**Klausurvorbereitung**

**Theorie**

**Statistik**

* **Lehre von Verfahren und Methoden zur Gewinnung, Erfassung, Analyse, Charakterisierung, Abbildung, Nachbildung und Beurteilung von beobachtbaren Daten über die Wirklichkeit** (Empirie
* **systematische Zusammenstellung von Zahlen und Daten zur Beschreibung von Zuständen, Entwicklungen und Phänomen**

**Beschreibende Statistik**

* **sammelt Daten bei allen Untersuchungseinheiten**, über die man Informationen erhalten will
* empirische Daten **durch Tabellen, Kennzahlen, (Maßzahlen oder Parameter) und Grafiken**
* **Beurteilung von Daten basierend auf Stichproben**
* Dient der Daten-Beobachtung / Datenbeurteilung
* **Konzentration auf das Wesentliche**

**Schließende Statistik**

* **Schluss von einer Teilmenge (Stichprobenergebnissen) auf das Eigenschaften der Grundgesamtheit (GG)**
* **Grund für Stichproben: enorme Größe der GG**
* **Stichprobe: Auswahl einer Teilmenge aus der GG** zu der Informationen gesammelt werden sollen
* **Datenbeurteilung (Analyse) durch Schlüsse auf Basis unvollständiger Daten**
* Dient dem **Beweis oder der Widerlegung von Hypothesen, die sich auf GG beziehen**
* Basiert auf Wahrscheinlichkeitsrechnung

**Auswahlverfahren zur Stichprobe oder Umfang zur Stichprobe?**

* Das **Auswahlverfahren zur Erhebung der Stichproben und Bestimmung des Stichprobenumfangs ist wichtiger als der Stichprobenumfang.**
* **Auswahlverfahren**: **Art und Weise, wie Elemente der Stichprobe zweckmäßig ausgewählt werden**
* Auswahlverfahren bestimmt Stichprobenumfang.
* **schlechte Auswahl zur Stichprobe liefert unabhängig vom Stichprobenumfang keine brauchbaren Erkenntnisse über die GG** (Stichprobe ist „verzerrt“ bzw. „nicht repräsentativ

**willkürliche Auswahl**

* **unkontrollierte Aufnahme eines Elements aus der GG in die Stichprobe**
* **kein Auswahlplan**
* **Interviewer sind frei in der Auswahl ihrer Interviewpartner**.

**zufällige Auswahl (Random-Auswahl)**

* **Zufallsgesteuert**
* **jedes Element aus GG gelangt mit gleicher Wahrscheinlichkeit** **in die Stichprobe**
* **Voraussetzung: Liste/Datei zu allen Elementen der GG**
* **ermöglicht Berechnung von Stichprobenfehler** (durch Wahrscheinlichkeitsrechnung)
* **Einfache Zufallsstichprobe**:
  + **Jede Stichprobe hat dieselbe Chance ausgewählt zu werden**
* **Geschichtete Zufallsstichprobe**
  + **Elemente der GG werden in Gruppen** (Schichten) **eingeteilt** (ein Element gehört genau zu einer Gruppe (Schicht))
  + **nach Gruppierung werden einfache Zufallsstichproben aus jeder Gruppe genommen**
* **Klumpenstichprobe**
  + **Einfache Zufallsauswahl aus zusammengefassten Elementen (Klumpen, Cluster)**
  + Stichprobe wird aus allen Elementen bestimmter Cluster genommen

**Quota-Auswahl**

* **bewusstes Auswahlverfahren**
* ausgewählte **Quotierungsmerkmale sollen in der Stichprobe dieselbe Verteilung wie in der Grundgesamtheit** erreichen.
* **Quotenpläne** **für ausgewählte Merkmale (Geschlecht, Alter, Beruf)** **ermöglichen dieselbe Verteilung in der Stichprobe wie in der GG**
* Verteilung der Quotierungsmerkmale in der GG muss bekannt sein

**TED-Umfrage im Fernsehen**

* **willkürliche Auswahl**
* **kein Auswahlplan**   
  (jeder Zuschauer kann an Umfrage einmal, mehrmals oder nicht teilnehmen

**Klassierung von Daten**

**Warum „bis unter“ zum Wertebereich bei klassierten Daten?**

„bis unter“

* **ermöglicht eindeutige Klassierung bei diskreten metrischen Merkmalen**
* **vermeidet Überschneidungen bei den Merkmalsausprägungen**   
  (Klassen 100 bis 200 € und 200 bis 300 € beinhalten beide einen Umsatz von 200,00 €)
* **vermeidet Lücken bei den Merkmalsausprägungen**   
  (Klassen 100 bis 199 € und 200 bis 300 beinhalten keinen Umsatz zwischen 199,01 und 199,99 €)

**Probleme bei Klassierung von Daten**

* **Zielkonflikt: Übersichtlichkeit vs. Informationsverlust**
* **Berechnung von exakten statistischen Kennzahlen** (z.B. **Mittelwerte**) ist bei klassierten Daten **nicht möglich**   
  **Mittelwerte, Quantile sind bei klassierten Daten nur Näherungswerte**
* **Näherungswerte können nur unter bestimmten Annahmen** (z.B. Gleichverteilung) **berechnet** werden

**offenen Randklassen**

* besitzen **keine Klassenuntergrenze** (z. B. „bis unter 50 kg“) oder **keine Klassenobergrenze** (z. B. „120 kg und schwerer“)
* **Berechnung von Klassenbreite und Klassenmitte nicht möglich**

**Berechnung der Klassenbreite bi für die Klasse „150 b.u. 180 cm“**

* **Klassenbreite bi** = **xk – xk-1** = **180 – 150 = 30 cm**

**Berechnung der Klassenmitte mi für die Klasse „150 b.u. 180 cm“**

* Klassenmitte **mi = (xk + xk-1) ½** = (180 + 150) / 2 = 330 / 2 = 165 cm

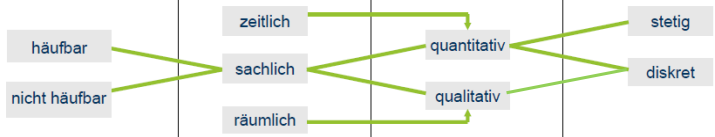
**Entscheidungen bei Klassierung**

* Anzahl der Klassen
* proportionale Klassenbreite bei allen Werten (alle gleich oder unterschiedliche Klassenbreite)
* obere oder untere Randklasse („bis unter 50 kg“, „120 kg und schwerer“
* Klassenobergrenzen / Klassenuntergrenzen

**Histogramm**

* **flächenproportionale Darstellung der Häufigkeiten von klassierten Daten**
* **x-Achse muss Skala mit geordneten Werten und gleichen Abständen zwischen den Werten** (z. B. 200, 400, 600, 800) sein
* **Rechtecke liegen auf x-Achse nebeneinander** (ohne Abstände)
* **Rechteckbreite = Klassenbreite bi**   
  (bei Klasse „150 b.u. 180“ wird die Rechteckbreite zwischen den Werten 150- 180 auf der x-Achse gezeichnet)
* **Rechteckhöhe** = absolute Klassenhäufigkeit hi / Klassenbreite bi 🡆 **hi / bi**

**Merkmale** **Merkmalstypen**



**Merkmale** sind

* **häufbar**
* **nicht häufbar**

**Merkmale** werden

* **zeitlich**
* **sachlich**
* **räumlich**

**abgegrenzt**

**Merkmale** haben **Merkmalstyp**

* **qualitativ**
* **quantitativ**

**Qualitative Merkmale = nicht metrische Merkmale**

* **Nicht metrische Skalen**
* **lassen sich nicht mit Zahlen messen**
* **sind immer diskret** (**haben nur eine abzählbare Menge möglicher Merkmalswerte**)
* **Codierung und Rangordnung möglich**
* **Keine Rechenoperationen** mit den Merkmalsausprägungen zulässig
* **Skalenniveau: Nominalskala und Ordinalskala (Rangskala)**  
  **Beispiele:**  
  Geschlecht (Nominalskala)  
  Güteklasse (Rangskala)  
  Umsatzklasse (Rangskala)  
  Wohnort (Nominalskala)  
  Kundenzufriedenheit, gemessen auf einer Skala von „1 = sehr zufrieden“ bis „5 = sehr unzufrieden“ (Rangskala)  
  Beruf (Nominalskala)  
  Steuerklasse (Rangskala)  
  Einkommensklasse (Rangskala)

