Statistik ist die Lehre von Verfahren und Methoden zur Gewinnung, Erfassung, Analyse, Charakterisierung, Abbildung, Nachbildung und Beurteilung von beobachtbaren Daten über die Wirklichkeit (Empirie).

WO BRAUCHT MAN STATISTIK?

In den empirischen Wissenschaften (auch Realwissenschaften bzw. Erfahrungswissenschaften genannt)

Naturwissenschaften, **Sozial**wissenschaften, Biologie/ Medizin (Biometrie), **Ingenieurs**wissenschaften (Technometrie), **Verhaltens**wissenschaften (Psychometrie)

➔ Gewinnung von neuen Erkenntnissen übr einen Ausschnitt der Realität. Dazu werden empirische Untersuchungen durchgeführt. Dabei fallen Daten an, die mit statistischen Methoden ausgewertet werden.

Was ist eine „Statistik"?

eine systematische Zusammenstellung von Zahlen und Daten

Wozu? ➔ Beschreiung bestimmter Zustände, Entwicklungen und Phänomenen

Ziel: Gewinnung von Informationen aus unübersichtlichen und/ode unstrukturierten und/oder großen Datenmengen

1. Datengewinnung durch amtliche Erhebung, Berichte, Umfragen und betriebliche Quellen

Datenerhebung = jede systematische Datengewinnung

Vorgang zur Ermitllung und zur Erfasung von Ausprägungen eines statistitschen Merkmals,

Primärerhebung= Erhebung neuer Daten nach Vorgaben, **Sekundärerhebung=** aus bereits vorhandenem Datenmaterial, **Vollerhebung=** Untersuchung statistischen Einheiten einer Gesamtheit

Teilerhebung n <N

2. Datenanalyse = Anwendung statistischer Verfahren zum Zweck des Erkenntnisgewinns

Datencharakterisierung = grafische und tabellarische Darstellung von Daten sowie Berechung von zusammenfassenden, den empirischen Sachverhalt beschreibenden Kennzahlen

3. Datenbeurteilung Die Beurteilung von Daten erfolgt durch

1. Schlüsse auf der Basis unvollständiger Daten, z. B. Schlüsse von der Stichprobeauf ihre **Grundgesamtheit**

2. All Allgemeiner: auf der Basis unsicherer Daten, unter Anwendung der Wahrscheinlichkeitsrechnung.

4. Datenaufbereitung = Ordnung, Zusammenfassung und Darstellung des erhobenen statistischen Datenmaterials in Datendateien, Tabellen und/oder geeigneten Grafiken.

Datenmissbrauch: Man sieht statistischen Ergebnissen nicht an, ob sie manipuliert wurden

Tabellarische Darstellung

Vorteil: liefert detailliertere Informationen, man kennt die genauen Werte das ist insbesondere bei Planungsaufgaben wichtig.

Nachteil: Tabellen sind schwerer zu lesen, man braucht Zeit, um die Information zu verarbeiten. Tabellen sind „langweilig".

Grafische Darstellung

Vorteil: Man kann sich sehr schnell ein Bild von den quantitativen Verhältnissen machen, man erkennt sehr schnell die wesentlichen Informationen (wenn das Diagramm gut gestaltet ist …).

Nachteil: Nur mit Mühe lassen sich genaue Werte ablesen.

Skalenart	Besonderheiten	zulässige Operationen	Beispiel für Merkmal	Beispiele für Operationen
Nominalskala	Merkmalsausprägungen sind diskret Die Werte unterliegen keiner Rangfolge und sind nicht vergleichbar	=, ≠	Geschlecht, Familienstand, Steuerklasse	Geschlecht von Claudia ≠ Geschlecht von Peter
Ordinalskala = Rangskala	Die Werte unterscheiden sich in ihrer Intensität und ordnen sich nach der Stärke dieser Intensität Ordnungsprinzip ist die Stärke bzw. der Grad der Intensität	=, ≠, <, >	Konfektionsgröße, Schulnoten, Windstärke	XXL > XL > L > M > S > XS
Intervallskala	Besitzt keinen natürlichen Nullpunkt, keine Verhältnisse können gebildet werden. Daten können alle Ausprägungen innerhalb eines Intervalls annehmen	=, ≠, <, >, +, -	Längendifferenzen, Temperatur in Celsius, IQ (Intelligenzquotient)	Länge (Klaus) - Länge (Claudia) < Länge (Peter) - Länge (Paula)
Verhältnisskala = Ratioskala	Besitzt natürlichen Nullpunkt Quotienten (das Verhältnis) gemessener Werte werden verglichen	=, ≠, <, >, +, - , x, /	Umsatz, Körpergröße, Einkommen	Der Umsatz ist um 7% gegenüber dem Vorjahr gestiegen oder doppelt so hoch wie…
Absolutskala	Ausprägungen absolut skalierter Merkmale sind Anzahlen und Stückzahlen	=, ≠, <, >, + - ,x, /	Zahl der Beschäftigten	150 Beschäftigte sind 3 mal so viel wie 50 Beschäftigte

	Stärken	Schwächen
R	•Sehr großer Funktionsumfang (weit über 2000 Pakete) •Sehr gut automatisier- und integrierbar (z.B. LaTeX, ODBC, MS …) •Sehr guter Community-Support sowie kostenpflichtiger Support über Drittanbieter •Umfangreiche Hilfe-Ressourcen frei verfügbar (Manuals, Tutorials •Alle gängigen Plattformen werden unterstützt (Windows, Linux...) •Zukunftssicher durch große, aktive Entwickler-Community	•Einarbeitung in die R-Syntax kann eine Einstiegshürde darstellen •Stabilität/Qualität wenig genutzter Pakete z.T. nicht auf dem hohen Niveau •Bei Verwendung sehr großer Datensätze wird leistungsfähige Hardware benötigt
SAS	•Schnelle Integration neuer statistischer Verfahren, sehr stabile und zuverlässige Routinen •Sehr gute Dokumentation und professioneller Support •Vielzahl von (kostenpflichtigen) Modulen und Schnittstellen, eigene Business Intelligence Software •Gut geeignet für Umgang mit großen Datensätzen •Umfangreiches hauseigenes Schulungsangebot	•Verschiedene, teils komplizierte (aber mächtige) Programmsprachen •Lizenzmodell verbunden mit hohen Kosten
SPSS	•Leicht erlernbar, Bedienung jedoch nicht immer intuitiv •Erweiterbar über kommerzielle Module •Umfangreiche Literatur vorhanden	•Versionen für Windows und MacOS •kurzes Update-Zyklus (1 Jahr) •schwierig automatisier- und integrierbar
STATA	•Großer Funktionsumfang - nahezu jede statistische Methode •Einfacher Einstieg durch GUI •Automatisierbar & mit alten Versionen kompatibel •Guter Support durch die STATA-Community, umfangr. Literatur •Lauffähig unter Windows, Mac, Unix •Im Vgl. zur kommerz. Konkurrenz vergleichsweise preiswert •Investitionssicherheit durch 3-jährigen Release-Zyklus	•Eher träge bei der Einarbeitung neuer Methoden (Versionsupdates) •Integration von und in andere Software ist umständlich •Beschränkung auf einen gleichzeitig geöffneten Datensatz

33

Merkmal	Menge der Merkmalsausprägungen	Messinstrument	Skala	Merkmalstyp
Familienstand	{ledig, verheiratet, verwitwet, geschieden}	Frage	Nominalskala	qualitatives Merkmal
Hotelgütekasse	{"*", "**", "***", "**", "*", "}	Fragebogen	Rangskala = Ordinalskala	qualitative Merkmale = Rangmerkmale
Klausurnote	{1,0 1,3 1,7 2,0 2,3 2,7 3,0 3,3 3,7 4,0 5,0}	Klausur		
Temperatur (°C)	$\mathbb{R}$	Thermometer	Metrische Skala = Intervallskala	Quantitative Merkmale = metrische Merkmale
Körpergröße	{x | x ∈ $\mathbb{R}$ und x > 0 }	cm-Maß	Metrische Skala = Verhältnisskala	
Kinderzahl	$\mathbb{N}_0$ ∪{0}	Frage	Metrische Skala = Absolutskala	

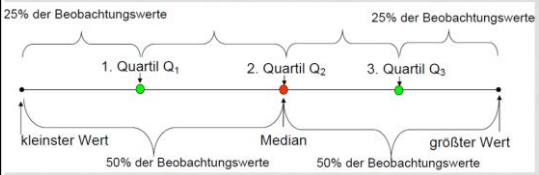
Metrische Merkmale (vgl. Folie 4)

diskret (z.B. Einwohnerzahl) — stetig (z.B. Körpergröße)

Klassierung	Klassierung	Klassierung	Klassierung
1. 0 – 19.999 2. 20.000 – 49.999 3. 50.000 – 99.999 4. 100.000 – 249.999 usw.	0 – 139 cm 100 – 159 cm 140 – 179 cm 160 – 199 cm usw.	0 – 100 cm 100 – 140 cm 140 – 160 cm 160 – 170 cm usw.	0 bis unter 100 cm 100 b.u. 140 cm 140 b.u. 160 cm 160 b.u. 170 cm usw.
	wg. Lücken	wg. Überschneidungen	richtig !!!

