Deskriptive Statistik

Deskriptive Statistik umfasst Methoden zur Sammlung, Organisation, Zusammenfassung und Darstellung von Daten. Ihr Hauptziel ist es, die grundlegenden Merkmale eines Datensatzes verständlich darzustellen, ohne dabei Schlussfolgerungen über eine größere Population zu ziehen.

Hauptkomponenten der Deskriptiven Statistik

1. **Lagemaße - Maßzahlen der zentralen Tendenz**:
    - **Mittelwert (arithmetisches Mittel)**: Summe aller Werte geteilt durch die Anzahl der Werte.
    - **Median**: Der mittlere Wert eines sortierten Datensatzes.
    - **Modus**: Der am häufigsten vorkommende Wert in einem Datensatz.
2. **Streuungsmaße**:
    - **Varianz**: Das durchschnittliche Quadrat der Abweichungen vom Mittelwert.
    - **Standardabweichung**: Die Quadratwurzel der Varianz, ein Maß für die Streuung der Daten um den Mittelwert.
    - **Spannweite (Range)**: Differenz zwischen dem größten und dem kleinsten Wert.
    - **Interquartilsabstand (IQR)**: Differenz zwischen dem oberen (75. Perzentil) und unteren Quartil (25. Perzentil).
3. **Lageparameter**:
    - **Perzentile**: Perzentile sind spezifische Quantile, die eine Datenverteilung in 100 gleich große Teile teilen. Sie sind nützlich, um zu verstehen, wie sich bestimmte Werte in einer Datenmenge relativ zu anderen Werten positionieren.
    - **Quantile**: Quantile sind Werte, die eine Datenmenge in gleich große Teile teilen. Sie sind ein Maß dafür, wie Daten in einem Datensatz verteilt sind. Typische Quantile sind das Median (50. Perzentil), das erste Quartil (25. Perzentil) und das dritte Quartil (75. Perzentil). Quantile helfen dabei, die Verteilung und Streuung von Daten zu verstehen, indem sie angeben, wie sich Daten in Prozentpunkten oder Fraktionen unterteilen lassen.
4. **Grafische Darstellungen**:

- **Histogramme**: Zeigen die Verteilung der Daten.
- **Boxplots**: Visualisieren die Verteilung der Daten und identifizieren Ausreißer.
- **Balkendiagramme**: Zeigen die Häufigkeit oder den Anteil von Kategorien.
- **Streudiagramme (Scatterplots)**: Zeigen die Beziehung zwischen zwei Variablen.

Induktive Statistik

Induktive Statistik (oder schließende Statistik) verwendet Stichprobendaten, um Schlussfolgerungen oder Vorhersagen über eine größere Population zu ziehen. Sie verwendet Wahrscheinlichkeitsmodelle, um die Unsicherheit bei diesen Schlussfolgerungen zu quantifizieren.

Hauptkomponenten der Induktiven Statistik

1. **Hypothesentests**:
   Hypothesentests sind statistische Verfahren, die verwendet werden, um zu prüfen, ob eine Annahme über eine Population oder Stichprobe auf der Grundlage der verfügbaren Daten unterstützt oder widerlegt werden kann. Sie helfen dabei, Schlussfolgerungen über eine bestimmte Hypothese zu ziehen, indem sie die Wahrscheinlichkeit berechnen, dass die beobachteten Daten unter der Annahme wahr sind (p-Wert).

   - **t-Tests**: Überprüfen, ob die Mittelwerte von zwei Gruppen unterschiedlich sind.
   - **Chi-Quadrat-Tests**: Überprüfen, ob es eine Assoziation zwischen zwei kategorialen Variablen gibt.
   - **ANOVA (Analysis of Variance)**: Überprüft, ob es signifikante Unterschiede zwischen den Mittelwerten mehrerer Gruppen gibt.

2. **Konfidenzintervalle**:
   Konfidenzintervalle sind Bereiche von Werten um einen Schätzwert herum, die angeben, wie sicher wir sind, dass der wahre Wert eines Parameters innerhalb dieses Intervalls liegt. Sie geben eine Spanne von Werten an, innerhalb derer wir erwarten können, dass der wahre Wert mit einer bestimmten Wahrscheinlichkeit (Konfidenzniveau) enthalten ist.

3. **Regression**:

- **Lineare Regression**: Modelliert die Beziehung zwischen einer abhängigen und einer unabhängigen Variablen.
- **Multiple Regression**: Erweiterung der linearen Regression mit mehr als einer unabhängigen Variablen.

4. **Bayes'sche Statistik**:
   Die Bayes'sche Statistik ist ein Ansatz, der sich auf die Verwendung von Bayes's Theorem zur Aktualisierung von Wahrscheinlichkeiten basiert, wenn neue Daten oder Informationen verfügbar werden. Es erlaubt uns, Wahrscheinlichkeiten für Hypothesen oder Modelle zu aktualisieren, indem wir vorhandenes Wissen (Prior) mit neuen Beobachtungen (Evidenz) kombinieren, um ein aktualisiertes Verständnis der Welt zu erhalten (Posterior).

deskriptiv vs induktiv

Die deskriptive Statistik ist der erste Schritt in der Datenanalyse, bei dem die Daten beschrieben und visualisiert werden. Die Inferenzstatistik geht darüber hinaus und verwendet diese beschreibenden Maßnahmen, um fundierte Schlussfolgerungen über die zugrunde liegende Population zu ziehen, wobei Unsicherheiten quantifiziert werden. Beide Methoden sind essentiell für eine umfassende Datenanalyse und ergänzen sich gegenseitig.

