**Statistik I Zusammenfassung**

**Deskriptive Statistik**

* **Grundbegriffe:**

**Statistische Erhebung – Grundsätzliche Überlegungen**

**Schritt 1:** wer soll befragt werden? (Festlegung der Gesamtheit der Untersuchungsobjekte)

**Schritt 2:** Welche Fragen sollen gestellt werden? (Festlegung der Eigenschaften der stat. Einheiten)

**Schritt 3:** Welche Antworten sind vorgesehen? (in welcher Form die Eigenschaften festgestellt werden sollen)

**Schritt 4:** Welche Antworten bzw. Ergebnisse werden festgehalten? (Feststellung der Merkmalsausprägungen der Untersuchungsobjekte)

**Statistische Einheit**: Einzelobjekt einer stat. Untersuchung

**Statistische Masse (Grundgesamtheit, Population):** Zusammenfassung der stat. Einheiten

**Abgrenzung und Identifikation der stat. Masse:** ob ein Objekt und in welcher Form es in der stat. Masse vorkommt.

**Kriterien zur Abgrenzung:** sachlich, räumlich, zeitlich

**Bestandsmasse:** stat. Masse, deren zeitliche Abgrenzung ein Zeitpunkt ist (Bevölkerung, Lagerbestand, Kassenbestand usw.)

**Bestandseinheiten:** ihnen ist ein Zeitraum, die Verweildauer, zugeordnet

**Ereignismasse:** stat. Masse, deren zeitliche Abgrenzung ein Zeitraum ist (Geburten, Todesfälle usw.)

**Ereigniseinheiten:** ihnen ist ein Zeitpunkt, das Eintreten des Ereignisses (Beginn, Ende) zugeordnet

* **Fortschreibungsformel:**

**Merkmal:** Beschreibungsmöglichkeit für stat. Einheiten (Merkmalsträger) wie Alter, Haarfarbe etc.

* **nicht häufbar:** eindeutige Zuordnung der Merkmalsausprägungen zu den Merkmalsträgern
* **häufbar:** Mehrdeutigkeit bei der Zuordnung, d.h. eine stat. Einheit kann mehrere Merkmalsausprägungen haben

**Merkmalsausprägung:** spezielle Eigenschaft eines Merkmals, z.B. „rot“ für das Merkmal Augenfarbe

**Liste der Merkmalsausprägungen:** Zusammenfassung aller bei der Untersuchung möglich auftretenden Merkmalsausprägungen; keine Wiederholungen

**Merkmalswert**: bei der Untersuchung beobachtete Merkmalsausprägung

**Liste der Merkmalswerte:** Auflistung aller Merkmalswerte; Wiederholungen (Häufungen) sind möglich

* **Merkmalsarten**

Grundtypen der Menge der Merkmalsausprägungen:

1. **Qualitative Merkmale**: Merkmalsausprägungen sind verbale Begriffe, die ohne eine ordnende Vergleichbarkeit nebeneinander stehen, Reihenfolge willkürlich, keine Einstufung möglich (Haarfarbe, Geschlecht usw.)
2. **Rangmerkmale**: Einteilung der stat. Masse in „Qualitätsstufen“, Vergleich zwischen 2 Merkmalen möglich, allerdings ist der Unterschied nicht qualifizierbar (Zuordnung von Weinen in „Tafelwein“, „Qualitätswein“ usw.)
3. **Quantitative Merkmale:** Merkmalsausprägungen reelle Zahlen, Feststellung des Merkmalswertes durch Abzählen etc., für den Vergleich stehen die gesamten reellen Zahlen zur Verfügung (Alter, Gewicht usw.)
	* **stetige Merkmale:** Merkmalswerte alle reelle Zahlen bzw. ein Intervall (Größe, Gewicht, Lebensdauer, Temperatur)
	* **diskrete Merkmale:** isolierte Zahlen (i.d.R. ganze Zahlen) → Kinderzahl, Semester zahl usw.
* **Skalen**

**Skalierung (Codierung):** für die Verarbeitung mit EDV-Anlagen werden den qualitativen und Rangmerkmalen reelle zahlen zugeordnet, heißt aber nicht, dass man darauf die Formeln der quantitativen Merkmalen verwenden darf.

* **Nominalskalen:** qualitative Merkmale
* **Ordinal-** **bzw. Rangskalen:** Rangmerkmale
* **Metrische bzw.** **Kardinalskalen:** quantitative Merkmale
	+ **Intervallskalen:** Nullpunkt willkürlich festgelegt, Vergleich durch Differenzen, aber keine Verhältnisse (Temperatur)
	+ **Verhältnisskalen**: Nullpunkt auf natürlich Weise oder durch Konvention vorgegeben, Vergleich durch Differenzen und Verhältnisse (Alter, Gewicht, Einkommen)
	+ **Absolutskalen:** neben Nullpunkt auch die Einheit natürlich vorgegeben (Stückzahl)
* **Transformation bei Skalen (Dimensionsänderungen)**

**Intervallskalen:**

**Beispiel:** Umrechnung von Fahrenheit (x-Wert) in Celsius (y-Wert) nach der Formel:

**Verhältnisskalen:**

Eine Verschiebung des Nullpunktes ist nicht möglich, da dieser fest vorgegeben ist.

