1 Deskriptive Statistik

Statistik und die von dieser Wissenschaft bereitgestellten Methoden sind stets notwendig, wenn im Rahmen empirischer Untersuchungen Daten erhoben und analysiert werden sollen. Die statistischen Methoden erfahren eine kontinuierliche Weiterentwicklung. Die Statistik ist von einer starken interdisziplinären Ausrichtung geprägt.

1.1 Grundlagen der Datenerhebung

1.1.1 Zum Begriff Statistik

Jeder von uns hat täglich mit Statistik zu tun, egal ob freiwillig oder unfreiwillig. Als Student ist man z.B. an seiner Durchschnittsnote interessiert. Die Durchschnittsnote ist nichts anderes als eine Statistik. Wie wird das Wetter morgen? Sollte man einen Regenschirm mitnehmen oder nicht? Wenn der Fernsehmeterologe sagt, mit 70%-iger Wahrscheinlichkeit wird es morgen regnen, so nutzt er für diese Aussage eine Statistik.

Wenn die Zeitungen berichten, bestimmte Fettsäuren seien schlecht für die Gesundheit, so sollte man beachten, dass statistische Verfahren zur Darstellung und Analyse der Ergebnisse verwendet wurden. Wir leben in einer Welt voller statistischer Manipulationen. Wir werden damit fast jeden Tag bombardiert. Um die diversen Behauptungen, Gegenargumente und Darstellungen richtig interpretieren zu können, müssen wir etwas über Statistik und dessen Anwendung wissen.

Wenn wir im Duden nach der Definition des Wortes Statistik nachschlagen, so finden wir mindestens zwei verschiedene Bedeutungen: Zunächst ist der Begriff „Statistik“ mit der Zusammenstellung von Zahlen und Daten verbunden. Diese Erklärung hat historische Wurzeln. Ursprünglich benutzte man die Statistik, um die Köpfe der Regierungen mit Daten zu versorgen, auf deren Basis sie regieren konnten. Solche Informationen in Form von Zahlen lassen sich sogar schon bei Aristoteles und seinem Buch „The matters of state“ finden. Tatsächlich haben die Wörter Staat und Statistik denselben sprachlichen Ursprung.

Seit frühester Zeit benutzen die meisten zivilisierten Staaten umfangreiche Statistiken für militärische und fiskalische Zwecke, um die Stärke und materielle Kraft einer Nation zu messen. Bereits in der Bibel finden sich solche Volkszählungen. Die Zusammenstellung von Zahlen für fiskalische Zwecke waren im alten Rom an der Tagesordnung. In ihrem Ursprung umfaßte die Wissenschaft der Statistik also Verfahren zur systematischen Erhebung, Klassifikation und Aufzählung von Informationen.

Die zweite Bedeutung des heutigen Wortes Statistik konzentriert sich auf die weiteren Aspekte der Aufbereitung, Analyse und Interpretation jeglicher Art von Daten. Man beschränkt sich längst nicht mehr auf quantitative Daten, wie z.B. die Anzahl von Haushalten, die besteuert werden können oder die Anzahl von Männern im wehrfähigen Alter. Auch qualitative Daten spielen mittlerweile eine große Rolle, wie z.B. die individuelle Entscheidung, wieviel von einem Gut konsumiert werden soll.

Der typische Ablauf einer statistischen Untersuchung umfasst folgende Punkte

(i) Datenerhebung (Beobachtung, Befragung, Experiment)

(ii) Datenaufbereitung, graphische Präsentation, Zusammenfassung (Darstellung der Zahlen in einer Tabelle oder Datenmatrix, alternativ: Darstellung mit Hilfe von Grafiken)

(iii) Datenanalyse, Abbildung des Sachverhalts in einem geeigneten (stochastischen) Modell, Schätzung des Modells, d.h. im allgemeinen Schätzung und Validierung seiner Parameter, Testen von Hypothesen, Ableitung von Ergebnissen bzw. Handlungsempfehlungen

Wichtig ist dabei die Unterscheidung zwischen deskriptiver (beschreibender), explorativer (suchender) und induktiver (schließender) Statistik.

1.1.2 Grundaufgaben der Statistik

Die Statistik besitzt drei Grundaufgaben im Rahmen der Datenanalyse. Jeder entspricht ein Teilgebiet.

Deskription (Beschreiben)

- Beschrieben oder dargestellt werden **Häufigkeiten** von *Ausprägungen* der betrachteten *Merkmale*.
- Graphische Datenaufbereitung (Diagramme, Verlaufskurven, Häufigkeitstabellen etc.), insbesondere relevant für die Präsentation umfangreichen Datenmaterials
- Gewinnung erster Eindrücke bzw. Ideen zur weiteren Analyse
- Datenvalidierung: Methoden der deskriptiven Statistik ermöglichen Erkennen von Fehlern im Datensatz (z.B. durch falsche Übertragung vom Fragebogen)
- keine Stochastik! (Rückschlüsse auf die Grundgesamtheit über Erhebungsdaten hinaus ist nicht möglich)

Exploration (Suchen)

- Auffinden von Strukturen in den Daten ohne stochastische Methoden
- Formulierung von Hypothesen für das den Daten zugrunde liegende stochastische Modell (wichtig für die Induktion)
- computerintensiv

Induktion (Schließen)

- Bereitstellung von Methoden, um Schlüsse (allgemeiner Art) auf die Grundgesamtheit ziehen zu können
- Schlüsse basieren auf wahrscheinlichkeitstheoretischem Modell (impliziert z.B. durch explorative Methoden)
- Vorgehensweise: Fragestellung → Formulierung eines stochastischen Modells (nicht datenbasiert) → Punktschätzung, Konfidenzintervalle, Tests

1.1.3 Grundbegriffe der Datenerhebung

- Ausgangspunkt einer empirischen Untersuchung: Erhebung und Analyse bestimmter Untersuchungsmerkmale.

Merkmalsträger
(stat. Objekt, stat. Einheit)

= Einzelobjekt einer statistischen Untersuchung (kleinste betrachtete Einheit)

– genau definierter Gegenstand, Vorgang, Person o.ä.

– Notation: ω_i, $i = 1, \ldots, n$

statistisches Merkmal

= Eigenschaft eines Merkmalsträgers

– Notation: X : Merkmal

$a_1, \ldots, a_k$: mögliche Merkmalsausprägungen

$x_1, \ldots, x_n$: beobachtete Ausprägungen bei den n betrachteten Merkmalsträgern

- **Grundgesamtheit**:

= Zusammenfassung aller relevanten Merkmalsträger

– Notation: $\Omega = \{\omega_1, \ldots, \omega_n\}$

– Bemerkung: Ω kann auch (abzählbar) unendlich sein
(z.B. Sterne im Universum, Menge aller Wassertropfen)

- **Merkmalsraum** (Zustandsraum):
  - = Zusammenfassung aller möglichen Merkmalsausprägungen
  - – Notation: $\mathbb{S} = \{a_1, \ldots, a_k\}$
  - – Bemerkung: $\mathbb{S}$ kann auch überabzählbar unendlich sein (z.B. Länge eines Werkstücks $\mathbb{S} = \mathbb{R}_+$)
- Damit lässt sich ein Merkmal beschreiben als eindeutige Abbildung $X : \Omega \to \mathbb{S}$.
- Bemerkungen:
  - (i) spätere Formalisierung von X als Zufallsvariable (Statistische Inferenz)
  - (ii) meist Beschränkung auf Teilgesamtheit (Teilpopulation) $T \subset \Omega$, spätere Formalisierung von T als Stichprobe
- **Kodierung des Zustandsraumes**:
  - = eineindeutige Abbildung $C : \mathbb{S} \to \mathbb{R}$ (i.d.R. $\mathbb{N}_0$)
  - – rechentechnisch effizienter

Multivariate Daten

- häufig: simultane Erhebung von zwei oder mehreren Merkmalen an den gleichen statistischen Einheiten

  $\Rightarrow$ Wir betrachten p Merkmale ($p \geq 1$) mit den Bezeichnungen $X_1, \ldots, X_p$.

  Der Zustandsraum ist dann das kartesische Produkt $\mathbb{S} = \mathbb{S}_1 \times \mathbb{S}_2 \times \ldots \times \mathbb{S}_p$.
- Die Daten bestehen aus $n \cdot p$ Informationen, die in einer $(n \times p)$-dimensionalen Datenmatrix $\mathbf{X}$ zusammengefasst werden:

$$\mathbf{X} = \begin{bmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{np} \end{bmatrix},$$

  wobei x_{ij} die Ausprägung von Merkmal X_j bei Objekt ω_i bezeichnet.
- Alternative Darstellungen von $\mathbf{X}$ sind

$$\mathbf{x}_i := \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix}, \quad i = 1, \ldots, n.$$

  Diese Darstellung entspricht der i-ten Zeile von $\mathbf{X}$, als Spaltenvektor (!) geschrieben, und beinhaltet die Ausprägungen aller p Merkmale für das i-te statistische Objekt.

$$\mathbf{x}_{(j)} := \begin{bmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{bmatrix}, \quad j = 1, \ldots, p.$$

  Diese Darstellung entspricht der j-ten Spalte von $\mathbf{X}$ und beinhaltet die Ausprägungen des j-ten Merkmals für alle n statistischen Objekte.

  Daraus folgt:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_n^\top \end{bmatrix} = \begin{bmatrix} \mathbf{x}_{(1)} & \ldots & \mathbf{x}_{(p)} \end{bmatrix}.$$

Identifikations- vs. Erhebungsmerkmale

- Identifikationsmerkmale
  - sachlich, räumlich, zeitlich
  - zur Klassifikation eines statistischen Objekts, ob es für die statistische Untersuchung relevant ist oder nicht
  - ergibt sich meistens aus sachlicher Beschreibung
  - Charakterisierung der statistischen Objekte
- Erhebungsmerkmale
  - eigentlich interessierende (und zu erhebende) Merkmale
  - Sie sollen die für die empirische Fragestellung notwendigen Informationen liefern.