100 b.u. 140 cm = {x | x ∈ IR, 100 < x < 140 cm}

5-Punkte-Zusammenfassung der geordneten statistischen Reihe:



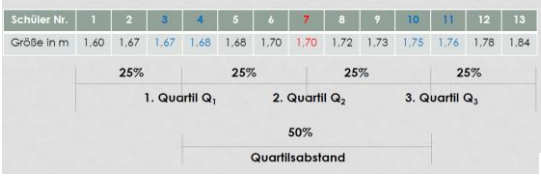
Der (Inter-)**Quartilsabstand** (*engl.: interquartile range*, **IQR**) bezeichnet die Differenz zwischen dem oberen und dem unteren Quartil Q_3-Q_1 und umfasst daher 50% der Verteilung.

Der Quartilsabstand wird als **Streuungsmaß** verwendet.

Beispiel:

Die Liste enthält von 13 Schülern die Körpergröße.
Die Merkmalsausprägungen (Beobachtungswerte) wurden nach der Größe geordnet.

Schüler Nr.	1	2	3	4	5	6	7	8	9	10	11	12	13
Größe in m	1,60	1,67	1,67	1,68	1,68	1,70	1,70	1,72	1,73	1,75	1,76	1,78	1,84



KLASSIERUNG BEI QUALITATIVEN MERKMALEN

Beispiel:

Frage: Beruf?

Antworten:

- Maurer
- Dachdecker
- Schreiner
- Fliesenleger

→ Berufsgruppe: Handwerker

Zielkonflikt: Übersichtlichkeit versus Informationsverlust

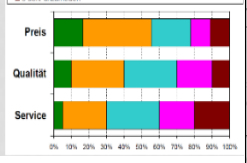
KLASSIERUNG BEI RANGMERKMALEN

Beispiel:

Frage: wie zufrieden sie mit einem Produkt?

Antworten:

- 1 sehr zufrieden
- 2 zufrieden
- 3 mehr oder weniger zufrieden
- 4 unzufrieden
- 5 sehr unzufrieden



ABLAUF EINER EMPIRISCHEN UNTERSUCHUNG

Definition (Phase 1): Definition des Informationsbedarfs, der Hypothesen, der Begriffe, der Untersuchungseinheiten, über die man Information haben will

Design-Entscheidungen (Phase 2): Abgrenzung der Grundgesamtheit, evtl. Stichprobenumfang

Erhebungsart:

Querschnitt-oder Längsschnittuntersuchung
Primärerhebung oder Sekundärerhebung
Vollerhebung oder Teilerhebung

Erhebungstechnik:

Befragung (persönlich, telefonisch, schriftlich oder online)
Beobachtung (offen oder verdeckt)
Dokumentenanalyse

„Konstruktion" von Messinstrumenten (Pretest der z.B. Fragebögen)

Ziel: das Risiko des Misserfolgs zu reduzieren und vorab Gründe für ein eventuelles Versagen zu finden

Entscheidungsspielraum wird eingeschränkt bzw. beeinflusst durch: Budget, Zeit, Thema/Aufgabenstellung

Interdependenzen (gegenseitige Abhängigkeit und Beeinflussung) zwischen den Design-Entscheidungen

Datenerhebung (Phase 3) - entsprechend der getroffenen Entscheidungen

Datenauswertung und –analyse (Phase 4)

= Vorbereitung der „maschinellen" Datenauswertung / –analyse(mit Software)

Dateiaufbau festlegen □ Datenimport □ Datenbereinigung □ Datenqualitätssicherung (Kontrolle auf Vollständigkeit und Plausibilität) Datenaufbereitung (Sortierung der Daten, Klassenbildung, …) □ Datenauswertung und Datenanalyse (univariateund multivariate Datenanalysen mit Anwendung geeigneter statistischer Methoden)□Einsatz von Statistik-Software

Dokumentation (Phase 5) = Dokumentation in Tabellen und Schaubildern und Interpretation der Ergebnisse

Beispiel für die Gliederung einer Ergebnisstudie:

Problemstellung: □Vorgehensweise, Beschreibung und Begründung aller Design-Entscheidungen□Hauptteil: Ergebnisse der empirischen Untersuchung□Folgerungen, Empfehlungen, Wertungen□Anhang: Fragebogen, Literatur-, Abbildungs-und Tabellenverzeichnis

Informationsbedarf → empirische (statistische) **Untersuchung. =** messen von Merkmalen bei ausgewählten Untersuchungseinheiten mit einem Messinstrument auf einer Skala.

Ergebnis: Messwerte = Merkmalswerte = Beobachtungswerte

Operationalisierung (siehe ebenfalls links unten)

Wenn ein Statistiker mit der Arbeit beginnt, muss er geklärt haben, was er

Ein **Datensatz** ist eine Gruppe von inhaltlich zusammenhängenden (zu einem Objekt gehörenden) Datenfeldern, z.B. Artikelnummer und Artikelname. Datensätze entsprechen einer logischen Struktur, die bei der Softwareentwicklung (z. B. im konzeptionellen Schema der Datenmodellierung) festgelegt wurde.

Histogramm

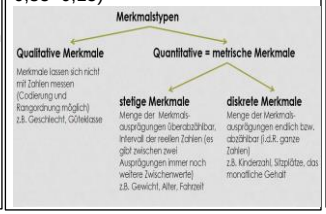
grafische flächenproportionale Darstellung der Häufogkeiten von klassierten Daten, - im Unterschied zum Säulendiagramm muss bei Histogramm die x-Achse immer eine Skala sein, deren Werte geordnet sind und gleiche Abstände haben - direkt **nebeneinanderliegende Rechtecke** (keine Abstände dazwisxhen) **von der Breite der jeweiligen Klasse** gezeichnet **(Breite der Rechtecke = Klassenbreite)**

Unter **Dependenzanalysen** werden Verfahren verstanden, mit denen Strukturen überprüft werden sollen. Zu diesen Verfahren gehören beispielsweise die Varianzanalyse, Regressionsanalyse, Diskriminanzanalyse, t-Tests, Chi-Quadrat-Tests etc. Ziel der Analysen ist es, Abhängigkeiten (Dependenzen) zwischen "abhängigen" und "unabhängigen" Variablen zu untersuchen.

Die **Normalverteilung** nimmt in der statistischen Theorie und in der Anwendung einen großen Raum ein. Viele Phänomene in der Natur und den Wirtschaftswissenschaften lassen sich entweder genau oder näherungsweise durch die Normalverteilung beschreiben. Verteilungen wie die Binomialverteilung konvergieren für hinreichend große Stichprobenumfänge gegen die Normalverteilung.

Die Normalverteilung stellt ein Modell für symmetrisch verteilte Zufallsvariablen dar. Sie ist durch zwei gut interpretierbare Parameter, μ, den Erwartungswert und σ2, die Varianz, vollständig charakterisiert.