**Skalenniveau: nicht metrische Skalen**

* **bei qualitativen Merkmalen**
* Keine Rechenoperationen zulässig
* nur Vergleichsoperationen zulässig = ≠ < >
* **Nominalskala**
* **Ordinalskala ( = Rangskala)**

|  |  |
| --- | --- |
| **Nominalskala** | **qualitatives Merkmal**  **nicht metrische Skala mit diskreten Merkmalen**  **diskret:**  **abzählbare Menge möglicher Merkmalsausprägungen**  Merkmal „Familienstand“ hat z. B. 3 Merkmalsausprägungen: „ledig“, „verheiratet“, „geschieden“) |
| * endliche Menge, * unterliegen keiner logischen Rangfolge * Zuordnung von Zahlen ist nur Kodierung der Merkmalsausprägungen * Daten sind nicht vergleichbar * nur Vergleichsoperationen =, ≠ möglich * **Gleichwertigkeit (Steuerklasse I** ≠ Steuerklasse IV |
| **Beispiele:**  Geschlecht, Familienstand, Steuerklasse, Postleitzahl,  Rechtsform eines Unternehmens, Wohnort, Beruf |

|  |  |
| --- | --- |
| **Ordinalskala** **(Rangskala)** | **qualitatives Merkmal**  **nicht metrische Skala mit diskreten Merkmalen**  **diskret:**  **abzählbare Menge der Merkmalsausprägungen**  Merkmal „Konfektionsgröße“ hat 3 Merkmalsausprägungen: „“M“, „L“, „XL“) |
| * endliche Menge von Ausprägungen, * hat Rangmerkmal (natürliche Rangfolge der Daten möglich) * **Anordnung, Rangfolge** * Ordnungsprinzip ist die Stärke/Grad der Intensität * Abstände zwischen den Ausprägungen können nicht interpretiert werden * Daten sind vergleichbar (z. B. Konfektionsgrößen XXL > XL > L) * **nur Vergleichsoperationen =, ≠, <, > möglich** |
| **Beispiele**  Punkte in einer Spieletabelle:  **1. Platz 20 Punkte, 2. Platz 10 Punkte**  natürliche Rangfolge durch die Punkte 🡆 keine Aussage zum Abstand zwischen 1. Und 2. Platz möglich   (1. Platz hat die doppelte Punktanzahl, jedoch muss das Team   auf dem 1. Platz nicht doppelt so gut sein  Konfektionsgröße  Schulnoten / Klausurnoten  Windstärke  Umsatzklasse eines Unternehmens  Bewertungsskala (zur Kundenzufriedenheit)  Einkommensklasse |

**Quantitative Merkmale = metrische Merkmale**

* **alle** **quantitativen Merkmale sind metrisch**!
* **Metrische Skalen sind Kardinalskalen (endliche (abzählbare) Anzahl (Menge) der Elemente)**
* Unterscheidung in
  + **stetige Merkmale**
    - **Menge der Merkmalsausprägungen überabzählbar**
    - **Intervall der reellen Zahlen** (weitere **Zwischenwerte zwischen 2 Ausprägungen**)   
      **Beispiele  
      Gewicht  
      Alter (wenn nicht Alter als ganzzahliger Wert in Jahren 🡆 dann ist es diskret)  
      Fahrzeit**
  + **diskrete Merkmale**
    - **Menge der Merkmalsausprägungen ist endlich bzw. abzählbar**   
      (i.d.R. **ganze Zahlen**)   
      **Beispiele**Kinderzahl  
      Sitzplätze  
      monatliches Gehalt

**Skalenniveau: metrische Skalen**

* **bei quantitativen Merkmalen**
* **Rechenoperationen sind zulässig (+ - \* /)**
* Skala **hat Maßeinheit und Nullpunkt**(muss kein natürlicher Nullpunkt sein   
  (z. B. Temperatur in °C ist Intervallskala 🡆 hat Nullpunkt, aber keinen natürlichen Nullpunkt)
* Metrische Skalen (Kardinalskalen) **können auf alle statistische Verfahren angewandt** werden
* **Intervallskala**
* **Verhältnisskala (Ratioskala)**
* **Absolutskala**

z. B.

**Häufigkeiten,**

**arithmetisches Mittel**,

**Streuungsmaße** wie:

Standardabweichung,

Varianz,

Spannweite

* Metrische Skalen (Kardinalskalen) **sind abwärtskompatibel**
  + jedes Verfahren, dass auf niedrigem Skalenniveau angewandt werden kann, darf auch auf höherem Skalenniveau angewandt werden

z. B. darf Median-Berechnung zu einer Ordinalskala auch auf einer  
 Kardinalskala (Intervallskala) angewandt werden

|  |  |
| --- | --- |
| **Intervallskala** | **quantitatives Merkmal**  **metrische Skala mit diskreten oder stetigen Merkmalen**  **diskret:**  **abzählbare Menge der Merkmalsausprägungen**  Merkmal „Kinderanzahl“ hat die drei Merkmalsausprägungen „1“, „2“, „3 und mehr“  **stetig:**  **Intervall reeller Zahlen** (weitere **Zwischenwerte zwischen 2 Ausprägungen**)  Merkmal „Temperatur in °C“ kann zwischen den Temperaturwerten auf der Skala weitere Zwischenwerte haben. |
| * **kein natürlicher Nullpunk**t **natürlicher Nullpunkt** ist **von Natur** aus gegeben und **nicht veränderbar**. Zur Einordnung der Skala die Frage stellen **Wie wurde der natürliche Nullpunkt bestimmt?  (natürlich oder willkürlich vom Menschen)** **nicht natürlicher (willkürlicher) Nullpunkt ist z. B. °C** es könnte auch jeder beliebiger anderer Punkt als 0°C sein **natürlicher Nullpunkt durch physikalische Größen** (z. B. Gramm) **oder durch logische Annahmen** * **Ausprägungen können auch Wert < 0 haben** * **gleichgroße Intervalle zwischen den Ausprägungen** (z. B. **Temperaturskala °C** enthält gleich große Intervalle von 1°C) * **gleichgroße Intervallabstände sind interpretierbar  (Temperatur in °C)** * **Reihenfolgen** und **quantifizierbare Abstände** möglich * **Differenzbildung** möglich (6°C – 2°C = 12°C – 8°C) * Daten können **unendlich viele Ausprägungen im Intervall** annehmen * **keine Verhältnisbildung möglich** * **Vergleichsoperationen =, ≠, <, > und Rechenoperation +, - möglich** * **keine Multiplikation oder Division  (da keine Verhältnisbildung möglich)** |
| **Beispiele**  Längendifferenzen, Temperatur in °C, IQ, Geburtsjahr    Diese Merkmale haben keinen natürlichen Null-Punkt   Einige dieser Merkmale können Werte < 0 haben  gleichgroße Intervalle, keine sinnvolle Verhältnisbildung möglich  (ausgenommen Temp. in ° C)  Mensch mit IQ 130 ist nicht doppelt so schlau wie Mensch mit IQ 65  Geb.-Jahr 1970 ist nicht 1,5 mal größer als Geb.-Jahr 1300 |

|  |  |
| --- | --- |
| **Verhältnisskala  (Ratioskala)** | **quantitatives Merkmal**  **metrische Skala mit diskreten oder stetigen Merkmalen**  **diskret:**  **abzählbare Menge der Merkmalsausprägungen**  Merkmal „Umsatz“ hat die drei Merkmalsausprägungen „10 Tsd“, „20 Tsd.“, „30 Tsd. und mehr“  **stetig:**  **Intervall reeller Zahlen** (weitere **Zwischenwerte zwischen 2 Ausprägungen**)  Merkmal „Temperatur in K“ kann zwischen den Temperaturwerten auf der Skala weitere Zwischenwerte haben. |
| * **natürlicher Nullpunk**t **natürlicher Nullpunkt** ist **von Natur** aus gegeben und **nicht veränderbar**. Zur Einordnung der Skala die Frage stellen **Wie wurde der natürliche Nullpunkt bestimmt?  (natürlich oder willkürlich vom Menschen)** **natürlicher Nullpunkt ist z. B. Temperatur in K natürlicher Nullpunkt durch physikalische Größen** (z. B. Gramm) **oder durch logische Annahmen** * **Werte der Verhältnisskala sind immer ≥ 0** * **Verhältnisbildung** (Quotientenbildung) möglich **Quotienten der gemessenen Werte werden verglichen** * **Vergleichsoperationen =, ≠, <, > und Rechenoperation +, -, \*, : möglich** * **Multiplikation oder Division möglich (da Verhältnisbildung möglich)** |
| **Beispiele**  Temperatur in °K, Körpergröße, Gewicht, Umsatz, Einkommen,  Geschwindigkeit, Preis, Länge, Breite, Alter, Klausur-Punkte    all diese Merkmale können nur Werte ≥ 0 haben  zu all diesen Merkmalen ist Verhältnisbildung möglich   (2faches Einkommen, Gewicht, Umsatz, Preis, Alter … |