**Beispiel:** α = 100 bei Umrechnung von m in cm

**Absolutskalen:** keine Transformation zulässig, da Nullpunkt und Einheit fest vorgegeben sind.

* **Häufigkeitsverteilungen**
* **Darstellung**

**- Erhebungsergebnis in allgemeiner Form**: für jede Einheit kann der zugehörige Merkmalswert abgelesen werden, n die Anzahl der stat. Einheiten

**- Urliste:**

Unterscheidung zwischen Urliste und Merkmalsausprägung

- **geordnete Urliste:**

der Ordnung der reellen Zahlen umsortiert, bei Rangmerkmal kann die Liste geordnet werden, wobei man dann entsprechend der natürlichen Reihenfolge der Merkmalsausprägungen vorgeht.

* **Stiel und Blatt-Darstellung**

****links stehen Zehnerziffern, rechts die Einerziffern der Merkmalswerte der Urliste, für bessere Sortierung **→** bei qualitativen Merkmalen ist eine Ordnung nicht möglich, aber man kann so umsortieren, dass identische Merkmalswerte benachbart sind.

* **Absolute Häufigkeit**

a ist die **Anzahl der Merkmalswerte der Urliste**, die mit a übereinstimmen (wie oft ein Merkmalswert vorkommt)

- **absolute Häufigkeitsverteilung:** Zusammenstellung der abs. Häufigkeiten h(a) für alle a M, also eine Abbl.

* **Relative Häufigkeit**

Anteil der abs. Häufigkeit einer Merkmalsausprägung an der Gesamtzahl der aufgeführten Merkmalswerte:

 bzw.

- **relative Häufigkeitsverteilung:** Zusammenstellung aller rel. Häufigkeiten p(a), a M, also eine Abbl.

* **Klassierung von Merkmalsausprägungen**

Zusammenfassung von Merkmalsausprägungen zu Klassen

**- Klassen:** Intervalle, Einteilung durch Angabe von unteren und oberen Klassengrenzen ([]) oder (( )), bei offenen nur eine Klassengrenze

- **Klassenbreite:** Differenz aus oberer und unterer Klassengrenze

- **Klassenbreite:** arithm. Mittel aus unterer und oberer Klassengrenze, Repräsentant der Klasse

- **Anzahl der Klassen:** Relation zwischen der Anzahl von Beobachtungswerten und der Anzahl von Klassen zu berücksichtigen, je größer die Klasse, desto mehr Information geht verloren

* **Summenhäufigkeiten**

- bei **Rangmerkmalen:** Liste der Merkmalsausprägungen in ihrer natürlichen Reihenfolge

- bei **Quantitativen Merkmalen:** ≤ - Beziehung

**- Absolute Summenhäufigkeit**

**- Relative Summenhäufigkeiten**

**- Summenhäufigkeit bei klassierten Daten**

1: obere Klassengrenze der Klasse I

2: Klassengrenze

* **Empirische Verteilungsfunktion**

bei einem quantitativen Merkmal kann für jede reelle Zahl x der Anteil aller Merkmalswerte festgestellt werden, die diese Merkmalsausprägung nicht überschreiten.

Funktion:

**Eigenschaften der Funktion:**

* von links nach rechts ansteigend (isoton)
* da zwischen 2 benachbarten Werten keine Merkmalswerte beobachtet wurden, ist die Funktion in diesem Bereich konstant, man erhält eine „Treppenfkt.“
* Höhe der Treppenstufen ist i.d.R. (stetige Merkmale), bei aufeinanderfallenden Merkmalswerten ist sie ein entsprechendes Vielfaches von
* an der Sprungstelle nimmt die Fkt. den höheren Wert an (rechtsseitig stetig)
* Minimalwert ist 0, Maximalwert ist 1
* kein Informationsverlust
* **Graphische Darstellungsmöglichkeiten**

- **Ziele**: größere Übersichtlichkeit, Einprägsamkeit und Attraktivität

* **Eindimensional**: Stab-, Linien- oder Säulendiagramm (Balkendiagramm) → auf der Abzisse werden die Merkmalsausprägungen aufgetragen, die Höhe spiegelt die abs. bzw. rel. Häufigkeiten wider (höhenproportional)
* **Zweidimensional**: durch Flächendiagramme (flächenproportional)
* **Dreidimensional**: volumenproportionale Darstellung → durch Volumina entsprechender Größe

- **Kreissektordiagramm:** Graph. Darstellung rel. Häufigkeiten durch sektorale Aufteilung einer Kreisfläche, die Größe der Sektoren proportional zur den rel. Häufigkeiten (Radius des Kreises ist proportional zur Wurzel aus der jeweiligen Anzahl stat. Einheiten bei Untersuchung verschiedener stat. Massen hinsichtlich desselben Merkmals)

- **Darstellung von Summenhäufigkeiten:** als Punkt, zur besseren Übersichtlichkeit werden die Punkte durch waagerechte Linien nach rechts fortgesetzt (ähnlich wie bei der empirischen Verteilungsfkt.)

- **Histogramm:** bei klassierten Merkmalen, Häufigkeiten für die jeweiligen Klassen auftragen und jede Klasse entspricht einem Intervall auf der Abzisse, das durch obere und untere Klassengrenze definiert ist. Flächendiagramm mit Rechtecken, eine Seite des Rechtecks das Intervall der Klasse.

 bei offenen Randklassen kann man das nicht machen, sofern nicht eine festlegbare Klassenober- bzw. Untergrenze existiert, wenn der Anteil dieser Randklasse an der stat. Masse gering ist, so kann man 0 als Höhe des Rechtecks wählen.

Bei einem Vergleich von zwei Häufigkeitsverteilungen, deren Grundgesamtheiten einen unterschiedlichen Umfang haben, so verwendet man die rel. Häufigkeitsdichten und legt die Histogramme übereinander.

