

Auswertung statistischer Daten 1

Dr. Elke Warmuth

Sommersemester 2018

Statistik

Untersuchungseinheiten u. Merkmale

Grundgesamtheit u. Stichprobe

Datenaufbereitung

Urliste, Strichliste, Häufigkeitstabelle, Stängel-Blatt-Diagramm

Klasseneinteilung

Relative Häufigkeiten und Häufigkeitsverteilung

Graphische Darstellungen von Häufigkeitsverteilungen

Vorbemerkungen:

- ▶ Stochastik = Wahrscheinlichkeitsrechnung + Statistik (beschreibende u. beurteilende)
- ▶ Plan:
 - ▶ Auswertung statistischer Daten
 - ▶ Modelle für Vorgänge mit zufälligem Ergebnis
 - ▶ Bedingte Wahrscheinlichkeit und Unabhängigkeit
 - ▶ Binomialverteilung
- ▶ Literatur: Warmuth, Elke; Warmuth, Walter: Elementare Wahrscheinlichkeitsrechnung, Vom Umgang mit dem Zufall. Stuttgart, Leipzig: B.G.Teubner, 1998

A deskriptive (beschreibende) Statistik

- ▶ Auswahl der Fragestellung
- ▶ Auswahl von Daten (Stichprobe)
- ▶ Aufbereitung der Daten
- ▶ Darstellung der Daten
- ▶ Arbeit in der Beobachtungsebene

B induktive (beurteilende) Statistik

- ▶ Wahl eines mathematischen (stochastischen) Modells
- ▶ Schlüsse im Modell
- ▶ Planung der Stichprobe
- ▶ Rückschlüsse von der Stichprobe auf die Grundgesamtheit
- ▶ Arbeit in der Modellebene

A und B hängen zusammen!

	Untersuchungseinheit	Merkmal	Ausprägung
1	Neugeborenes	Geschlecht	m, w
2	Neugeborenes	Größe (in cm)	..., 49, 50, 51, ...
3	Schülerin	Sportnote	1, 2, ..., 6
4	Arbeitnehmerin	Steuerklasse	1, 2, ..., 6
5	Schülerin	Entfernung zur Schule (in km)	0,1; 0,2; ...
6	Wahlberechtigter	Zweitstimme	SPD, CDU, ...
7	Würfelwurf	Augenzahl	1, 2, ..., 6
8	zweifacher Würfelwurf	Augensumme	2, 3, ..., 12

Merkmalsarten:

- ▶ qualitativ
 - ▶ nominal (Geschlecht, Steuerklasse, Partei)
 - ▶ ordinal (Schulnote)
- ▶ quantitativ (Größe, Entfernung, Augenzahl, Augensumme)

Weitere Unterscheidungen bei quantitativen Merkmalen:

► Intervallskalen

- Beispiele: Temperatur in °C, Intelligenzquotient
- Differenzen haben Sinn
- Transformationen der Form $y = ax + b$ mit $a > 0$ sind zulässig
z.B. $y = 1,8x + 32$ mit y in °F und x in °C
- Nullpunkt bleibt nicht fest ($0^{\circ}\text{C} \neq 0^{\circ}\text{F}$)
- Quotient aus Messwerten ändert sich bei Skalenwechsel:

$$\frac{y_1}{y_2} = \frac{ax_1 + b}{ax_2 + b} \neq \frac{x_1}{x_2}$$

20°C ist es nicht doppelt so warm wie 10°C.

► Verhältnisskalen

- Beispiele: Temperatur in °K, Gewicht, Geschwindigkeit
- Quotienten haben Sinn
- Transformationen der Form $y = ax$ mit $a > 0$ sind zulässig
z.B. $y = 1,8x + 32$ mit y in °F und x in °C
- fester Nullpunkt existiert

Grundgesamtheit (Population)

ist die Menge aller denkbaren Beobachtungseinheiten einer Untersuchung.

endlich (alle Schülerinnen einer Schule) oder unendlich (alle Würfelwürfe)

Stichprobe

ist eine (endliche) Teilmenge der Grundgesamtheit. Hat diese Teilmenge n Elemente, so heißt sie Stichprobe vom Umfang n .