|  |  |
| --- | --- |
| **Absolutskala** | **quantitatives Merkmal**  **metrische Skala mit diskreten oder stetigen Merkmalen**  **diskret:**  **abzählbare Menge der Merkmalsausprägungen**  Merkmal „Einwohnerzahl“ hat die drei Merkmalsausprägungen „10 Tsd“, „20 Tsd.“, „30 Tsd. und mehr“  **stetig:**  **Intervall reeller Zahlen** (weitere **Zwischenwerte zwischen 2 Ausprägungen**)  Merkmal „Anzahl der Beschäftigten“ kann zwischen den Werten auf der Skala weitere Zwischenwerte haben. |
| * **natürlicher Nullpunk**t **natürlicher Nullpunkt** ist **von Natur** aus gegeben und **nicht veränderbar**. Zur Einordnung der Skala die Frage stellen **Wie wurde der natürliche Nullpunkt bestimmt?  (natürlich oder willkürlich vom Menschen)** **natürlicher Nullpunkt ist z. B. Temperatur in K natürlicher Nullpunkt durch physikalische Größen** (z. B. Gramm) **oder durch logische Annahmen** * **Werte der Absolutskala sind immer ≥ 0** * prinzipiell wie Verhältnisskala, nur mit ganzzahligen Werten * alles, was ganzzahlig zählbar ist * Anzahl und Stückzahl, Häufigkeiten * **Verhältnisbildung** (Quotientenbildung) möglich **Quotienten der gemessenen Werte werden verglichen** * **Vergleichsoperationen =, ≠, <, > und Rechenoperation +, -, \*, : möglich** * **Multiplikation oder Division möglich (da Verhältnisbildung möglich)** |
| **Beispiele**  Einwohneranzahl, Beschäftigtenanzahl  Klausur-Punkte (wenn nur ganze Punkte vergeben werden)   all diese Merkmale können nur Werte ≥ 0 haben  zu all diesen Merkmalen ist Verhältnisbildung möglich   (3fache Anzahl der Beschäftigten, 100 Punkte sind doppelt soviel wie   50 Punkte … |

**Grundgesamtheit**

* die **Menge aller möglichen Erhebungseinheiten**

**Stichprobe**

* eine **n-elementige Teilmenge der Grundgesamtheit mit N Elementen** (**Merkmalsträgern**)

**Merkmal**

* **relevante Eigenschaften der Merkmalsträger**
* Statistische Variable  
  **Beispiele**:  
  Familienstand  
  Alter  
  Größe  
  Gewicht  
  Umsatz  
  Einkommen

**Merkmalsausprägung**

* **mögliche Ausformungen eines Merkmals**
* **Merkmalswerte (Variablenwerte) / Beobachtungswerte in einer Werteliste**  
  Beispiele:  
  {„ledig, verheiratet, geschieden“}  
  {„zufrieden“, „eher unzufrieden“, „unzufrieden“}  
  {„1 Std.“, 2 Std.“, „3 Std.“}  
  {10.000, 20.000, 30.000 }

**Merkmalsträger**

* **Einzelnes Objekt einer statistischen Untersuchung, Träger der Informationen  
  Untersuchungseinheit, Erhebungseinheit**   
  Beispiele:  
  Menschen  
  Kunden

**Merkmalswerte**

* **einzelner Messwert / Beobachtungswert (Ergebnis) aus Merkmalsausprägungen** (Werteliste)  
  **Beispiele:**  
  „ledig“ aus den Merkmalsausprägungen {„ledig, verheiratet, geschieden“}  
  „zufrieden“ aus den Merkmalsausprägungen {„zufrieden“, „eher unzufrieden“, „unzufrieden“}

**Statistische Masse**

* **Menge aller Merkmalsträger**, die
  + **mit dem Untersuchungsziel in Verbindung** stehen,
  + **mindestens eine übereinstimmende Eigenschaft** haben,
  + sich exakt **sachlich, räumlich, zeitlich abgrenzen lassen**

**Häufigkeiten und Häufigkeitsverteilungen**

**Datendokumentation**

* Formen der Datendokumentation
* Einzelwerte (Einzelbeobachtungen)
  + ungeordnete Reihe
    - Urliste
    - Rohdaten
    - Primärdaten
  + erzeugt INPUT-Blase

**Urliste**

* direktes Ergebnis einer Datenerhebung
* Daten der Urliste müssen aufbereitet werden

**Vorteile:**

* enthält alle Beobachtungswerte
* keine Auslassungen
* keine Übertragungsfehler
* keine verlorene Information

**Nachteile:**

* sehr großer unübersichtlich und nicht auswertbarere Datenbestand
* unkorrigierten Urliste kann offensichtliche Fehler (Zahlendreher, unplausible Daten) enthalten

**Aufbereitung der Daten einer Urliste**

* **durch Häufigkeitsverteilungen**

**Häufigkeitsverteilungen für nicht klassierte Daten**

1. **Sortieren der Daten** 
   * **geordnete Reihe** nach irgendeiner Ordnung ( z. B. **alphabetische Ordnung der Merkmalsträger oder Größenordnung der Merkmalsausprägung**)
2. **Verdichten der sortierten Daten auf Merkmalsausprägungen und zählen wie oft diese vorkommen** 
   * **Häufigkeitsverteilung:** **geordnete Menge von Wertepaaren (Merkmalsausprägung und zugehörige Häufigkeit)**
3. **Darstellung der sortierten Häufigkeitsverteilungen nach Merkmalsausprägungen in Häufigkeitstabelle**

**Häufigkeitsverteilungen für klassierte Daten**

1. **Einteilung der Werte in Klassen** (klassierte Daten )
   * keine Sortierung erforderlich
2. **Verdichten der klassierten Daten** 
   * **Häufigkeitsverteilung für klassierte Daten** (klassierte Verteilung)
3. **Darstellung der klassierten Daten in Häufigkeitstabelle für klassierte Daten**

**absolute Häufigkeit**

* **Anzahl** des Auftretens einer bestimmten **Merkmalsausprägung**

**relative Häufigkeit**

* **Verhältnis der absoluten Häufigkeit zur Summe der Einzelhäufigkeiten** (n)

**Summenhäufigkeiten**

* **nur für Rangmerkmale und metrische Merkmale** sinnvoll
* **absolute Summenhäufigkeit** (**absolute kumulierte Häufigkeit**)  
  **H(xi) = h(x1) + h(x1) + … + h(xi) 🡆 ∑h(xi) = n**
* **relative Summenhäufigkeit** (**relative kumulierte Häufigkeit**)  
  **F(xi) = f(x1) + f(x1) + … + f(xi) 🡆 ∑f(xi) = 1**

**Übung:**

**Häufigkeitsverteilung zum Wohnort**

Merkmal: Wohnort

Merkmalsträger: befragte Menschen

Merkmalsausprägungen: „A“, „B“, „C“, „D“, „k. A.“

Merkmalswert: z. B. ist „A“ ein Merkmalswert zu genannten Ausprägungen

Grundgesamtheit: alle Antworten

Stichprobe: in die Statistik einfließende Antworten

Statistische Masse: alle befragten Menschen

**Antworten**:

**B** **C** **A** **B** **C** **B B B** **A A** **D** **k.A.** **A** **B B** **A** **k.A**. **A** **B B** (k.A. = keine Antwort)

* Verdichten der Häufigkeitsverteilung und Darstellen in einer Häufigkeitstabelle

Ablauf

1. Tabelle mit Merkmalsausprägungen „A“, „B“, „C“, „D“, „k. A.“ erstellen
2. Antworten zählen und in die Spalte zur absoluten Häufigkeit eintragen
3. Summe zur absoluten Häufigkeit bilden
4. relative Häufigkeit in Tabelle eintragen
5. wenn gefragt, absolute und relative Summenhäufigkeit in Tabelle eintragen

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Wohnort  xI | absolute Häufigkeit  h(xi) | absolute  Summenhäufigkeit  H(xi) | relative Häufigkeit  f(xi) | relative  Summenhäufigkeit  F(xi) | relative Häufigkeit  f(xi)  zu gültigen Antworten |
| **A** | 6 | 6 | 0,30 | 0,30 | 0,333 |
| **B** | 9 | 15 | 0,45 | 0,75 | 0,500 |
| **C** | 2 | 17 | 0,10 | 0,85 | 0,111 |
| **D** | 1 | 18 | 0,05 | 0,90 | 0,056 |
| **k. A.** | 2 | 20 | 0,10 | 1,00 |  |
| **SUMME** | **20** |  |  |  |  |
|  |  |  |  |  |  |

Hinweis:

gültige Antworten = 18 (Antworten, die nicht „k. A“ beinhalten.