Bei nicht übereinstimmender Klassenbreite ist die Höhe der Rechtecke nicht einfach proportional zur Häufigkeit, sondern zum Verhältnis aus Häufigkeit und Klassenbreite. Die Rechtecke sind demnach umso höher, je größer die Anzahl der Beobachtungen ist, die in einen Bereich vorgegebener Länge fallen. Damit ist die Höhe des Rechteckes davon abhängig, wie dicht die Beobachtungen beieinander liegen.

Häufigkeitsdichte entspricht der Höhe des Rechtecks in einem Histogramm.

Wenn die Klassenbreiten aller Klassen übereinstimmen, so entspricht das Flächendiagramm auch einem Säulendiagramm und man kann die Ordinate mit rel. oder abs. Häufigkeiten versehen.

- **Höhe des Rechtecks:**

- **Häufigkeitspolygon:** Polygonzug, der durch die Verbindung der Mittelpunkte der oberen Rechteckkanten entsteht

Bei dem Histogramm der Häufigkeitsverteilung besteht die Möglichkeit, die aufsummierte Häufigkeit mehrerer Klassen dadurch zu bestimmen, dass man die Fläche über den Klassen und unter der oberen Begrenzung durch das Histogramm bestimmt.

- **Absolute bzw. relative Summenhäufigkeitsfunktion:** einzelne Punkte bei den Klassengrenzen, da nur für diese eine Summenhäufigkeit definiert ist, näherungsweise durch Geradenstücke verbunden, man geht davon aus, dass die Merkmalswerte zwischen den Klassengrenzen gleichmäßig verstreut liegen. (ist nicht die empirische Verteilungsfunktion, aber eine Näherung dieser)

- **Summenhäufigkeitsfunktion:** Häufigkeit bis zu einem Wert z

→ *Zusammenhang zwischen Histogramm und Summenhäufigkeitsfunktion, denn die Fläche des Histogramms links von einem Punkt z ergibt den Wert der Summenhäufigkeitsfunktion an der Stellt z, wenn die Gesamtfläche des Histogramms als Flächeneinheit verwendet wird.*

* **Lage- und Streuungsparameter**

- **Ziele:** Reduzierung der Daten auf markante Vergleichsgrößen und Festlegung von Kerngrößen einer Grundgesamtheit, die durch Stichproben geschätzt werden können (Parameter)

- **Lageparameter:** Kennzahl über die Lage der Beobachtungswerte, Kern der Verteilung

- **Streuungsparameter:** Kennzahl über die Streuung der Beobachtungswerte, Beurteilung der Positionierung der einzelnen Werte zum Zentrum

* **Lageparameter**
1. **Modus oder Modalwert**

- die **am häufigsten auftretende Merkmalsausprägung**, kann bei Rang-, quantitativen und qualitativen Merkmalen bestimmt werden, wenn mehrere häufigste Werte vorliegen, dann liegen auch mehrere Modalwerte vor.

bei klassierten Daten ist die Modale Klasse die Klasse mit der größten Häufigkeitsdichte.

1. **Zentralwert oder Median**

- teilt die geordnete stat. Masse in der Mitte in zwei gleichgroße Teile

* **n ungerade**:
* **n gerade:**
	+ Fallunterscheidung:
		-
		-
* **bei einer empirischen Verteilungsfunktion:**
	+ Fallunterscheidung:
		- **F nimmt den Wert 0.5 an**: n ist gerade mit und

man erhält:

* **F nimmt den Wert 0.5 nicht an:** n ist gerade mit oder ungerade und ist die Stelle, an der erstmals der Wert 0.5 überschritten wird:

1. **α-Quantil**

Die Urliste wird in den Proportionen α zu 1 – α geteilt. Eine Fallunterscheidung ist wieder notwendig: wenn der „Teilungspfeil“ genau einen Merkmalswert trifft (dann hat man das Quantil) oder auf eine Lücke (dann bestimmt man das Mittel aus den benachbarten Werten).

oberes Quartil: α = 0.25

unteres Quartil: α = 0.75

* **bei einer empirischen Verteilungsfunktion:**
	+ **es gibt ein x mit F(x) = α**:

* **es gibt kein x mit F(x) = α**:

1. **(Feinberechneter) Zentralwert**

Bei einem klassierten Merkmal mit Summenhäufigkeitsfunktion SF heißt Zentralwert,

* **falls eindeutig festgelegt ist**:

* **falls**  **nicht auf einer Klassengrenze liegt:**
	+ eine Einfallsklasse muss bestimmt werden, in die der zentralwert fällt:
1. **Arithmetisches Mittel**

Die genaue Lage aller Merkmalswerte hat Einfluss auf das arith. Mittel.

* **allgemein:**
* **absolute Häufigkeitsverteilung:**
* **relative Häufigkeitsverteilung:**
* **klassierte Merkmale:**
* **Streuungsparameter**

- Kenngröße für die Variation der Werte (Volarität), Beschreibung der Position der einzelnen Werte und Bezug auf das Zentrum, nur für quantitative Merkmale, Ergänzung zu Lageparametern

1. **Spannweite**

Differenz aus dem größten und dem kleinsten Beobachtungswert (Ausreißer legen die Spannweite fest, während die anderen Werte unberücksichtigt bleiben)

**bei klassierten Merkmalen:**

1. **Quartilsabstand**

Differenz aus dem oberen und unteren Quartil (man umgeht ein wenig die Ausreißerabh.)