Beispiele: aus jeder Klasse eine Schülerin, 20 Würfelwürfe

Auswahlverfahren:

- ▶ nicht-zufällige Auswahlverfahren: abhängig vom Verhalten der Person, die die Auswahl durchführt
- ▶ zufällige Auswahlverfahren: Prozess der Ziehung wird vorab festgelegt, nicht vom Verhalten des Interviewers abhängig. Es können Wahrscheinlichkeiten dafür angegeben werden, dass eine Beobachtungseinheit in die Stichprobe gezogen wird.
 - ▶ einfache Zufallsstichprobe: Jede Teilmenge vom Umfang n aus einer Grundgesamtheit vom Umfang N hat die Wahrscheinlichkeit

$$\frac{1}{\binom{N}{n}}$$

gezogen zu werden.

- ▶ andere Zufallsstichproben, z.B. geschichtete Stichprobe

Vorsicht vor dem schillernden Begriff „repräsentative Stichprobe“

Suggestion: Strukturgleichheit zwischen Stichprobe und Grundgesamtheit

Wikipedia: Der Ausdruck „repräsentative Stichprobe“ ist kein Fachbegriff.

Versuch einer Definition (Gabler, Quatember (2012)):

Eine Stichprobe heißt für eine Gesamtheit hinsichtlich einer interessierenden Verteilung oder eines interessierenden Parameters dieser Verteilung repräsentativ, wenn diese Verteilung oder der Parameter damit (zumindest annähernd) unverzerrt geschätzt werden können und bei dieser Schätzung auch eine vorgegebene Genauigkeitsanforderung eingehalten wird.

Urliste

$x_1, x_2, \dots, x_n$ enthält die Ausprägungen des Merkmals X in einer Stichprobe vom Umfang n in der Reihenfolge ihres Auftretens.

Beispiel: Merkmal Familienstand in einer Stichprobe vom Umfang 18

Strichliste und Häufigkeitstabelle

Merkmalswert	ledig	verheiratet	verwitwet	geschieden	LP	andere
Häufigkeit						
Häufigkeit	4	7	1	3	2	1

LP – eingetragene Lebenspartnerschaft

$H_n(x_i)$ – **absolute** Häufigkeit des Beobachtungswertes x_i

Beispiel Lebendgeborene und Gestorbene je 1000 Einwohner in Deutschland 2015 ¹⁾ nach Bundesländern

Lebendgeb.	9,3	9,3	10,9	7,7	9,8	11,1	9,3	8,3
Gestorbene	10,0	10,5	9,8	12,4	11,7	9,9	10,8	12,7
Lebendgeb.	8,5	9,0	8,7	7,6	9,0	7,8	8,3	8,3
Gestorbene	11,7	11,5	11,6	13,5	13,4	14,4	11,8	13,3

1) Quelle:

http://www.statistik-portal.de/Statistik-Portal/de_jb01_jahrta3.asp,
02.01.2018

Einwohner auf der Basis des Zensus 2011

Stängel-Blatt-Diagramm

Gestorbene						Lebendgeborene				
					7	7	6	8		
					8	3	5	7	3	3
		9	8		9	3	3	8	3	0 0
	8	5	0		10	9				
8	6	5	7	7	11	1				
		7	4		12					
	3	4	5		13					
		4			14					

Beispiel Lebendgeborene je 1000 Einwohner in Deutschland 2015

Lebendgeb.	9,3	9,3	10,9	7,7	9,8	11,1	9,3	8,3
Lebendgeb.	8,5	9,0	8,7	7,6	9,0	7,8	8,3	8,3

Klasseneinteilung

Wertebereich der Merkmalsausprägungen wird in disjunkte Intervalle (Klassen) zerlegt und ein Element des Intervalls (meist die Klassenmitte) wird als Repräsentant der Klasse für weitere Rechnungen verwendet.