**Empirische Verteilungsfunktion F(x)**

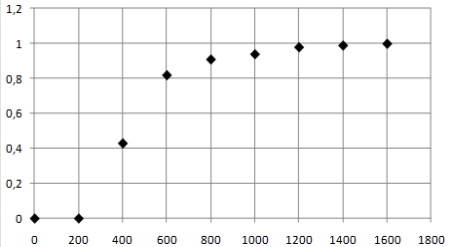
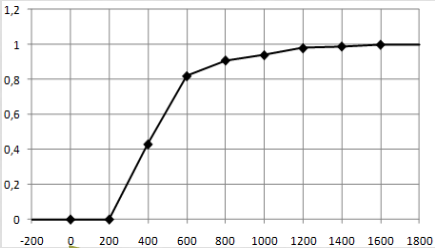
* **enthält** die **gesamte Information der Daten**
* nur ursprüngliche Reihenfolge geht verloren
* **ist** (relative) **Summenhäufigkeitskurve**
* gibt für jede reelle Zahl x den **Anteil der Merkmalsträger** an, für die das Merkmal X einen (Ausprägungs)-Wert xi annimmt, der kleiner oder gleich x ist

**relative Summenfunktion / Verteilungsfunktion F(x)**

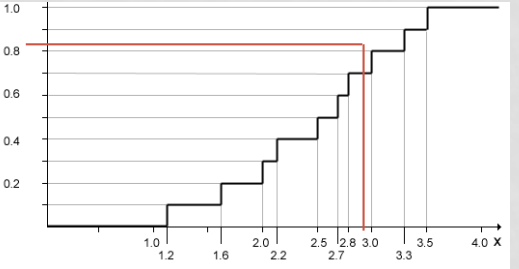
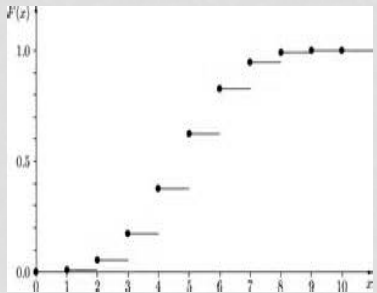
* **Wertebereich: 0 ≤ F(x) ≤ 1**
* **F(x) steigt oder ist konstant** (ist monoton nichtfallend)
* F(x) ist eine **Treppenfunktion mit Sprungstellen bei x1 , x2 , ..., xi**
* Die **Größe der Sprünge** beträgt **fi = F(xi ) - F(xi-1 )**  
  F(xi ): Wert aus letzter Summenfunktion,   
  F(xi-1 ): Wert aus vorausgehender Summenfunktion

**theoretische Verteilung ist „idealisierte“ Verteilung**

**Empirische Verteilungsfunktion bei klassierten Daten**



Die Sprünge beginnen bei der oberen Klassengrenze



Interpretation zum Wert „3.0“ auf der x-Achse:

0,8 = 80% der Merkmalsträger (Untersuchungseinheiten) haben einen Notendurchschnitt ≤ 3,0

* **y-Achse** enthält die **Werte zur relativen Summenfunktion f(xi)**
* **x-Achse enthält die Werte zu xi**
* **Wichtig in der Anwendung:** **Achsenbeschriftung f(xi) für y-Achse und xi für x-Achse nicht vergessen!**

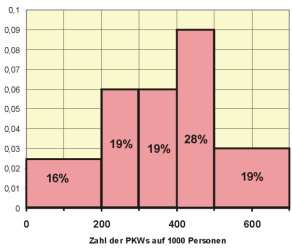
**Diagramme (grafische Darstellung) zur Häufigkeitsverteilung**

**Wahlentscheidung:**

* **Form der grafischen Darstellung**
* **Achsenmaßstab**
* **Evtl. nur Ausschnitt darstellen**; Achtung: Manipulationen (optische Täuschung) möglich  
  Diagramm richtig interpretieren, d. h. relative Häufigkeit ermitteln und Diagramm hierzu interpretieren

**typische Darstellungsformen:**

* **Säulendiagramm**
  + **höhenproportionale Darstellungsform einer Häufigkeitsverteilung**
  + **x-Achse**: **vertikale, nicht aneinandergrenzende Säulen** (mit beliebiger Breite) **für Beobachtungswert (Merkmalswert) xi**
  + **y-Achse: absolute oder relative Häufigkeit pro xi-Wert**
  + **für wenige Ausprägungen**
* **Stabdiagramm / Liniendiagramm**
  + **wie Säulendiagramm** nur mit schmalen bzw. sehr schmalen Säulen
* **Balkendiagramm**
  + **wie Säulendiagramm** nur mit horizontalen Balken (transporniert)
  + **x-Achse**: **absolute oder relative Häufigkeit pro xi-Wert**
  + **y-Achse: horizontale, nicht aneinandergrenzende Balken** **für Beobachtungswert (Merkmalswert) xi**
  + sehr gut zur **Darstellung von Rangfolgen** geeignet
* **Kreisdiagramm**
  + jeder Kreissektor enthält einen Teilwert
  + gesamter Kreis = Summenhäufigkeit
  + max. 7 Teilwerte
  + xi-Werte im Uhrzeigersinn der Größe nach sortieren
  + für die Darstellung von diskreten Daten (besonders für Nominal und Ordinalskalen)
* **Histogramm (nur bei klassierten Daten)**
  + **flächenproportionale Darstellung der Häufigkeiten von klassierten Daten**
  + **x-Achse: geordnete Werten mit gleichen Abständen zwischen den Werten**   
    (z. B. 200, 400, 600, 800)
  + **Rechtecke liegen auf x-Achse nebeneinander** (ohne Abstände)
  + **Rechteckbreite = Klassenbreite bi**   
    (bei Klasse „150 b.u. 180“ wird die Rechteckbreite zwischen den Werten 150- 180 auf der x-Achse gezeichnet)
  + **y-Achse = Rechteckhöhe** = absolute **Klassenhäufigkeit** hi / **Klassenbreite** bi 🡆 **hi / bi**
  + **absolutes Histogramm**: **Summe der Flächeninhalte aller Rechtecke** = **Summe der Einzelhäufigkeiten (**n)



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Klasse | Anzahl PKW / 1000 Personen  xi | Anzahl der Länder  h(xi) | Anteil der  Länder  f(xi) | Klassen-  breite  bi | Rechteck-  höhe  ri = hi / bi |
| 1 | 0 b. u. 200 | 5 | 0,1563 | 200 | 0,025 |
| 2 | 200 b. u. 300 | 6 | 0,1875 | 100 | 0,060 |
| 3 | 300 b. u. 400 | 6 | 0,1875 | 100 | 0,060 |
| 4 | 400 b. u. 500 | 9 | 0,2812 | 100 | 0,090 |
| 5 | 500 b. u. 700 | 6 | 0,1875 | 200 | 0,030 |
| SUMME |  | 32 |  |  |  |

**Hinweis:**

* die **Skala auf der X-Achse muss geordnete Werte mit gleichem Abstand** haben   
  (im Bsp. 200, 400, 600, 800)
* **letzter Skalenwert auf x-Achse ≥ Summe zu allen Klassenbreiten**   
  (im Bsp. 200 + 100 + 100 + 100 + 200 ≥ 700)
* Aber: **Rechtecke = Klassenbreite**   
  (im Bsp. 200, 100, 100, 100, 200)

**Lageparameter (Lokationsmaße)**

* **beschreiben** die **“Lage”** der Elemente **der Grundgesamtheit / Stichprobe** **bezüglich der Messskala**

**allgemeine Lageparameter**

**Allgemeine Mittelwerte:**

* **Modus (Modalwert) x̅D**
* **Median x̅Z**
* **arithmetisches Mittel x̅**
* **Quantil x̃p**

**Modus (Modalwert) x̅D**

* + die am **häufigsten auftretende Merkmalsausprägung (maximale Häufigkeit)**.
  + **hauptsächlich für nominale Merkmale, aber auch andere diskrete Merkmalstypen**
  + bei **klassierten Daten: Klassenmitte der Klasse mit den größten Häufigkeiten 🡆 Modalklasse   
    Klassenmitte** ½ \* (k + k-1) = ½ \* (Klassenobergrenze + Klassenuntergrenze)  
    Bsp. Klasse **„200 b. u. 300“** = **½ \* (300 + 200) = ½ \* 500 = 250**
  + **mehrere Modalwerte**, wenn es **mehrere Merkmalsausprägungen mit der gleichen maximalen Häufigkeit** gibt (Multimodale Verteilungen)  
    **x̅D1, x̅D2, x̅D3,**

**Median x̅Z**

* + sehr **robustes Lokationsmaß** 🡆 **wenig anfällig gegen Datenausreißer**
  + Median beschreibt die Verteilung besser als Mittelwert ( Ausreißer haben auf Median keinen Einfluss)
  + **Wert in der Mitte der geordneten Reihe** **Wichtig:** **Werte müssen geordnet sein**
  + **Mindestens 50% der Werte liegen links und mindestens 50% rechts des Medians** (ggf. Median mit eingerechnet).
  + für ordinale und metrische Merkmale
  + für **gerade Werte von n**  
      
    Bsp.  
    n = 14 (Summe der Einzelhäufigkeiten)  
      
    1, x2, x3, x4, x5, x6, **x7**, **x8**, x9, x10, x11, x12, x13, x14  
      
    wenn   
    x7 = 45 (7. Element in den Ausprägungen) und   
    x8 = 20 (8. Element in den Ausprägungen)  
      
    **x̅Z** = ½ \* (45 + 20) = ½ \* 65 = 32,5
  + **bei ordinal skalierten Werten** (z. B. **Zeugnisnoten**) **müssen bei geradem n die Merkmalswerte zu den betreffenden Ausprägungen identisch sein**  
    1 2 **3 3** 4 5 Median wird aus 3. und 4 Element ermittelt und hat den Wert 3  
     ½ \* (3 + 3) ) 3  
      