1. **Mittlere absolute Abweichung (vom Zentralwert)**
* **allgemein:**
* **aus einer Häufigkeitsverteilung:**
* **klassierte Merkmale:**

1. **Varianz**

Arithm. Mittel der quadrierten Abweichungen der Werte vom Mittelwert

* **Korrigierte Stichprobenvarianz:**
* **Standardabweichung:**
* **Varianz einer Verteilung:**
* **Varianz der Verteilung eines klassierten Merkmals:**
* **Lage- und Streuungsparameter bei Transformation der Merkmalswerte/-ausprägungen**

Es sei eine **Transformation mit:**

Die stat. Reihe geht über in mit:

 (für Modalwert, Zentralwert und arith. Mittel geht dieselbe Formel)

* **mittlere absolute Abweichung:**
* **Varianz**
* **Variationskoeffizient der Verteilung**

→ **Nutzen und graphische Darstellung: siehe Skript!!!**

* **Konzentration von Merkmalswerten**

**Ziel:** Beschreibung, graphische Darstellung und Messung von Ungleichheiten, z.B. bei der Verteilung von „Besitz“, gegeben ist eine „Besitz“-Verteilung in Form einer nichtnegativen geordneten Urliste

* **Lorenzkurve**

- graphisches Hilfsmittel zur Verdeutlichung von Konzentrationsphänomenen, Gegenüberstellung des Anteils an der stat. Masse und des Anteils an der Merkmalssumme der k stat. Einheiten mit den kleinsten Merkmalswerten

**Anteil an der Merkmalssumme, der auf die k stat. Einheiten mit den kleinsten Merkmalswerten entfällt:**

**Anteil an der gesamten stat. Masse:**

→ daraus ergeben sich (-Tupel, die man in eine Koordinatensystem einträgt, beginnend mit dem Ursprung!

**Eigenschaften der Lorenzkurve:** beginnt in (0, 0) und endet in (1, 1), steigt monoton, konvex

Die Diagonale ist die Bezugskurve zur Lorenzkurve. Sie gibt den Zustand wieder, in dem die Merkmalssumme völlig gleichmäßig verteilt ist. Wenn alle Beobachtungswerte gleich sind, dann sind , die Lorenzkurve entspricht also der Diagonalen. **Extremfall** tritt ein, wenn die gesamte Merkmalssumme in einer stat. Einheit vereint ist:

Bei großem n erhält man nahezu die Katheten eines rechtwinkligen Dreiecks, also die Eckpunkte

(0, 0), (1, 0), (1, 1). Je mehr die Lorenzkurve von der Diagonalen entfernt ist, desto größer ist die Konzentration.

* **Ermittlung der Lorenzkurve aus der absoluten und relativen Häufigkeitsverteilung:**

**Merkmalssumme:**

**Anteil der k niedrigsten Merkmalsausprägungen an der Merkmalssumme:**

**Anteil dieser Merkmalsausprägungen an der stat. Masse:**

* **Ermittlung der Lorenzkurve bei klassierten Merkmalen:**

Weder noch könne gebildet werde, man hat relative bzw. absolute Häufigkeiten, anstatt kann man folgende Werte verwenden:

Für ergibt sich dann:

Bei relativen Häufigkeiten dasselbe Vorgehen!

* **Gini-Koeffizient**

Anteil der Fläche zwischen der Diagonalen und der Lorenzkurve an Gesamtfläche unterhalb der Diagonalen, also die Abweichung der Lorenzkurve von der Diagonalen.



**Fläche :**

**Gini-Koeffizient:**

**Maximalwert des Gini-Koeffizienten:**

**Normierter Gini-Koeffizient:**

Es gilt: und

**→** Unterschiedliche Lorenzkurven können zum selben Gini-Koeffizienten führen, er ist ein Maß für de relative Konzentration, aber nicht für die absolute.

* **Weitere Konzentrationsmaße**

**Konzentrationskoeffizient:**

gibt an, welchen Anteil der Merkmalssumme die g letzten Merkmalswerte der geordneten stat. Reihe in sich vereinen. Das Vorgehen ist dasselbe, nur die geordnete Liste wird in umgekehrter Reihenfolge abgearbeitet. Die Kurve wird als **Paretokurve** bezeichnet.

**Herfindahl-Index:**

H ist die Summe der quadrierten individuellen Anteile an der Merkmalssumme. Je größer H, desto größer die Konzetrantion.

* **Mehrdimensionale Merkmale**

Es liegt nicht mehr ein einzelnes Merkmal vor b : S → M, sondern k ≥ 2 Merkmale

 . Diese Werte lassen sich zu einem k-Tupel zusammenfassen.

Bei zweidimensionalen Merkmalen erhält man für jede stat. Einheit ein Paar von Beobachtungswerten. Man ordnet sie lexikographisch, d.h. zunächst werden die Wertepaare nach der ersten Zahl geordnet und wenn die ersten Komponenten bei 2 Paaren übereinstimmen, so richtet sich die Reihenfolge nach dem zweiten Merkmal.

**Absolute Häufigkeit der Merkmalskombination**

**Summe der absoluten Häufigkeiten:**

**Relative Häufigkeit der Merkmalskombination** :

→ daraus ergibt sich eine zweidimensionale Häufigkeitsverteilung, tabellarische Darstellung einer zweidimensionalen Häufigkeitsverteilung von nominalskalierten Merkmalen heißt Kontingenztabelle. Sind beide Merkmale mind. ordinalskaliert, so nennt man sie die Korrelationstabelle.

→ **Bei klassierten Merkmalen dasselbe Vorgehen!**

**Graphische Darstellung von zweidimensionalen Merkmalen:** durch Stabdiagramm, Histogramm, Streuungsdiagramm

* **Zusammenhang zweier Merkmale**

**Randverteilung:**

Häufigkeit einer Merkmalsausprägung unter Vernachlässigung des jeweils anderen Merkmals. Dies erhält man, wenn man die Häufigkeiten der entsprechenden Zeile oder Spalte summiert. Sie werden in Randzeile- oder spalte eingetragen. Die Summe der Randzeilen- oder spalten muss n bzw. 1 bei relativen Häufigkeiten ergeben.