Klasse	[7, 0; 8, 0) 7 bis unter 8	[8, 0; 9, 0) 8 bis unter 9	[9, 0; 10, 0) 9 bis unter 10	[10, 0; 11, 0) 10 bis unter 11	[11, 0; 12, 0) 11 bis unter 12
Kl.mitte	7,5	8,5	9,5	10,5	11,5
Häufigkeit	3	5	6	1	1

- ▶ Klasseneinteilung muss eindeutig sein
- ▶ Verlust an Information vs. Gewinn an Übersichtlichkeit
- ▶ Faustregeln für Klassenanzahl k : $k \approx \sqrt{n}$, $k \approx 5 \log n$,
 $5 \leq k \leq 20$
- ▶ gleiche Klassenbreite vorteilhaft, aber nicht immer erreichbar
Beispiel: Merkmal Einkommen
- ▶ Wahl der Klasseneinteilung beinhaltet
Manipulationsmöglichkeiten

Relative Häufigkeiten und Häufigkeitsverteilung

Ziel: Vergleich von Stichproben mit unterschiedlichem Umfang

$x_1, x_2, \dots, x_s$ – mögliche Ausprägungen des Merkmals X

n – Umfang der Stichprobe

Relative Häufigkeit des Wertes x_i

$$h_n(x_i) = \frac{H_n(x_i)}{n}, i = 1, \dots, s.$$

Bei Klasseneinteilung: x_i – Repräsentant der Klasse, $H_n(x_i)$ – Häufigkeit der Klasse

Häufigkeitsverteilung

Merkmalsausprägung	x_1	x_2	$\dots$	x_s
rel. Häufigkeit	$h_n(x_1)$	$h_n(x_2)$	$\dots$	$h_n(x_s)$

Relative Häufigkeiten und Häufigkeitsverteilung

Beispiel Familienstand

x_i	ledig	verheiratet	verwitwet	geschieden	LP	andere
$h_{18}(x_i)$	$\frac{4}{18}$	$\frac{7}{18}$	$\frac{1}{18}$	$\frac{3}{18}$	$\frac{2}{18}$	$\frac{1}{18}$

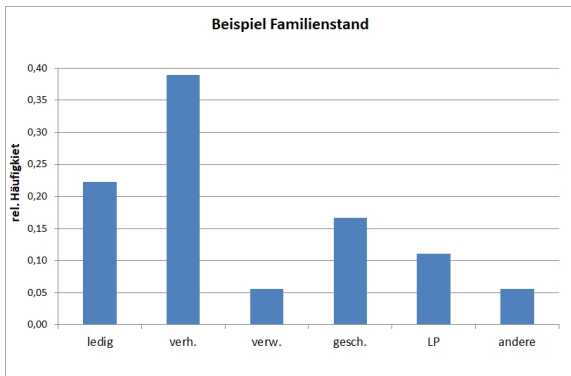
Beispiel Lebendgeborene je 1000 Einwohner nach BL

x_i	7,5	8,5	9,5	10,5	11,5
$h_{16}(x_i)$	$\frac{3}{16}$	$\frac{3}{16}$	$\frac{8}{16}$	$\frac{1}{16}$	$\frac{1}{16}$

Die Häufigkeitsverteilung gibt an, welcher Anteil aller Beobachtungen auf die einzelnen Merkmalswerte entfällt. Die Summe der relativen Häufigkeiten ist (bis auf Rundungsfehler) 1.