    1 2 **3 4** 5 6 🡆 es existiert **kein Median, da es keine Note 3,5 gibt**
  + für **ungerade Werte von n**Bsp.  
    n = 13 (Summe der Einzelhäufigkeiten)  
      
    1, x2, x3, x4, x5, x6, **x7**, x8, x9, x10, x11, x12, x13  
      
    wenn   
    x7 = 45   
    **x̅Z** = 45
* **Median bei klassierten Daten**
* **Median für metrische Daten in Klassen ist Näherungswert**   
  (**exakte Merkmalsausprägung des Medians nicht bestimmbar, da Verteilung in den Klassen unbekannt** (Klassenbreite bezieht sich auf ein Intervall (von/bis), es ist unbekannt, wie die Ausprägungen verteilt sind
* Annahme: Beobachtungswerte sind in den Klassen gleich verteilt
* **Achtung!! Fehler im Näherungswert möglich**

k = Einfallsklasse 🡆 Klasse mit F(x) ≥ 50% (relativer Summenhäufigkeit ≥ 0,5)

xk-1 = Untergrenze der Einfallsklasse

xk = Obergrenze der Einfallsklasse

Fk-1 = relative Summenhäufigkeit der Klasse, die unterhalb der Einfallsklasse liegt   
 (Anteilswert

fk = relative Häufigkeit der Einfallsklasse (Anteilswert)

Bsp.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Klassen-Nr.  i | Größen-  klasse  xi | hi | fi | Fi |  |
| 1 | 100 b. u. 150 | 40 | 0,4 | **0,4** |  |
| 2 | **150 b. u. 170** | 40 | **0,4** | 0,8 | Einfallsklasse k |
| 3 | 170 b. u. 200 | 20 | 0,2 | 1,0 |  |
| **Summen** |  | **100** | **1,0** |  |  |

**arithmetisches Mittel x̅ (Mittelwert, oder Durchschnitt)**

* **für beliebige metrische Merkmale**
* **nur für quantitative Merkmale** sinnvoll
* **empfindlich gegen Ausreißer** (Vorsicht bei **schiefen Verteilungen!**)

**Summe der Abweichungen der Einzelwerte vom arithmetischen Mittel = 0**  
 Bsp. 3, 5, 7 = ∑ = 15 🡆 x̅ = 5 🡆 (3 – 5) + (5 – 5) + (7 – 5) = 0

**arithmetisches Mittel aus Häufigkeitstabellen**

**absolute Häufigkeit relative Häufigkeit**

hi absolute Häufigkeit

fi releative Häufigkeit

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| xi | hi | xi \* hi | fi = hi / n | xi \* fxi |
| 10 | **10** | 100 | **0,17** | 1,70 |
| 15 | **20** | 300 | **0,33** | 4,95 |
| 25 | **30** | 750 | **0,50** | 12,5 |
| **SUMMEN** | **n = 60** | **1150** | **1,0** | **19,15** |
|  |  | **x̅ = 1150 / 60 = 19,16** |  | **x̅ = 19,15** |

**n = Summe der Einzelhäufigkeiten**

**arithmetisches Mittel bei klassierten Daten**

* Näherungswert, da Verteilung in den Klassen unbekannt
* Annahme: Beobachtungswerte liegen jeweils in der Klassenmitte

**absolute Häufigkeit relative Häufigkeit**

hi absolute Häufigkeit

fi releative Häufigkeit

mi Klassenmitte = ½ \* ( ki + ki-1 )

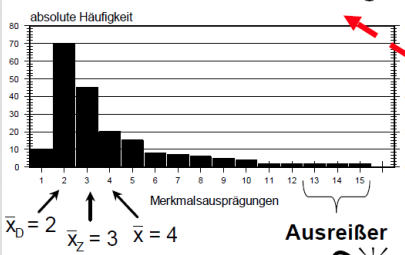
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Klassennr.  i | Körpergewicht  (kg)  xi | **hi** | **mi** | mi \* hi | **fi = hi / n** | mi \* fi |
| 1 | 41 - 50 | **20** | **45,5** | 910,0 | **0,40** | 18,2 |
| 2 | 51 - 60 | **15** | **55,5** | 832,5 | **0,30** | 16,65 |
| 3 | 61 - 70 | **10** | **65,5** | 655,0 | **0,20** | 13,1 |
| 4 | 71 - 80 | **4** | **75,5** | 302,0 | **0,08** | 6,04 |
| 5 | 81 - 90 | **1** | **85,5** | 85,5 | **0,02** | 1,71 |
| **SUMMEN** | | **n = 50** |  | **2785,0** | **1,0** | **55,7** |
|  | |  |  | **x̅ = 2785 / 50 = 55,70** |  | **x̅ = 55,7** |

**Neigung / Schiefe**

**rechtsschiefe** (linkssteile) **Häufigkeitsverteilung**

x̅D < **x̅Z < x̅**

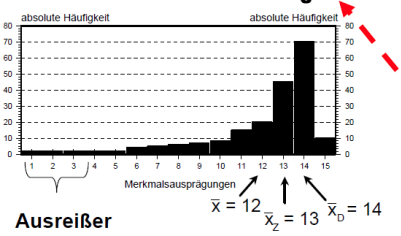
Modus < **Median < Mittelwert**



**linksschiefe** (rechtssteile) **Häufigkeitsverteilung**

x̅D > **x̅Z > x̅**

Modus > **Median > Mittelwert**



**unimodale symmetrische Häufigkeitsverteilung**:

x̅D ≈ x̅Z ≈ < x̅

Modus ≈ Median ≈ arithmetisches Mittel

**x̅D Modus = Wert zur Ausprägung mit der größten Häufigkeit** (maximale Häufigkeit)

**x̅Z Zentralwert = Mindestens 50% der Werte liegen links rechts des Medians**

**x̅ Mittelwert**

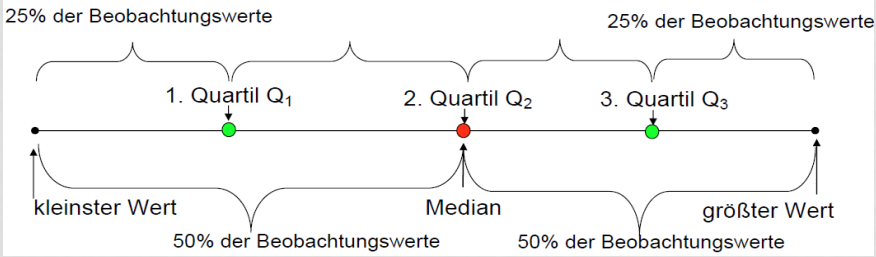
**Quantile**

* **Lagemaß**
* **teilen Verteilung in Abschnitte gleicher Häufigkeit**

**Q1, Q3 = 25%-Quantil = Quartil**

**Q2 = 50%-Quantil = Median**

**5-Punkte-Zusammenfassung der geordneten statistischen Reihe**



* Median teilt einen nach Größe sortierten Datensatz in der Mitte
* links und rechts vom Median liegen gleich viele Beobachtungswerte
* Unterteilung der linken und rechten Hälfte nach gleicher Vorschrift wie beim Median

🡆 4 gleich große Bereiche, die in 3 Quartils aufgeteilt werden.   
 **25%** aller geordneten Beobachtungswerte sind **kleiner als das 1. Quartil.**   
 **50%** aller geordneten Beobachtungswerte sind **kleiner als das 2. Quartil**.   
 **75%** aller geordneten Beobachtungswerte sind **kleiner als das 3. Quartil.**

**Streuungsparameter**

* keine Information über Daten-Streuung in den Lageparametern
* arithmetische Mittel x̅ und Median (Zentralwert) x̅Z verdecken oft Ungleichheiten in der Streuung
* Durchschnittsberechnung ist nicht immer sinnvoll 🡆 Durchschnitt kann nicht immer alles beschreiben
* Streuungsparameter beschreiben die Daten-Streuung

**erforderliche Eigenschaften an die Kennzahl zur Messung der Streuung**

* **geeigneten Lageparameter (Bezugspunkt) wählen, um den die Werte streuen**
* **alle Beobachtungswerte werden berücksichtigt**
* **Streuung = 0: alle Werte sind gleich** 🡆 Streuungsparameter = 0
* je **größer die Streuung**, **umso größer der Streuungsparameter**
* Streuungsparameter ist unabhängig von der Anzahl der Beobachtungswerte n

**Inter-Quartilsabstand IQR**

* **Streuungsmaß**
* **von Ausreißern unabhängig**
* umfasst **50% der Verteilung** 🡆 **Differenz zwischen oberem und dem unterem Quartil**   
  **Q3 -Q1**
* **Breite des mittleren Bereichs an, in dem ca. 50% aller Werte liegen**
* **Abstand zwischen Q1 und Q3-Quartil**

Bsp.