Wachstumsfaktor 2006/2007: 0,5; Wachstumsfaktor 2007/2008: 1,7 Wachstumsfaktor 2006 bis 2008: 0,5*1,7 = 0,85, das entspricht einer Wachstumsrate von –15% (1-0,85=0,15)



Aus dem Wort MISSISIPPI durch umordnen verschiedene beliebige Wörter bilden M=1x, I=4x,S=4x,P=2x

$$P_{mW} = \frac{11!}{1!4!4!2!} = \frac{39.916.800}{1.152} = 34.650$$

Einschluss-Ausschluss-Verfahren (s. Folien 22 und 28 Modul 7)

$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|$$

$25 = 14 + 10 + \mathbf{C} - 5 - 6 - 1 + 2$

$\mathbf{C} = 25 - 14 - 10 + 5 + 6 + 1 - 2 = 11$

a) Worüber informiert …
   a. … die Standardabweichung?
   b. … das Quartil Q3?
   c. … der Variationskoeffizient?
   d. … der Interquartilsabstand?
   e. … das 5%-Quantil?
   f. … der Median?
   g. … der Modus?
   h. … die Spannweite?
   i. … das Quartil Q1?

a) Über die (absolute Streuung der einer verteilung **b)** Welcher Wert von 75% unterschritten und 25% überschritten wird **c)** v informiert über die relative Streuung einer Verteilung **d)** Spannweite der mittleren 50% der Beobachtungswerte einer Verteilung **e)** Informiert über den Wert der von 5% unterschritten und 95% überschritten wird (der Beobachtungswerte **f)** X_z - 50% der Werte, Mittelerer Wert g) X_d - Wert der am häufigsten vorkommt **h)** Maximaler Abstand, größter Wert minus kleinster Wert **i)** Siehe Q3

A = {3; 4; 5; 6; 7; 8}
B = {-1; 0; 1; 2; 3; 4}
C = {1; 2; 3; 4; 6; 8; 12; 24}
D = {5; 10; 15; 20; 25; …}
E = {-4; -3; -2; -1; 0; 1; 2; 3; 4}

$A \cup B$={−1; 0; 1; 2; 3; 4; 5; 6; 7; 8}

$B \cap C$={1;2;3;4}

$(C/A)/B$={12;24}

$D \cup E$={−4; −3; −2; −1; 0; 1; 2; 3; 4; 5; 10; 15; 20; 25;…}

$A \cap D^c = A/D$={3;4;6;7;8}

Gegeben sind folgende Mengen A, B, C, D, E
A={$x \in \mathbb{N} \mid 3 \leq x \leq 8$}
B={$x \in \mathbb{Z} \mid -2 < x \leq 4$}
C={$x \in \mathbb{N} \mid x$ *ist Teiler von* 24}
D={$x \in \mathbb{N} \mid x$ *ist Vielfaches von* 5}
E={$x \in \mathbb{Z} \mid |x| < 5$}

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

6) In einer Region wird untersucht, wer von verheirateten Paaren regelmäßig in die Kirche geht. Es hat sich ergeben, dass 40 % der Männer und 50 % der Frauen regelmäßige Kirchgänger sind. Geht eine Frau in die Kirche, so beträgt die Wahrscheinlichkeit 0,3, dass ihr Mann auch hingeht.

Bedingte Wahrscheinlichkeit: Es seien M und F die Ereignisse, dass ein Mann bzw. eine Frau in die Kirche geht. Dann ist P(M) = 0,4 und P(F) = 0,5, außerdem P(M|F) = 0,3.

a) (3 Punkte) Die Wahrscheinlichkeit, dass Mann und Frau in die Kirche gehen, ist dann nach dem Multiplikationssatz P(M ∩ F) = P(M|F)*P(F) = 0,3·0,5 = 0,15.

b) (3 Punkte) Die Wahrscheinlichkeit, dass eine Frau in die Kirche geht, wenn ihr Mann dies auch tut, ist P(F|M) = P(M∩F)/P(M) = 0,15/0,4 = 0,375.

c) (4 Punkte) Die Wahrscheinlichkeit, dass wenigstens einer der beiden Ehepartner regelmäßiger Kirchgänger ist, ist nach dem Additionssatz P(M ∪ F) = P(M) + P(F) – P(M ∩ F) = 0,4 + 0,5 – 0,15 = 0,75.

Spannweite w: $w = x_{max} - x_{min}$

(Inter)Quartilsabstand: $Q_A = IQR = Q_3 - Q_1$

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

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

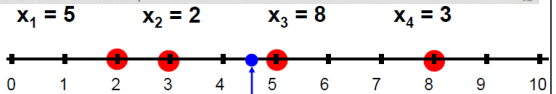
mittlere absolute Abweichung: $d = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$

(Das erste absolute zentrierte Moment gegenüber dem Mittelwert → wird oft zugunsten der Varianz umgangen, welche analytisch leichter zu behandeln ist)

Variationskoeffizient: $v = \frac{s}{\bar{x}}$

(dimensionslose Größe) 32

$x_1 = 5$ $x_2 = 2$ $x_3 = 8$ $x_4 = 3$



$\bar{x} = \frac{1}{4}\cdot(5+2+8+3) = \frac{18}{4} = 4{,}5$

Berechnung der Varianz

$s^2 = \frac{1}{4}\cdot((5-4{,}5)^2 + (2-4{,}5)^2 + (8-4{,}5)^2 + (3-4{,}5)^2) =$

$\frac{1}{4}\cdot(0{,}25 + 6{,}25 + 12{,}25 + 2{,}25) = \frac{21}{4} = 5{,}25$

i	x_i	h_i	$x_i\,h_i$	$\bar{x}$	$x_i - \bar{x}$	$(x_i-\bar{x})^2 h_i$
1	1	5	5	3,28	-2,28	25,992
2	2	8	16	3,28	-1,28	13,107
3	3	14	42	3,28	-0,28	1,098
4	4	16	64	3,28	0,72	8,294
5	5	5	25	3,28	1,72	14,792
6	6	2	12	3,28	2,72	14,797
Σ		50	164	$\bar{x}$=164/50=3,28		78,08

Varianzberechnung

$$s^2 = \frac{1}{50}\sum_{i=1}^{6}(x_i - \bar{x})^2 * h_i = \frac{78{,}08}{50} = 1{,}562$$

Berechnungstabelle mit Hilfsgrößen

Filiale Nr. i	x_i Verkaufsfläche (1000qm)	y_i Filialumsatz (Mio €)	$x_i \cdot y_i$	x_i^2	y_i^2
1	3	30	90	9	900
2	2	10	20	4	100
3	6	40	240	36	1.600
4	5	20	100	25	400
Summe	16	100	450	74	3.000

$$r_{xy} = \frac{(\frac{1}{n}\sum x_i \cdot y_i) - \bar{x}\cdot\bar{y}}{\sqrt{(\frac{1}{n}\sum x_i^2) - \bar{x}^2}\cdot\sqrt{(\frac{1}{n}\sum y_i^2) - \bar{y}^2}} = \frac{\frac{1}{4}\cdot 450 - 4\cdot 25}{\sqrt[2]{\frac{1}{4}\cdot 74 - 4^2}\cdot\sqrt[2]{\frac{1}{4}\cdot 3000 - 25^2}} = \frac{112{,}5 - 100}{\sqrt{2{,}5}\cdot\sqrt{125}} = \frac{12{,}5}{1{,}581\cdot 11{,}180} = \frac{12{,}5}{17{,}676} = +0{,}707$$

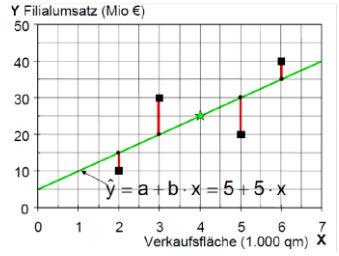
Korrelationsrechnung

$R^2 = (-0{,}95)^2 = 0{,}903$

90% der Varianz der erreichten Punkte lassen sich erklären durch die Varianz der Anzahl der verpassten Vorlesungen. Die restlichen 10% lassen sich nur durch andere Einflussfaktoren erklären.

Interpretation des Bestimmtheitsmaßes: nur 28% der Varianz der Produktionskosten werden erklärt durch die Varianz der Produktionsmengen, d.h., nur 28% der monatlichen Kostenunterschiede werden dadurch erklärt, dass die monatlichen Produktionsmengen unterschiedlich waren. Die restlichen 72% lassen sich nur durch andere Einflussgrößen erklären.

Regressionssrechnung



Modus, Median, arithmetisches Mittel, Schiefe

x_i	$h(x_i)$	$f(x_i)$ (%)	$H(x_i)$	$F(x_i)$ (%)
1	1	8,3%	1	8,3%
2	6	50,0%	7	58,3%
3	2	16,7%	9	75,0%
4	2	16,7%	11	91,7%
9	1	8,3%	12	100,0%
Summe	12	100,0%	-	-

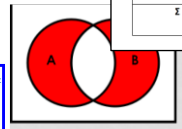
$\bar{x}_D = 2$ $\bar{x}_Z = 2$

$\bar{x} = \frac{1}{12}(1\cdot 1 + 2\cdot 6 + 3\cdot 2 + 4\cdot 2 + 9\cdot 1) = \frac{36}{12} = 3$

wegen $\bar{x}_Z < \bar{x}$ rechtsschiefe Verteilung

M / G	weiblich	männlich	Σ	
Produkt A	4	3	7	58,3%
Produkt B	2	3	5	41,7%
Σ	6	6	12	100%
	50%	50%		

M / G A	weiblich		männlich		Σ	
	bis unter 40 Jahre alt	40 Jahre und älter	bis unter 40 Jahre alt	40 Jahre und älter		
Produkt A	3	1	1	2	7	58,3 %
Produkt B	1	1	1	2	5	41,7 %
Σ	4	2	2	4	12	100 %
	33,3%	16,7%	16,7%	33,3%		



Symmetrische Differenz von A und B

Mengenoperationen:

∩	Schnittmenge
U	Vereinigung
\	Mengendifferenz
c	Komplementbildung

a) $\sum_{i=1}^{m} h(x_i) = n$

b) $\sum_{i=1}^{m} f(x_i) = 1$ bzw. 100%

c) $H(x_m) = n$, $F(x_m) = 1$ bzw. 100%

Beispiel 3.2:

Einer Urne mit 6 roten und 4 grünen Kugeln werden gleichzeitig 5 Kugeln entnommen.