* **Gegenseitige Beeinflussung zweier Merkmale**

**Bedingte Häufigkeit:**

 ist die reduzierte stat. Masse unter der Bedingung a.

**Relative Häufigkeit der Merkmalsausprägung „b unter der Bedingung a“:**

**Bedingte Lage- und Streuungsparameter:**

* **Arithmetisches Mittel:**
* **Varianz:**

Ein Einfluss von Merkmal 1 auf das Merkmals 2 spiegelt sich dadurch wider, dass die bedingte Häufigkeitsverteilung von Merkmals 2 unter der Bedingung a von der Merkmalsausprägung a beeinflusst wird. Es liegt kein Einfluss von Merkmal 1 auf Merkmal 2 vor, wenn für alle

**Unabhängigkeit:**

Wenn gilt:

**Abhängigkeit:**

Wenn gilt:

* **Überprüfen auf Unabhängigkeit**
1. Berechnung der bedingten relativen Häufigkeitsverteilung von Merkmal 2 für jede Merkmalsausprägung von Merkmal 1 als Bedingung und Überprüfung auf Übereinstimmung. Wenn ja liegt Unabhängigkeit vor.
2. Analog zu 1. mit vertauschten Merkmalen.
3. Überprüfung der Gleichung:

für alle a und b. Falls eine Kombination (a, b) gefunden ist, bei der die Gleichung verletzt ist, sind die Merkmale abhängig und die Überprüfung ist abgebrochen.

* **Untersuchung der Abhängigkeit**

**Ziel**: Messung der Abweichung von Unabhängigkeit

- für unabhängige Merkmale (relativ): siehe oben

- für die absoluten Häufigkeiten:

Man kann ermitteln, wie die Häufigkeitsverteilung aussehen müsste, wenn die Merkmale unabhängig wären. Diese Tabelle nennt man **Indifferenztafel oder Indifferenztabelle**. Der Unterschied der beiden Tabellen dokumentiert die Abweichung der tatsächlichen Häufigkeitsverteilung von der bei Unabhängigkeit. Die **Abweichung ist im Feld (a, b):**

* **Chi-Quadrat:**

Maßzahl für die Abweichung von Unabhängigkeit, die durch Summation der relativen quadrierten Abweichungen der beobachtenden Merkmalsausprägungen von den Werten bei Unabhängigkeit entsteht (je größer chi-quadrat ist, desto größer sind die rel. Abweichungen in den einzelnen Feldern usw.):

wichtig: Bei Verdoppelung von n verdoppelt sich auch .

* **Kontingenzkoeffizient nach Pearson:**

Der Maximalwert von steigt mit n an, also kann er Werte größer als 1 annehmen. Die Methode dies zu korrigieren ist der **Kontingenzkoeffizient nach Pearson:**

es gilt:

→ je größer C, desto stärker die Abweichung von Unabhängigkeit. Der Maximalwert von C ergibt sich bei „absoluter Abhängigkeit“.

* **Absolute Abhängigkeit:**

liegt vor, wen jede Merkmalsausprägung a von Merkmal 1 nur in Kombination mit einer ganz bestimmten Merkmalsausprägung b von Merkmals 2 beobachtet wurde und umgekehrt. Dann ist es zu vermuten, dass C maximal wird.

Es gilt dann:

**Für den Kontingenzkoeffizienten nach Pearson gilt dann**:

r, s sind die Anzahlen der beobachteten Merkmalsausprägungen von Merkmal 1 bzw. 2.

* **Korrigierter Kontingenzkoeffizienten nach Pearson (bei 2 verschiedenen stat. Massen):**

es gilt:

* **Lineare Regression**

Bei Abhängigkeit zwischen 2 Merkmalen stellt sich die Frage über die Art und Stärke der Abhängigkeit. Wenn beide Merkmale quantitativ sind, bestehen die Tupel als reellen Zahlen, woraus man ein Streuungsdiagramm (Punktwolke) erhält.

Man nimmt an, dass den Beobachtungspaaren ein linearer Zusammenhang zugrunde liegt, den man mit folgender Funktion darstellen kann:

* **Prinzip der linearen Regression oder Methode der kleinsten Quadrate:**

Angenommen, m und b wären bekannt, so ließe sich zu jedem x-Wert der „Trendwert“ ermitteln:

Daraus ergäbe sich dann auch die Störgröße:

→ Bestimme m und b so, dass die Summe der quadrierten Störgrößen minimiert wird:

* **Bestimmen von m und b:**

→Diese Methode lässt sich auch bei nichtlinearen Zusammenhängen anwenden. Man muss zuerst die Daten transformieren (meistens mit ln), dann rechnet man damit und am Ende transformiert man wieder rückgängig. (siehe Aufgaben!)

* **Korrelationsrechnung:**

Es ist nicht immer sinnvoll, den Zusammenhang zwischen 2 Merkmalen mittels linearer Regression zu berechnen. Mit der Korrelationsrechnung will man die „Güte“ eines unterstellten linearen Zusammenhangs messen. Dazu vergleicht man die Varianz der „Trendwerte“ mit der Varianz der y- Werte.