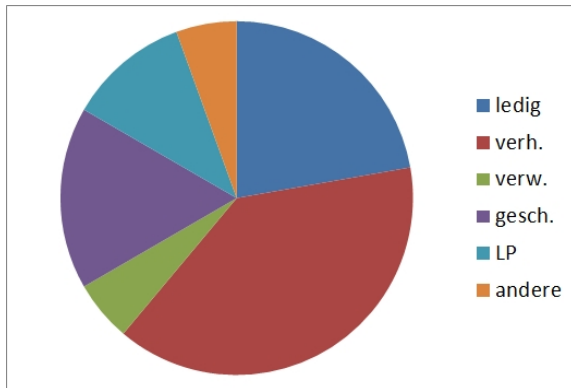
Strecken- oder Stabdiagramm

Auf der x -Achse werden die Werte $x_1, x_2, \dots, x_s$ eingetragen. Über jedem x_i wird eine Strecke (ein Stab) der Länge $h_n(x_i)$ oder $H_n(x_i)$ errichtet.

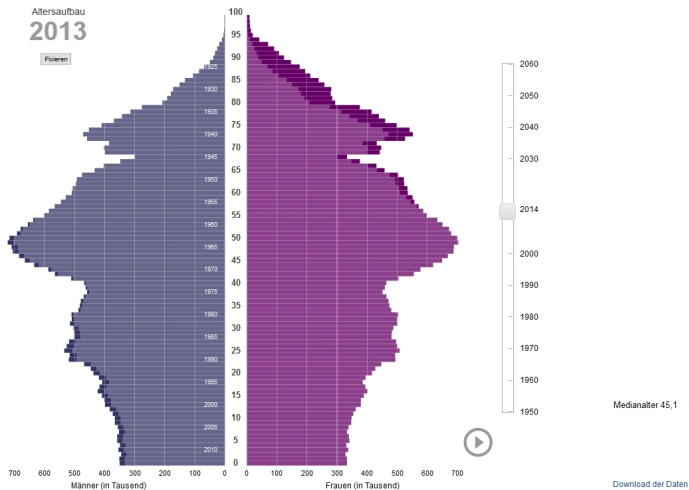


Kreisdiagramm

Die relativen Häufigkeiten $h_n(x_i)$ werden als „Tortenstücke“ veranschaulicht. Zum Wert x_i gehört der Winkel $h_n(x_i) \cdot 360^\circ$.

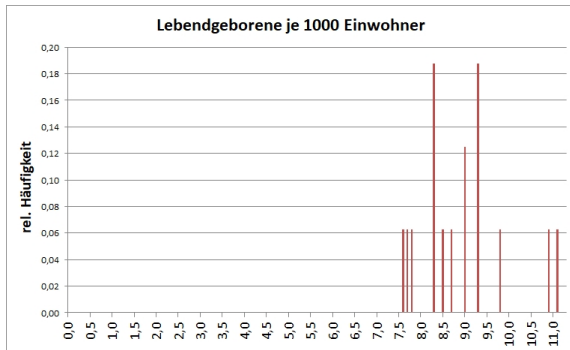


Graphische Darstellungen von Häufigkeitsverteilungen



<https://service.destatis.de/bevoelkerungspyramide>

Stabdiagramm zum Beispiel Lebendgeborene ohne Klasseneinteilung

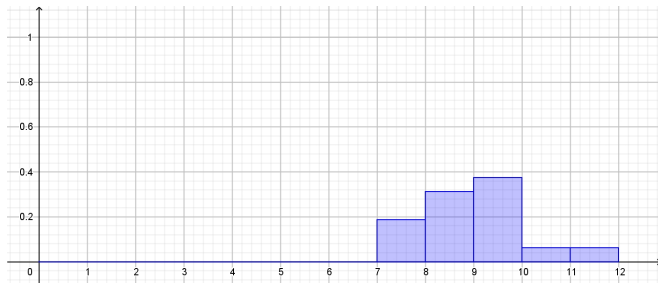


unübersichtlich bei vielen verschiedenen Beobachtungswerten
→ Klasseneinteilung und Histogramm als zugehörige Darstellung

Histogramm

Über jedem Intervall der Klasseneinteilung wird ein Rechteck errichtet, dessen Flächeninhalt gleich der relativen Häufigkeit der Klasse ist. Die gesamte Fläche des Histogramms hat den Flächeninhalt 1.