Wie ist die Wahrscheinlichkeit, dass genau 2 der Kugeln rot sind?

n=10, n_1 = 6, n_2 = 4, k = 5

Schritt 1: (Ergebnisraum)

$$|\Omega| = K_{oW} = \binom{n}{k} = \binom{10}{5} = \frac{10!}{5!\,(10-5)!} = 252$$

Schritt 2: (Menge der günstigen Ereignissen: 2 Kugeln müssen aus der Menge der roten Kugeln und der Rest aus der Menge der grünen Kugeln stammen)

$$|E| = K_{oW} = \binom{6}{2}\binom{4}{3} = \frac{6!}{2!\,4!}\cdot\frac{4!}{3!\,1!} = 60$$

Schritt 3: (Wahrscheinlichkeit)

$$p = \frac{|E|}{|\Omega|} = \frac{60}{252} = \frac{15}{64} = 0{,}24$$

43

Filiale Nr. i	x_i Verkaufsfläche (1000qm)	y_i Filialumsatz (Mio €)	$x_i \cdot y_i$	x_i^2	y_i^2
1	3	30	90	9	900
2	2	10	20	4	100
3	6	40	240	36	1.600
4	5	20	100	25	400
Summe	16	100	450	74	3.000

$$a = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{74\cdot 100 - 16\cdot 450}{4\cdot 74 - 16^2} = \frac{7400 - 7200}{296 - 256} = \frac{200}{40} = 5$$

$$b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{4\cdot 450 - 16\cdot 100}{4\cdot 74 - 16^2} = \frac{1800 - 1600}{296 - 256} = \frac{200}{40} = 5$$

Klassen-Nr. i	Größen-klassen (cm)	h_i	f_i (%)	H_i	F_i (%)
1	100 b.u. 150	40	40%	40	40%
2	150 b.u. 170	40	40%	80	80%
3	170 b.u. 200	20	20%	100	100%
Summe		100	100%	-	-

Einfallsklasse: k = 2

$$\bar{x}_Z = x^*_{k-1} + (x^*_k - x^*_{k-1})\cdot\frac{(0{,}5 - F_{k-1})}{f_k}$$

$$\bar{x}_Z = 150 + (170 - 150)\cdot\frac{(0{,}5 - 0{,}4)}{0{,}4} = 150 + 20\cdot(\frac{0{,}1}{0{,}4}) = 155\ (\text{cm})$$

Klasse Nr. i	Umsatzklasse (Mio €)	Anzahl Unternehmen h_i	Anteil f_i (%)	H_i	F_i (%)	Klassenmitte m_i
1	0 b.u. 1	60	30%	60	30%	0,5
2	1 b.u.2	80	40%	140	70%	1,5
3	2 b.u.5	40	20%	180	90%	3,5
4	5 b.u.10	10	5%	190	95%	7,5
5	10 b.u.20	10	5%	200	100%	15
Σ		200	100%			

Varianzberechnung bei klassierten Daten

Merkmalsausprägung und zugehörige Häufi

$x_j \rightarrow h(x_j)$ j = 1, … , m — absolute Häufigkeit

$x_j \rightarrow f(x_j) = \frac{h(x_j)}{n}$ j = 1, … , m

$f(x_j) = \frac{h(x_j)}{n} \cdot 100\ (\%)$ — relative Häufigkeit

Beispiel 1.1:

Wie viele "Wörter" (auch unsinnige) können aus 5 Buchstaben zustande kommen aus einem Alphabet vom Umfang 26?

n = 26, k = 5 (k < n → Variationen mit Wiederholung)

$$|\Omega| = V_{mW} = n^k = 26^5 = 11.881.376$$

Beispiel 2.1:

Auf wie viele Arten können sich 5 Personen auf 5 freie (unterscheidbare) Plätze verteilen?

n = 5, k = 5 (n = k → Variation ohne Wiederholung)

$$|\Omega| = V_{oW} = n! = 5! = 120\ (Arten)$$

Beispiel 3.1:

Experiment: gleichzeitiges Ziehen von 2 Kugeln aus der Urne U_6 (Urne mit 6 Kugeln) ohne Zurücklegen.

Wie viele Paare sind möglich wenn man die Kugeln durchnummeriert?

Ergebnisraum: **Ω = {{1,2}, {1,3}, …., {5,6}}**

Mächtigkeit des Ergebnisraumes:

$$|\Omega| = K_{oW} = \binom{n}{k} = \frac{n!}{k!\,(n-k)!} = \binom{6}{2} = \frac{6!}{2!\,(6-2)!} = \frac{6!}{2!*4!} = \frac{6*5*4*3*2}{2*4*3*2} = 15$$

Beispiel 4.1:

10 Sportlerinnen nehmen an 3 Wettbewerben teil, bei denen es jeweils genau eine Siegerin gibt. Auf wie viele Arten können die Preise verteilt werden?

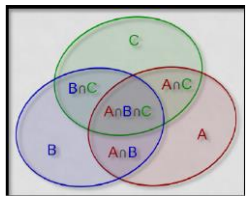
n = 10, k = 3

$$|\Omega| = K_{mW} = \binom{n+k-1}{k} = \frac{(n+k-1)!}{k!\,(n-1)!} = \frac{12!}{3!\,9!} = 220$$

Interpretation des Bestimmtheitsmaßes: nur 28% der Varianz der Produktionskosten werden erklärt durch die Varianz der Produktionsmengen, d.h., nur 28% der monatlichen Kostenunterschiede werden dadurch erklärt, dass die monatlichen Produktionsmengen unterschiedlich waren. Die restlichen 72%

Die grafische Darstellung der 5-Punkte-Zusammenfassung heißt **Box-and-Whisker-Plot**

Die 5-Punkte-Zusammenfassung besteht aus: Minimum, Q1, Median, Q3, Maximum



Gegenwahrscheinlichkeit Bsp.: 1-5/6*5/6 bei 2 Würfen genau einmal die 5

Beispiel 2.3:

Wie ist die Wahrscheinlichkeit, beim viermaligen Werfen eines Würfels lauter verschiedene Augenzahlen zu erzielen?

n = 6, k = 4

Schritt 1: (Ergebnisraum)

$$|\Omega| = V_{mW} = n^k = 6^4 = 1.296$$

Schritt 2: (Menge der günstigen Ereignissen)

$$|E| = V_{oW} = \binom{n}{k}k! = \frac{n!}{(n-k)!} = \frac{6!}{(6-4)!} = 6*5*4*3 = 360$$

Schritt 3: (Wahrscheinlichkeit)

$$p = \frac{|E|}{|\Omega|} = \frac{360}{1.296} = \frac{5}{18} = 0{,}278$$

Sei **A** einer **n**-elementige Menge **A = {1,…,n}**, aus der man nacheinander **k** Elemente auswählt (**k** Ziehungen). Dabei unterscheidet man mit und ohne Zurücklegen und mit und ohne Berücksichtigung der Reihenfolge (Anordnung).

Es ergeben sich **vier Grundmuster der Kombinatorik**:

1. mit Reihenfolge, mit Zurücklegen
$$|\Omega| = V_{mW} = n^k$$
2. mit Reihenfolge, ohne Zurücklegen
$$k < n: \ |\Omega| = V_{oW} = \binom{n}{k}k! = \frac{n!}{(n-k)!}$$
$$k = n: \ |\Omega| = V_{oW} = n!$$
3. ohne Reihenfolge, ohne Zurücklegen (Binomialkoeffizient)
$$|\Omega| = K_{oW} = \binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$
4. ohne Reihenfolge, mit Zurücklegen
$$|\Omega| = K_{mW} = \binom{n+k-1}{k} = \frac{(n+k-1)!}{k!\,(n-1)!}$$

Für zwei endliche disjunkte Mengen A und B gilt (**Summenregel**):

$$|\mathbf{A \cup B}| = |\mathbf{A}| + |\mathbf{B}| - |\mathbf{A \cap B}|$$

Beispiel:

100 Studenten haben u.a. Kurse A, B und C belegt. 65 Studenten haben Kurs A belegt, 32 Kurs B, 18 Kurs C, 15 A und B, 9 B und C, 7 A und C, 3 haben alle drei Kurse belegt.

