Statistik SS 2012

Deskriptive Statistik

Bernhard Spangl[1]

[1]Institut für angewandte Statistik und EDV
Universität für Bodenkultur

March 1, 2012

Was ist Statistik?

- Statistik kann helfen, Zusammenhänge zu verdeutlichen, Risiken abzuschätzen, Entscheidungsprozesse zu unterstützen und zu verbessern ("give me a number!")
- Statistik ist in der Lage, *unter gewissen Annahmen, Unsicherheit zu quantifizieren*
- Was Statistik nicht kann
  - Substanzwissenschaftliche Fragen klären, z.B. *weshalb* ein bestimmtes Medikament besser wirkt als ein anderes Medikament. Statistik kann nur helfen zu erkennen, ob und gegebenenfalls bei welcher Gruppe von Patienten ein bestimmtes Medikament bessere Wirkung zeigt.
  - Erkennen von geschickter Datenmanipulation (soll leider vorkommen)

Grundaufgaben der Statistik

- *Deskription* (Beschreibung):
  Beschreibung, graphische Aufbereitung,
  Komprimierung der Daten in Tabellen
  usw. ohne Stochastik
- *Exploration* (Suchen nach Mustern und Strukturen)

} Datenanalyse ohne Stochastik

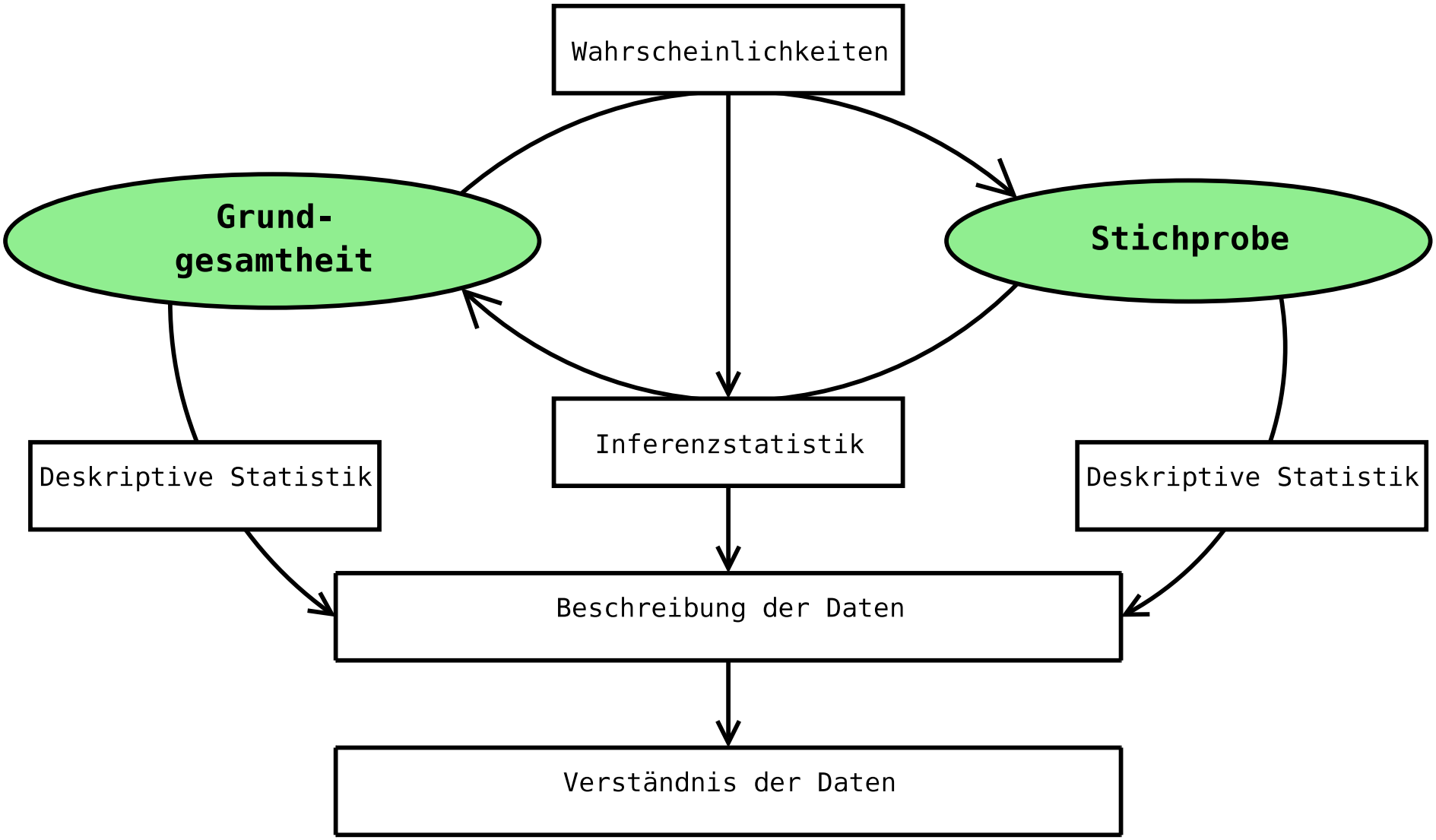
- *Inferenz*: Schluss von Stichprobe auf Grundgesamtheit

} Datenanalyse mit Stochastik, Wahrscheinlichkeitsrechnung nötig

Grundaufgaben der Statistik

- **Wahrscheinlichkeitsrechnung**: Gesetze bekannt, mit welchen Wahrscheinlichkeiten beobachte ich gewisse Ereignisse?
- **Deskriptive Statistik**: Verdichtung von Daten zur einfacheren Informationsgewinnung.
- **Inferenzstatistik**: Ein bestimmtes Ereignis ist eingetreten (Daten der Stichprobe), welche Rückschlüsse können wir über die datenerzeugenden Prozesse ziehen?