* **Bestimmtheitsmaß:**

Verhältnis aus der Varianz der -Werte und der Varianz der y-Werte. Beschreibt den Anteil an der Varianz der y-Werte, der sich bei linearer Regression aus der Varianz der x-Werte begründen lässt, also um den Teil der Varianz, den man auch erhalten würde, wenn der lineare Zusammenhang eingehalten wäre, also die Störgrößen gleich 0 wären.

Varianz der Trendwerte:

Varianz der y-Werte:

**Bestimmtheitsmaß:**

* **Kovarianz:**
* **Varianz der Summe:**
* **Kovarianz von absoluten bzw. relativen Häufigkeiten:**

bei Unabhängigkeit ist

* **Korrelationskoeffizient:**

Maß für den linearen Zusammenhang zweier Merkmale:

**Das Bestimmtheitsmaß ist das Quadrat des Korrelationskoeffizienten:**

→ es gilt:
**1.** alle Beobachtungswerte liegen auf einer steigenden Geraden, .

**2.**  alle Beobachtungswerte liegen auf einer fallenden Geraden,

* **Spearmannscher Rangkorrelationskoeffizient:**

Bei Rangmerkmalen hat der Korrelationskoeffizient keine direkte Bedeutung. Man ordnet die Merkmalswerte der stat. Reihe ihrer natürlichen Reihenfolge und ersetzt den Wert durch eine „Rangziffer“.

Aus Merkmalspaaren werden Rangziffernpaare

Dann erhält man den **Korrelationskoeffizienten:**

**Die 2. Formel** (falls kein Merkmalswert doppelt auftritt und damit die Rangziffern jeweils die Werte 1,…,n durchlaufen):

**Anmerkung:**

gegenläufig

unabhängig

identisch

* **Einführung in die Zeitanalyse**

**Zeitreihe:** Eine Folge beobachtet Werte einer Größer zu – in der Regel äquidistanten – Zeitpunkten bzw. für aufeinanderfolgende – in der Regel gleichlange – Zeiträume für

Die Werte setzen sich additiv aus ihren Komponenten zusammen:

mit:

* **Methode der gleitenden Durchschnitte:**

Saison- und Störkomponente werden ausgeschaltet, indem man die gleitende Durchschnitte über den Zeitraum der Periodenlänge bildet. Man erhält die **glatte Komponente**:

**Gleitender Durchschnitt ungerader Ordnung 2k + 1:**

die Werte lassen sich bilden für

**Gleitender Durchschnitt gerader Ordnung 2k:**

**z.B.** man berechnet das arith. Mittel aus

aber die zeitlichen Reihen wäre dadurch versetzt. Deswegen bildet man das arith. Mittel von:

und das arith. Mittel von beiden Werten:

**Allgemeines k:**

* **Saisonfigur**
1. **Konstante Saisonfigur**: saisonale Schwankungen wirken sich von Periode zu Periode in absolut gleicher Höhe aus.
2. **Nicht konstante Saisonfigur**: saisonale Schwankungen sind proportional zur glatten Komponente.
* **Berechnung einer konstanten Saisonfigur:**

**Annahme:**

→

Die Abweichungen innerhalb einer Periode werden trotzdem schwanken, dies eliminiert man, indem man den **Mittelwert der Abweichungen bildet:**

Sei und die Periodenlänge der Saison:

Diesen Wert muss man berichtigen, indem man das arith. Mittel von allen bildet:

**Berechnung der Saisonfigur:**

**Saisonbereinigte Zeitreihe:** es werden von der originalen Zeitreihe der jeweils zugehörige Wert der Saisonfigur abgezogen.

* **Berechnung der Saisonindexziffern:**

**Annahme:**

→

Auch diese Schwankungen werden durch die Mittelwertbildung eliminiert:

Sei :

Auch diesen Wert muss man berichtigen, indem man das arith. Mittel von allen bildet:

**Saisonindexziffer:**

**Wahrscheinlichkeitstheorie**

* **Grundbegriffe**

**Konstruktion des Wahrscheinlichkeitsmodells eines Zufallsprozesses:**

* Ω (Grundgesamtheit) besteht aus Elementarereignissen: (**endlich oder abzählbar unendlich)**
* Für jedes gibt es eine Wahrscheinlichkeit für das Eintreten:
* Ein zusammengesetztes Ereignis A ist eine Teilmenge von
* Wahrscheinlichkeit für das Eintreten von A:

**Häufigkeitsverteilung eines stetigen Merkmals:**

* „Glättung“ der Messwerte: Klassierung der Daten und Verwendung der Summenhäufigkeitsfunktion zur Bestimmung von Häufigkeiten, die auf Wahrscheinlichkeiten übertragen werden. Wahrscheinlichkeit eines Ereignisses ≤ α:
* Wahrscheinlichkeit von einem bestimmten Ereignis ist:
* aus folgt:
	+ **A endlich:**
	+ **A abzählbar unendlich:**

**Rechenregeln von Häufigkeiten und Wahrscheinlichkeiten:**

**Modellierung eines Zufallsprozesses:**

1. Feststellung der Elementarereignisse und Zusammenfassung einer Grundmenge Ω
2. Kombination von Ereignissen: ein (zusammengesetztes) Ereignis
3. Zusammenstellung aller Ereignisse: System von Teilmengen von Ω, A(Ω)
4. Wahrscheinlichkeitsmaß: zu jeden Ereignis A wird eine Wahrscheinlichkeit zugeordnet
5. Für die Ereignisse und zugehörige Wahrscheinlichkeiten müssen die Rechenregeln gelten.
* **Wahrscheinlichkeitsraum**

Ein Wahrscheinlichkeitsraums (mathematisches Modell zur Beschreibung von Zufallsexperimenten) besteht aus 3 Komponenten:

1. **Grundgesamtheit Ω (Menge der Elementarereignisse):**
2. **Mengensystem der Ereignisse (σ-Algebra):**
	* **mögliche σ-Algebren:**
3. **Wahrscheinlichkeitsmaß P:**

**Bemerkung:** Falls

|  |  |
| --- | --- |
| jedes ω einzeln beschrieben |  LaPlace’scher W-Raum: |

* **Zufallsvariable**

**Variable:** Größe, die verschiedene Werte annehmen kann: x,y, …

**Zufallsvariable:** Variable, deren Wert vom Ausgang eines Zufallsexperiments abhängt: X, Y, … oder auch: eine Funktion , wenn für jede Borelsche Menge R die Urbildmenge ist, d.h. bei wenn X messbar ist. X kann diskret, stetig oder eine Mischform sein (s. Bsp. Abfüllanlange)

**Beschreibung:**

* Zufallsprozess, modelliert durch einen Wahrscheinlichkeitsraum
* Jeder Ausgang des Zufallsexperiments ist ein Elementarereignis
* Wert von Zufallsvariable hängt ab vom Ausgang
* Man betrachtet zu jedem den zugehörigen Wert der Zufallsvariablen, d.h. die Funktion (bei reellwertigen Zufallsvariablen)

**Die Wahrscheinlichkeit für einen Wert in einer Teilmenge**

**→** man betrachtet die Urbildmenge bei der Teilmenge R der reellen Zahlen:

Zu der Funktion :

1. **Ω endlich oder abzählbar unendlich:**
2.

es gibt Teilmengen, für die keine Wahrscheinlichkeit angegeben werden kann, deshalb beschränkt man sich auf:

* + 1. σ-Algebra der Borelschen Menge
		2. R aus σ-Algebra der Borelschen Menge

**Fragen/Eigenschaften:**

1. **Wann ist ?**

→ Hängt ab von R und X.

1. **Wann ist eine Borelsche Menge ?**

→ Potenzmenge von Ω: immer erfüllt,

wenn σ-Algebra der Borelschen Mengen: dann ist

1. **Abbildung X ist messbar**, wenn für jede Teilmenge , die aus der kommt, die Urbildmenge auch im Mengensystem der .

**Wahrscheinlichkeitsmaß (induziert durch X):**

**Regeln:**

1. aus 3. folgt:
2. ist eindeutig festgelegt durch:

**→ Durch die Verteilungsfunktion ist die Wahrscheinlichkeitsverteilung der Zufallsvariable X eindeutig festgelegt.** ( monoton steigend, , , von rechts stetig: )

* **Spezielle Verteilungen: Diskrete Verteilungen**

**Allgemein**: X heißt diskret, wenn es Zahlen gibt mit

* **Bernoulli-Verteilung:**

Bei Beschreibung von Grundgesamtheiten, deren Elemente eine Eigenschaft haben (, wenn die Eigenschaft hat) oder nicht haben (, wenn .

* **Hypergeometrische Verteilung (Ziehen ohne Zurücklegen):**

N – Umfang der stat. Reihe

M – # der Einheiten mit der Eigenschaft E in der stat. Masse (

p – Anteil der Einheiten mit der Eigenschaft E in der stat. Masse

n – Stichprobenumfang

X – # der Einheiten mit der Eigenschaft E in der Stichprobe (Zufallsvariable)

* **Binomialverteilung (Ziehen mit Zurücklegen):**

→ X heißt binomialverteilt mit

p – relative Häufigkeit bzw. Wahrscheinlichkeit für das Ziehen eines Elementes mit der Eigenschaft E bei einmaligem Ziehen

 – Wahrscheinlichkeit für das Ziehen eines Elementes mit der Eigenschaft E bei m-maligem Ziehen

n – Stichprobenumfang

**Vorteile** der Binom. Verteilung gegenüber der hypergeom. Verteilung:

* N geht nicht in die Verteilung ein
* Berechnung von Potenzen ist einfacher als von Binomialkoeffizienten
* Die Werte sind für alle definiert und nicht nur für Brüche M/N

**→ :**

der Unterschied zwischen den beiden Verteilungen ist so gering, dass die Binomialverteilung als Näherung für die hypergeometrische verwendet werden darf.

* **Poisson-Verteilung:**

Bei Modellierung der Anzahl von Ereignissen bei einem Zufallsprozess.

X heißt Poisson-verteilt mit

 – prozessspezifischer Parameter

 X – Anzahl der Ereignisse

**→ :**

Poisson-Verteilung ist Näherung der Binomialverteilung für großes n und kleines p. Die Poisson-Verteilung entsteht aus der Binomialverteilung bei Grenzübergang:



* **Stetige Verteilungen**

**Definition:**

Sei X eine Zufallsvariable mit der Verteilungsfunktion heißt **stetig**, wenn es eine Funktion gibt mit:

→ Die Verteilungsfunktion einer stetigen Zufallsvariable ist stetig ( heißt **Dichtefunktion von X**).