Wie viele Studenten haben keinen der Kurse A, B oder C belegt.

(Anzahl Studenten ohne Kurse A, B, C) =

100 – (65 + 32 + 18 – 15 – 9 – 7 + 3) = 100 – 87 = 13

Noch allgemeiner gilt die **Summenregel** für drei Mengen:

$$|\mathbf{A \cup B \cup C}| = |\mathbf{A}| + |\mathbf{B}| + |\mathbf{C}| - |\mathbf{A \cap B}| - |\mathbf{A \cap C}| - |\mathbf{B \cap C}| + |\mathbf{A \cap B \cap C}|$$

P(+|krank)=0,95
P(-|krank)=0,05
P(+|gesund)=0,10
P(-|gesund)=0,90
P(krank)=0,03, daraus folgt, dass P(gesund)=0,97 ist.

(Formel s. Modul 7 Folien 46, 48-49)

$$P(krank|+) := \frac{P(krank \cap +)}{P(+)} = \frac{P(krank) * P(+|krank)}{P(krank) * P(+|krank) + P(gesund) * P(+|gesund)}$$
$$= \frac{0{,}03 * 0{,}95}{0{,}03 * 0{,}95 + 0{,}97 * 0{,}10} = \frac{0{,}0285}{0{,}0285 + 0{,}097} = \frac{0{,}0285}{0{,}1255} = 0{,}227$$

Eine „**Statistik**“ ist eine systematische Zusammenstellung von Zahlen und Daten zur Beschreibung von Zuständen, Entwicklungen und Phänomen.
Beispiele: Häufigkeitsverteilungen, Zeitreihenvergleich und -analyse, Zusammenhangs- und Abhängigkeitsanalysen, statistische Kennzahlen zur Beschreibung von Verteilungen, Zusammenhängen.
Im Rahmen der "**Beschreibende Statistik**"(deskriptive Statistik) sammelt man Daten bei allen Untersuchungseinheiten, über die man Informationen erhalten will. Die beschreibende Statistik hat zum **Ziel**, empirische Daten durch Tabellen, Kennzahlen (auch: Maßzahlen oder Parameter) und Grafiken übersichtlich darzustellen und zu ordnen. Dies ist vor allem bei umfangreichem Datenmaterial sinnvoll, da dieses nicht leicht überblickt werden kann.
Im Rahmen der "**Schließenden Statistik**" (induktive Statistik) wählt man aus der Grundgesamtheit, über die man Informationen haben will, eine Teilmenge = Stichprobe aus. Die **schließende Statistik** wird zu einem wesentlichen Teil zum Beweis oder zur Widerlegung von vorher aufgestellten Behauptungen, den Hypothesen, die sich auf definierte Grundgesamtheit bezieht, eingesetzt. Nur bei den Einheiten der Stichprobe erhebt man Daten, die man dann mit statistischen Methoden auswertet. Von den Stichprobenergebnissen versucht man, auf die Eigenschaften der Grundgesamtheit zu schließen.
Bei einer Teilerhebung muss man Entscheidungen fällen über den Stichprobenumfang und das Auswahlverfahren. Welche Entscheidung ist wichtiger?
Auswahlverfahren ist wesentlich wichtiger. Schlechte Stichprobe kann noch so groß sein, liefert keine brauchbaren Erkenntnisse über die Grundgesamtheit.
Bsp.: Grundgesamtheit: Einwohner in DE (ca, 82Mio), Stichprobe: Studierende in DE (ca. 2 Mio) ➔ Stichprobe verzerrt und nicht repräsentativ.
willkürliche Auswahl: (=Auswahl aufs Geratewohl oder convenicene sample) gibt es keinen Auswahlplan. Die Interviewer sind frei in der Auswahl ihrer Interviewpartner. Daher suchen sie sich die Personen aus, die für sie am bequemsten zu erreichen sind. Das führt meist zu einer verzerrten Stichprobe.
Bei einer "**zufälligen Auswahl**" ist die Auswahl zufallsgesteuert, d.h., jede Einheit der Grundgesamtheit (über die man Informationen erhalten will) muss mit **gleicher Wahrscheinlichkeit** in die **Stichprobe** gelangen können. Dies setzt voraus, dass eine Liste/Datei aller Einheiten der Grundgesamtheit vorliegt. Die Interviewer sind nicht frei in der Auswahl ihrer Interviewpartner. Sie bekommen feste Zielpersonen vorgegeben. Nur bei der Zufallsauswahl (=**Random-Auswahl**) lässt sich mit Hilfe der Wahrscheinlichkeitsrechnung der Stichprobenfehler berechnen.
Quota-Auswahl: bewusstest Auswahlverfah-

Bei der **TED-Umfrage** im Fernsehen liegt eine **willkürliche Auswahl** (der Bewohner eines Landes) vor. Es gibt keinerlei Auswahlplan, jeder kann, wenn er Lust hat sich an der Umfrage beteiligen. Es gibt Personen, die grundsätzlich an einer solchen Umfrage niemals teilnehmen würden, andere versuchen mehrmals ihre Meinung zu äußern.
Welche zwei Probleme hat man bei der Klassierung von Daten? 1. Übersichtlichkeit – Informationsverlust, für klassierte Daten können keine exakten statistischen Kennzahlen (z.B. Mittelwerte) berechnet werden. Näherungswerte können nur unter bestimmten Annahmen (z.B. Gleichverteilung in den Klassen) berechnet werden.
offenen Randklassen“ = entweder die Klassenuntergrenze nicht angegeben (b.u. 100kg) oder Klassenobergrenze icht angegeben (200kg und schwerer), Klassenbreite und Klassenmitte kann nicht berechnet werden. **b.u. = bis unter**
quantitatives Merkmal = Zahl messbar, zählbar (metrisch)
stetige Merkmale = Menge der Merkmalsausprägung überabzählbar, Intervall der reelen Zahlen (es gibt zwischen zwei Ausprägungen immer noch weitere Zwischenwerte z.B.: Gewicht, Alter, Fahrzeit, Eink.satz, TempInKelvin
diskrete Merkmale = Menge der Merkmalsausprägungen endlich bzw. abzählbar (i.d.R. ganze Zahlen) z.B.: Kinderzahl, Sitzplätze, monatliches Gehalt, Steuerklasse, Geschlecht, soziale Schicht, Schulnote, Klausurpunkte, Einwohnerzahl, Semesterzahl, Handelsklasse (Obst)
qualitatives Merkmal = lassen sich nicht messen
Nominalskalen: Hersteller, Wagenfarbe, Unternehmensrechtsform, Wohnort, Beruf, Steuerklasse, Geschlecht, soziale schicht,
Ordinalskalen bzw. Rangskalen: Umsatzklassen eines Unternehmens, Einkommensklasse(steuersatz), Kundenzufriedenheit (von 1 bis 5), Körpergewicht, Schulnote (1-6), Handelsklasse (z.B.: Obst)
Verhältnisskalen: Geschwindigkeit eines Fahrzeuges, Produktpreis, Unternehmensumsatz, Hubraumgröße in ccm, Kraftstoffverbrauch, Beschleunigung 0auf100, Kaufpreis,TempInKelvin
Intervallskalen: Semesterzahl, Geburtsjahr, Erstzulassung, Temperatur in celcius
Absolutskala: Motorleistung in PS, Zylinderzahl, Mitarbeiterzahl, Fachsemesterzahl, Windstärke m/Sek,Klausurpunkte, Einwohnerzahl