Grundaufgaben der Statistik



Einheiten und Grundgesamtheit

- **Statistische Einheiten**: Objekte, an denen interessierende Größen erfasst werden
- **Grundgesamtheit**: Menge aller für die Fragestellung relevanten statistischen Einheiten
- **Stichprobe**: tatsächlich untersuchte Teilmenge der Grundgesamtheit
- **Merkmal**: interessierende Größe, *Variable*
- **Merkmalsausprägung**: konkreter Wert des Merkmals für eine bestimmte statistische Einheit

Merkmalstypen

- *Diskretes Merkmal*
  endlich oder abzählbar unendlich viele verschiedene Werte
  Beispiele: Geschlecht, Kinderanzahl,...
- *Stetiges Merkmal*
  alle Werte eines Intervalles können angenommen werden
  Beispiele: Zeitdauern, Größe, Gewicht,...
- *Gruppierte Daten, Häufigkeitsdaten*: stetiges Merkmal X
  Wertebereich wird in Gruppen (Klassen, Kategorien) eingeteilt.
  Beispiele: Gehalt in Gehaltsklassen, Alter in Altersklassen
  Bemerkung: Gruppierung dient auch dem Datenschutz!

Skalen

„Skalenniveau“ charakterisiert Informationsgehalt der gemessenen / beobachteten Werte bzw. Ausprägungen eines Merkmals.

- **Nominalskala**: keine innere Ordunung der Ausprägungen. „Kodierung“ von Kategorien in Zahlen 1,2,... dient *nur* zur Unterscheidung und effizienten Speicherung, $1 < 2$ oder $1 + 2$, $2 - 1$, $1 : 3$ etc. machen keinen Sinn.
  Beispiele: Geschlecht (w, m), Nationalität (deutsch, österreichisch, ...), Verwendungszweck (Beruf, privat), Tageszeitung, Partei
- **Ordinalskala**: innere Ordnung, aber Differenzen können nicht interpretiert werden. Werte bzw. Kategorien lassen sich ordnen, arithmetische Operationen nicht sinnvoll
  Beispiele: Schulnoten (1, 2, 3, 4, 5), Temperatur (sehr kalt, kalt, warm)

Skalen

Metrische Skala oder Kardinalskala:

- **Intervallskala**: kein Nullpunkt, Differenzen können Interpretiert werden, Quotienten nicht.
  Beispiele: Temperatur in Grad Celsius
- **Verhältnisskala**: Nullpunkt inhaltlich interpretierbar, alle arithmetischen Operationen sinnvoll.
  Beispiele: Geldbetrag, Wohnfläche, Temperatur in Grad Kelvin

Skalen

Table: Sinnvolle Berechnungen für Daten verschiedener Skalen

	sinnvoll interpretierbare Berechnungen			
Skalenart	auszählen	ordnen	Differenzen bilden	Quotienten bilden
nominal	ja	nein	nein	nein
ordinal	ja	ja	nein	nein
intervall	ja	ja	ja	nein
verhältnis	ja	ja	ja	ja

- nominal & ordinal: heißen auch *kategorisch* oder *qualitativ*
- metrisch: heißen auch *quantitativ*

(Univartiate) Deskriptive Statistik: Ziele

- Darstellung, Beschreibung, explorative Analyse *eines* Merkmals
- Zusammenfassung der Daten auf möglichst wenige Kennzahlen (aber *meistens mehr als 1 Kennzahl*!)
- Veranschaulichen durch einfache Grafiken
- Auch geeignet zum Vergleich in verschiedenen „*Schichten*“ oder „*Gruppen*“ der Grundgesamtheit

Rohdaten und Ausprägungen

- **Rohdaten**: (Urliste, Primärdaten)
  n Beobachtungen $x_1, x_2, \ldots, x_n$ für Merkmal X
  Reihenfolge hat in der Regel keine Bedeutung
- **Ausprägungen**: *Verschiedene* vorkommende Werte von X
  $a_1, a_2, \ldots, a_k$, $k \leq n$
  Falls möglich (ab Ordinalskala) nach Größe sortiert:
  $a_1 < a_2 < \ldots < a_k$

Oft gilt:

- **Qualitative Daten**: $k << n$
- **Quantitative Daten**: $k \approx n$

Häufigkeiten

Einfachste Kompression, bis auf Reihenfolge kein Informationsverlust:

Absolute Häufigkeit: $h_j = h(a_j)$
Anzahl der x_i aus $x_1, \ldots, x_n$ mit $x_i = a_j$

Relative Häufigkeit: $f_j = f(a_j) = h_j/n$

Häufigkeitsverteilung: Die Mengen $h_1, \ldots, h_k$ und $f_1, \ldots, f_k$ heißen absolute bzw. relative Häufigkeitsverteilung von x.