Explorative Datenanalyse (EDA)

Die **Explorative Datenanalyse (EDA)** spielt eine zentrale Rolle in der Datenanalyse, indem sie den Übergang zwischen der deskriptiven und der Inferenzstatistik erleichtert. Sie dient dazu, erste Einblicke in die Daten zu gewinnen, Muster zu erkennen und Hypothesen zu generieren, die dann in der inferentiellen Statistik formell getestet werden können. Hier ist eine detaillierte Erklärung der EDA im Kontext der deskriptiven und Inferenzstatistik:

Ziele der Explorativen Datenanalyse (EDA)

1. **Verständnis der Daten**: EDA hilft dabei, ein grundlegendes Verständnis der Struktur, Verteilung und Eigenschaften der Daten zu erlangen.
2. **Erkennung von Mustern und Zusammenhängen**: Identifizieren von Trends, Beziehungen zwischen Variablen und anderen relevanten Mustern in den Daten.

3. **Identifikation von Anomalien**: Erkennen von Ausreißern, fehlenden Werten und anderen Unregelmäßigkeiten in den Daten.
4. **Hypothesengenerierung**: Entwickeln erster Hypothesen, die in späteren, formelleren Analysen getestet werden können.

Methoden der Explorativen Datenanalyse (EDA)

Grafische Methoden

1. **Histogramme**: Zeigen die Verteilung der Daten und helfen, die Form der Verteilung zu verstehen.
2. **Boxplots**: Visualisieren die Verteilung, Zentralwerte und Streuung sowie potenzielle Ausreißer.
3. **Streudiagramme (Scatterplots)**: Zeigen die Beziehung zwischen zwei kontinuierlichen Variablen.
4. **Heatmaps**: Visualisieren die Intensität von Datenpunkten in einer Matrixform, oft verwendet für Korrelationen.
5. **Balkendiagramme und Kreisdiagramme**: Zeigen die Häufigkeit oder Anteile von kategorialen Daten.

Numerische Methoden

1. **Zusammenfassende Statistiken**: Berechnung von Mittelwert, Median, Modus, Standardabweichung, Varianz, Quartilen und Perzentilen.
2. **Korrelation**: Untersuchung der Beziehungen zwischen Variablen, z.B. mittels Pearson-Korrelation oder Spearman-Rangkorrelation.
3. **Kreuztabellen**: Analyse der Häufigkeiten von zwei oder mehr kategorialen Variablen.

Anwendung der Explorativen Datenanalyse (EDA)

- EDA verwendet ähnliche Werkzeuge wie die deskriptive Statistik, geht jedoch einen Schritt weiter, indem sie Muster und Beziehungen in den Daten untersucht.
- Durch das Erkennen von Anomalien und Unregelmäßigkeiten hilft EDA, Daten für die deskriptive Analyse zu bereinigen und zu strukturieren.
- EDA generiert Hypothesen, die später durch formale inferenzstatistische Methoden getestet werden können.

- Sie hilft, potenzielle unabhängige Variablen für Regressionsmodelle zu identifizieren und zu prüfen, ob die Annahmen für inferenzstatistische Tests erfüllt sind.

Ablauf einer Explorativen Datenanalyse

1. **Datensammlung**: Sammlung der relevanten Daten.
2. **Datenbereinigung**: Identifikation und Behandlung fehlender Werte, Entfernung von Ausreißern.
3. **Erste Visualisierungen**: Erstellen von Histogrammen und Boxplots, um die Verteilung der Daten zu verstehen.
4. **Zusammenfassende Statistiken**: Berechnung von Mittelwerten, Medians, Standardabweichungen usw.
5. **Untersuchung von Zusammenhängen**: Erstellung von Streudiagrammen und Berechnung von Korrelationen zwischen Variablen.
6. **Hypothesengenerierung**: Basierend auf den beobachteten Mustern und Beziehungen, werden erste Hypothesen formuliert.

Beispiel für Ziele der EDA

1. Verständnis der Daten:

Du hast einen Datensatz mit den monatlichen Verkaufszahlen eines Online-Shops. Ein erstes Histogramm zeigt dir, dass die Verteilung der Verkäufe nicht normalverteilt ist, sondern eine positive Schiefe aufweist (mehr niedrigere Verkaufszahlen als hohe).

2. Erkennung von Mustern und Zusammenhängen:

Ein Streudiagramm zeigt dir eine positive Korrelation zwischen den Werbeausgaben und den Verkaufszahlen, was darauf hindeutet, dass höhere Werbeausgaben mit höheren Verkäufen einhergehen.

3. Identifikation von Anomalien:

Ein Boxplot der Verkäufe im Monat Januar zeigt einen extrem hohen Wert, der als Ausreißer identifiziert wird. Dieser Ausreißer könnte auf einen Datenerfassungsfehler oder ein besonderes Ereignis wie einen großen Rabatt hindeuten.

4. Hypothesengenerierung:

Basierend auf den Mustern im Streudiagramm entwickelst du die Hypothese, dass die Verkaufszahlen signifikant mit den Werbeausgaben zusammenhängen und möchtest diesen Zusammenhang später durch eine Regressionsanalyse testen.