1.1.4 Erhebungsarten

Erhebung: Bezeichnung für die Beschaffung der benötigten Informationen (Merkmale) über die statistischen Einheiten bzw. die Gewinnung der Daten

- **Beobachtung**: dient zur Datengewinnung bei Festhalten von zeitlichen Vorgängen sowie bei Erfassen von Sachverhalten, die nicht gesteuert werden. Systematisierung durch Beobachtungsprotokoll (zum Festhalten des Beobachteten) als Erhebungsinstrumentarium. Verdeckte Beobachtung: Beobachter gibt sich nicht zu erkennen. Teilnehmende Beobachtung: Beobachter nimmt aktiv am Geschehen teil.
- **Befragung**: mündliche, schriftlich, telefonisch, via Internet, mit/ohne Interviewer. Grundlage: Fragebogen. (Keine direkte Beobachtung möglich, die Befragten müssen diese Merkmale selbst bei sich „beobachten“)
- **Experiment**: Erzeugung der Daten durch Simulation der interessierenden Situationen. Anwendung meistens in Naturwissenschaften aber auch Sozialwissenschaften, Psychologie, Wirtschaftswissenschaften (Entscheidungssituationen). (Interessierende Situation wird künstlich erzeugt, um direkte oder indirekte (über Befragung) Beobachtung zu ermöglichen.)

Datenform

- **Querschnittanalyse**: Querschnitt durch eine Population zu festem Zeitpunkt
  (z.B. Einkommen von Studenten)
- **Längsschnittanalyse**: ein Objekt wird über längere Zeit hinweg beobachtet (Eine solche Folge von Beobachtungen heißt *Zeitreihe*, z.B. Arbeitslosenquote in Deutschland)
- **Panel**: Längsschnittanalyse für (Teil-)Population
  (z.B. klinische Studien zum Wikungsgrad eines neuen Medikaments)

Umfang

- **Vollerhebung (Totalerhebung)** Untersuchung aller statistischen Einheiten der Grundgesamtheit. Beispiel: Volkszählung. Vorteil: Wir erhalten **vollständige Information über Grundgesamtheit**, mögliche Unsicherheiten aufgrund fehlender Informationen sind somit ausgeschlossen. Nachteil: Antwortverweigerungen, Kosten, evt. unmöglich z.B. bei zerstörender Materialprüfung oder bei hypothetischen Grundgesamtheiten.

- **Teilerhebung (Stichprobe)** Untersuchung beschränkt sich auf eine Teilgesamtheit (verschiedene Auswahlmöglichkeiten: Quotenauswahl = repräsentativer Querschnitt, Zufallsauswahl (rein oder geschichtet); siehe Stichprobentheorie). Diese soll für die Grundgesamtheit repräsentativ sein. Problem: Merkmalsausprägungen für Grundgesamtheit unbekannt, daher Stichprobenfehler durch Ungewissheit der Repräsentativität. Vorteil: geringere Kosten, generell weniger Aufwand.

Herkunft

- **Primärerhebung**. Die Erhebung wird speziell im Hinblick auf die aktuelle Fragestellung durchgeführt.
- **Sekundärerhebung**. Es wird auf Daten aus anderen Erhebungen zu ähnlichen Fragestellungen zurückgegriffen, z.B. auf bereits vorhandene Originaldaten aus dem statistischen Jahrbuch.
- **Tertiärerhebung**: Für die Untersuchung stehen nur noch bereits transformierte oder komprimierte Daten zur Verfügung, z.B. Mittelwerte.

1.1.5 Kategorisierung statistischer Merkmale

Statistische Merkmale lassen sich nach verschiedenen Gesichtspunkten unterscheiden. Siehe Tabelle 1.

Unterscheidung nach ...		
... Quantifizierbarkeit der Ausprägungen	**qualitative (kategoriale) Merkmale**: • nur zuordbar • Bsp. Wohnort, Name	**quantitative (metrische) Merkmale**: • mess- oder zählbar • Bsp. Alter, Körpergröße
... Anzahl der Ausprägungen*	**diskrete Merkmale**: • höchstens abzählbar unendlich viele mögliche Ausprägungen • Bsp. Gehaltsklassen, Kaufverhalten	**stetige Merkmale**: • überabzählbar (unendlich) viele mögliche Ausprägungen • Bsp. Geschwindigkeit, Gewicht
... Direktheit der Informationsgewinnung	**manifeste/beobachtbare Merkmale**: • können direkt erhoben werden • Bsp. Abiturnote	**latente Merkmale**: • Operationalisierung über Indikatoren/Items notwendig • Bsp. Bildungsgrad, Kreativität, Nutzen
... Skalenniveau	siehe unten	

Tabelle 1: Klassifikationsmöglichkeiten von Merkmalen.

***Bemerkung**: Es gibt Zwischenformen, sogenannte *quasi-stetige Merkmale*. Diese diskreten Merkmale haben so viele mögliche Ausprägungen, dass man sie wie stetige Merkmale behandeln kann. Beispiele: Alter, Einkommen. Weiterhin können stetige Merkmale durch Kategorisierung der Merkmalsausprägungen (Klassenbildung) in diskrete Merkmale transformiert werden.

- Messen ist die Zuordnung von Zahlen oder Symbolen zu den Ausprägungen von Merkmalen

  - gemäß festgelegter Regeln, die sicherstellen, dass
  - die Messwerte die gleichen Beziehungen zueinander aufweisen, wie die Ausprägungen der Merkmale.
  - Das entspricht dem Anlegen einer Skala und dem Ablesen des Zahlenwertes.

- Eine **Skala** ist ein Hilfsmittel zur Messung der Merkmalsausprägungen in Zahlenform.
- im einfachsten Fall: Kodierungstabelle. Beispiel: männlich $\rightarrow 0$, weiblich $\rightarrow 1$.
- komplexer: Physikalische Merkmale bekommen ihre Ausprägung als Zahlenwert zugeordnet. Beispiel: Länge, Temperatur mit entsprechenden Skalen von Lineal, Thermometer.
- Jede Skala definiert zum *empirischen Relativ* (Menge der Ausprägungen eines Merkmals = Zustandsraum) das sogenannte *numerische Relativ* (Menge der reellen Zahlen, die den Ausprägungen zugeordnet werden können).
- numerisches Relativ $\subseteq \mathbb{R} \rightarrow$ Alle auf $\mathbb{R}$ definierten Operationen sind im numerischen Relativ durchführbar.
- Achtung: Nicht alle Operationen im numerischen Relativ einer Skala machen im korrespondierenden empirischen Relativ Sinn
- daher: Unterscheidung von Skalen nach den Eigenschaften, die vom numerischen Relativ auf das empirische Relativ übertragbar (und dort sinnvoll interpretierbar) sind. Die wichtigsten Skalentypen sind in Tabelle 2 dargestellt. Weitergehende Charakterisierungen fasst Tabelle 3 zusammen.

Skalentyp	zulässige Transformationen	invariant bleiben unter zulässigen Transformationen	Beispiele
Nominalskala	jede eineindeutige Funktion	Eindeutigkeit der Meßwerte	Numerierung von Fußballspielern, Kontonummern, Matrikelnummern
Ordinalskala	jede streng monoton steigende (isotone) Funktion	Rangordnung der Meßwerte	Richtersche Erdbebenskala, Schulnoten, Hunger
Intervallskala	jede positiv lineare (affine) Funktion $y = ax + b, a, b \in \mathbb{R}, a > 0$	Verhältnisse der Intervalle zwischen Meßwerten	Temperatur (Celcius, Fahrenheit), Nutzen
Verhältnisskala	jede Ähnlichkeitsfunktion $y = ax, a \in \mathbb{R}_+$	Verhältnisse von Meßwerten	Länge, Masse, Zeit, Winkel, Temperatur (Kelvin), Preise
Absolutskala	nur die Identitätsfunktion $y = x$	Meßwerte	Häufigkeit, Wahrscheinlichkeit

Tabelle 2: Die wichtigsten Skalentypen

	Verschiedenartigkeit	natürl. Reihenfolge	Interpretierbarkeit der Verhältnisse der Differenzen	natürl. Nullpunkt	natürl. Einheit
Nominalskala	ja	nein	nein	nein	nein
Ordinalskala	ja	ja	nein	nein	nein
Intervallskala	ja	ja	ja	nein	nein
Verhältnisskala	ja	ja	ja	ja	nein
Absolutskala	ja	ja	ja	ja	ja

Tabelle 3: Charakterisierungen der wichtigsten Skalentypen

1.2 Univariate deskriptive Statistik

1.2.1 Häufigkeiten

- Ausgangspunkt ist die **Urliste** (auch Primärdaten, Rohdaten), d.h. die Werte $x_1, \ldots, x_n$ eines Merkmals X von n Untersuchungseinheiten (statistischen Objekten).

  Beispiele:

  - Alter in Jahren: $21, 25, 22, 23, \ldots$,
  - Konfession (kodiert: rk = 1, ev = 2, ...): $1, 1, 2, 1, 3, 4, 4$,
  - Mietpreis Euro/qm: $15.2, 20, 17.50$,
  - Umsatz in Filialen (in Euro)...