Visualisierung: Stab-, Balken-, Punkt-, Kreisdiagramm

Häufigkeiten

- Wenn statt der Urliste nur die Ausprägungen $a_1, \ldots, a_k$ und die Häufigkeiten $f_1, \ldots, f_k$ bzw. $h_1, \ldots, h_k$ vorliegen, sprechen wir von *Häufigkeitsdaten*.
- Klassenbildung, gruppierte Daten:
  Bei metrischen oder stetigen Merkmalen oft Gruppierung der Urliste durch Bildung geeigneter Klassen.
  - **Erhebung**: Anonymisierung, einfacher & schneller, ...
  - **Beschreibung**: inhaltlich gute Zusammenfassung der Daten

Daten

```
> source("micro.R")
> micro[1:13, ]
```

	diameter	resistence	color
1	0.5	very sensitive	yellow
2	4.1	sensitive	yellow
3	4.4	inbetween	yellow
4	5.6	resistent	yellow
5	6.8	very resistent	yellow
6	7.2	very sensitive	yellow
7	7.7	resistent	yellow
8	7.8	inbetween	yellow
9	8.2	resistent	yellow
10	9.5	very resistent	yellow
11	9.2	very sensitive	yellow
12	9.9	sensitive	yellow
13	11.9	inbetween	yellow

Häufigkeiten

- Gesamtanzahl: $n = 40$
- absolute Häufigkeiten h_i:

```
> table(micro$resistence)
```

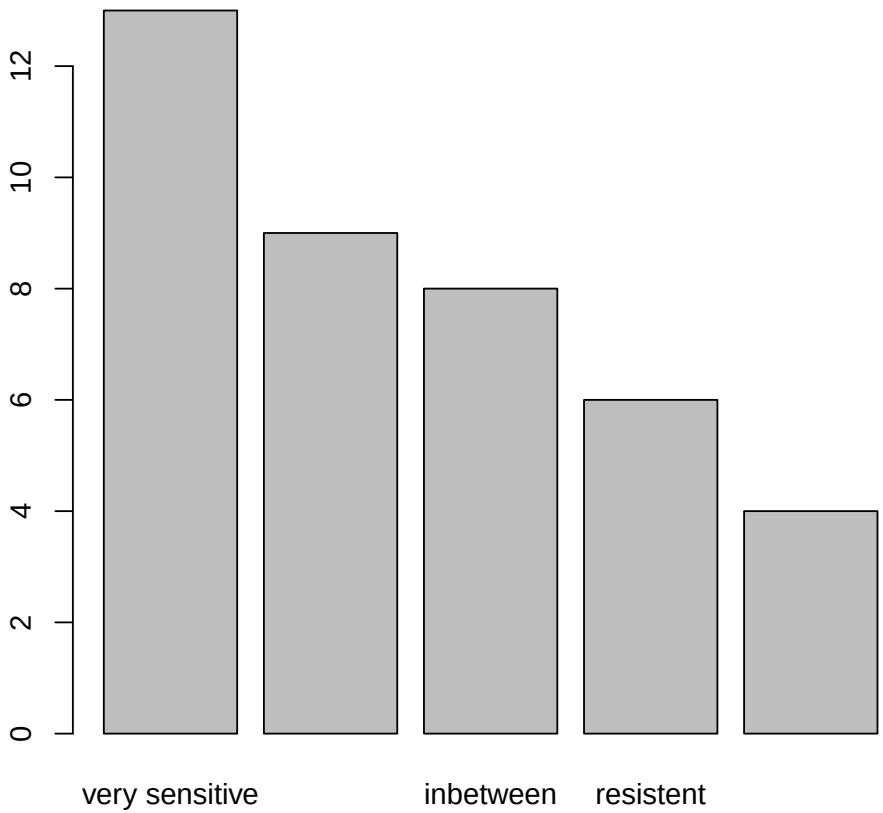
```
very sensitive      sensitive      inbetween      resistent
            13              9              8              6
very resistent
             4
```

- relative Häufigkeiten $f_i = h_i/n$:

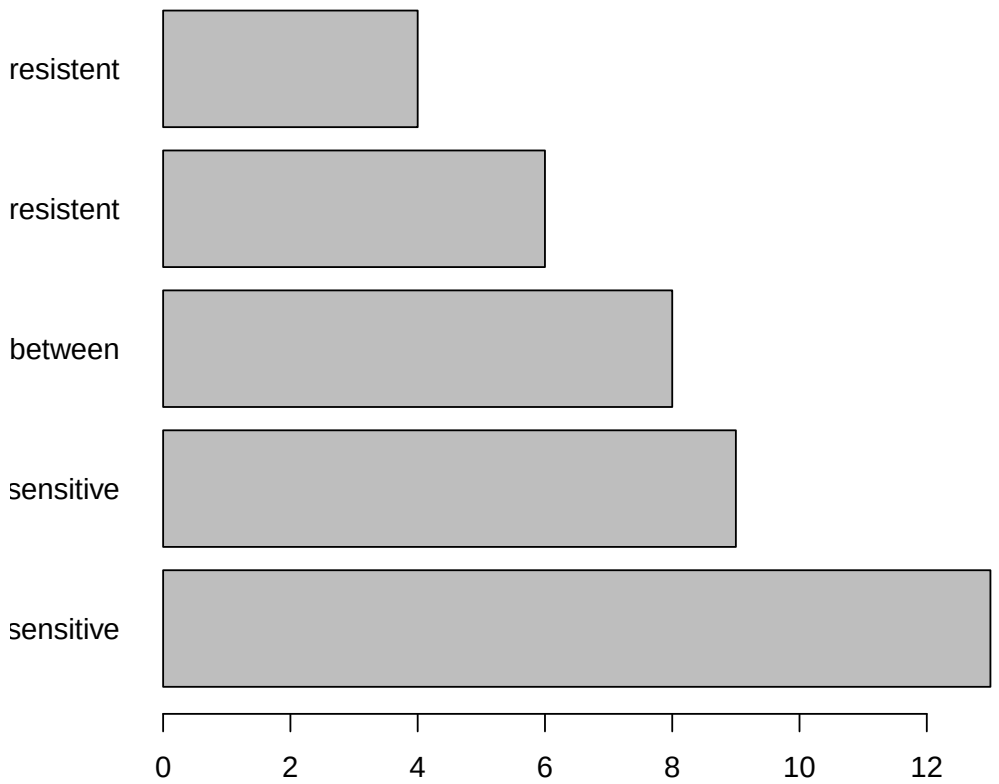
```
> table(micro$resistence)/40
```

```
very sensitive      sensitive      inbetween      resistent
         0.325          0.225          0.200          0.150
very resistent
         0.100
```