Zur 1.9: Angenommen die Würfel sind weiß und rot. Ergebnis des weißen Würfels ist von dem des roten Würfels unabhängig.
P (Augensumme=4) = P ((rot=1 und weiß=3) oder (rot=2 und weiß=2) oder (rot=3 und weiß=1)) =
= P(rot=1) * P(weiß=3) + P(rot=2) * P(weiß=2) + P(rot=3) * P(weiß=1) =
= 1/6*1/6 + 1/6*1/6 + 1/6*1/6 = 3/36 = 1/12
Zur 1.10: Ziehen mit Zurücklegen, 1. Pfadregel. 3.a)

i	Merkmalsausprägung X_i	$h(x_i)$	$f(x_i)$ (%)
1	1	1	8,3%
2	2	6	50,0%
3	3	2	16,7%
4	4	2	16,7%
5	9	1	8,3%
	Summe	12	100,0%

b)
X_D=2 X_Z=2 X=1/12x(1x1+2x6+..=36/12

Form der Verteilung wg. $\bar{x}_7 < \bar{x}$: rechtsschief

c) w=9-1=8

$$s^2 = \frac{1}{12}\cdot((1-3)^2\cdot 1+(2-3)^2\cdot 6+(3-3)^2\cdot 2+(4-3)^2\cdot 2+(9-3)^2\cdot 1) = \frac{1}{12}\cdot(4+6+0+2+36) = 4$$

d) Der Wert der 13. Person kann beliebig groß sein, ohne dass sicher der oben berechnete Median bei Hinzunahme ändert. Das arithmetische Mittel ändert sich nur dann nicht, wenn der Wert der 13. Person 3 (=x) ist. In allen anderen Fällen verändert sich bei Hinzunahme das arithmetische Mittel. **4.a)**

i	x_i	Y_i	x_i^2	x_iY_i	Y_i^2
1	16	30	256	480	900
2	2	88	4	176	7.744
3	3	65	9	195	4.225
4	8	50	64	400	2.500
5	0	90	0	0	8.100
Σ	29	323	333	1.251	23.469

$$a = \frac{333*323-29*1.251}{5*333-29^2} = \frac{107.559-36.272}{1.665-841} = \frac{71.280}{824} = 86{,}504$$

$$b = \frac{5*1.251-29*323}{824} = \frac{6.255-9.367}{824} = \frac{-3.112}{824} = -3{,}7767$$

Regressionsfunktion $\hat{Y} = \hat{Y}(x) = a + b*x = 86,50 – 3,78*x$ 4.b) Zeichnung **c)** a = 86,50 Punkte möglich unabhängig von der Anzahl verpasster Vorlesungen b = – 3,78 variabler Abstieg/Rückfall der Punktezahl, mit jeder verpassten Vorlesung sinkt die Punktezahl um ca. 3 bis 4 Punkte. **d)** Prognose: $\hat{Y}(0)$ = 86,50 – 3,78*0 = 86,50 (Punkte)
Prognose: $\hat{Y}(1)$ = 86,50 – 3,78*1 = 82,72 (Punkte) **e)**

$$r = \frac{\frac{1}{5}*1.251-5,8*64,6}{\sqrt{\frac{1}{5}*333-5,8^2}*\sqrt{\frac{1}{5}*23.469-64,6^2}} = \frac{}{\sqrt{}} = -0,95$$

Sehr starke/s negative/s Korrelation / Zusammenhang

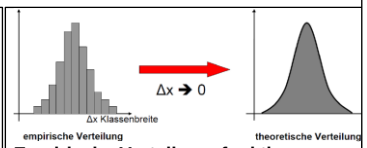
GRUNDBEGRIFFE DER STATISTIK
Merkmalsträger = Einzelnes Objekt einer statistischen Untersuchung, Träger der Informationen, für die man sich interessiert.
Statistische Masse = Menge aller Merkmalsträger, die mit dem Untersuchungsziel in Verbindung stehen,
Merkmal = Im Rahmen der statistischen Erhebung relevante Eigenschaften der Merkmalsträger ➔ Statistische Variable
Merkmalsausprägung = Grundsätzlich mögliche Ausformungen eines Merkmals ➔ Wert der Variable, Beobachtungswert
Operationalisierung“ definiert, wie man den Begriff konkret misst. Die Operationalisierung ist besonders wichtig bei Begriffen ohne direkten empirischen Bezug (so genannte „**latente Variable**“), z.B. Kundenzufriedenheit, Teamfähigkeit, Intelligenz, Werbewirkung **Bsp**.: : Intelligenz kann operational durch die Anzahl der Lösungen von Intelligenzaufgaben in einem konkreten Intelligenztest definiert werden.
Ein Hersteller von Schokoladenwaren möchte Informationen über die Verbrauchsgewohnheiten von Jugendlichen in Süddeutschland haben. Potentielle Untersuchungseinheiten sind hier Menschen. Die Zielsetzung erfordert folgende sachliche Abgrenzung: Jugendliche (Begriff ist zu operationalisieren, was verstehen wir unter „Jugendlichen“, eine mögliche Definition: „Personen im Alter zwischen 10 und 18 Jahren“.) Räumliche Abgrenzung: Süddeutschland (Begriff ist zu operationalisieren: z.B. Bayern + Baden-Württemberg). Zeitliche Abgrenzung: Hierzu gibt es in der Aufgabe keine Hinweise, denkbar wäre Jahresebene.

Datendokumentation

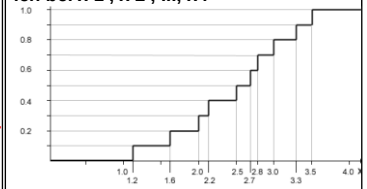
Einzelwerte (Einzelbeobachtungen)➔ ungeordnete Reihe (Urliste ,Rohdaten, Primärdaten
➔ Urliste im Bereich = Ergebnis einer **Datenerhebung.**
Schritt a: Sortieren
Schritt b: Verdichten in Häufigkeitsverteilung
Schritt c: Darstellen als eine Häufigkeitstabelle

j	x_j	Anzahl $h(x_j)$	Anteil $f(x_j)$	Anteil in % $f(x_j)$ (%)
1	ledig	6	0,30	30
2	verheiratet	9	0,45	45
3	geschieden	2	0,10	10
4	verwitwet	3	0,15	15
	Summe	20	1,00	100

Bsp für Geschlecht x Artikel (erstgenanntes nach rechts zweitgenanntes nach unten

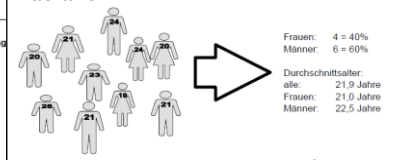


Empirische Verteilungsfunktion
Die empirische Verteilungsfunktion F(x) gibt für jede beliebige reelle **Zahl x** den Anteil der Merkmalsträger an, für die das **Merkmal X** einen **Wert x** i annimmt, der kleiner oder gleich **x** ist
Wertebereich: 0 ≤ F(x) ≤ 1, F(x) ist eine Treppenfunktion mit Sprungstellen bei x 1 , x 2 , ..., x i



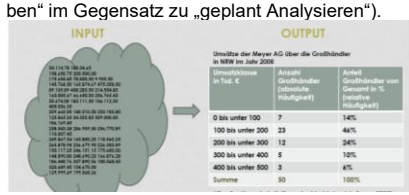
Die Abbildung zeigt die empirische Verteilungsfunktion für das Merkmal Abiturnoten. Greift man auf der x Achse den Wert 3 heraus, so lässt sich der dazugehörige y Wert 0.8 wie folgt interpretieren: 80 % der Abiturienten haben im schlechtesten Fall den Notendurchschnitt 3 bekommen.
EIGENSCHAFTEN DER HÄUFIGKEITSVERTEILUNGEN
Lage (mehrere Berge nebeneinander)**, Streuung = Wölbung, Schiefe**
GRAFISCHE DARSTELLUNG DER HÄUFIGKEITSVERTEILUNG
Pro: ein anschauliches Bild der Daten
Ziel : das Wesentliche der Verteilung aufzuzeigen, Achtung, Manipulationen sind denkbar,
Bsp. Für grafische Darstellungsformen:
Säulendiagramm, (geeignet. für wenige Ausprägungen) Stabdiagramm(SD mit sehr schmalen Säulen), Balkendiagramm (Darstellung von Rangfolgen), Kreisdiagramm (max 7 Teilwerte, sonst unübersichtlich, keine Nullwerte/negativen Werte), **Histogramm**(Verteilung von klassifizierten Daten)
abhängige/Unabhängige Variablen
Diese Variable verändert sich in Abhängigkeit von einer oder mehreren unabhängigen Variablen. Sie wird auch Reaktionsvariable (endogene Variable) genannt, weil sie eine Reaktion auf Veränderungen der unabhängigen (exogenen) Variable aufzeigt. Beispiel: In einem Experiment wird in einem Raum die Temperatur verändert. Personen in diesem Raum geben an, wie wohl sie sich bei den unterschiedlichen Temperaturen fühlen. Die Raumtemperatur ist hier die unabhängige Variable. Die abhängige Variable (die Reaktionsvariable) ist das angegebene Wohlbefinden der Befragten.