- **geordnete Urliste** (auch **Ordnungsstatistik**): $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ ist bei metrischen Merkmalen informationsträchtig

- **Ziel**: übersichtliche, zusammenfassende Darstellung; Informationsreduktion

- **Idee**: Durchsuchung der Urliste nach vorkommenden Zahlenwerten bzw. Ausprägungen $a_1 < a_2 < \ldots < a_k$

- **Frage**: Wie kann man das rechentechnisch implementieren? Betrachte hierzu folgendes Beispiel:
  Sei $\mathbb{S} = \{1, 2, 3\}$ und $T = \{\omega_1, \ldots, \omega_6\} \subset \Omega$ mit folgenden beobachteten Merkmalsausprägungen

  $$X(\omega_1) = 1,\ X(\omega_2) = 1,\ X(\omega_3) = 2,\ X(\omega_4) = 3,\ X(\omega_5) = 3,\ X(\omega_6) = 3.$$

  Die rechentechnische Bestimmung der absoluten Häufigkeiten erfolgt mit Hilfe von $|\mathbb{S}| \cdot |T|$ vielen Booleschen Funktionen, die für jeweils ein $\omega_i \in T$ und ein $a_j \in \mathbb{S}$ entscheiden, ob $X(\omega_i) = a_j$ ist oder nicht (wahr-falsch-Entscheidung, hier codiert mit 0 (=falsch) und 1 (=wahr)). Für das Beispiel erhalten wir

$a_j \backslash \omega_i$	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	Summe:
1	1	1	0	0	0	0	2
2	0	0	1	0	0	0	1
3	0	0	0	1	1	1	3

Definition 1.1. *Die absolute Häufigkeit h_j der Ausprägung a_j, d.h. die Anzahl der $x_i, i = 1, \ldots, n$, mit $x_i = a_j$, ist definiert als*

$$h_j = h(a_j) = \sum_{i=1}^{n} 1_{\{a_j\}}(x_i), \quad \textit{mit} \quad 1_{\{a_j\}}(x_i) = \begin{cases} 1, & \textit{falls } x_i = a_j, \\ 0, & \textit{sonst}, \end{cases} \tag{1.1}$$

- Eigenschaften der absoluten Häufigkeit:

  (i) $h_j \in \{0, 1, \ldots, n\}$,

  (ii) $\sum_{j=1}^{k} h(a_j) = n$

  Beweis: $$\sum_{j=1}^{k} h(a_j) = \sum_{j=1}^{k} \sum_{i=1}^{n} 1_{\{a_j\}}(x_i) = \sum_{i=1}^{n} \underbrace{\sum_{j=1}^{k} 1_{\{a_j\}}(x_i)}_{=1} = \sum_{i=1}^{n} 1 = n$$

- **Problem**: keine Berücksichtigung des Stichprobenumfangs n: z.B. $h(a_j) = 10, n = 20, h(a_j) = 10, n = 200$ dieselben numerischen Werte, aber relativ zum Stichprobenumfang betrachtet ist der Anteil der Häufigkeiten unterschiedlich.

  Definition 1.2. *Die relative Häufigkeit f_j der Ausprägung a_j ist definiert als*

  $$f_j = f(a_j) = \frac{h_j}{n}. \tag{1.2}$$

- Eigenschaften der relativen Häufigkeit:

  (i) $f(a_j) \in [0, 1]$,

  (ii) $\sum_{j=1}^{k} f(a_j) = 1$.

- Die Häufigkeiten $h_1, \ldots, h_k$ bzw. $f_1, \ldots, f_k$ fasst man in einer *Häufigkeitstabelle* zusammen.

- Die Ausprägungen $a_1, \ldots, a_k$ zusammen mit den Häufigkeiten bezeichnet man als *Häufigkeitsdaten*.

	j	a_j	h_j	f_j
	1	a_1	h_1	f_1
	⋮	⋮	⋮	⋮
	k	a_k	h_k	f_k
Summe:	-	-	n	1

Graphische Darstellungen von Häufigkeiten

- Aussagen über die Verteilung der (relativen) Häufigkeiten auf einzelne Merkmalsausprägungen (d.h. die Häufigkeitsdaten) sind optisch/graphisch einfacher als bloße Zahlendarstellung.

- **Stabdiagramm**: Trage über $a_1, \ldots, a_k$ jeweils einen zur x-Achse senkrechten Strich (Stab) mit Höhe $h_1, \ldots, h_k$ (oder $f_1, \ldots, f_k$) ab.

- **Säulendiagramm**: wie Stabdiagramm, aber mit Rechtecken statt Strichen.

- **Balkendiagramm**: wie Säulendiagramm, aber mit vertikal statt horizontal gelegter x-Achse.

- **Kreisdiagramm**: Flächen der Kreissektoren proportional zu den Häufigkeiten: Winkel des Kreissektors $j = f_j \cdot 360°$.

- **Problem**: Diese Darstellungsformen sind nur geeignet, wenn die Anzahl k der möglichen Ausprägungen nicht zu groß ist. Insbesondere für metrische Merkmale mit vielen verschiedenen Werten sind sie daher ungeeignet. Zwei einfache Darstellungsformen für metrische Merkmale mit vielen Ausprägungen sind das Stamm-Blatt-Diagramm (Stem-Leaf) und das Histogramm.

Klassenbildung (Gruppierung):

- **Problem bei metrischen Merkmalen**: oft nur wenige Werte der Urliste identisch, k fast so groß wie n.
- Alternative: Klassierung (= Bildung geeigneter Klassen) und Betrachtung der Häufigkeiten für die gruppierten Daten.
- k Klassen der Form $[c_0, c_1), [c_1, c_2), \ldots, [c_{k-1}, c_k)$, wobei c_{j-1} die untere und c_j die obere Klassengrenze der j-ten Klasse bezeichnen.
- Klassenbreite: $d_j = c_j - c_{j-1} \qquad j = 1, \ldots, k$
- Klassenmitte: $m_j = (c_j + c_{j-1})/2$
- absolute Häufigkeit der Klasse j:

$$h_j = \sum_{a_i \in [c_{j-1}, c_j)} h(a_i) = \sum_{i=1}^{n} 1_{[c_{j-1}, c_j)}(x_i),$$

  wobei wieder die Indikatorfunktion

$$1_{[c_{j-1}, c_j)}(x_i) = \begin{cases} 1, & \text{falls } x_i \text{ in die } j\text{-te Klasse fällt,} \\ 0, & \text{sonst.} \end{cases}$$

  verwendet wird.

- relative Häufigkeit der Klasse j: $f_j = h_j/n$

Stem-Leaf-Diagramm (Stamm-Blatt-Diagramm)

- ist eine semigraphische Darstellungsform für metrische Merkmale
- **Schritt 1**: Teile den Datenbereich in Intervalle gleicher Breite $d = 0.5$ oder 1 mal einer Potenz von 10 ein. Trage die erste(n) Ziffer(n) der Werte im jeweiligen Intervall links von einer senkrechten Linie der Größe nach geordnet ein. Dies ergibt den *Stamm*.
- **Schritt 2**: Runde die beobachteten Werte auf die Stelle, die nach den Ziffern des Stamms kommt. Die resultierenden Ziffern ergeben die *Blätter*. Diese werden zeilenweise und der Größe nach geordnet rechts vom Stamm eingetragen.
- Vorteil: Stem-Leaf-Diagramme enthalten (bis auf Rundung) die Werte der Urliste und somit einen guten Einblick in die Datenstruktur für explorative Analyse.
- Nachteil: unübersichtlich für große Datensätze

Histogramm

- **Voraussetzung**: mindestens ordinalskaliertes Merkmal
- **Vorgehensweise**: Für die Gruppierung wählt man als Klassen benachbarte Intervalle $[c_0, c_1), [c_1, c_2), \ldots, [c_{k-1}, c_k)$. **Klassenbreite**: $d_j = c_j - c_{j-1}, j = 1, \ldots, k$.
- Da das Auge primär die Fläche der Rechtecke bzw. Säulen wahrnimmt, wird das Histogramm so konstruiert, daß die Fläche über den Intervallen gleich oder proportional zu den absoluten bzw. relativen Häufigkeiten ist (**Prinzip der Flächentreue**). Es gilt: Fläche = Breite × Höhe, bzw. Höhe = Fläche / Breite. Daher ergibt sich als abzutragende Höhe (gleich oder proportional zu) h_j/d_j bzw. f_j/d_j.

$$\text{„Blockhöhe (abzutragende Höhe)“} = \widetilde{f}(x) = \begin{cases} f_j/d_j & x \in [c_{j-1}, c_j) \\ 0 & \text{sonst} \end{cases} \qquad \text{für} \quad j = 1, \ldots, k.$$

- **Empfehlungen**: Klassenbreiten d_j sollten gleich groß sein. Relative Häufigkeiten als Höhe wählen. Vermeidung offener Randklassen.
- **Eigenschaften**: Bei sehr kleiner Klassenbreite erhält man sehr unruhige Histogramme, die im Extremfall dem Stabdiagramm ähnlich werden. Bei sehr großer Klassenbreite wird das Histogramm weniger Sprünge aufweisen, im Extremfall nur ein einziges Rechteck (wenn alle Daten in einer Klasse liegen). Faustregel für Anzahl der Klassen (und damit Wahl der Klassenbreite) etwa $k = [\sqrt{n}], k = 2[\sqrt{n}], k = [10 \log_{10} n]$
- **Problem**: Form / Aussehen des Histogramms ist nicht nur von der Klassenbreite abhängig sondern auch vom Anfangspunkt
- Idee: Man berechnet Histogramme mit gleicher Klassenbreite für verschiedene Anfangspunkte und über sie den Durchschnitt. Das führt zum sogenannten ASH (average shifting histogram) und der Methode des WARPings.
- **WARPing** (weighted averaging of rounded points): Man betrachtet ein Histogramm mit Anfangspunkt x_0 und den Intervallen $[x_0, x_0+h), [x_0+h, x_0+2h), [x_0+2h, x_0+3h), \ldots, h = d_i$ (gleiche Klassenbreiten). Angenommen man will S neue Klassen bilden, die jeweils um $l/(S+1), l \in \{0, 1, \ldots, S\}$ nach rechts verschoben sind. Im Ergebnis erhalten wir $S + 1$ Werte (S neue Werte durch die Verschiebungen) für die Häufigkeit für jedes x. Aus diesen Werten bildet man den Durchschnitt und nimmt diesen Wert als „Schätzer“ (Begriff wird später noch eingeführt) für die Häufigkeit eines x-Wertes. Die graphische Darstellung über alle x-Werte wird als **ASH** bezeichnet.
- **Problem**: praktisch nur computergestützt berechenbar.
- interessant: Für $S \to \infty$ besitzt das ASH keine Abhängigkeit mehr vom Anfangspunkt x_0 und das ASH geht über von einer stückweise konstanten Funktion (Treppenfunktion) zu einer stetigen Funktion. Dieses asymptotische Verhalten kann jedoch auch durch die Verwendung von Kerndichteschätzern erreicht werden.

