Skalen, Häufigkeiten, Verteilungen, Maßzahlen

Grundlage jeder Untersuchung (Studie) sind Messungen und / oder Beobachtungen von Merkmalen an Merkmalsträgern (statist. Einheiten).

		Beispiel
Merkmalsträger:	Personen, Objekte	Patienten
Merkmal (Variable):	bestimmte Eigenschaft eines Merkmalsträgers	Blutgruppe
Merkmalsausprägung:	Werte, die ein Merkmal annehmen kann	(0, A, B, AB)
Skalenniveau	Typisierung der Merkmalsausprägung	nominal

Die Analyse eines so erhaltenen Datensatzes beginnt in der Regel mit der Beschreibung der einzelnen Merkmale (Variablen) und ihrer Merkmalsausprägungen. Diese univariate Datenbeschreibung liefert zuerst Erkenntnisse über die Art der Daten (*qualitativ* oder *quantitativ*), über deren Skalenniveau (*nominal, ordinal*, *metrisch*) und über die Eigenschaft *diskret* oder *stetig* . Hierbei ist anzumerken, dass stetige Daten, wie z.B. Alter oder Gewicht, in der Regel nur diskret gemessen werden (Alter z.B. in Jahren, Gewicht in kg).

Art	Skalenniveau	Eigenschaft
quantitativ	nominal / dichotom	diskret
	ordinal	diskret
qualitativ	metrisch	diskret
		stetig

Beispiele für Skalenniveaus:

Nominal:
- Farbe
- Geschlecht
- Blutgruppe
- Ja / Nein (dichotom)
- Familienstand

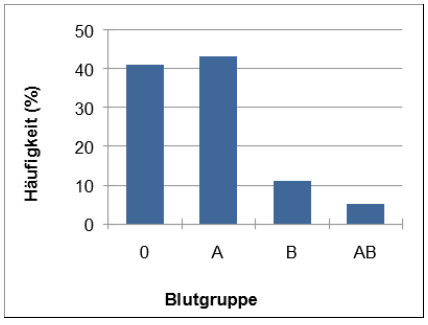
Ordinal:
- Rangliste Sportler
- Schulnoten
- Schmerzskala
- Hotelkategorien

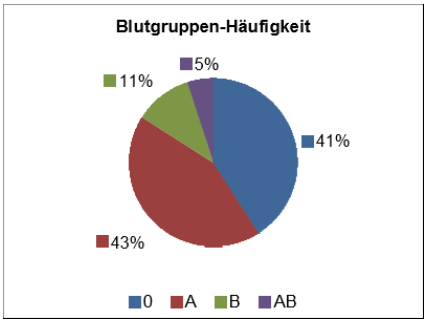
Metrisch:
- Temperatur
- Größe
- Gewicht
- Blutdruck
- Alter
- DMFT-Wert
- Zahl der Geburten

Im nächsten Schritt interessiert die Häufigkeitsverteilung der Merkmalsausprägungen. Folgende Nomenklatur soll hier verwendet werden:
h = Häufigkeit, rh = relative Häufigkeit, kh = kumulierte Häufigkeit, krh = kumulierte relative Häufigkeit

Absolute (h) und relative Häufigkeit (rh) **nominaler Daten** lassen sich tabellarisch darstellen und z.B. durch Balkendiagramme oder Kreisdiagramme visualisieren. Ein Beispiel:
Merkmal: Blutgruppe:

Blutgruppe	h	rh%
0	1025	41
A	1075	43
B	275	11
AB	125	5
Gesamt	2500	100

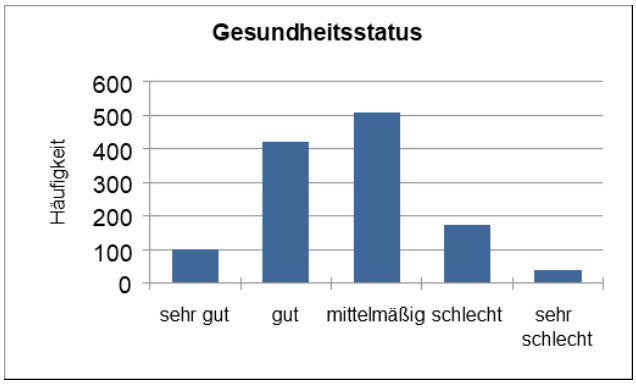


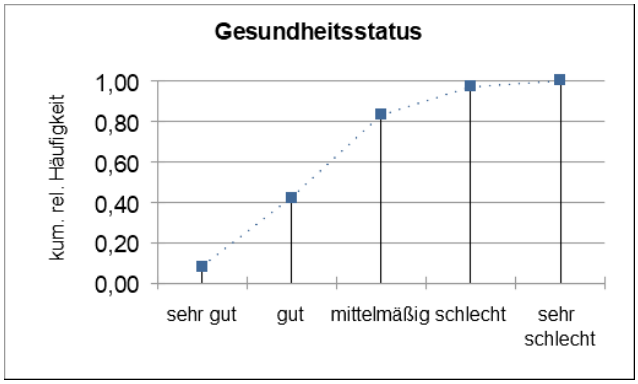


Ordinale Daten lassen sich in eine Rangordnung bringen, z.B. gut, mittel, schlecht, oder bei Schulnoten 1, 2, 3, 4, 5. Ordnet man die Häufigkeiten den auf- oder absteigenden Rängen zu, so macht es Sinn, neben h und rh auch kh oder krh zu berechnen und ggf. grafisch darzustellen mittels empirischer Verteilungsfunktion (Summenhäufigkeitsfunktion). Hierzu ein Beispiel:
Merkmal: Einschätzung der eigenen Gesundheit

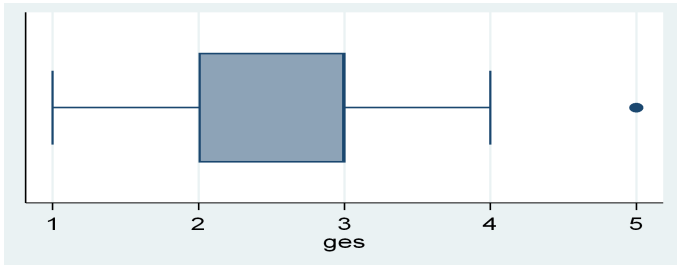
Gesundheit	Gesundheit	h	rh	krh
1	sehr gut	99	0,08	0,08
2	gut	420	0,34	0,42
3	mittelmäßig	506	0,41	0,83
4	schlecht	173	0,14	0,97
5	sehr schlecht	37	0,03	1,00
Gesamt		1235	1	

(Quelle: G&G 11/22)



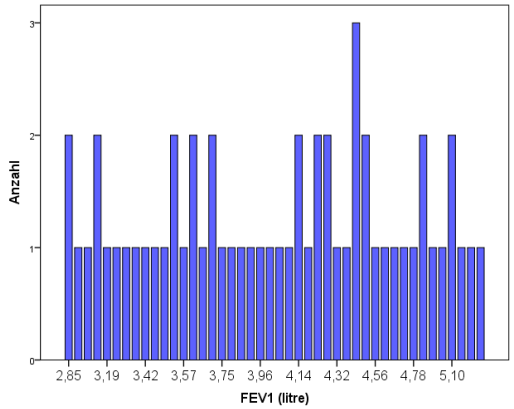


So lässt sich z.B. sagen, dass rund 40% aller Befragten ihre Gesundheit als gut oder sehr gut einschätzen und 17% als schlecht oder sehr schlecht.
Für die aufsteigend geordneten Werte kann man zudem das Minimum = 1, das Maximum = 5 und die Quartile $Q_1 = 2$, $Q_2 = 3$ und $Q_3 = 3$ berechnen (**Tukey's 5 numbers**) und damit einen Box-Plot darstellen.



Interquartilsabstand IQA (IQR Interquartilsrange)
$IQA = Q_3 - Q_1 = 1$ (Breite der Box)
Länge der Whisker oben: $Q_3 + 1{,}5 \cdot IQA = 4{,}5$
Länge der Whisker unten: $Q_1 - 1{,}5 \cdot IQA = 0{,}5$
Werte, die diese Grenzen überschreiten oder unterschreiten sind "Ausreißer".
Ist das Minimum größer als der berechnete untere Whisker, gilt das Minimum für den Whisker.
Ist das Maximum kleiner als der berechnete obere Whisker, gilt das Maximum für den Whisker.

Stetige metrische Daten treten je nach Meßgenauigkeit meist einzeln auf. Für eine Häufigkeitsverteilung ist daher eine Klasseneinteilung notwendig. Die Anzahl der Klassen richtet sich nach der Anzahl der Werte n und kann mit $\sqrt{n}$ oder für große n mit $10 \cdot \log(n)$ abgeschätzt werden. Die Anzahl der Klassen sollte 20 möglichst nicht überschreiten, da die Grafik sonst zu unübersichtlich wird. Hierzu ein Beispiel:
Merkmal: FEV1 (forciertes expiratorisches Volumen in 1 sek. (Liter/s))



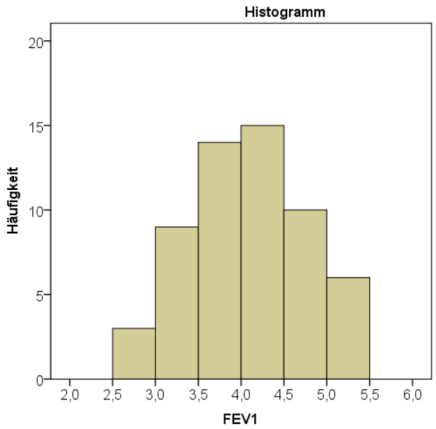
von 57 Studenten (Quelle: M. Bland)