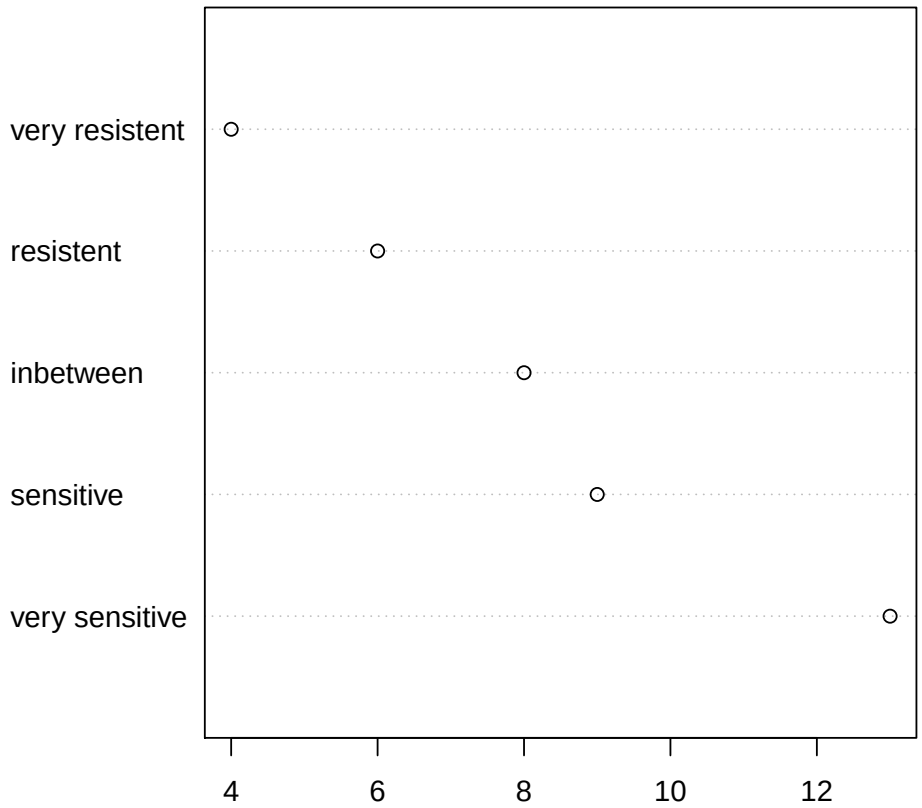
Balkendiagramm



Horizontales Balkendiagramm



Punktdiagramm



Stabdiagramm

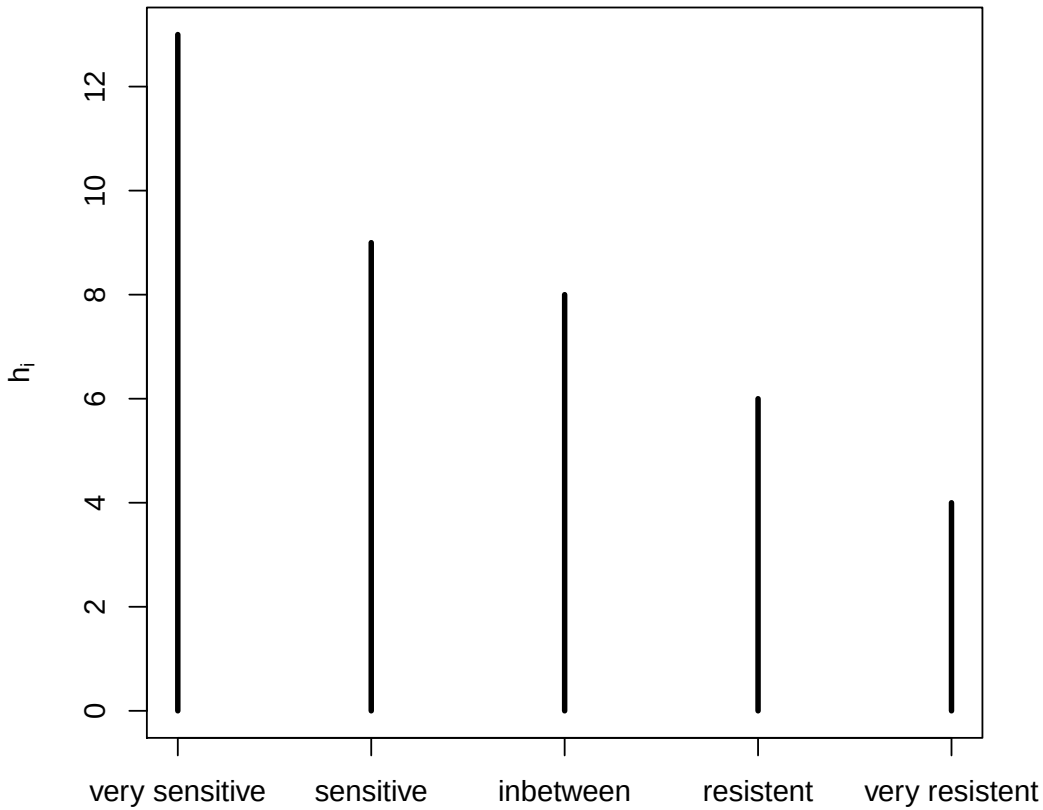


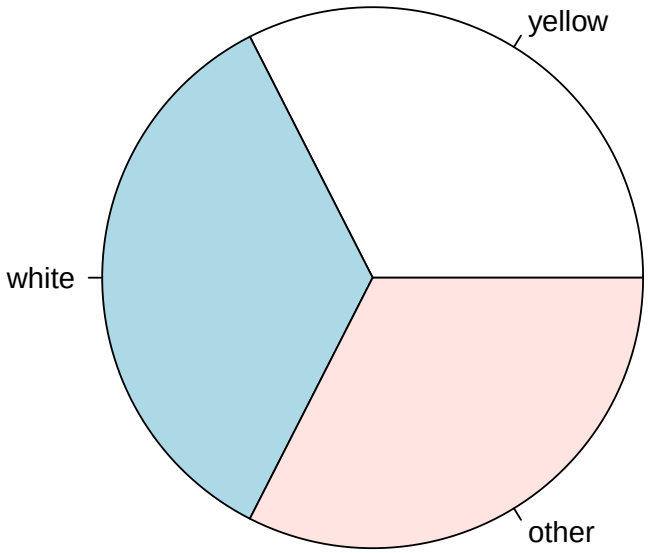
Diagramme für Häufigkeiten

- Aus inhaltlicher Sicht gibt es keine großen Unterschiede zwischen den verschiedenen Diagrammtypen, und über Geschmack läßt sich nicht streiten.
- Horizontale Diagramme sind meist leichter zu lesen als vertikale, Punkte dominieren optisch nicht so stark wie Balken.
- Stabdiagramme werden oft verwendet, um zu signalisieren, daß nur eine einzige Zahl als Ausprägung vorkommen kann.

Kreis- (Torten-)diagramme

- Flächen der Kreissektoren proportional zu den Häufigkeiten: Winkel des Kreissektors $j = f_j \cdot 360°$.
- Wohl am stärksten überschätztes Diagramm: Für mehr als $2 - 3$ Klassen sind Kresidiagramme *sehr schlecht lesbar* (insbesonders in Perspektive), es gibt fast nichts, was nicht ein Balkendiagramm besser darstellen würde.
- Grund: Menschen können Längen viel besser schätzen als Winkel.
- Ausnahme: absolute Mehrheiten.

Kreisdiagramm



Histogramme

- Bei quantitativen Merkmalen, die klassiert wurden, können die Kategorien in einem Balkendiagramm an der "richtigen" Position auf der x-Achse eingezeichnet werden.
- Da von der nicht klassierten Variablen in der Regel alle Werte auf der x-Achse angenommen werden können, werden die Abstände zwischen den Balken weggelassen.
- Das Resultat ist ein Histogramm der Daten.

Histogramme

Formale Konstruktion:

- Klassen $[c_0, c_1)$, $[c_1, c_2)$, $\ldots$, $[c_{k-1}, c_k)$
- Breite der Klassen $d_j = c_j - c_{j-1}$
- Rechtecke der Höhe h_j/d_j (absolute Häufigkeiten) bzw. f_j/d_j (relative Häufigkeiten).
  Fläche jedes Rechteckes ist dann gleich der Häufigkeit der Klasse, die Gesamtfläche des Histogramms ist n bzw. 1.
- Falls alle d_j denselben Wert haben, wird in der Regel nicht durch die Klassenbreite dividiert, dann können die Häufigkeiten direkt an der Höhe der Rechtecke abgelesen werden.