Kerndichteschätzer

- Unerwünschte Eigenschaften des Histogramms (neben der Abhängigkeit vom Anfangspunkt x_0): Das Histogramm weist jedem Punkt x innerhalb einer Klasse dieselbe Häufigkeit zu. Das Histogramm besitzt an der Grenze zweier benachbarter Klassen eine Sprungstelle.
- Idee des Histogramms: Die Häufigkeit einer Ausprägung x ist (Anzahl der statistischen Einheiten, die in die Klasse, in der x liegt, fallen) / (Klassenbreite $d \cdot$ Gesamtzahl n statistischer Einheiten).
- Idee der Kerndichteschätzung: Man betrachtet nicht mehr die Klasse, in der x liegt, sondern eine Klasse bzw. ein Intervall (symmetrisch) um x, also $[x-h, x+h)$. Von x verschiedenen Beobachtungen können unterschiedliche Gewichte zugewiesen werden - in Abhängigkeit von ihrer Entfernung von x. Wir können also schreiben
$$\widetilde{f}(x) = \frac{1}{2nh} \sum_{i=1}^{n} 1_{[x-h,x+h)}(x_i).$$
- Dieser Ausdruck weist also jeder Beobachtung x_i, deren Abstand zu x (dem Argument, dessen relative Häufigkeit wir schätzen/abtragen wollen) nicht größer als h ist, den Wert $1/2$ zu. Dies lässt sich mit Hilfe der **Kernfunktion**
$$K(u) = \begin{cases} \frac{1}{2} & \text{für } -1 \le u \le 1, \\ 0 & \text{sonst} \end{cases}$$
darstellen

- Dann ist
$$\frac{1}{h}K\left(\frac{x-x_i}{h}\right) = \begin{cases} \frac{1}{2h} & \text{für } x_i - h \leq x \leq x_i + h, \\ 0 & \text{sonst} \end{cases} \tag{1.3}$$
ein über x_i zentriertes Rechteckfenster mit Fläche 1 und Breite $2h$.

- Dann lässt sich $\hat{f}(x)$ auch in der Form
$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-x_i}{h}\right) \tag{1.4}$$
schreiben, wobei $K(\cdot)$ eine beliebige Kernfunktion ist.

Charakterisierung von Verteilungen

- **Symmetrie**: Eine Verteilung heißt **symmetrisch**, wenn es eine Symmetrieachse gibt, so daß die linke und die rechte Hälfte der Verteilung annähernd spiegelbildlich zueinander sind. Die Daten liegen also gleichmäßig um ein Zentrum.

- Deutlich unsymmetrische Verteilungen heißen **schief**. Eine Verteilung ist **linkssteil** (oder **rechtsschief**), wenn der überwiegende Anteil von Daten linksseitig konzentriert ist.

- **Gipfel**: mind. lokales Maximum der Häufigkeitsverteilung, d.h. Merkmalsausprägung(en) mit maximaler Häufigkeit (im Vergleich zu den anderen Ausprägungen), von der aus die Häufigkeiten flacher zu den Randbereichen hin verlaufen.

- **Modalität**: Eine Verteilung heißt *unimodal* (eingipfelig), wenn sie nur einen Gipfel besitzt. Tritt ein zweiter deutlicher Gipfel auf, so heißt die Verteilung *bimodal* (zweigipfelig). Treten noch mehrere Nebengipfel auf, so heißt die Verteilung *multimodal* (mehrgipfelig). Tritt meistens auf, wenn die Daten eines Merkmals unterschiedlichen Teilgesamtheiten entstammen.

Kumulierte Verteilung

- Darstellungsziel: Wieviel Prozent (Welcher Anteil) der Daten unterschreiten/überschreiten einen bestimmten Wert. Bei welchem Wert liegen z.B. 50% der Daten darunter? Bemerkung: Diese Art von Fragestellung ist nur sinnvoll bei mindestens ordinalskalierten Merkmalen.

- Absolute kumulierte Häufigkeitsverteilung ist definiert durch
$$H(x) = \sum_{a_i \leq x} h(a_i) = \sum_{i:a_i \leq x} h_i = \sum_{a_i \leq x}\sum_{j=1}^{n} 1_{\{a_i\}}(x_j)$$

- Relative kumulierte Häufigkeitsverteilung ist definiert durch
$$F(x) = H(x)/n = \sum_{a_i \leq x} f(a_i) = \sum_{i:a_i \leq x} f_i \tag{1.5}$$

- relative kumulierte Häufigkeitsverteilung = empirische Verteilungsfunktion

- Beide Funktionen sind monoton wachsende **Treppenfunktionen**, die an den Ausprägungen $a_1, \ldots, a_k$ um die entsprechende absolute bzw. relative Häufigkeit nach oben springen. Sie sind in den Sprungstellen **rechtsseitig stetig**, d.h. der obere Wert, die sog. Treppenkante, ist der zugehörige Funktionswert.

- $H(x) = 0$ bzw. $F(x) = 0$ für alle $x < a_1$, $H(x) = n$ bzw. $F(x) = 1$ für alle $x > a_k$.

- $0 \leq F(x) \leq 1$, $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to+\infty} F(x) = 1$.

Theorie: Lineare Interpolation

- Absolute Häufigkeiten für klassierte Daten eines Merkmals X mit k Klassen sind folgendermaßen definiert
$$h_j = h(c_{j-1} \leq X < c_j), \quad j = 1, \ldots, k,$$
mit Klassengrenzen $c_0 < c_1 < \ldots c_k$.
- Annahme: Merkmalswerte sind innerhalb jeder der Klassen gleichverteilt, d.h. es wird unterstellt, dass sich die Beobachtungen gleichmäßig über den Bereich der jeweiligen Klasse erstrecken.
- Idee: Approximation durch Polygonzug (d.h. eine stetige, stückweise lineare Funktion)
- Dazu wird die lineare Interpolation verwendet: Gegeben sind jeweils zwei Punkte $(c_{j-1}, F(c_{j-1}))$ und $(c_j, F(c_j))$. Diese Punkte sollen durch eine lineare Funktion $y = ax + b$ verbunden werden. Problem: a und b sind unbekannt. Idee: Löse folgendes lineares Gleichungssystem:

(I)	$F(c_{j-1})$	$=$	$ac_{j-1} + b$
(II)	$F(c_j)$	$=$	$ac_j + b$
$(II) - (I)$	$F(c_j) - F(c_{j-1})$	$=$	$a(c_j - c_{j-1})$

Wir erhalten
$$a = \frac{F(c_j) - F(c_{j-1})}{c_j - c_{j-1}}$$
und durch Einsetzen in (I)
$$b = F(c_{j-1}) - \frac{F(c_j) - F(c_{j-1})}{c_j - c_{j-1}} c_{j-1}$$
und damit für $x \in [c_{j-1}, c_j)$:
$$\begin{aligned} F(x) &= \frac{F(c_j) - F(c_{j-1})}{c_j - c_{j-1}} x + F(c_{j-1}) - \frac{F(c_j) - F(c_{j-1})}{c_j - c_{j-1}} c_{j-1} \\ &= F(c_{j-1}) + \frac{x - c_{j-1}}{c_j - c_{j-1}} (F(c_j) - F(c_{j-1})) \end{aligned}$$

1.2.2 Maßzahlen von Verteilungen

Ziel: Formale Quantifikation der Eigenschaften von Verteilungen in komprimierter Form durch numerische Werte. Das entspricht einer radikalen Komprimierung!

Lageparameter

Ziel: Beschreibung des Zentrums (=Schwerpunkt) einer Verteilung durch einen numerischen Wert

Modalwert (Modus)

- x_{mod}: Ausprägung mit der größten Häufigkeit
  - nichtklassierte Daten: $x_{mod} = \{a_j | \, h_j = \max_{a_i} h_i\}$
  - klassierte Daten: Idee: Berücksichtigung der abzutragenden Höhen $\widetilde{f}_j$ der benachbarten Klassen der Modalklasse, damit ebenfalls Berücksichtigung der Klassenbreite
$$x_{mod} = c_{j-1} + \frac{\widetilde{f}_j - \widetilde{f}_{j-1}}{2\widetilde{f}_j - \widetilde{f}_{j-1} - \widetilde{f}_{j+1}} (c_j - c_{j-1}),$$
mit j : Modalklasse

- Eigenschaften:
  - Der Modus ist eindeutig, falls die Häufigkeitsverteilung ein eindeutiges Maximum besitzt.
  - stimmt auf jeden Fall mit (mind.) einer Merkmalsausprägung (oder Klasse) überein.
  - robust gegen Ausreißer.

Median

- Für ungerades n ist der Median x_{med} die **mittlere Beobachtung der geordneten Urliste** und für gerades n ist der Median x_{med} das arithmetische Mittel (s.u.) der beiden in der Mitte liegenden Beobachtungen, d.h.

$$x_{med} = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{für } n \text{ ungerade,} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right), & \text{für } n \text{ gerade.} \end{cases} \tag{1.6}$$

- Median für klassierte Daten:

$$x_{med} = c_{j-1} + \frac{0.5 - F(c_{j-1})}{F(c_j) - F(c_{j-1})}(c_j - c_{j-1}),$$

  wobei j die sogenannte Medianklasse bezeichnet, in der der Median liegt.

- Voraussetzung: gegebene natürliche Rangfolge, daher ordinalskaliertes Merkmal.
- Eigenschaften: mind. 50% der Daten sind kleiner oder gleich. mind. 50% der Daten sind größer oder gleich. Robustheit. Anschaulichkeit (einfache Interpretierbarkeit).
- Bemerkung: Zusammenhang zur empirischen Verteilung. $F(x_{med}) = 0.5$ bzw. kleinster Wert u mit $F(u) \geq 0.5$, $x_{med} = \min\{u : F(u) \geq 0.5\}$.

