Typische Konfidenzniveaus sind:

- 95% ($\gamma = 0.95$): Das Konfidenzintervall enthält den wahren Wert in 95 von 100 Fällen.
- 99% ($\gamma = 0.99$): Mit 99% Wahrscheinlichkeit liegt der wahre Wert im berechneten Intervall.

Ein höheres Konfidenzniveau führt zu einem breiteren Intervall, da man sicherstellen möchte, dass der wahre Wert mit größerer Wahrscheinlichkeit innerhalb des Intervalls liegt.

3.1 Interpretation des Konfidenzintervalls

Das Konfidenzintervall stellt den Bereich dar, in dem der wahre Parameter der Grundgesamtheit mit einer Wahrscheinlichkeit von γ liegt. Beispielsweise bedeutet ein 95%-Konfidenzintervall, dass wir zu 95% sicher sind, dass der wahre Mittelwert der Population in diesem Intervall liegt.

In der Praxis wird das Konfidenzintervall oft zur Einschätzung der Genauigkeit von Schätzungen verwendet. Je größer das Konfidenzintervall, desto unsicherer ist die Schätzung des Parameters.

4 Analyse des Datensatzes und Berechnung des Konfidenzintervalls

4.1 Der Datensatz

Der Datensatz, in unserem Beispiel `Daten_TV.txt`, enthält numerische Messwerte einer bestimmten Variablen (z.B. TV-Zuschauerzahlen). Um die statistischen Parameter (Mittelwert, Varianz) zu berechnen, wird dieser Datensatz im Code eingelesen und ausgewertet. Die Anzahl der Datenpunkte n, der Mittelwert μ sowie die Varianz σ^2 werden berechnet, um anschließend das Konfidenzintervall zu bestimmen.

4.2 Berechnung des Mittelwerts und der Varianz

Der Mittelwert μ eines Datensatzes gibt den Durchschnitt der Werte an und wird folgendermaßen berechnet:

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

wobei x_i die einzelnen Datenpunkte und n die Anzahl der Datenpunkte ist. Im Python-Code sieht diese Berechnung so aus:

```
mittelwert = np.sum(A) / n
```

Die **Varianz** σ^2 gibt an, wie stark die Daten um den Mittelwert streuen. Eine höhere Varianz deutet auf eine größere Streuung hin. Die Varianz wird mit der Formel:

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu)^2$$

berechnet. Diese Formel liefert die erwartungstreue Schätzung der Varianz für eine Stichprobe. Im Python-Code wird sie so umgesetzt:

```
varianz = np.sum((A - mittelwert) ** 2) / (n - 1)
```

4.3 Berechnung des Konfidenzintervalls

Das Konfidenzintervall basiert auf der Berechnung von μ, σ^2 und einer Konstanten c, die vom Konfidenzniveau γ abhängt. Für eine Normalverteilung kann das Konfidenzintervall durch die Formel:

$$\left(\mu - c \cdot \frac{\sigma}{\sqrt{n}}, \mu + c \cdot \frac{\sigma}{\sqrt{n}} \right)$$

berechnet werden, wobei $\sigma = \sqrt{\sigma^2}$ die Standardabweichung und n die Stichprobengröße ist. Der Wert von c wird aus der Standardnormalverteilung bestimmt, um die entsprechende Fläche unter der Kurve abzudecken.

Im Python-Code wird das Konfidenzintervall wie folgt berechnet:

```
oben = mu + c * np.sqrt(varianz / n)
unten = mu - c * np.sqrt(varianz / n)
```

Dieses Intervall gibt den Bereich an, in dem der wahre Mittelwert der Population mit einer Wahrscheinlichkeit von γ liegt.

4.4 Interpretation des Konfidenzintervalls

Das Konfidenzintervall stellt den Bereich dar, in dem der wahre Parameter der Grundgesamtheit mit einer Wahrscheinlichkeit von γ liegt. Beispielsweise bedeutet ein 95%-Konfidenzintervall, dass wir zu 95% sicher sind, dass der wahre Mittelwert der Population in diesem Intervall liegt.

In der Praxis wird das Konfidenzintervall oft zur Einschätzung der Genauigkeit von Schätzungen verwendet. Je größer das Konfidenzintervall, desto unsicherer ist die Schätzung des Parameters. Wenn das Konfidenzintervall beispielsweise sehr breit ist, deutet dies auf eine größere Unsicherheit bezüglich des wahren Werts hin. Ein schmaleres Intervall hingegen bedeutet, dass die Schätzung präziser ist.

5 Wesentliche Teile des MATLAB-Codes

Nachfolgend werden die zentralen Teile des MATLAB-Codes präsentiert, der zur Berechnung des Mittelwerts, der Varianz und des Konfidenzintervalls aus einem Datensatz dient.

5.1 Mittelwert und Varianz

Der Mittelwert μ und die Varianz σ^2 werden mit den MATLAB-Befehlen `mean()` und `var()` berechnet. Dies sind Standardbefehle zur Berechnung von Mittelwert und Varianz einer Stichprobe.

```
% Daten laden
data = load('Daten_TV.txt');

% Berechnung des Mittelwerts
n = length(data);
mu = mean(data);

% Berechnung der Varianz
varianz = var(data, 1); % Erwartungstreue Varianz
```

5.2 Berechnung des Konfidenzintervalls

Um das Konfidenzintervall zu berechnen, benötigen wir das Konfidenzniveau γ sowie die Konstante c, die aus der Standardnormalverteilung entnommen wird. MATLAB bietet hierfür die Funktion `norminv()`, mit der der entsprechende z-Wert für das Konfidenzniveau ermittelt wird.

```
% Konstante c für das gegebene Konfidenzniveau gamma bestimmen
gamma = 0.99; % Beispiel für 99%-Konfidenzintervall
c = norminv((gamma + 1) / 2);

% Konfidenzintervall berechnen
sigma = sqrt(varianz);
unten = mu - c * sigma / sqrt(n);
oben = mu + c * sigma / sqrt(n);
```

```
% Ausgabe des Konfidenzintervalls
fprintf('Konfidenzintervall: [%f, %f]\n', unten, oben);
```

Das berechnete Intervall gibt an, in welchem Bereich der wahre Mittelwert der Grundgesamtheit mit 99% Wahrscheinlichkeit liegt.

6 Zusammenfassung

In dieser Arbeit wurden die Grundlagen der Normalverteilung, Standardnormalverteilung, t-Student-Verteilung und Chi²-Verteilung sowie deren Anwendung im statistischen Kontext erläutert. Ein besonderer Fokus lag auf der Berechnung von Konfidenzintervallen, die eine zentrale Rolle in der statistischen Schätztheorie spielen.

Das Konfidenzniveau γ beschreibt die Wahrscheinlichkeit, mit der das Intervall den wahren Mittelwert der Grundgesamtheit enthält. Höhere Konfidenzniveaus führen zu breiteren Intervallen, was mehr Sicherheit auf Kosten der Genauigkeit bedeutet.

Anhand des gegebenen Python- und MATLAB-Codes wurde demonstriert, wie Mittelwert, Varianz und das Konfidenzintervall aus einem Datensatz berechnet werden können. Die Anwendung der Trapezregel zur numerischen Integration wurde ebenfalls im Code implementiert, um die Konstante c für das Konfidenzintervall zu bestimmen.

Dieses Wissen ist von zentraler Bedeutung für viele Bereiche der Statistik und der Datenanalyse, insbesondere bei der Arbeit mit kleinen Stichproben oder bei unbekannten Varianzen der Grundgesamtheit.