**Hinweis: X-Elemente müssen sortiert sein (hier nach Größe aufsteigend) !**

**Falls w in der Klausur auf ungeordnete Werte berechnet werden soll, zuerst die X-Werte sortieren!**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | **Q1 = ½ \* (x6/2 + x6/2+1)**  **Q1 = ½ \* (x3 + x4)** | | | | **x̅**  **Q2 = x(n + 1)/2**  **Q2 = x(13 + 1)/2**  **Q2 = x14** | | | **Q3 = ½ \* (x7+3 + x7+4)**  **Q3 = ½ \* (x10 + x11)** | | | |  |
| Schüler-Nr  (Element-Nr)  i | 1 | 2 | **3** | **4** | 5 | **6** | **7** | 8 | 9 | **10** | **11** | 12 | 13 |
| Größe (m)  xi | 1,60 | 1,67 | **1,67** | **1,68** | 1,68 | 1,70 | **1,70** | 1,72 | 1,73 | **1,75** | **1,76** | 1,78 | 1,84 |
|  | | |  | **Inter-Quartilsabstand IQR** | | | | | | |  |  | |

**Q1 = ½ \* (1,67 + 1,68) = 1,675**

Q2 = 1,7

**Q3 = ½ \* (1,75 + 1,76) = 1,755**

**Inter-Quartilsabstand IQR = 1,755 – 1,675 = 0,08**

**Spannweite w**

* für **ordinale und metrische Merkmale**
* **Ausdehnung der Werte (Breite des Streubereichs einer Häufigkeitsverteilung)**
* **Gesamtbreite**, in dem alle Werte liegen
* **Vom kleinsten und größten Wert abhängig**

**w = xmax – xmin**

Bsp.

**Hinweis: X-Elemente müssen sortiert sein (hier nach Größe aufsteigend) !**

**Falls w in der Klausur auf ungeordnete Werte berechnet werden soll, zuerst die X-Werte sortieren!**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Schüler-Nr  (Element-Nr)  i | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| Größe (m)  xi | 1,60 | 1,67 | 1,67 | 1,68 | 1,68 | 1,70 | 1,70 | 1,72 | 1,73 | 1,75 | 1,76 | 1,78 | 1,84 |
| Spannweite w | **1,60** |  | | | | | | | | | | | **1,84** |

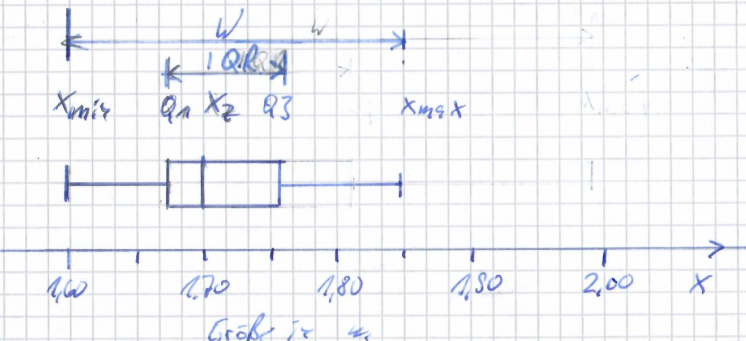
**w = xmax – xmin = 1,84 – 1,60 = 0,24**

**Boxplot**

* **grafische Darstellung** zur 5-Punkte-Zusammenfassung
* **5 Punkte** im Diagramm:

1. **Minimum**
2. **Q1-Quartil**
3. **Median / Zentralwert (Q2-Quartil)**
4. **Q3-Quartil**
5. **Minimum**

* Informationen aus dem Boxplot
  + **Lokalisation (Lage des Median)**
  + Streuungsmaße:
    - **Spannweite** = Ausdehnung des Boxplots (Differenz w = xmax – xmin)
    - **Quartilsabstand** = Ausdehnung der Box (Differenz IQR = Q3 – Q1 )
  + **Schiefe: Vergleich der beiden Hälften der Box oder Längen der Whisker**)
  + **Ausreißer**



**Varianz s²**

* **arithmetische Mittel der Abweichungsquadrate**
* wichtiger Streuungsparameter
* **nur für metrische Merkmale**
* **Mittelwert und Varianz** bzw. Standardabweichung **hängen eng zusammen**
* **Ausgangswert für die Streuungsparameter**:
  + **Standardabweichung s**
  + **Variationskoeffizient**
* **Bezugspunkt** ist das **arithmetische Mittel x̅**
* **Summe der quadratischen Abweichungen ∑(x – x̅)²**

**n Summe der absoluten Häufigkeiten**

xi Merkmalsausprägung

Bsp.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| i |  | xi | x̅ = 18 / 4 = 4,5 | (xi – x̅)² |
| 1 | 5 | 4,5 | 0,25 |
| 2 | 2 | 4,5 | 6,25 |
| 3 | 8 | 4,5 | 12,25 |
| 4 | 3 | 4,5 | 2,25 |
| **n = 4** | **∑** | **18** | **-** | **21,56** |

**Varianz s² aus Häufigkeitstabellen**

hi absolute Häufigkeit der Ausprägung i

**n Summe der absoluten Häufigkeiten**

xi Merkmalsausprägung

Bsp.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| i | **xi** | **hi** | Hi | fi | Fi | xi \* hi | **x̅ = ∑(xi \* hi) / n**  **x̅ 164 / 50**  **x̅ = 3,28** | (xi – x̅)² \* hi | (xi – x̅)² \* fi |
| **1** | **1** | **5** | 5 | 0,10 | 0,10 | 5 | **3,28** | 25,99 | 0,520 |
| **2** | **2** | **8** | 13 | 0,16 | 0,26 | 16 | **3,28** | 13,11 | 0,262 |
| **3** | **3** | **14** | 27 | 0,28 | 0,54 | 42 | **3,28** | 1,10 | 0,022 |
| **4** | **4** | **16** | 43 | 0,32 | 0,86 | 64 | **3,28** | 8,29 | 0,166 |
| **5** | **5** | **5** | 48 | 0,10 | 0,96 | 25 | **3,28** | 14,79 | 0,295 |
| **6** | **6** | **2** | 50 | 0,04 | 1,00 | 12 | **3,28** | 14,80 | 0,296 |
| **∑** | | **n = 50** |  |  |  | **164** | **-** | **78,08** | **1,561** |

**Varianz s² aus Häufigkeitstabellen bei klassierten Daten**

hi absolute Häufigkeit der i-ten Klasse

**n Summe der absoluten Häufigkeiten**

mi Klassenmitte der i-ten Klasse

Bsp.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| i | Größe in cm  Klasse xi | **hi** | **mi** = ½ (hki + hki+1) | mi \* hi | x̅ = ∑(mi \* hi) / n  x̅ = 4970 / 30  **x̅ = 165,67** | (mi – x̅)² \* hi |
| 1 | 150 b. u. 160 | **9** | **155** | 1395 | **165,67** | 1024,64 |
| 2 | 160 b. u. 170 | **12** | **165** | 1980 | **165,67** | 5,39 |
| 3 | 170 b. u. 180 | **7** | **175** | 1225 | **165,67** | 609,34 |
| 4 | 180 b. u. 190 | **2** | **185** | 370 | **165,67** | 747,30 |
| **∑** | | **n = 30** |  | **4970** |  | **2386,67** |

**Standardabweichung S**

* **Stärke der Aussagekraft zum arithmetischen Mittel** (Mittelwert) **x̅**
* **kleine Standardabweichung** = Beobachtungswerte nahe am Mittelwert (kleine Streuung).
* **große Standardabweichung = Beobachtungswerte sind weit um den Mittelwert gestreut**
* bei normalverteilten Daten liegen ca. 95% der Beobachtungswerte im Intervall [x̅ - 2s, x̅ + 2s]

s² Varianz

Bsp.

8,91

**Varianzkoeffizient v**

s Standardabweichung

x̅ arithmetisches Mittel (Mittelwert)

Bsp.

**Korrelation und Regression**

**Analyse mehrerer Merkmale**

* in empirischen Analysen werden **i. d. R. mehrere Merkmale gemessen**, z.B. X, Y, Z
* **univariate Analyse:** wenn **Merkmale einzeln ausgewertet** werden
* **multivariate Analyse:** wenn **mehrere Merkmale gemeinsam ausgewertet** werden

**Analyse von Zusammenhängen**

* **Teilgebiet der multivariaten Statistik** (mehrere **Merkmale** werden **gleichzeitig untersucht**)
* Fragestellungen:
  + Besteht ein Zusammenhang bzw. eine Abhängigkeit zwischen Merkmalen (Variablen)?
  + Wie stark ist dieser Zusammenhang / die Abhängigkeit?
  + Wie lässt sich die Stärke (die Intensität) des Zusammenhangs bzw. die Abhängigkeit zwischen mehreren Merkmalen messen?
  + Lässt sich der Zusammenhang darstellen?
  + Lassen sich die beobachteten Werte einer Variable X durch die Werte einer oder mehrerer anderen Variablen Y (Y1 , Y2 , …) näherungsweise bestimmen?