Quantile

Jeder Wert x_p, mit $0 < p < 1$, für den mindestens ein Anteil p der Daten $\leq x_p$ und mindestens ein Anteil $1 - p$ der Daten $\geq x_p$ ist, heißt *p-Quantil:*

$$\frac{\text{Anzahl}\,(x\text{-Werte} \leq x_p)}{n} \geq p \quad \text{und} \quad \frac{\text{Anzahl}\,(x\text{-Werte} \geq x_p)}{n} \geq 1 - p.$$

Alternativ: x_p ist der kleinste x-Wert, für den $F(x) \geq p$ gilt, d.h.

$$F(x) < p \text{ für } x < x_p \quad \text{und} \quad F(x_p) \geq p.$$

Damit gilt für das *p-Quantil*

- bei nichtklassierten Daten:

$$x_p = x_{(\lfloor np \rfloor + 1)}, \qquad \text{wenn } np \text{ nicht ganzzahlig, wobei } \quad \lfloor x \rfloor := \max_{k \in \mathbb{Z}, k \leq x}(k)$$
$$\text{die zu } x \text{ nächstkleinere ganze Zahl bezeichnet;}$$

$$x_p \in [x_{(np)}, x_{(np+1)}], \qquad \text{wenn } np \text{ ganzzahlig.}$$

- bei klassierten Daten:

$$x_p = c_{j-1} + \frac{p - F(c_{j-1})}{F(c_j) - F(c_{j-1})}(c_j - c_{j-1}), \quad p \in (0,1), \quad \text{mit } j \text{ Quantilklasse}$$

- *Speziell:* $x_{0.5}$ = Median, $x_{0.25}$ = unteres Quartil, $x_{0.75}$ = oberes Quartil
- komprimierte Visualisierung einer Verteilung mit Hilfe von Quantilen durch **Boxplots**

Boxplot

- $x_{0.25}$ = Anfang der Box.
- $x_{0.75}$ = Ende der Box. Damit gibt der Interquartilsabstand $d_Q = x_{0.75} - x_{0.25}$ die Länge der Box an.
- Der Median wird durch einen Punkt (oder eine durchgezogene Linie) in der Box markiert.
- Zwei Linien („Zäune“ bzw. „whiskers“) außerhalb der Box gehen bis zu $x_{\min}$ und $x_{\max}$.

Konstruktion des modifizierten Boxplots

- Begrenzung der Box durch unteres und oberes Quartil $x_{0.25}$ und $x_{0.75}$. Damit gibt der Interquartilsabstand d_Q die Länge der Box an.
- Der Median $x_{0.5}$ wird als durchgezogene Linie eingezeichnet.
- Für den oberen Zaun ergibt sich

$$z_o = \min\{x_{0.75} + 1.5 \cdot d_Q; \max_i\{x_1, \ldots, x_n\}\},$$

  für den unteren Zaun ergibt sich

$$z_u = \max\{x_{0.25} - 1.5 \cdot d_Q; \min_i\{x_1, \ldots, x_n\}\}.$$

- Gegebenenfalls werden alle Beobachtungen außerhalb der Zäune als Punkte eingezeichnet.
- Bemerkung: Es gibt eine weitere Modifikation des Boxplots, in der alle Beobachtungen die größer als $x_{0.75} + 3 \cdot d_Q$ bzw. kleiner als $x_{0.25} - 3 \cdot d_Q$ sind als Sterne eingezeichnet werden. Diese Punkte benötigen oft eine besondere Behandlung (Ausreißer?, falsche Daten?).

Beispiel:

Stadt	Bevölkerung (in 10 000)	Ordnungsstatistik
New York	778	$x_{(15)}$
Chicago	355	$x_{(14)}$
Los Angeles	248	$x_{(13)}$
Philadelphia	200	$x_{(12)}$
Detroit	167	$x_{(11)}$
Baltimore	94	$x_{(10)}$
Houston	94	$x_{(9)}$
Cleveland	88	$x_{(8)}$
Washington D.C.	76	$x_{(7)}$
Saint Louis	75	$x_{(6)}$
Milwaukee	74	$x_{(5)}$
San Francisco	74	$x_{(4)}$
Boston	70	$x_{(3)}$
Dallas	68	$x_{(2)}$
New Orleans	63	$x_{(1)}$

Tabelle 4: Die 15 größten US-Städte in 1960

Arithmetisches Mittel

- Das arithmetische Mittel wird aus der Urliste durch

$$\bar{x} = \frac{1}{n}(x_1 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

berechnet. Für Häufigkeitsdaten mit Ausprägungen $a_1, \ldots, a_k$ und relativen Häufigkeiten $f_1, \ldots, f_k$ gilt

$$\bar{x} = a_1 f_1 + \ldots + a_k f_k = \sum_{j=1}^{k} a_j f_j.$$

Bei Schichtenbildung gilt

$$\bar{x} = \frac{1}{n}(n_1\bar{x}_1 + \ldots + n_r\bar{x}_r) = \frac{1}{n}\sum_{j=1}^{r} n_j\bar{x}_j.$$

- Eigenschaften: Schwerpunkteigenschaft, d.h.

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \underbrace{\sum_{i=1}^{n} x_i}_{=n\bar{x}} - n\bar{x} = 0.$$

Harmonisches Mittel:

$$\bar{x}_H = \frac{\sum_{i=1}^{n} g_i}{\sum_{i=1}^{n} \frac{g_i}{x_i}}, \quad g_i : \text{Gewicht der } i\text{-ten Beobachtung}$$

Beispiel: Ein Wanderer legte einen Weg von zwei Kilometern Länge zurück. Den ersten Kilometer ging er mit einer Geschwindigkeit von 6 km pro Stunde, den zweiten mit einer solchen von 4 km pro Stunde. Wie groß war seine Durchschnittsgeschwindigkeit?
Lösung:

- ges: Durchschnittsgeschwindigkeit
- Lös: $v = \frac{s}{t}$, v : Geschwindigkeit, s : Weg, t : Zeit
  - Durchschnittsgeschwindigkeit = Gesamtweg / Gesamtzeit, $\overline{v} = \frac{s_{gesamt}}{t_{gesamt}}$

$$\begin{aligned} s_{gesamt}\ (\hat{=} \sum g_i) &= 1\text{km} + 1\text{km} = 2\text{km} \\ t_{gesamt}\ (\hat{=} \sum \frac{g_i}{x_i}) &= \underbrace{\frac{1km}{6km \cdot h^{-1}}}_{\text{benötigte Zeit für 1.km}} + \underbrace{\frac{1km}{4km \cdot h^{-1}}}_{\text{benötigte Zeit für 2.km}} = \frac{2+3}{12}h = \frac{5}{12}h \end{aligned}$$

  - $\rightarrow \overline{v} = \frac{2km}{5/12h} = \frac{24}{5}\frac{km}{h} = 4.8\frac{km}{h}$

- Die Durchschnittsgeschwindigkeit des Wanderers beträgt $\overline{v} = 4.8$ km/h.

Geometrisches Mittel:

$$x_G = \sqrt[n]{x_1 \cdot \ldots \cdot x_n}$$

mittleres Entwicklungstempo:

$$i_G = \sqrt[n]{i_1 \cdot \ldots \cdot i_n}, \quad i_t = \frac{x_t}{x_{t-1}},\ t = 1, \ldots, n$$

Beispiel: Die Warenausfuhr des deutschen Außenhandels betrug 2002 510.008 Mrd. Euro. Sie stieg 2004 gegenüber 2002 um 25.1486% und betrug 2007 731.544 Mrd. Euro.

(a) Wie hat sich die Warenausfuhr im Zeitraum von 2002 bis 2007 im Mittel jährlich entwickelt?

Angenommen diese durchschnittliche Entwicklung der Warenausfuhr im Zeitraum 2002 bis 2007 setzt sich in den nächsten Jahren fort,

(b) wie hoch wird die Warenausfuhr im Jahre 2009 voraussichtlich sein?

(c) In welchem Jahr wird die Warenausfuhr 900 Mrd. Euro überschreiten?

Lösung: **a)**

- geg: X_t : Warenausfuhr im Jahre t, $t = 2002, 2003, \ldots$, $\mathbb{S} = \mathbb{R}$
- ges: mittlere jährliche Wachstumsrate der Warenausfuhr $\overline{x}_g$ der Jahre 2002 bis 2007
- Lös: Aufpassen: Grundlage für geometrisches Mittel sind die Wachstumsraten!!!
  - Wachstumsrate $i_t = \frac{x_t}{x_{t-1}} \rightarrow \overline{x}_g = \sqrt[n]{i_1 \cdot \ldots \cdot i_n}$
  - $i_1 \cdot \ldots \cdot i_n = \frac{x_1}{x_0} \cdot \frac{x_2}{x_1} \cdot \ldots \cdot \frac{x_n}{x_{n-1}} = \frac{x_n}{x_0}$

$$\rightarrow \overline{x}_g = \sqrt[5]{\frac{x_{2007}}{x_{2002}}} = \sqrt[5]{\frac{731.544}{510.008}} = 1.0748$$

  - Die durchschnittliche jährliche Wachstumsrate beträgt 7.48%.

b)

- Prognosewert: $x^*_{n+T} = x_n \cdot \overline{x}_g^T$
- hier:

$$\begin{aligned} x^*_{2007+2} &= 731.544 \cdot 1.0748^2 \\ &= 845.076. \end{aligned}$$

- Im Jahr 2009 wird die Warenausfuhr voraussichtlich 845.076 Mrd. Euro betragen.

c) (Bestimmung der Zeitdauer)

- geg: $x_n, x_{n+T}, \overline{x}_g$
- ges: T
- Lös:

$$\begin{aligned} x_{n+T} &= x_n \cdot \overline{x}_g^T \\ \Leftrightarrow \overline{x}_g^T &= \frac{x_{n+T}}{x_n} \\ \Leftrightarrow T \log(\overline{x}_g) &= \log\left(\frac{x_{n+T}}{x_n}\right) \\ \Leftrightarrow T \log(\overline{x}_g) &= \log(x_{n+T}) - \log(x_n) \quad | : \log(\overline{x}_g) \\ \Leftrightarrow T &= \frac{\log(x_{n+T}) - \log(x_n)}{\log(\overline{x}_g)} \end{aligned}$$

- Einsetzen:

$$\begin{aligned} x_{2007+T} &= 900.000 \\ x_{2007} &= 731.544 \\ \overline{x}_g &= 1.0748 \\ \to T &= \frac{\log(900.000) - \log(731.544)}{\log(1.0748)} \\ &= 2.8729 \\ &\approx 3. \end{aligned}$$

- 2010 werden 900 Mrd. Euro voraussichtlich überschritten.