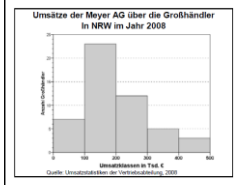
Deskriptive (beschreibende Statistik= Betrachtung der Daten an sich. Datencharakterisierung (siehe links). Die gewonnene Daten werden verdichtet bzw. so dargestellt, dass das Wesentliche deutlich hervortritt. **Darstellungsformen**: Tabellen, grafische Darstellungen, charakteristische Maßzahlen



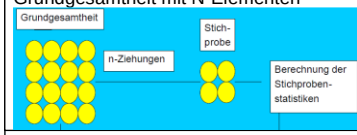
induktive (schließende) Statistik ➔ Der Schluss vom Teil aufs Ganze Wahrscheinlichkeitsrechnung, weitere Synonyme: analytische oder inferentielle Statistik.

dient dazu, aus den erhobenen Fakten Schlüsse auf die Ursachenkomplexe zu ziehen, die zu diesen Daten geführt haben.
Die Einteilung in **deskriptive** und **induktive Statistik** wurde verwendet, um die unterschiedliche Zielsetzung der in diesen beiden Bereichen verwendeten Methoden herauszustellen („Beschreiben“ im Gegensatz zu „geplant Analysieren“).





FachTerminologie:
Untersuchungseinheit = statistische Einheiten (Träger der Information = z.B.: *einzelne Großhändler, einzelne Person*
Grundgesamtheit = Menge aller relevanten Einheiten (***alle*** *Großhändler der Meyer AG in NRW im Jahre 2008)*
Merkmal = Variable = *Umsatz im Jahre 2008*
Merkmalsausprägung = mögliche Ausformung eines Merkmales *z.B.: AG, KG, OHG*
Merkmalswert:
Umfang einer Gesamtheit = Anzahl ihrer Einheiten (Elemente) = *Anzahl der Großhändler*
Information der Tabelle = (klassierte) Häufigkeitsverteilung mit absoluten und relative (Klassen-) Häufigkeiten
Stichprobe = eine n-elementige Teilmenge der Grundgesamtheit mit N-Elementen



Ein **Merkmal X** mit m **Merkmalsausprägungen** (x1, …, xm) wird bei n Untersuchungseinheiten gemessen. Beantworten Sie für die entsprechende Häufigkeitsverteilung die folgenden Fragen:

a) Wie groß ist $\sum_{i=1}^{m} h(x_i)$

b) Wie groß ist $\sum_{i=1}^{m} f(x_i)$

c) Welchen Wert hat H(x_m), F(x_m)?

a) $\sum_{i=1}^{m} h(x_i)$=n

b) $\sum_{i=1}^{m} f(x_i)$=1 bzw. 100%

c) H(x_m)=n, F(x_m)=1 bzw. 100%

Lageparameter

Lageparameter beschreiben die "Lage " der Elemente der Grundgesamtheit bzw. der Stichprobe in Bezug auf die Messskala.

- $\bar{x}_D$ ▪ **Modus**
- $\bar{x}_Z$ ▪ **Median**
- $\bar{x}$ ▪ **arithmetisches Mittel**
- $\tilde{x}_p$ ▪ **Quantil**

Modus = ist die am häufigsten auftretende **Merkmalsausprägung** (**maximale Häufigkeit**). Bei klassierten Werten ist er die Mitte der Klasse mit den größten Häufigkeiten. Mehrere maximale Häufigkeiten = Multimodale Verteilung

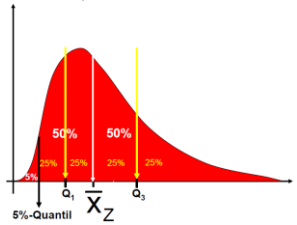
Median (Zentralwert) (Zentralwert)
der Wert, der genau in der Mitte liegt. ➔ Zahlen sortieren, bei ungerader Anzahl ist es genau die Mitte, bei gerader bildet man den **Mittelwert** der beiden inneren Zahlen, dies ist der **Median**. Bei den Zeugnisnoten 1 2 3 4 5 6 existiert kein Median, denn 3,5 alsZeugnisnote ist nicht üblich Aber : 1 2 3 3 4 5 hat den **Median 3**. Für metrische Daten in Klassen , kann die exakte Merkmalsausprägung des Medians nicht bestimmt werden ➔ **Näherungswerte für Median** wobei k= Einfallsklasse (Klasse mit F(x) = 50%) ➔ siehe Tabelle oben. Der Median beschreibt die Verteilung besser als der Mittelwert, Ausreißer haben auf den Median keinen Einfluss.
Arithmetisches Mittel
Berechnung über die relative Häufigkeit: einzelne Verteilung durch n | einzelne Häufigkeiten berechnen (**n** = **gesamtanzahl**), mit Werten Multiplizieren und addieren. ➔ Note 1, 5 Schüler = 5/n und anschließend x*1 + x*2 +……
ARITHMETISCHES MITTEL BEI KLASSIERTEN DATEN
Berechnung mithilfe der Klassenmitte, diese bestimmen und entweder mit absoluten Häufigkeiten rechnen oder mit relativen (vorhin beschrieben)
Klassenmitte = unteres Ende + oberes Ende durch 2
f)
$R2=(-0,95)2=0,903$

90% der Varianz der erreichten Punkte lassen sich erklären durch die Varianz der Anzahl der verpassten Vorlesungen. Die restlichen 10% lassen sich nur durch ander Einflussfaktoren erklären.

NEIGUNG / SCHIEFE
Rechtsschiefe (linkssteile) Häufigkeitsverteilung:
Modus < Median < arithmetisches Mittel
es gibt Ausreißer im rechten Bereich.
linksschiefe (rechtssteile) Häufigkeitsverteilung:
Modus > Median > arithmetisches Mittel
Ausreißer im linken Bereich
unimodale symmetrische Häufigkeitsverteilung:
Modus ≈ Median ≈ arithmetisches Mittel

QUANTILE = Lagemaß in der Statistk, Teilen

eine Verteilung in Abschnitte gleicher Häufigkeit
Quartilabstand = gibt die Breite des mittleren Bereichs an, in dem ca. 50% der Werte liegen.



Problem der Lageparameter:
Keinerlei Indikator für die Streuung der Daten. Das arithmetische Mittel (der Durchschnitt) und auch der Median verdecken oft eine große Ungleichheit.

STREUUNGSPARAMETER

Die Statistik bietet Möglichkeiten, die Streuung näher zu untersuchen und mit Hilfe derStreuungsparametern die Streuung zu beschreiben. Der Quartilabstand wird als **Streuungsmaß** verwendet

Spannweite (Variationsbreite) **W =** Ausdehnung der Werte (Maß für die Breite des Streubereichs einer Häufigkeitsverteilung)

Zusammenfassung:
Der Median teilt einen nach Größe sortierten Datensatz in der Mitte □ links und rechts vom Median liegen gleich viele Beobachtungswerte. Unterteilt man die linke und die rechte Hälfte nach gleicher Vorschrift, wie man den Median bestimmt, so erhält man 4 gleich große Bereiche, die durch drei Quartils aufgeteilt werden.
25 % aller geordneten **Beobachtungswerte** sind kleiner als das 1.
50 % aller geordneten **Beobachtungswerte** sind kleiner als das 2.
75 % aller geordneten **Beobachtungswerte** sind kleiner als das 3.
Zwischen dem 1. und 3. Quartil liegen 50 % aller **Beobachtungswerte** Dieser Bereich wird auch **Quartilsabstand** genannt.
MODUS BEI KLASSIERTEN DATEN
Modus normal bestimmen und dann die Klassenmitte bestimmen, Ergebnis ist der gefragte Modus
MEDIAN BEI KLASSIERTEN DATEN
X(klein-K) = Einfallsklasse = Modus der klassierten Daten

$$\bar{x}_Z = x^*_{k-1} + (x^*_k - x^*_{k-1}) \cdot \frac{(0,5 - F_{k-1})}{f_k}$$

VARIANZ

In der beschreibenden Statistik nennt man das arithmetische Mittel der Abweichungsquadrate die Varianz wichtiger Streuungsparameter
Voraussetzung : metrisches Merkmal
Ausgangswert für weitere folgende Streuungsparameter
•Standardabweichung
•Variationskoeffizient
VARIANZ
Handelt es sich bei den zu untersuchendenD aten um die Grundgesamtheit (Population), dann wird mit 1/n gewichtet:

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

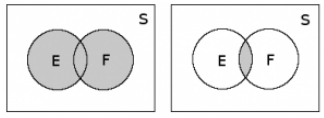
Wird hingegen eine Stichprobe (Teil einer Population) so wird mit 1 /(n-1) gewichtet:

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

Varianz berechnen:
(Wert der Variable - Durschnitt ale Werte) ins Quadrat * absolute oder relative Häufigkeiten

STANDARDABWEICHUNG (Streuung)