Histogramme

- Anzahl der Klassen: schwieriges Problem, in der Regel sollten immer mehrere Werte ausprobiert werden.
- Faustregel: $k \approx \sqrt{n}$
- "Schöne" Grenzen erhöhen Lesbarkeit.

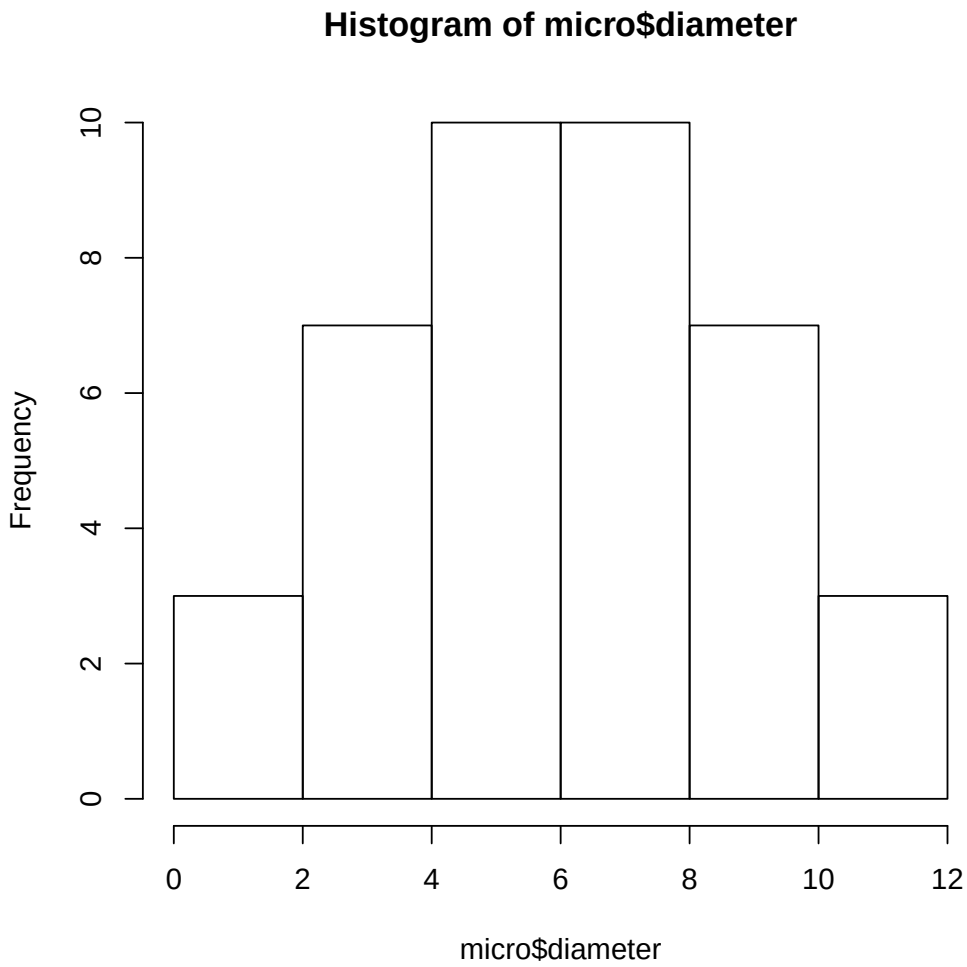
Klassierung

- Die Klassierung von quantitativen Merkmalen kann extrem stark von den Klassengrenzen abhängen.
- Oft enthalten Daten gerundete Werte, bzw. werden "“schöne”" Werte statt der tatsächlichen Werte angegeben.
- Diese werden aber oft auch als "“natürliche”" Klassengrenzen empfunden.
- In der Praxis empfiehlt es sich, Klassengrenzen zu variieren und die Auswirkungen auf das Resultat zu beobachten.

Häufigkeitstabelle

	H	h	%
0-2	3	0.075	7.5
2-4	7	0.175	17.5
4-6	10	0.250	25.0
6-8	10	0.250	25.0
8-10	7	0.175	17.5
10-12	3	0.075	7.5

Histogramm


Histogram of micro$diameter
Frequency
0
2
4
6
8
10
0
2
4
6
8
10
12
micro$diameter

Anzahl der Modi

Unimodal: Histogramm hat nur ein Maximum.

Bimodal: Histogramm hat zwei lokale oder globale Maxima.

Multimodal: Histogramm hat mehr als zwei lokale oder globale Maxima.