1.2.3 Streuungsparameter

- Lageparameter geben eine zentrale Tendenz in den Daten an. Allerdings ist diese Information zur Charakterisierung der Daten meistens nicht ausreichend.
- Beispiel: Betrachte zwei Gruppen von jeweils 5 Daten:

$$\begin{array}{ll} \text{Gruppe 1:} & 17, 18, 20, 22, 23 \\ \text{Gruppe 2:} & 5, 10, 20, 30, 35 \end{array}$$

  Für beiden Gruppen gilt $\bar{x}_1 = \bar{x}_2 = 20$. Aber: Die Daten besitzen eine unterschiedliche Streuung.

- Daher: Betrachtung von Streuungsmaßen zur Messung der Variabilität in den Daten.
- **Spannweite:**

$$SP = x_{(n)} - x_{(1)} = x_{max} - x_{min}$$

- **Quantilsabstand:**

$$x_{1-p} - x_p$$

- **Interquartilsabstand:**

$$d_Q = x_{0.75} - x_{0.25}$$

- Idee: Betrachte mittlere Abweichung (d.h. Abweichung für alle Beobachtungen $x_1, \ldots, x_n$) von einem (zentralen) Lageparameter, z.B. dem arithmetischen Mittel.
- Mögliche Ansätze: *mittlere absolute Abweichung* (ADev = mean absolute deviation)

$$ADev := \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|.$$

  und *(empirische) Varianz*

$$\tilde{s}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

  bzw *Stichprobenvarianz*

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

  Die *Standardabweichung* ist die Wurzel aus der Varianz

$$\tilde{s} = +\sqrt{\tilde{s}}.$$

- Problem der Varianz: große Abweichungen werden durch die Quadrierung überproportional stark bewertet.
- Häufige Fragestellung in der Statistik: Bestimme einen unbekannten (zentralen) Lageparameter, so dass die Streuung minimal wird.
- Problem der ADev: Betragsfunktion ist numerisch ungünstig, insb. im Hinblick auf Optimierungsprobleme.
- Vorteil der ADev: stärkere Robustheit gegenüber Ausreißern im Vergleich zur Varianz.
- Vorteil der Varianz: quadratische Funktion → erste Ableitung ist lineare Funktion.

Satz 1.3. *Die empirische Varianz besitzt folgende Eigenschaften*

(i) Verschiebungssatz: Es gilt

$$\tilde{s}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

(ii) Transformationsregel: Für $y_i = ax_i + b$ gilt

$$\tilde{s}_y^2 = a^2\tilde{s}_x^2 \quad bzw. \quad \tilde{s}_y = |a|\tilde{s}_x$$

(iii) Streuungszerlegung: Für r disjunkte statistische Massen $E_1, \ldots, E_r$, deren jeweilige arithmetische Mittel bzw. mittlere quadratische Abweichungen mit $\bar{x}_1, \ldots, \bar{x}_r$ bzw. $\tilde{s}_1^2, \ldots, \tilde{s}_r^2$ bezeichnet sind, berechnet sich die mittlere quadratische Abweichung für die Gesamtmasse folgendermaßen:

$$\tilde{s}_{Ges}^2 = \frac{1}{n}\sum_{j=1}^{r} n_j\tilde{s}_j^2 + \frac{1}{n}\sum_{j=1}^{r} n_j(\bar{x}_j - \bar{x}_{Ges})^2$$

wobei $n_j = |E_j|$ und $\bar{x}_{Ges} = \frac{1}{n}\sum_{j=1}^{r} n_j\bar{x}_j$.

Beweis. (i) Es gilt

$$\begin{aligned}\tilde{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - 2\bar{x}\underbrace{\sum_{i=1}^{n} x_i}_{=n\bar{x}} + n\bar{x}^2\right)\\ &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\bar{x}^2 + \bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2.\end{aligned}$$

(ii) Für $y = ax + b$ erhalten wir

$$\begin{aligned}\tilde{s}_y^2 = Var(ax+b) &= \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(ax_i + b - (a\bar{x}+b))^2\\ &= \frac{1}{n}\sum_{i=1}^{n}(ax_i - a\bar{x})^2 = a^2\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = a^2\tilde{s}_x^2.\end{aligned}$$

(iii) Bei einer Erhebung in r Schichten erhalten wir

$$\begin{array}{ll} E_1: & x_{11}, \ldots, x_{1n_1} \\ E_2: & x_{21}, \ldots, x_{2n_2} \\ \vdots & \vdots \\ E_r: & x_{r1}, \ldots, x_{rn_r} \end{array}$$

mit $\bar{x}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ji}$ und $\tilde{s}_j^2 = \frac{1}{n_j} \sum_{i=1}^{n_j} (x_{ji} - \bar{x}_j)^2$.

Das (gewichtete) arithmetische Mittel über alle r Schichten lautet $\bar{x} = \frac{1}{n} \sum_{j=1}^{r} n_j \bar{x}_j$. Damit erhalten wir für die Gesamtvarianz

$$\begin{aligned} \tilde{s}^2 &= \frac{1}{n} \sum_{j=1}^{r} \sum_{i=1}^{n_j} (x_{ji} - \bar{x})^2 \\ &= \frac{1}{n} \sum_{j=1}^{r} \sum_{i=1}^{n_j} (x_{ji} - \bar{x}_j + \bar{x}_j - \bar{x})^2 \\ &= \frac{1}{n} \sum_{j=1}^{r} \sum_{i=1}^{n_j} \left((x_{ji} - \bar{x}_j)^2 + 2(x_{ji} - \bar{x}_j)(\bar{x}_j - \bar{x}) + (\bar{x}_j - \bar{x})^2 \right) \\ &= \frac{1}{n} \left(\sum_{j=1}^{r} n_j \tilde{s}_j^2 + \sum_{j=1}^{r} n_j (\bar{x}_j - \bar{x})^2 + 2 \sum_{j=1}^{r} (\bar{x}_j - \bar{x}) \underbrace{\left(\sum_{i=1}^{n_j} (x_{ji} - \bar{x}_j) \right)}_{=0} \right) \\ &= \underbrace{\frac{1}{n} \sum_{j=1}^{r} n_j \tilde{s}_j^2}_{\text{Streuung (Varianz) innerhalb der Schichten}} + \underbrace{\frac{1}{n} \sum_{j=1}^{r} n_j (\bar{x}_j - \bar{x})^2}_{\text{Streuung (Varianz) zwischen den Schichten}} \end{aligned}$$

□

1.3 Multivariate deskriptive Statistik

In vielen Anwendungen werden mehrere interessierende Merkmale gleichzeitig erhoben. Neben der Verteilung der einzelnen Merkmale interessiert

- die gemeinsame Verteilung
- die Stärke des Zusammenhangs zwischen den einzelnen Variablen
- die Richtung des Zusammenhangs (falls sinnvoll interpretierbar)

1.3.1 Zweidimensionale Häufigkeiten

- geg: (diskrete) Merkmale $X \in \{\underbrace{a_1, \ldots, a_k}_{\mathbb{S}_X}\}, Y \in \{\underbrace{b_1, \ldots, b_\ell}_{\mathbb{S}_Y}\}$
- Sei $(x_i, y_j) := \{(X = a_i) \cap (Y = b_j)\}$, $i = 1, \ldots, k, j = 1, \ldots, \ell$

Definition 1.4. *Die absolute Häufigkeit h_{ij} gibt an, wie oft das Ausprägungstupel (x_i, y_j) in der Stichprobe auftritt. Sie ist definiert als*

$$h_{ij} := h(x_i, y_j) = \sum_{r=1}^{n} 1_{\{a_i\}}(x_r) 1_{\{b_j\}}(y_r), \; i = 1, \ldots, k, j = 1, \ldots, \ell$$

- Eigenschaften:
  - $h_{ij} \in \{0, \dots, n\}$
  - $\sum_{i=1}^{k} \sum_{j=1}^{\ell} h_{ij} = n$

  Definition 1.5. *Die relative Häufigkeit f_{ij} gibt den Anteil des Ausprägungstupels (x_i, y_j) an allen Beobachtungen an. Sie ist definiert als*

$$f_{ij} = f(x_i, y_j) = \frac{1}{n} h_{ij}.$$

- Eigenschaften:

  (i) $0 \leq f_{ij} \leq 1$,

  (ii) $\sum_{i=1}^{k} \sum_{j=1}^{\ell} f_{ij} = 1$.