**Eigenschaften/Forderungen von/an :**

* aus:

folgt: (Gesamtfläche unter f ist 1)

* Ist in diff’bar, dann gilt:

**Hinreichende Bedingung:**

Sei f bis auf endlich viele Stellen stetig mit den Eigenschaften

1. Existiert

→ f ist Dichte einer Zufallsvariable mit der Verteilungsfunktion

**Beispiele für stetige Verteilungen:**

1. **Geometrische Verteilung**: Gleichverteilung im Intervall [a, b] und Trapezverteilung
2. **Exponentialverteilung:**
3. **Normalverteilung** mit Parametern
4. **Standardnormalverteilung** mit
* **Lage- und Streuungsparameter**

Wie in der deskriptiven Stasi: Beschreibung von Wahrscheinlichkeitsverteilung durch Kenngrößen:

* Bei diskreten Verteilungen: Wahrscheinlichkeiten durch relative Häufigkeiten
* Bei stetigen Verteilungen: mit der Dichte (analog zum Histogramm)
* **Lageparameter:**
1. **Modalwert:** Merkmalsausprägung mit maximaler Häufigkeit

**Bei diskreten Verteilungen:** Der (die) Wert(e) einer ZV mit der größten Wahrscheinlichkeit

**Bei stetigen Verteilungen:** Maximalstelle(n) der Dichtefunktion

1. **Median:** Zentralwert, „Halbierung“ der Verteilung

 ist Median (Zentralwert), wenn

 ist Median von

**Bei diskreten Verteilungen:**

**Fall 1:** es gibt kein x mit :

Sprungstelle, an der das Niveau 0.5 übersprungen wird → Median eindeutig

**Fall 2:** es gibt ein x mit :

Es gibt ein Intervall Mediane sind alle Punkte des abgeschlossenen Intervalls

**Bei stetigen Verteilungen:**

**Bestimmung des Medians mit der Verteilungsfunktion:**

F ist stetig und nimmt jeden Wert zwischen 0 und 1 an. Jedes x mit heißt Median. Jedes x mit heißt α-Quantil.

**Beispiele:** siehe Skript S. 18-22!!!

Für eine ZV X mit symmetrischer Dichtefunktion ist der Symmetriepunkt μ Median

1. **Erwartungswert**

Arithmetisches Mittel aus relativer Häufigkeitsverteilung (allgemein):

**Bei diskreter ZV:**

**Erwartungswert:**

Wenn für die Reihe absolut konvergiert:

 **stetige ZV mit Dichtefunktion f:**

* **Streuungsparameter**
* **Varianz:**

Analog zur deskriptiven Stasi.

**Bei diskreten ZV X** (falls diese Reihe konvergiert und muss gegeben sein):

**Bei stetigen ZV mit Dichte f und** (falls dieses Integral existiert):

|  |  |  |
| --- | --- | --- |
|  ParameterVerteilungen | Erwartungswert  | Varianz  |
| Bernoulli-Verteilung |  |  |
| Binomialverteilung |  |  |
| Poisson-Verteilung |  |  |
| Exponentialverteilung |  |  |
| Gleichverteilung auf |  |  |
| Normalverteilung |  |  |

* **Momente höherer Ordnung**

Verallgemeinerung von Erwartungswert und Varianz.

* **k-tes Moment** (Verallgemeinerung des Erwartungswerts):

**Bei diskreten X:**

Heißt k-tes Moment von X, falls die Reihe absolut konvergiert.

**Bei stetigen X** (mit Dichtefunktion f)**:**

Heißt k-tes Moment von X, falls das Integral existiert. K-tes Moment ist Erwartungswert von .

* **k-tes zentrales Moment** (Verallgemeinerung der Varianz)

**Bei diskreten X:**

Heißt k-tes zentrales Moment von X, falls die Reihe absolut konvergiert.

**Bei stetigen X** (mit Dichtefunktion f)**:**

Heißt k-tes zentrales Moment von X, falls das Integral absolut existiert.

* **Funktion und Transformation von Zufallsvariablen**
* **Funktion von ZV**

ZV sind häufig Argument einer Funktion. Sei und . Zu Wert der ZV. Dieser Wert wird in g eingesetzt und man erhält: , was der Komposition der beiden Funktionen entspricht: .

 ist ZV, wenn messbar ist. Nimmt X nur Werte in einem Teilbereich D von ℝ an, so muss g auch nur dür die Argumente aus D definiert sein.

**Bei diskreten ZV** (mit Wahrscheinlichkeitsverteilung **:**

**Bei stetigen ZV** (mit Dichte :

Verteilungsfunktion von ist:

→ Falls X stetig ist, dann ist dennoch nicht notwendig stetig.

* **Transformation von ZV:**

Die Funktion g ist invertierbar. Es gibt eine Umkerfunktion . Nimmt X nur Werte im D an, so muss g auch nur von auf D invertierbar sein.

**Bei diskreten ZV:**

g ist bijektiv, g ist Transformation, falls

1. Gibt es **ein** eindeutig bestimmt und es gilt:
2. Gibt es **kein** , so ist:

**Bei stetigen ZV:**

g ist bijektiv, d.h. es gibt zu g eine Inverse .

X stetig mit Dichte Ist g diff’bar mit bis auf endlich viele Stellen y, so gilt:

Die letzte Formel erhält man, wenn man substituiert ().

Nimmt X nur Werte in einem Teilbereich D der reellen Zahlen an, so muss g auch nur in diesem Teilbereich invertierbar sein.

* **Erwartungswert:**

Es gilt:

**Bei diskreten ZV:**

**Bei stetigen ZV** (mit Dichte ):

 muss nicht existieren.

Aus der Existenz von folgt nicht die Exstenz von und umgekehrt.

**Folgerungen:**

1. **Mit gilt:**

(analog für diskrete ZV)

1. **Mit**

(analog für stetige ZV)

1. **Varianz von**

Vllt mal selbst nachrechnen ☺!!!

1. **Standardisierung von X:**

Damit kann jede ZV X durch lineare Transformation in eine ZV Y transformiert werden mit Erwartungswert 0 und Varianz 1. Der Typ der Verteilung bleibt bei der Transformation im Allgemeinen **nicht** erhalten (außer bei Normalverteilung).