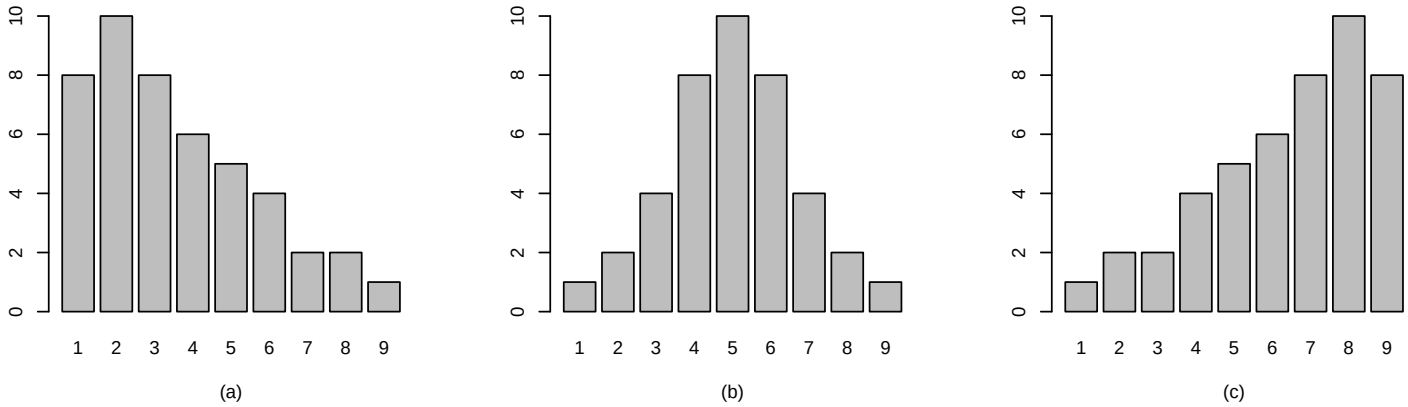
Achtung: Modalität hängt oft von Anzahl der Klassen ab!

Symmetrie und Schiefe

symmetrisch $\Leftrightarrow$ Rechte und linke Hälfte der Verteilung sind annähernd zueinander spiegelbildlich

linkssteil (rechtsschief) $\Leftrightarrow$ Verteilung fällt nach links deutlich steiler und nach rechts langsamer ab

rechtssteil (linksschief) $\Leftrightarrow$ Verteilung fällt nach rechts deutlich steiler und nach links langsamer ab



Eine linkssteile (a), symmetrische (b) und rechtssteile Verteilung (c)

Beschreibung von Stichproben

- Neben Grafiken wie Histogrammen zur Beschreibung von Verteilungen sind oft auch numerische Beschreibungen notwendig, sogenannte *statistische Kennzahlen*.
- Die absoluten oder relativen Häufigkeiten der Ausprägungen codieren die Stichprobe *ohne Informationsverlust* (bis auf die Reihenfolge).
- Für kategorische Größen mit wenigen Kategorien sind die Häufigkeiten der Ausprägungen auch die üblichste numerische Beschreibung der Daten.
- Bei numerischen Größen gibt es in der Regel viele verschiedene Ausprägungen, eine Möglichkeit der Zusammenfassung ist die Klassierung der Daten.
- Es gibt jedoch noch eine Reihe weiterer Kennzahlen für numerische Größen.

Einfache Kennzahlen

Numerische Stichzahlen werden üblicherweise durch 2 Typen von Kennzahlen beschrieben:

Lage (Lokation): Wo liegt das Zentrum der Verteilung?

Streuung: Wie stark schwanken die Beobachtungen rund um das Zentrum?

Für beide Typen gibt es mehrere verschiedene Kennzahlen, die unterschiedliche Fragen beantworten.

Lage: arithmetisches Mittel $\bar{x}$

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

Für Häufigkeitsdaten gilt:

$$\bar{x} = \frac{1}{k} \sum_{j=1}^{k} a_j h(a_j) = \sum_{j=1}^{k} a_j f(a_j)$$

Bei klassierten Daten ist dies Näherung für $\bar{x}$.
Eigenschaften: $\sum(x_i - \bar{x}) = 0$, $\sum(x_i - \bar{x})^2$ minimal

Lage: arithmetisches Mittel $\bar{x}$

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

Für Häufigkeitsdaten gilt:

$$\bar{x} = \frac{1}{k} \sum_{j=1}^{k} a_j h(a_j) = \sum_{j=1}^{k} a_j f(a_j)$$

Bei klassierten Daten ist dies Näherung für $\bar{x}$.
Eigenschaften: $\sum(x_i - \bar{x}) = 0$, $\sum(x_i - \bar{x})^2$ minimal

Lage: Median x_{med}

Der Median ist der mittleren Wert der *geordneten* Stichprobe

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

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

Bei klassierten Daten verwendet man als Näherung die Mitte der Klasse, die die kumulierte relative Häufigkeit 0.5 enthält.
Eigenschaften: Hälfte der x_i kleiner, andere Hälfte größer.
$\sum |x_i - x_{med}|$ minimal, ab Ordinalskala verwendbar.

Lage: Median x_{med}

Der Median ist der mittleren Wert der *geordneten* Stichprobe

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

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

Bei klassierten Daten verwendet man als Näherung die Mitte der Klasse, die die kumulierte relative Häufigkeit 0.5 enthält.
Eigenschaften: Hälfte der x_i kleiner, andere Hälfte größer.
$\sum |x_i - x_{med}|$ minimal, ab Ordinalskala verwendbar.