Die Standardabweichung ist ein Maß dafür, wie hoch die Aussagekraft des Mittelwertes ist . Eine kleine Standardabweichung bedeutet, alle Beobachtungswerte liegen nahe am Mittelwert (kleine Streuung).
bei normalverteilten Daten liegen ca. 95% der Beobachtungswerte im Intervall $[x-2s,\ x+2s]$
Wichtige Begriffe
Der **Variationskoeffizient v** informiert über die relative Streuung einer Verteilung.
Das **5%-Quantil** informiert über den Wert, der von 5% der Beobachtungswerte einer Verteilung unterschritten wird und von 95% der Beobachtungswerte überschritten wird.
Varianz, **Standardabweichung** und **Variationskoeffizient** sind **Streuungsparameter**. Mit ihnen lässt sich die **Streuung** einer **Häufigkeitsverteilung** charakterisieren. Varianz und Standardabweichung informieren über die **absolute Streuung**, der **Variationskoeffizient** über die relative Streuung.
Aufgabe: Sie lesen in einer Studie über die Einkommensverteilung einer Berufsgruppe: $\bar{x}Z$ = 30.000 €, $\bar{x}$ = 40.000 €, Q1 = 25.000 €, Q3 = 45.000 €. Welche Informationen erhalten Sie aus diesen 4 statistischen Kennzahlen über die Einkommensverteilung? Erhalten Sie auch Informationen über die Streuung der Verteilung?
Antwort Die Einkommensverteilung ist rechtsschief, da $\bar{x}Z < \bar{x}$.
Es gibt Ausreißer im oberen Einkommensbereich. 25% der Personen verdienen weniger und 75% mehr als 25.000 €. 75% verdienen weniger und 25% mehr als 45.000 €. Ein Einkommen von 30.000 € wird von 50% über- und von den anderen 50% unterschritten. Der Interquartilsabstand IQR = Q3 - Q1 = 20.000 € gibt die Spannweite bei den mittleren 50% an und informiert somit über die Streuung des "mittleren" Teils der Einkommensverteilung.

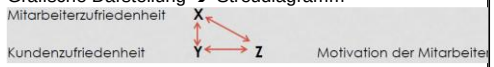


Korrelation und Regression

univariate Analyse: Merkmale werden einzeln ausgewertet
multivariate Analyse: Merkmale werden gemeinsam ausgewertet (Analyse von Zusammenhängen)
Korrelationsanalyse: ob 2 Variablen linear zusammenhängen und Stärke
Rangkorrelationsanalyse:
Zusammenhang zweier ordinalskalierter Merkmale mit Hilfe von Rangzahlen. Praktische Bedeutung da einfache Berechnung.
Kontingenzanalyse: Zusammenhangsanalyse (auf Basis einer Häufigkeitstabelle, je Stärker der Unterschied zwischen den Häufigkeiten, desto gößer ist Zusammenhang bzw. die Abhängigkeit zwischen den Merkmalen

2 Arten von Analysen.

Zusammenhangsanalyse(Interdependanzanalyse):
Wechselwirkung der Variablen untereinander wird untersucht, Zusammenhangsmaß oder auch Assoziativmaß gibt in der Statistik die Stärke und gegebenenfalls die Richtung des Zusammenhang
Zusammenhangsanalyse zwischen zwei metrischen Merkmalen X und Y
Zusammenhangs analyse ➔ Korrelationsanalyse
Zusammenhangs maß ➔ Korrelationskoeffizient 1 ≤ r xy ≤ +1
Grafische Darstellung ➔ Streudiagramm



Abhängigkeitsanalyse (Dependenzanalyse) unterschieden zwischen unabhängigen und abhängigen Merkmalen, Anwender weiß oder vermutet welche Merkmale auf andere Merkmale einwirken (können) .
Abhängigkeitsanalyse zwischen zwei metrischen Merkmalen X und Y
Abhängigkeits analyse ➔ Regressionsanalyse
Abhängigkeits maß ➔ Regressionsfunktion $\hat{y} = a + b*x$
Grafische Darstellung ➔ Regressionsgerade in Streudiagramm

Rangkorrelationskoeffizient: Maß für die Stärke des Zusammenhangs zweier ordinalskalierter Merkmale
Kontingenzkoeffizient: ein Maß für die Stärke des Zusammenhangs von min. zweier nominaler oder ordinaler Merkmale. Er basiert auf dem Vergleich von tatsächlich ermittelten Häufigkeiten zweier Merkmale
Phi-Koeffizient: Maß für Stärke des Zusammenhangs dichotomer Merkmale
Streudiagramm (oder Streuungsdiagramm): Ein Streudiagramm ist die graphische Darstellung von beobachteten Wertepaaren zweier Merkmale.
Aus Lage und Form der dargestellten Punktwolke lassen sich die Stärke und die Richtung des Zusammenhangs der Merkmale ablesen. Das Streudiagramm liefert erste Hinweise über eine mögliche Abhängigkeit zwischen Merkmalen.
Korrelation ➔ zahlenmäßiger statistischer Zusammenhang zwischen zwei Merkmalen X und Y.
positive Korrelation: beide Merkmale entwickeln sich gleichförmig
negative Korrelation: beide Merkmale entwickeln sich gegenläufig
kausaler Zusammenhang: zwischen den Merkmalen existiert eine Ursache-Wirkung-Beziehung, Veränderung von abhängigem Merkmal Y eindeutig auf Veränderung von X zurückzuführen.
Eine Korrelation sagt nichts über einen kausalen Zusammenhang aus und auch nicht über eine Kausalitätsrichtung.
Scheinkorrelation = kein kausaler Zusammenhang

KORRELATIONSKOEFFIZIENT

statistische Kenzahl die über die **Stärke** und die **Richtung** des linearen Zusammenhangs informiert ➔ dimensionslos
Korrelationskoeffizient r = 0 ➔ kein linearer Zusammenhang
r>0: x=hoch, y=hoch; **r<0**: x=hoch, y=niedrig
r=-1: extrem starker negative linearer Zusammenhang (Punktewolke liegen auf einer Geraden mit negative Steigung, sehr starke negative Korrelation / Zusammenhang
Bestimmtheitsmaß = Bei einer einfachen linearen Regression (nur eine unabhängige Variable) entspricht das **Bestimmtheitsmaß** dem Quadrat des Korrelationskoeffizienten **r** nach Pearson
R^2 = 0,903 = 90,3% bedeutet 90,3 % der Varianz der erreichten Punkte lassen sich erklären durch die Varianz der Anzahl der verpassten Vorlesungen. Die restlichen 10% lassen sich nur durch andere Einflussfaktoren erklären. Je näher der Wert an 100% liegt, desto eher ist er für Prognosen zu gebrauchen.

WAHRSCHEINLICHKEITSTHEORIE

Ein (Zufalls) Experiment ist ein beliebig oft (unter identischen Bedingungen) wiederholbarer Vorgang, dessen Ergebnis „vom Zufall abhängt", d.h. nicht exakt vorhergesagt werden kann.
Experiment ➔ die Erhebung eines Merkmals an einem Merkmalsträger
Elementarereignisse➔ die Merkmalsausprägungen
Stichprobe vom Umfang n ➔ die n malige Wiederholung des Experiments
Ereignisraum Ω ➔ auch Ergebnismenge oder Merkmalraum genannt (oder auch Stichprobenraum)
Die Anzahl der Ergebnisse der Menge **Ω** nennt man **Mächtigkeit | Ω |= n.** Ω kann endlich , abzählbar oder sogar ü berabzählbar unendlich sein.
Um exakte Voraussagen über die Begrenzung unserer Möglichkeiten zu treffen, brauchen wir einen Maß für die Sicherheit (oder Unsicherheit). Ein solches Maß ist die **Wahrscheinlichkeit p**
axiomatische Begründung
1. Die Wahrscheinlichkeit für das Eintreten eines Ereignisses A ist immer eine reelle Zahl zwischen 0 und 1: 0 ≤ p(A) ≤ 1
2. Das sichere Ereignis Ω hat die Wahrscheinlichkeit 1
3. Die Wahrscheinlichkeit einer Vereinigung abzählbar vieler disjunkter Ereignisse ist gleich der Summe der Wahrscheinlichkeiten der einzelnen Ereignisse ➔ σ Additivität (, Additivität
Wahrscheinlichkeiten
1. mit Reihenfolge, mit zrücklegen: **n hoch k**
2. mit Reihenfolge, ohne zurücklegen: **Fakultät n = n!**
Schritt 1: Ergebnisraum bestimmen **n hoch k**
Schritt 2: Menge der günstigen Ergebnisse bestimmen **n!**
Schritt 3: Wahrscheinlichkeit **|E| durch Omega** (alle Möglichkeiten)