**Zusammenhangsanalyse (Korrelationsanalyse / Interdependenzanalyse)**

* **Untersuchung der Wechselwirkung zwischen den Variablen untereinander**
* **Zusammenhangsmaß (Assoziationsmaß)**: **Stärke und Richtung eines Zusammenhangs**
* **Korrelationskoeffizient** ist das **Zusammenhangsmaß in der Korrelationsanalyse**-1 ≤ rxy ≤ +1
* **Grafische Darstellung** mittels **Streudiagramm**
* **Zusammenhangsanalyse (Korrelationsanalyse)** **zwischen metrischen Merkmalen X und Y**
  + prüft, ob ein **linearer Zusammenhang** zwischen den Variablen **besteht**
  + **wie stark der lineare Zusammenhang** ist.
* **Kontingenzanalyse (Assoziationsanalyse)**
* ist **Zusammenhangsanalyse (Korrelationsanalyse), die auf Kontingenztabelle** (Häufigkeitstabelle) **basiert**
* **Je größer der Unterschied zwischen den Häufigkeiten**, **umso** **stärker** ist der **Zusammenhang** / die **Abhängigkeit** zwischen den Merkmalen

**Zusammenhangsanalyse bei nicht metrischen Merkmalen**

**Rangkorrelationskoeffizient nach Spearmann**

* **Maß zur Stärke des Zusammenhangs** zweier **ordinalskalierter Merkmale (**Rangmerkmale)
* Spearmansche Rangkorrelationskoeffizient **nutzt Ränge statt Beobachtungswerte**
* **Spezialfall des Pearsons Korrelationskoeffizient:** **Daten werden vor Berechnung des Korrelationskoeffizient in Ränge konvertiert**
* benötigt **keine Annahme, zu einer linearen Beziehung zwischen den Variablen**
* **robust gegenüber Ausreißern**

**Kontingenzkoeffizient C**

* Maß zur **Stärke des Zusammenhangs zweier nominaler oder ordinaler Merkmale**
* basiert auf **Vergleich von ermittelten Häufigkeiten mit zu den Merkmalen erwarteten Häufigkeiten**
* **Anwendung bei beliebig großen Kreuztabellen**

**0 ≤ C ≤ 1**

**Phi-Koeffizient (Vierfelder-Korrelationskoeffizient)**

* **Stärke des Zusammenhangs zweier dichotomer Merkmale**
* **basiert auf Kontingenztafel** (enthält **gemeinsame Häufigkeitsverteilung der Merkmale**)

**Analyse von Abhängigkeiten zwischen zwei metrischen Merkmalen X und Y**

* gerichteter Zusammenhang zwischen zwei (oder mehreren) Merkmalen
* Unterscheidung in
  + **unabhängige Merkmale** (Variablen = **Regressor** = **Einflussgröße X**) und
  + **abhängige Merkmale** (Variablen = **Regressand** =  **Zielgröße Y)**

**Abhängigkeitsanalyse (Regressionsanalyse / Dependenzanalyse )**

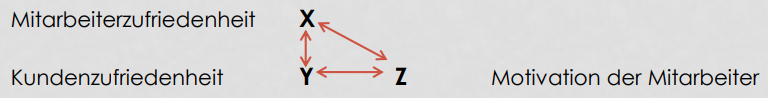
* Abhängigkeitsmaß wird über Regressionsfunktion ŷ = a + b \* x ermittelt

ŷ abhängige Variable (Regressand = Zielgröße)  
 a erster Regressionskoeffizient  
 b zweiter Regressionskoeffizient  
 x unabhängige Variable (Regressor = Einflussgröße)

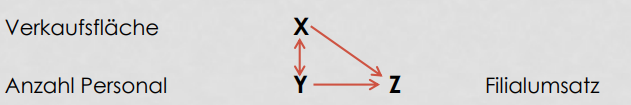
* **Grafische Darstellung** mittels **Streudiagramm** und **Regressionsgerade**

**Multivariate Analysemethoden (gemeinsame Analyse zu mehreren Merkmalen)**

**Zusammenhangsanalyse (Korrelationsanalyse / Interdependenzanalyse)**



**Abhängigkeitsanalyse (Regressionsanalyse) / Dependenzanalyse**



**Streudiagramm**

* **graphische Darstellung von Wertepaaren zweier Merkmale** (**unabhängige Variable X** (= **Einflussgröße**) und **abhängige Variable Y** (= **Zielgröße**))
* **Kreuzpunkte zu den Wertepaaren** werden in kartesisches Koordinatensystem eingetragen 🡆 ergibt **Punktwolke**
* Lage und Form der **Punktwolke geben Hinweis auf Stärke und Richtung des Zusammenhangs** .
* **erste Hinweise über mögliche Abhängigkeit zwischen den Merkmalen**

**Korrelation (Zusammenhangsanalyse zwischen 2 metrischen Merkmalen X und Y)**

* Ist **linearer (gerichteter) Zusammenhang**
* Ist **zahlenmäßiger statistischer Zusammenhang zwischen zwei metrischen Merkmalen** X und Y
* **nur bei** **metrischen Merkmalen**
* **Positive Korrelation: beide Merkmale X und Y entwickeln sich gleichförmig:**
* bei höheren Werten von X hat auch Y höhere Werte
* **wenn Werte von X steigen, steigen auch die Werte von Y**
* **Negative Korrelation: beide Merkmale X und Y entwickeln sich gegenläufig**
* bei höheren Werten von X hat Y niedrigere Werte
* **wenn Werte von X steigen, fallen die Werte von Y**
* **Kausaler Zusammenhang**: zwischen zwei Merkmalen X und Y existiert ein **Ursache-Wirkungs-Beziehung**
  + **Veränderung eines abhängigen Merkmals Y ist auf Veränderung des unabhängigen Merkmals X zurückzuführen**
  + **Korrelation** sagt **nichts über den kausalen Zusammenhang** und **nichts über die Kausalitätsrichtung** aus.  
      
    **Korrelation ≠ Kausalität**
* Korrelation ist für Kausalität notwendig
  + **nicht jede Korrelation ist** ein **kausaler Zusammenhang**, aber
  + **jeder kausale Zusammenhang basiert auf einer Korrelation**

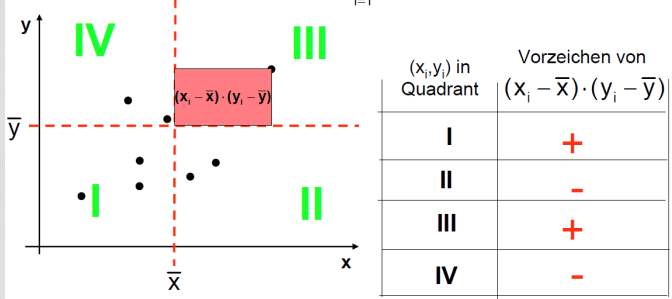
**Korrelationskoeffizient rx,y nach Pearson**

* **Korrelationskoeffizient rx,y -1 ≤ rxy ≤ +1**
* **dimensionsloses Maß** (statistische Kennzahl) **für Grad des linearen Zusammenhangs**
* **Stärke und Richtung des linearen Zusammenhangs**
* für **2 mind. intervallskalierte Merkmale** X und Y mit **positiver Kovarianz cov(x,y) und positiver Standardabweichung s** für 2 Merkmale

* **Korrelationskoeffizient r > 0: Positiver Zusammenhang**
  + **hohe Werte in einem Merkmal** (Variablen) **X** gemeinsam **mit hohen Werten in einem anderen Merkma**l (Variablen) **Y**
* **Korrelationskoeffizient r < 0: Negativer Zusammenhang**
  + **hohe Werte in einem Merkmal** (Variablen) **X mit niedrigen Werten in einem anderen Merkmal** (Variablen) **Y**
* **Korrelationskoeffizient r = +1: extrem starker positiver Zusammenhang**
  + selten
  + Punktewolke liegt auf einer Geraden mit positiver Steigung (links unten nach rechts oben)
* **Korrelationskoeffizient r = -1: extrem starker negativer Zusammenhang**
* selten
* Punktewolke liegt auf einer Geraden mit negativer Steigung (von links oben nach rechts unten)
* **Korrelationskoeffizient r = 0: kein Zusammenhang**

**Kovarianz cov(x,y)**

oder



Bsp.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Filiale Nr.  i |  | Verkaufs-  fläche  (1000 m²)  xi | Umsatz  (Mio EUR)  yi | xi² | yi² | x \* y |
| 1 |  | 3 | 30 | 9 | 900 | 90 |
| 2 |  | 2 | 10 | 4 | 100 | 20 |
| 3 |  | 6 | 40 | 36 | 1600 | 240 |
| 4 |  | 5 | 20 | 25 | 400 | 100 |
| **n = 4** | **∑** | **16** | **100** | **74** | **3000** | **450** |
| **x̅ = 4 x̅² = 16** | | | | | | | |
| **y̅ = 25 y̅² = 625** | | | | | | | |

**Interpretation des Korrelationskoeffizienten**

**-1 bis -0,8 „starke“ negative** Korrelation,

**-0,8 bis -0,5 „mittlere“ negative** Korrelation

**-0,5 bis 0 „schwache“ negative** Korrelation

**0 keine** Korrelation

**0 bis 0,5 „schwache“ positiv**e Korrelation

**0,5 bis 0,8 „mittlere“ positive** Korrelation

**0,8 bis 1 „starke“ positive** Korrelation

Im Bsp. liegt eine mittlere positive Korrelation (rx,y = 0,707) vor.