Lage: Form der Verteilung

Symmetrisch: $\bar{x} \approx x_{med} \approx x_{mod}$

Linkssteil: $\bar{x} > x_{med} > x_{mod}$

Rechtssteil: $\bar{x} < x_{med} < x_{mod}$

Einfluß von Ausreißern

Eine einzige Beobachtung kann den Mittelwert beliebig verändern:

$$x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4, x_5 = 5$$

$$\bar{x} = x_{med} = 3$$

$$x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4, x_5 = 50$$

$$\bar{x} = 12, \qquad x_{med} = 3$$

Einfluß von Ausreißern

Eine einzige Beobachtung kann den Mittelwert beliebig verändern:

$$x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4, x_5 = 5$$

$$\bar{x} = x_{med} = 3$$

$$x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4, x_5 = 50$$

$$\bar{x} = 12, \qquad x_{med} = 3$$

Lage: Getrimmtes Mittel

Um den Einfluß von Randdaten auf den Mittelwert einzuschränken, werden diese manchmal weggelassen.
Für das k% getrimmte Mittel werden die k/2% kleinsten Daten und k/2% größten Daten aus der Stichprobe entfernt, und danach der Mittelwert berechnet.

Beispiel:

$$1, 2, 3, 4, 50$$
$$2, 3, 4$$

$$\bar{x}_{trim} = (2 + 3 + 4)/3 = 3$$

Lage: Getrimmtes Mittel

Um den Einfluß von Randdaten auf den Mittelwert einzuschränken, werden diese manchmal weggelassen.
Für das k% getrimmte Mittel werden die k/2% kleinsten Daten und k/2% größten Daten aus der Stichprobe entfernt, und danach der Mittelwert berechnet.
Beispiel:

$$1, 2, 3, 4, 50$$

$$2, 3, 4$$

$$\bar{x}_{trim} = (2 + 3 + 4)/3 = 3$$

Lage: Getrimmtes Mittel

Um den Einfluß von Randdaten auf den Mittelwert einzuschränken, werden diese manchmal weggelassen.
Für das k% getrimmte Mittel werden die k/2% kleinsten Daten und k/2% größten Daten aus der Stichprobe entfernt, und danach der Mittelwert berechnet.
Beispiel:

$$1, 2, 3, 4, 50$$
$$2, 3, 4$$

$$\bar{x}_{trim} = (2 + 3 + 4)/3 = 3$$

Lage: Getrimmtes Mittel

Um den Einfluß von Randdaten auf den Mittelwert einzuschränken, werden diese manchmal weggelassen.
Für das k% getrimmte Mittel werden die k/2% kleinsten Daten und k/2% größten Daten aus der Stichprobe entfernt, und danach der Mittelwert berechnet.
Beispiel:

$$1, 2, 3, 4, 50$$
$$2, 3, 4$$

$$\bar{x}_{trim} = (2 + 3 + 4)/3 = 3$$

Lokationsmaße

```
> mean(micro$diameter)    # Mittelwert
[1] 5.9
> mean(micro$diameter, trim=0.05)    # 10% getrimmtes Mittelwert
[1] 5.925
> median(micro$diameter)    # Median
[1] 6
```

Streuung

Information über die Lage einer Verteilung ohne begleitende Information zur Streuung *ist wertlos*. Wie weit sind die Beobachtungen von der Lokation der Verteilung entfernt?
Bei Mittelwert und Median haben wir bereits gesehen, daß diese zwei verschiedene Abstandsmaße minimieren:

$$\text{Mittelwert} \leftrightarrow \text{Quadratabstand}$$
$$\text{Median} \leftrightarrow \text{Absolutabstand}$$

Daraus ergeben sich zwei populäre Kennzahlen für die Streuung einer Verteilung.

Varianz und Standardabweichung

Lagemaß: Mittelwert
Distanz: Quadratabstand $(x_i - \bar{x})^2$
Streuungsmaß: Mittelwert der Quadratabstände

Varianz:

$$s^2 = \frac{1}{n-1}\left[(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2\right] = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Standardabweichung:

$$s = \sqrt{s^2}$$

Eigenschaften: nur für quantitative Merkmale, Verschiebungssatz, Transformationsregel, s^2 sind sehr ausreisserempfindlich wegen $(\ldots)^2$.

Varianz und Standardabweichung

- *Verschiebungssatz* zum schnelleren Rechnen (nicht am Computer!):

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

- *Transformationsregel*:

Für $y_i = a + bx_i$ gilt:

$$s_y^2 = b^2 s_x^2 \quad \text{bzw.} \quad s_y = |b| s_x \,.$$

Median der Absolutdistanzen

Lagemaß: Median
Distanz: Absolutabstand $|x_i - x_{med}|$
Streuungsmaß: Median der Absolutabstände

MAD (median absolut deviation):

$$\text{median}(|x_i - x_{med}|)$$

Einfache Interpretation: Die Hälfte der Stichprobe ist weniger als der MAD vom Median entfernt, die andere Hälfte weiter weg.
Eigenschaften: nur für quantitative Merkmale.

Streuungsmaße

```
> var(micro$diameter)

[1] 8.707179

> sd(micro$diameter)    # Streuung (standard deviation)

[1] 2.950793

> mad(micro$diameter)

[1] 2.9652
```

Gruppenvergleich