Beispiel Lebendgeborene je 1000 Einwohner



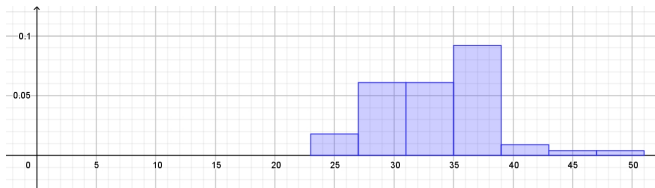
Beispiel Silvesterlauf 2017

Bautzener Silvesterlauf – BLV Rot-Weiß 90									
            Impressum Übersicht Allgemein Sportarten									
Frauen 5,3 km									
Total	AK	Stnr	Name	Verein	Strecke	AK	Zeit	Geb.-jahr	Nation
1	1	2187	Schröter, Annett	Heideläufer	5,3 km	W40-44	00:23:10	1977	DEU
2	1	2070	Masur, Sophia	The Magnificent SIX	5,3 km	W20-29	00:24:48	1990	DEU
3	2	2219	Marschner, Darja	Niederkaina	5,3 km	W40-44	00:26:39	1973	DEU
4	2	2899	Kögler, Katharina	Langburkursdorf	5,3 km	W20-29	00:26:43	1995	DEU
5	1	2939	Gnauck, Gesine	Aeroteam	5,3 km	W30-34	00:27:35	1985	DEU
6	1	2138	Tanzmann, Maria	Wilthen	5,3 km	WJU 18	00:27:57	2000	DEU
7	3	2152	Kulke, Andrea	Kraftwerker Bautzen	5,3 km	W40-44	00:28:02	1976	DEU
8	1	2135	Ebert, Petra	ski & bike Team Schirgiswalde	5,3 km	W45-49	00:28:02	1969	DEU
9	1	2001	Knebel, Simone	Bischofswerda	5,3 km	W50-54	00:28:22	1966	DEU
10	2	2063	Lukas, Doreen	Go In Aktiv Leben	5,3 km	W30-34	00:28:29	1984	DEU
11	4	2150	Warmuth, Nadine	Post SV Dresden	5,3 km	W40-44	00:29:12	1977	DEU
12	1	2147	Lehmann, Helene	Singwitz	5,3 km	WJU 14	00:29:29	2005	DEU
13	3	2898	Kögler, Charlotte	Zielgruppe	5,3 km	W20-29	00:29:42	1995	DEU
14	5	2869	Herrmann, Kerstin	Gaußig	5,3 km	W40-44	00:29:42	1977	DEU
15	3	2200	Hauptmann, Marie	Bautzen	5,3 km	W30-34	00:29:49	1984	DEU

Strecke: 5,3 km Cross

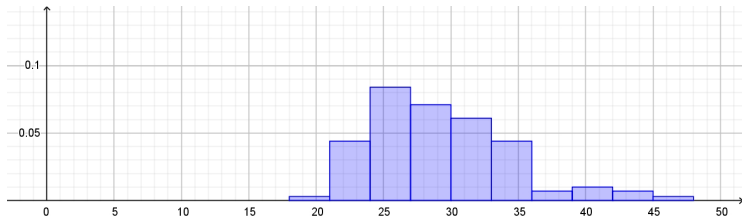
Frauen

- ▶ 57 Starterinnen
- ▶ schnellste Laufzeit: 23:10 Minuten
- ▶ langsamste Laufzeit: 50:52 Minuten
- ▶ 7 Klassen der Länge 4 Minuten
- ▶ Säulenhöhe = $\frac{\text{rel. Häufigkeit}}{4}$



Männer

- ▶ 99 Starter
- ▶ schnellste Laufzeit: 18:56 Minuten
- ▶ langsamste Laufzeit: 47:27 Minuten
- ▶ 10 Klassen der Länge 3 Minuten
- ▶ Säulenhöhe = $\frac{\text{rel. Häufigkeit}}{3}$



Graphische Darstellungen von Häufigkeitsverteilungen