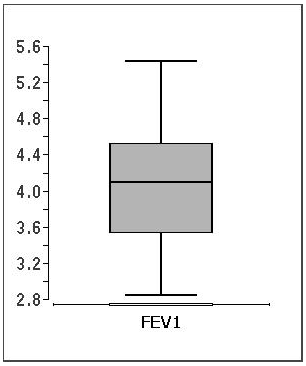
2.85	3.19	3.50	3.69	3.90	4.14	4.32	4.50	4.80	5.20
2.85	3.20	3.54	3.70	3.96	4.16	4.44	4.56	4.80	5.30
2.98	3.30	3.54	3.70	4.05	4.20	4.47	4.68	4.90	5.43
3.04	3.39	3.57	3.75	4.08	4.20	4.47	4.70	5.00	
3.10	3.42	3.60	3.78	4.10	4.30	4.47	4.71	5.10	
3.10	3.48	3.60	3.83	4.14	4.30	4.50	4.78	5.10	

Häufigkeitsverteilung der Einzelwerte. Bei genauerer Messung würde man Gleichverteilung erwarten.

Geschätzte Anzahl der Klassen: Wurzel (57) = 7,6 . Empirische Überlegungen ergeben für einen Startpunkt von 2,5 und einen Endpunkt von 5,5 eine Spannweite von 3,0 und bei Verwendung von 6 Klassen eine Klassenbreite von 0,5 {Spannweite / 6 = (5,5-2,5) / 6 = 0,5} . Häufig wird die untere Intervallgrenze zum Intervall und die obere Intervallgrenze zum nächsten Intervall gezählt: [a,b);[b,c);[c,d) ....usw. Man erhält so aus den obigen Daten folgende Tabelle mit Histogramm als grafische Darstellung:

FEV1	h	rh	krh
2.5 - 3.0	3	0,053	0,053
3.0 - 3.5	9	0,158	0,211
3.5 - 4.0	14	0,246	0,456
4.0 - 4.5	15	0,263	0,719
4.5 - 5.0	10	0,175	0,895
5.0 - 5.5	6	0,105	1,000





Histogramm und Box-Plot (aus den Originaldaten) zeigen eine symmetrische, unimodale Verteilung. Weitere wichtige Verteilungsformen sind z.B. linksschief, rechtsschief und bimodal.

Die Tukey's 5 numbers für diese FEV1 - Daten lauten:
Min. = 2,85, Q_1 = 3,54, Q_2 = 4,1, Q_3 = 4,53 und Max. = 5,43
Die Whisker sind hier bei Minimum und Maximum.

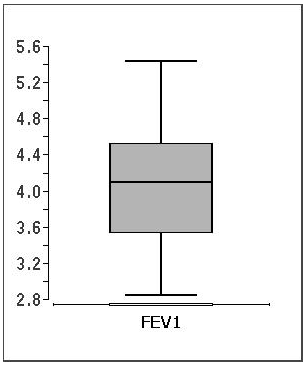
Max. --->

Q3 --->

Q2 --->

Q1 --->

Min .--->



Mit Tukey's 5 numbers lassen sich Häufigkeitsverteilungen gut beschreiben. Für die aufsteigend geordneten Werte charakterisiert der Median die "Mitte". Er teilt die Daten in zwei Hälften. Die eine Hälfte ist kleiner, die andere größer als der Median. Handelt es sich um eine ungerade Anzahl von Daten, ist der Median gleich dem Wert einer Messung /Beobachtung. Bei gerader Anzahl ist es der Mittelwert der beiden zentralen Werte.
Beispiel **ungerade** Anzahl: 2, 3, 5, 6, 8 Median Q_2 = 5
gerade Anzahl: 2, 3, 5, 6 Median Q_2 = (3+5)/2 = 4

Einfache Berechnung der Quartile $Q_1 = \tilde{x}_{0,25}$, $Q_2 = \tilde{x}_{0,5}$ und $Q_3 = \tilde{x}_{0,75}$:

Sei **n** die Anzahl der Messungen, α = 0,25 ; 0,50 ; 0,75 für Q_1, Q_2 und Q_3 entsprechend und **k** die laufende Nummer der aufsteigenden Reihenfolge. Wenn n·α keine ganze Zahl ist, dann ist k die auf n·α folgende ganze Zahl und $\tilde{x}_\alpha = x_{(k)}$ Für das obige Beispiel mit ungerader Anzahl findet man:

Q_1: aus n·α = 5·0,25 = 1,25 folgt Q_1 = „2. Wert in der Reihe“ = 3
Q_2: aus n·α = 5·0,50 = 2,5 folgt Q_2 = „3. Wert in der Reihe“ = 5
Q_3: aus n·α = 5·0,75 = 3,75 folgt Q_3 = „4. Wert in der Reihe“ = 6

Wenn n·α eine ganze Zahl k ist, dann ist $\tilde{x}_\alpha = \frac{1}{2}(x_{(k)} + x_{(k+1)})$

Für das obige Beispiel mit gerader Anzahl findet man für den Median:
Q_2: aus n·α = 4·0,50 = 2 folgt Q_2 = (3 + 5) / 2 = 4
Erkennbar berücksichtigt der Median nur den Rangplatz der aufsteigenden Reihenfolge, nicht die Meßwerte selbst. Für statistische Analysen bedarf es jedoch anderer Kenngrößen.

Das bekannteste Lagemaße ist der **arithmetische Mittelwert** $\bar{x}$ oder der Durchschnitt, auch einfach das "Mittel" genannt. Im Gegensatz zum Median werden hier alle Beobachtungs- oder Meßdaten $\mathbf{x_i}$ mit gleichem Gewicht genutzt. Er wird für statistische Berechnungen oder Gruppenvergleich verwendet. Allerdings beeinflussen Extremwerte den Mittelwert mehr als den Median. Bei symmetrischen Verteilungen sind beide nahezu gleich. Bei schiefen Verteilungen unterscheiden sie sich.

Diskrete metrische Daten sind Zähldaten, wie z.B. die Anzahl der Geburten von 125 zufällig ausgewählten Frauen in einer Region. Die Quartile berechnen sich hier wie folgt:

Geburten	h	rh	krh
0	59	0.47	0.47
1	44	0.35	0.82
2	14	0.11	0.94
3	3	0.02	0.96
4	4	0.03	0.99
5	1	0.01	1.00
Gesamt	125	1	

Q_1 = min { Geb | F(Geb) ≥ 0,25 } somit $Q_1 = 0$
Q_2 = min { Geb | F(Geb) ≥ 0,5 } somit $Q_2 = 1$
Q_3 = min { Geb | F(Geb) ≥ 0,75 } somit $Q_3 = 1$

Dabei sind Geb = Geburten und
F = krh (Summenhäufigkeitsfunktion)

Arithmetisches Mittel - Einzeldaten

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

Beispiel (FEV1 (Liter/s) von 57 männlichen Studenten): $\bar{x} = 4{,}061$

Arithmetisches Mittel - gehäufte Daten

$$\bar{x} = \frac{1}{n} \cdot \sum_{j=1}^{k} x_j \cdot h_j = \sum_{j=1}^{k} \frac{h_j}{n} \cdot x_j$$

Beispiel (Geburten): $\bar{x}$ = 0,47·0 + 0,35·1 + 0,11·2 + 0,02·3 + 0,03·4 + 0,01·5 = 0,816

Arithmetisches Mittel - klassierte Daten

$$\bar{x} = \frac{1}{n} \cdot \sum_{j=1}^{k} x_j^* \cdot h_j = \sum_{j=1}^{k} \frac{h_j}{n} \cdot x_j^*$$

Beispiel (Klasseneinteilung von FEV1), dabei sind x_j^* = Klassenmitte der j-ten Klasse und
h_j / n = relative Häufigkeiten (rh) der j-ten Klasse.
$\bar{x}$ = 0,053·2,75+0,158·3,25+0,246·3,75+0,263·4,25+0,175·4,75+0,105·5,25 = 4,082
Man erhält einen approximativen Mittelwert, der Mittelwert der Originaldaten beträgt 4,061

Der Mittelwert kennzeichnet die Mitte einer Verteilung, von der die Einzelwerte $\mathbf{x_i}$ mehr oder weniger stark abweichen (streuen). Das wichtigste Streuungsmaß, das die Abweichung der Einzelwerte vom Mittelwert quantifiziert ist die **Varianz** - die mittlere quadratische Abweichung:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2$$ oder auch für Berechnungen per Hand $$s^2 = \frac{1}{n-1} \cdot \left(\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2 \right)$$

Dividiert wird durch (n-1) , siehe Beitrag SRS_1.

Besser interpretierbar ist die Wurzel aus der Varianz, die **Standardabweichung** $\sqrt{s^2} = \mathbf{s}$, da sie die gleiche Dimension besitzt wie die Einzelwerte.
Bei annähernd normalverteilten Daten liegen etwa 67% der Werte im Intervall ($\bar{x}$ - s ; $\bar{x}$ + s) und etwa 95% im Intervall ($\bar{x}$ - 2s ; $\bar{x}$ + 2s) .