- gemeinsame Verteilung:
  - Die gemeinsame absolute Häufigkeitsverteilung ist die Abbildung

$$h_{XY} : \mathbb{S}_X \times \mathbb{S}_Y \to \mathbb{N}_0.$$

  - Die gemeinsame relative Häufigkeitsverteilung ist die Abbildung

$$f_{XY} : \mathbb{S}_X \times \mathbb{S}_Y \to \mathbb{Q} \quad (\mathbb{R}).$$

- Randhäufigkeiten (auch Randverteilungen (RV))
  - Ziel: Beschreibung der Häufigkeitsverteilung eines der beiden Merkmale (ohne Berücksichtigung der konkreten Ausprägungen des anderen Merkmals)
  - Ansatz: Zusammenfassung der Ausprägungen des anderen Merkmals.
  - Notation:

$$f_{i\bullet} := \sum_{j=1}^{\ell} f_{ij}, \quad h_{i\bullet} := \sum_{j=1}^{\ell} h_{ij}, \quad f_{\bullet j} := \sum_{i=1}^{k} f_{ij}, \quad h_{\bullet j} := \sum_{i=1}^{k} h_{ij}$$

- Zusammenfassende Darstellung der (zweidimensionalen) gemeinsamen Häufigkeitsverteilung und der Randverteilungen in einer (zweidimensionalen) Kontingenztabelle:

	Merkmal Y					
Merkmal X	b_1	$\dots$	b_j	$\dots$	b_ℓ	$RV(X)$
a_1	f_{11}	$\dots$	f_{1j}	$\dots$	$f_{1\ell}$	$f_{1\bullet}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
a_i	f_{i1}	$\dots$	f_{ij}	$\dots$	$f_{i\ell}$	$f_{i\bullet}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
a_k	f_{k1}	$\dots$	f_{kj}	$\dots$	$f_{k\ell}$	$f_{k\bullet}$
$RV(Y)$	$f_{\bullet 1}$	$\dots$	$f_{\bullet j}$	$\dots$	$f_{\bullet \ell}$	$f_{\bullet\bullet} = 1$

- Problem: Bei metrischen Merkmalen sind k, ℓ sehr groß (überabzählbar viele „Kategorien“). Die Darstellung der gemeinsamen Häufigkeitsverteilung mit Hilfe einer Kontingenztabelle ist dann nicht mehr sinnvoll. Eine Alternative ist die sogenannte Korrelationstabelle:

i	Merkmal X	Merkmal Y
1	x_1	y_1
$\vdots$	$\vdots$	$\vdots$
i	x_i	y_i
$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n

- Graphische Darstellungen für zweidimensionale Häufigkeiten:
  - gruppiertes oder zweidimensionales Balkendiagramm (bei zwei kategorialen Merkmalen),
  - zweidimensionales Histogramm (bei zwei mind. ordinalskalierten Merkmalen),
  - zweidimensionaler Kerndichteschätzer (bei zwei metrischen Merkmalen),
  - Scatterplot (Streudiagramm) (bei zwei metrischen Merkmalen),
  - gruppierte Boxplots (ein metrisches und ein kategoriales Merkmal).
- Bedingte Häufigkeiten
  - Ziel: Beschreibung der Häufigkeitsverteilung eines Merkmals bei Vorliegen einer konkreten Ausprägung des anderen Merkmals.
    Beispiel: Verteilung von X für $Y = b_j$ bzw. kurz $X|Y = b_j$. Man betrachtet dann nur die j-te Spalte der Kontingenztabelle.
  - Problem: Es gilt
    $$\sum_{i=1}^{k} f_{ij} = f_{\bullet j}, \quad 0 \leq f_{\bullet j} \leq 1, \quad 0 \leq f_{ij} \leq f_{\bullet j},$$
    d.h. der Wertebereich der relativen Häufigkeiten in der j-ten Spalte ist abhängig von $f_{\bullet j}$. Die Idee besteht darin, ein Maß $\nu_{i|j}$ zu finden, für dass $\sum_{i=1}^{k} \nu_{i|j} = 1$ statt $\sum_{i=1}^{k} f_{ij} = f_{\bullet j}$ gilt. Ansatz: Adjustierung auf Teilgesamtheit $h_{\bullet j}$ statt auf Stichprobenumfang n.

    Definition 1.6. *Die bedingte relative Häufigkeitsverteilung von X unter der Bedingung $Y = b_j$, kurz $X|Y = b_j$ ist definiert durch*
    $$f_{X|Y}(x_i|y_j) = f_{i|j} = \frac{h_{ij}}{h_{\bullet j}} = \frac{f_{ij}}{f_{\bullet j}}, \quad i = 1, \ldots, k.$$
    Äquivalent gilt für $Y|X = a_i$:
    $$f_{Y|X}(y_j|x_i) = f_{j|i} = \frac{h_{ij}}{h_{i\bullet}} = \frac{f_{ij}}{f_{i\bullet}}, \quad j = 1, \ldots, \ell.$$
  - Für zwei diskrete Merkmale kann die bedingte Häufigkeitsverteilung (hier $X|Y$) in einer Kontingenztabelle darstellt werden:

	Merkmal Y					
Merkmal X	b_1	$\ldots$	b_j	$\ldots$	b_ℓ	
a_1	$f_{1\|1}$	$\cdots$	$f_{1\|j}$	$\cdots$	$f_{1\|\ell}$	
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	
a_i	$f_{i\|1}$	$\cdots$	$f_{i\|j}$	$\cdots$	$f_{i\|\ell}$	
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	
a_k	$f_{k\|1}$	$\cdots$	$f_{k\|j}$	$\cdots$	$f_{k\|\ell}$	
$\sum_{i=1}^{k} f_{i\|j}$	1	$\ldots$	1	$\ldots$	1	

  - Eigenschaften:

(i) $0 \leq f_{i|j} \leq 1 \quad$ bzw. $\quad 0 \leq f_{j|i} \leq 1$,
(ii) $\sum_{i=1}^{k} f_{i|j} = 1 \quad$ bzw. $\quad \sum_{j=1}^{\ell} f_{j|1} = 1$,
(iii) Im allgemeinen gilt $f_{i|j} \neq f_{j|i}$. Gleichheit gilt genau dann, wenn $f_{\bullet j} = f_{i\bullet}$.

– Folgerung: Die bedingten Häufigkeiten entsprechen eindimensionalen relativen Häufigkeiten. Dies ermöglicht beispielsweise eine äquivalente Anwendung graphischer Darstellungen oder Lageparameter für univariate Häufigkeiten.

- Das Konzept der Unabhängigkeit
  - Idee: Zwei Merkmale sind unabhängig, wenn die (beobachtete) Ausprägung des einen Merkmals keinen Einfluss auf die (beobachtete) Ausprägung des anderen Merkmals besitzt.
  - Folgerung: Dann sind alle bedingten Häufigkeiten (gegeben das andere Merkmal) gleich den Randhäufigkeiten: $f_{i|j} = f_{i\bullet}$.
  - Abhängige Merkmale: O.B.d.A. Kenntnis von Y beeinflusst Häufigkeit von X. Bei perfekter Abhängigkeit ist in jeder Spalte nur eine Zelle besetzt, d.h. es gilt $f_{ij} = f_{\bullet j}$.
  - Unabhängige Merkmale: Kenntnis der konkreten Ausprägung von Y ändert nichts an der Häufigkeitsverteilung von X, d.h. $f_{i|1} = f_{i|2} = \ldots = f_{i|\ell} = f_{i\bullet}$. $X = a_i$ hat stets dieselbe Häufigkeit, unabhängig davon, welche Ausprägung Y annimmt. Für unabhängige Merkmale gilt
$$f_{i|j} = \frac{f_{ij}}{f_{\bullet j}}, \quad f_{i|j} = f_{i\bullet}$$
und damit $f_{ij} = f_{i\bullet} f_{\bullet j}$.

1.3.2 Zusammenhangsmaße für nominalskalierte Merkmale

Die Anordnung der Merkmalsausprägungen ist bei nominalskalierten Merkmalen willkürlich. Daher kann bei ihnen lediglich die Stärke des Zusammenhanges untersucht werden, nicht jedoch die Richtung (Assoziation).

Pearsons χ^2-Statistik

- Idee: Vergleich der beobachteten Zellhäufigkeit h_{ij} einer $(k \times \ell)$-dimensionalen Kontingenztabelle mit denen unter Unabhängigkeit zu erwartenden Zellhäufigkeiten
$$e_{ij} := \frac{h_{i\bullet} h_{\bullet j}}{n} = n f_{i\bullet} f_{\bullet j}.$$
Die intendierten Maßeigenschaften lauten:
  - großer Zusammenhang zwischen X und Y (Dies entspricht einer starken Diskrepanz zwischen h_{ij} und e_{ij}) $\Longrightarrow$ hoher Wert,
  - geringer Zusammenhang zwischen X und Y (Dies entspricht einer kleinen Diskrepanz zwischen h_{ij} und e_{ij}) $\Longrightarrow$ niedriger Wert
- Ansatz:
  - Diskrepanzabbildung durch Differenz,
  - Vermeidung negativer Werte durch Quadrieren,

  Ergebnis: χ^2-Koeffizient.

  Definition 1.7. *Der χ^2-Koeffizient ist definiert durch*
$$\chi^2 := \sum_{i=1}^{k} \sum_{j=1}^{\ell} \frac{(h_{ij} - e_{ij})^2}{e_{ij}} = n \left(\sum_{i=1}^{k} \sum_{j=1}^{\ell} \frac{h_{ij}^2}{h_{i\bullet} h_{\bullet j}} - 1 \right)$$

Satz 1.8. *Für den Wertebereich des χ^2-Koeffizienten gilt*

$$\chi^2 \in [0, \underbrace{n(\min\{k,\ell\}-1)}_{\geq 1}].$$

Beweis. Da aufgrund ihrer Definition für die absolute Häufigkeit ein Wertebereich von $0 \leq h_{ij} \leq n$ gegeben ist und dieser Wertebereich definitionsbedingt auch für die Randhäufigkeiten $h_{i\bullet}$ und $h_{\bullet j}$ gilt, ist χ^2 stets nicht-negativ.

Für den Wertebereich von χ^2 überlegt man sich folgendes: Der χ^2-Koeffizient ist ein Zusammenhangsmaß, d.h. er besitzt als Extremfälle die (vollständige) Unabhängigkeit und die (vollständige) Abhängigkeit zweier Merkmale. Unter Unabhängigkeit gilt für alle $k \cdot \ell$ Kombinationen von Merkmalsausprägungen

$$h_{ij} = \frac{h_{i\bullet}h_{\bullet j}}{n} = e_{ij}, \quad i = 1,\ldots,k,\ j = 1,\ldots,\ell.$$

Der χ^2-Koeffizient wäre in diesem Fall gleich null.