**Regressionsanalyse (Abhängigkeitsanalyse / Dependenzanalyse) für gerichteten Zusammenhang**

**Regressionsanalyse** (für gerichteten Zusammenhang)

* ist **Abhängigkeitsanalyse zwischen 2 Merkmalen** (unabhängige Variable X, und abhängige Variable Y)
* **basiert auf Regressionsfunktion ŷ = a + b \* x** (ermittelt Abhängigkeitsmaß)

**einfache Regressionsanalyse**

* **2 metrische Größen:** 
  + **Zielgröße Y** ist Regressand (**abhängig**es Merkmal)
  + **Einflussgröße X** ist Regressor (**unabhängig**es Merkmal)
  + **zum Erstellen eines Vorhersagemodells**
  + **Quantifizierung** der **Stärke des Zusammenhangs**
  + **ermittelt xj, die keinen Zusammenhang mit y** haben
  + ermittelt **Teilmengen xi , … , xj , die redundante Information über y** enthalten.

**multiple Regressionsanalyse**

* **Verallgemeinerung der einfachen linearen Regression mit k Regressoren**   
  (welche die abhängige Variable Y erklären)
* mit **mehreren Einflussgrößen X1, X2, … Xk und einer Zielgröße Y**.
* Zweck:
  + zum Erstellen eines Vorhersagemodells,
  + Quantifizierung der Stärke des Zusammenhangs

**Regressionsfunktion**

**Voraussetzungen**

* X und Y sind quantitative Merkmale
* X → Y (Zusammenhang X → Y)

**ŷ = a + b \* x**

**Vorbereitung**

1. Überprüfung, ob Abhängigkeitsanalyse sinnvoll
2. Datenerhebung für X und Y
3. **Daten-Visualisierung im Streudiagramm** (qualitative Abhängigkeitsanalyse)
4. Auswahl Funktionstyp (hier: Beschränkung auf lineare Funktion)
5. **Berechnung der Regressionsfunktion** (nach Methode der kleinsten Quadrate)

**Bestimmung der Regressionsfunktion ŷ = a + b\*x nach der Methode der kleinsten Quadrate:**

* **OSL** 🡆 **Methode der kleinsten Quadrate**
* Regressionskoeffizienten a und b (Kurvenparameter) so bestimmen, dass Summe der quadratischen Abweichungen (Residuen **ui = (y – ŷ)²** ) von der Kurve zu den beobachteten Punkten minimal ist:

**Regressionskoeffizient a**

Formel im Nenner ist in beiden Gleichungen identisch

**Regressionskoeffizient b**

Bsp.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Filiale Nr.  i |  | Verkaufs-  fläche  (1000 m²)  xi | Umsatz  (Mio EUR)  yi | xi² | yi² | x \* y |
| 1 |  | 3 | 30 | 9 | 900 | 90 |
| 2 |  | 2 | 10 | 4 | 100 | 20 |
| 3 |  | 6 | 40 | 36 | 1600 | 240 |
| 4 |  | 5 | 20 | 25 | 400 | 100 |
| **n = 4** | **∑** | **16** | **100** | **74** | **3000** | **450** |
| **y̅ = 25 (100 : 4)** | | | | | | |

**Formel im Nenner ist in beiden Gleichungen identisch**

**Interpretation der Regressionsfunktion**

**Fragen für die Schätzung:**

* **Lohnt sich eine Erweiterung der Verkaufsfläche um 1.000 qm**?
* **Wie stark ist der Zusammenhang zwischen** Verkaufsfläche (**Einflussfaktor** (unabhängige Variable / Regressor) X) **und** Umsatz (**Zielgröße** (abhängige Variable / Regressand Y)?
* **Wie stark ist der Umsatz von der Verkaufsfläche abhängig**?
  + **Wie gut erklärt Regressionsfunktion die Abhängigkeit zwischen Umsatz und Verkaufsfläche**? (🡆 **Erklärungskraft des Modells**)

Für die Antwort wird ein Gütemaß (Bestimmtheitsmaß R² in %) benötigt

* Gütemaß für die Schätzung der Parameter
* Wie gut ist die „goodness of fit“ (Anpassungsgüte)
* **Wie gut beschreibt die Regressionsfunktion die Abhängigkeit**
* wie gut erklärt die Varianz in den Verkaufsflächen die Variant (Unterschiede) bei den Filialumsätzen?

**Prognosewerte und Residuen**

* **Zerlegung der Gesamtabweichung** **(yi – y̅)**
  + **unerklärte Abweichung** **(yi – ŷi)**
  + **erklärte Abweichung** **(ŷi – y̅I)**
* **Gesamtabweichung**  **(yi – y̅)**

**Differenz aus yi und Mittelwert y̅**

* **unerklärte Abweichung** (Residuum) **(yi – ŷi)   
  Differenz aus yi und Regressionsgeraden (Residuum ui)**  
  Regressionsgerade: yi – ŷi   
  Differenz aus yi und Regressionsfunktion   
  **ui = yi – ŷi = yi – ( a + b \* x)**
* **erklärte Abweichung** **(ŷi – y̅I)**

Differenz aus Regressionsgeraden und Mittelwert x̅

(**a + b \* x) – y̅**

Gesamtabweichung unerkl. Abweichung + erkl. Abweichung

**(yi – y̅) = (yi – ŷi) + (ŷi – y̅I)**

Interpretation der Abweichungszerlegung anhand des Ausprägungswerts yi = 40

Ausprägungswert yi = 40

Ausprägungswert xi = 6

Mittelwert y̅ = 25

Regressionskoeffizient a = 5

Regressionskoeffizient b = 5

Regressionsfunktion ŷi = 5 + 5 \* 6 = 35

Residuum ui = yi - ŷi

Gesamtabweichung unerkl. Abweichung + erkl. Abweichung

40 – 25 = (40 – 35) + (35 – 25)

15 = 5 + 10

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | **a = 5**  b = 5 |
| Filiale Nr.  i |  | Verkaufs-  fläche  (1000 m²)  xi | Umsatz  (Mio EUR)  yi | xi² | yi² | x \* y | ŷi = **a** + **b \* x**  ŷi = **5** + **5 \* x** | |
| 1 |  | 3 | 30 | 9 | 900 | 90 | **20** | |
| 2 |  | 2 | 10 | 4 | 100 | 20 | **15** | |
| 3 |  | 6 | 40 | 36 | 1600 | 240 | **35** | |
| 4 |  | 5 | 20 | 25 | 400 | 100 | **30** | |
| **n = 4** | **∑** | **16** | **100** | **74** | **3000** | **450** | **100** | |
|  | | | | **ŷ̅i = 25** 100 / 4 | | | | |

**Varianz der Residuen**

**Bestimmtheitsmaß R²**

* **Gütemaß der linearen Regression**
* **Anteil der Varianz der abhängigen Variable Y, der durch die Varianz der unabhängigen Variable X erklärt wird**   
  (wie gut erklärt die unabhängige Variable X die Varianz der abhängigen Variable Y)
* Bei **einfacher linearer Regression** (**eine unabhängige Variable X** ):  
  **R² = (rxy)²** (Korrelationskoeffizienten)
* **Verhältnis aus Streuung der Prognosewerte ŷ und der Gesamtstreuung der y-Werte**
* nutzt das Konzept der Varianzzerlegung
  + Zerlegung der Varianz des abhängigen Merkmals Y in erklärte und nicht erklärte Varianz (Residualvarianz)

Bsp.

x = Verkaufsfläche

y = Umsatz

rXY = 0,707

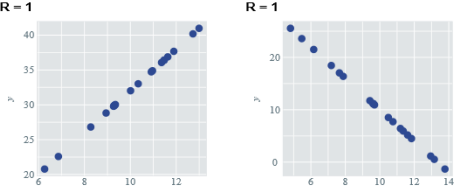
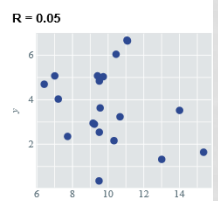
R2 = 0,707² = 0,50 = 50 %

50% der Varianz in den Umsätzen (Umsatzvarianz) lassen sich durch Varianz der Verkaufsflächen erklären. die übrigen 50% sind durch andere Einflussfaktoren zu erklären.

**Bestimmtheitsmaß R² anhand des Streudiagramms ableiten**

Je dichter die Datenpunkte im Streudiagramm auf einer Linie liegen, desto höher ist R².

Streuen die Datenpunkte ohne Zusammenhang im Raum, liegt das R² nahe 0.



**Probleme in der linearen Regression und Korrelation**

* Nur lineare Zusammenhänge
* Ausreißer (einzelne Fälle) können Regression und Korrelation stark beeinflussen