Für den Fall der vollständigen Abhängigkeit überlegt man sich folgendes. Sei o.B.d.A. $k \leq \ell$. Dann gilt: Für jedes $i \in \{1,\ldots,k\}$ existiert ein $j_i \in \{1,\ldots,\ell\}$, so dass

$$h_{ij_i} = h_{i\bullet} = h_{\bullet j_i} \quad \text{und} \quad h_{it} = 0\ \forall t \neq j_i.$$

Weiterhin gilt, dass $j_r \neq j_s$ für alle $r \neq s$.

Damit ergibt sich, dass für jedes $i \in \{1,\ldots,k\}$

$$\sum_{j=1}^{\ell} \frac{h_{ij}^2}{h_{i\bullet}h_{\bullet j}} = \frac{h_{ij_i}^2}{h_{i\bullet}h_{\bullet j_i}} = 1$$

und damit

$$\sum_{i=1}^{k}\sum_{j=1}^{\ell} \frac{h_{ij}^2}{h_{i\bullet}h_{\bullet j}} = k.$$

Für χ^2 gilt damit

$$\begin{aligned}\chi^2 &= n\left(\sum_{i=1}^{k}\sum_{j=1}^{\ell}\frac{h_{ij}^2}{h_{i\bullet}h_{\bullet j}} - 1\right)\\ &= n(k-1).\end{aligned}$$

□

- Problem: Der Wertebereich des χ^2-Koeffizienten hängt von der Dimension der Kontingenztabelle und dem Stichprobenumfang ab. Eine Alternative ist der Kontingenzkoeffizient:

  Definition 1.9. *Der Kontingenzkoeffizient K ist definiert als*

  $$K := \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

- Folgerung: Für den Wertebereich gilt $0 \leq K \leq K_{\max}$, mit $K_{\max} := \sqrt{\frac{M-1}{M}}$ und $M := \min\{k,\ell\}$.

  Definition 1.10. *Der korrigierte Kontingenzkoeffizient K^* ist definiert als*

  $$K^* := \frac{K}{K_{\max}}.$$

- Folgerung: Es gilt

$$0 \leq K^* \leq \frac{\sqrt{\frac{M-1}{M}}}{\sqrt{\frac{M-1}{M}}} = 1.$$

- Eigenschaften der Maße χ^2, K und K^*:

  (i) Sie messen lediglich die Stärke des Zusammenhanges, nicht die Richtung.

  (ii) Mißt man den Zusammenhang von X und Y über zwei Teilpopulationen, so lässt sich damit die Stärke des Zusammenhanges für die Teilpopulationen vergleichen.

  (iii) Abhängigkeit vom Stichprobenumfang.

  (iv) Alle Maße benutzen nur das Nominalskalenniveau von X und Y. Sie sind daher invariant gegenüber der Vertauschung von Zeilen bzw. Spalten.

- Das Konzept der Chancen bzw. relativen Chancen

  – Idee: Betrachte das Verhältnis der relativen Häufigkeiten zwischen Teilpopulationen. Interpretation als relative Chance (oder relatives Risiko, in Abhängigkeit vom Sachverhalt).

  – Ausgangspunkt: bedingte Häufigkeiten

  Definition 1.11. *Unter einer Chance („odds“) versteht man o.B.d.A. das Verhältnis zwischen dem Auftreten von $Y = b_{j_1}$ und $Y = b_{j_2}$ in einer Teilpopulation $X = a_i$. Die (empirische) bedingte Chance für festes $X = a_i$ ist bestimmt durch*

$$\gamma(b_{j_1}, b_{j_2}|a_i) = \frac{h_{ij_1}}{h_{ij_2}}.$$

  Das Verhältnis zwischen den Chancen zweier Teilpopulationen $X = a_{i_1}$ und $X = a_{i_2}$ wird als relative Chance („odds ratio“) bezeichnet

$$\gamma(j_1, j_2|i_1, i_2) = \gamma_{Y|X}(b_{j_1}, b_{j_2}|a_{i_1}, a_{i_2}) = \frac{\frac{h_{i_1j_1}}{h_{i_1j_2}}}{\frac{h_{i_2j_1}}{h_{i_2j_2}}} = \frac{h_{i_1j_1}h_{i_2j_2}}{h_{i_1j_2}h_{i_2j_1}}$$

  – Eigenschaften: Es gilt $\gamma \in (0, \infty)$. Für unabhängige Merkmale gilt $\gamma = 1$. Für abhängige Merkmale gilt

    * $\gamma > 1$ Chancen in Population $X = a_{i_1}$ besser als in Population $X = a_{i_2}$
    * $\gamma < 1$ Chancen in Population $X = a_{i_1}$ schlechter als in Population $X = a_{i_2}$
    * $\gamma = 1$ Chancen in beiden Populationen gleich

  – Bemerkung: odds und odds ratio sind alternative „Wahrscheinlichkeitsmaße“ („Schätzer“ für das Eintreten eines Ereignisses)

1.3.3 Zusammenhangsmaße für ordinalskalierte Merkmale

- Eigenschaft ordinalskalierter Merkmale: (natürliche) Rangfolge der möglichen Merkmalsausprägungen

- Idee: Betrachte für jede Beobachtung der Merkmale (X, Y) die Rangfolge für jede Komponente

- Sei $rg(X_i)$ der Rang der X-Komponente bei der i-ten beobachteten statistischen Einheit (Das entspricht sozusagen der Platzzahl, die der Wert bei größenmäßiger Anordnung aller Werte erhält.), äquivalent $rg(Y_i)$.

- Bindungen (Ties): Haben zwei oder mehrere Beobachtungen die gleiche Ausprägung des Merkmals X oder Y, so liegt eine sog. Bindung vor. Als Rang der einzelnen Beobachtungen wird dann der Mittelwert der zu vergebenden Ränge genommen.

  Definition 1.12. *Der Rangkorrelationskoeffizient von Spearman lautet*

$$r_{SP} = \frac{\sum_{i=1}^{n}(rg(x_i) - \bar{rg}_X)(rg(y_i) - \bar{rg}_Y)}{\sqrt{\sum_{i=1}^{n}(rg(x_i) - \bar{rg}_X)^2 \sum_{i=1}^{n}(rg(y_i) - \bar{rg}_Y)^2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(rg(x_i)rg(y_i)) - \bar{rg}_X\bar{rg}_y}{s_{rg_X}\, s_{rg_Y}},$$

  wobei die Mittelwerte der Ränge gegeben sind durch

$$\begin{aligned} \bar{rg}_X &= \frac{1}{n}\sum_{i=1}^{n} rg(x_i) = \frac{1}{n}\sum_{i=1}^{n} i = \frac{n+1}{2}, \\ \bar{rg}_Y &= \frac{1}{n}\sum_{i=1}^{n} rg(y_i) = \frac{1}{n}\sum_{i=1}^{n} i = \frac{n+1}{2}. \end{aligned}$$

  Bei Abwesenheit von Bindungen ergibt sich die einfachere Formel

$$r_{SP} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

  mit $d_i = rg(x_i) - rg(y_i)$.

- Eigenschaften:

  (i) $-1 \le r_{SP} \le 1$,

  (ii) $r_{SP} > 0 \Longrightarrow$ gleichsinniger monotoner Zusammenhang (X groß, wenn Y groß, X klein, wenn Y klein),

  (iii) $r_{SP} < 0 \Longrightarrow$ gegensinniger monotoner Zusammenhang (X groß, wenn Y klein, X klein, wenn Y groß),

  (iv) $r_{SP} \approx 0 \Longrightarrow$ kein monotoner Zusammenhang.

- Bemerkung: Der Rangkorrelationskoeffizient entspricht dem Bravais-Pearson-Korrelationskoeffizienten für die Rangzahlen $rg(X_i)$ und $rg(Y_i)$.

1.3.4 Zusammenhangsmaße für kardinalskalierte (metrische) Merkmale

- Ansatz: Betrachte ein Maß für die gemeinsame Streuung zweier kardinalskalierter Merkmale.

  Definition 1.13. *Die (empirische) Kovarianz als Maß für die gemeinsame Streuung zweier metrischer Merkmale X und Y ist definiert als*

$$\tilde{s}_{XY} := Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}.$$

- Eigenschaften:

  (i) $\tilde{s}_{XY} \in [0, \infty)$, d.h. es handelt sich um eine nicht-normierte Charakteristik,

  (ii) $\tilde{s}_{XX} = \tilde{s}_X^2$,

  (iii) lineare Transformationen: $V = aX + b$ und $W = cY + d$ führen zu

$$Cov(V,W) = a \cdot c \cdot Cov(X,Y).$$

(iv) $\tilde{s}_{XY} = \tilde{s}_{YX}$ (symmetrisch)

- Problem: Die Kovarianz ist nicht normiert und nicht dimensionslos. Die Stärke des Zusammenhanges kann nur zwischen Stichproben vergleichend angegeben werden.

Definition 1.14. *Der Bravais-Pearson-Korrelationskoeffizient ist definiert als*

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}$$

- Eigenschaften:

(i) $-1 \leq r_{XY} \leq 1$

(ii) Interpretation:
- $r_{XY} = -1$: exakter negativer linearer Zusammenhang (Gerade mit negativer Steigung)
- $r_{XY} \in (-1, 0)$: negativer linearer Zusammenhang
- $r_{XY} = 0$: kein linearer Zusammenhang (entspricht z.B. einer kreisähnlichen Punktwolke im Streudiagramm)
- $r_{XY} \in (0, 1)$: positiver linearer Zusammenhang
- $r_{XY} = 1$: exakter positiver linearer Zusammenhang (Gerade mit positiver Steigung)

(iii) r_{XY} ist dimensionslos und ein Maß für den linearen Zusammenhang

(iv) $r_{XY} = r_{YX}$ (symmetrisch),

(v) Translationsäquivarianz: $\tilde{X} = aX + b, \tilde{Y} = cY + d$ dann ist $r_{\tilde{X}\tilde{Y}} = r_{XY}$.