Stochastik für die Naturwissenschaften

Dr. C.J. Luchsinger

2. Beschreibende Statistik (descriptive Statistics)

Literatur Kapitel 2

* Storrer: Kapitel 29 - 31
* Stahel: Kapitel 1-3
* Statistik in Cartoons: Kapitel 1-2

2.1 Merkmale und Skalen (Scales)

In diesem ersten Teil werden wir eine wichtige Klassifikation von Datentypen vornehmen. Es ist davon auszugehen, dass die meisten von Ihnen im Umgang mit Daten automatisch und gefühlsmässig richtig handeln würden. Ich nehme in dieser Vorlesung aus Zeitgründen und weil es intellektuell nicht sehr anspruchsvoll ist dieses Kapitel nur kurz zusammenfassend durch. Es ist aber wichtig: lesen Sie es also bitte im Verlauf der ersten Semesterwoche durch und lösen Sie aufmerksam die entsprechenden Aufgaben in den Übungen.

In Umfragen/Untersuchungen kann man von verschiedenen Personen zum Beispiel unter anderem auch das **Geschlecht** registrieren. Es wird bei Menschen entweder männlich oder weiblich sein. In EDV-Systemen wird diese Information oft nicht ausgeschrieben als "männlich" oder "weiblich" abgelegt, sondern man codiert diese Information mit 0 und 1 (bzw. 1 und 0) und gibt in der Beschreibung der Statistik an, ob 0 bzw. 1 männlich oder weiblich bedeutet.

Sie haben dann eine Kolonne mit lauter Folgen wie

$$\ldots, 0, 1, 1, 0, 0, 1, 1, 1, 0, 1 \ldots \tag{1}$$

Eine weitere Zahlenfolge in der gleichen Erhebung kann zum Beispiel **Schweregrad einer Krankheit** sein. Eine Krankheit ist bei den Personen 0 = nicht vorhanden, 1 = in der leichtesten Form vorhanden, 2 = in einer mittelschweren Form vorhanden oder gar 3 = in schwerer Form vorhanden. Die entsprechenden Zahlenfolgen sehen dann also zum Beispiel auszugsweise folgendermassen aus:

$$\ldots, 0, 0, 3, 0, 2, 1, 3, 0, 0, 1 \ldots \tag{2}$$

Eine weitere Zahlenfolge in der gleichen Erhebung kann zum Beispiel die **Körpertemperatur der Person** sein. Diese wird hier in Celsius angegeben und sieht folgendermassen aus:

$$\ldots, 36, 36, 37, 39, 41, 36, 37 \ldots \tag{3}$$

Eine weitere Zahlenfolge in der gleichen Erhebung kann zum Beispiel das **Gewicht der Person** sein. Dieses wird hier in kg angegeben und sieht folgendermassen aus:

$$\ldots, 80, 66, 73, 104, 82, 51, 73 \ldots \tag{4}$$

Vergleichen Sie die 4 Datentypen miteinander. Überlegen Sie sich, ob zum Beispiel die Differenzen von 2 Zahlen eine konkrete Bedeutung haben oder gar das Verhältnis; (wo) ist es sinnvoll, das arithmetische Mittel zu bestimmen?

Eine erste Ordnung von Merkmalen sieht folgendermassen aus (Storrer Seite 4):

Des weiteren kann man sich überlegen, welche (Rechen-)Operationen wo sinnvoll sind. Dazu unterscheiden wir 4 Arten von Merkmalen (Storrer Seite 8):

Nominalskala	Keine Rangordnung; Rechenoperationen sinnlos (qualitatives Merkmal)
Ordinalskala	**Rangordnung festgelegt**; Rechenoperationen sinnlos (qualitatives Merkmal)
Intervallskala	Rangordnung festgelegt; **Abstände zwischen Zahlen haben Bedeutung**, Nullpunkt willkürlich, Bildung von Verhältnissen (und Prozenten) nicht sinnvoll (quantitatives Merkmal)
Verhältnisskala	Rangordnung festgelegt; Abstände zwischen Zahlen haben Bedeutung, **Nullpunkt absolut**, Bildung von Verhältnissen (und Prozenten) sinnvoll (quantitatives Merkmal)

Lesen Sie jetzt (nach der Vorlesung) Kapitel 29 von Storrer durch. Dazu in der Vlsg ein paar Bemerkungen zum Begriff "abzählbar unendlich" (countably infinite) (vgl Storrer p 3).

https://schweizermonat.ch/sag-mir-wo-die-zahlen-sind-wo-sind-sie-geblieben/
https://schweizermonat.ch/stichprobenauswahl/

2.2 Darstellung von Versuchsergebnissen

Wir folgen hier den Ausführungen in Storrer ab p 10: DISKRET (discrete): Wir betrachten Prüfungsnoten (welche ev. sehr umfangreich sein können - vgl diese Vlsg mit über 600 Studis). Anstelle einer ungeordneten *Urliste* oder *Rohdaten* wollen wir die Daten übersichtlicher darstellen. Eine erste Zwischenstufe kann eine *Strichliste* sein. Anstelle von (n=) 40 (bzw über 600 Zahlen) haben wir jetzt lediglich eine Folge von 11 Zahlen. Es sind dies die *absoluten Häufigkeiten* (absolute frequencies) des betrachteten Merkmalswerts. Abstrakt können wir dies formalisieren in der folgenden Art: die möglichen Werte seien

$$\omega_1, \omega_2, \ldots, \omega_k.$$

In unserem Beispiel ist $\omega_1 = 1, \omega_2 = 1.5, \ldots, \omega_{11} = 6$ und somit $k = 11$. Die absolute Häufigkeit des Merkmalswerts ω_i bezeichnen wir mit H_i (hier also $H_1 = 1, \ldots, H_{11} = 3$). Es gilt dann immer

$$\sum_{i=1}^{k} H_i = n.$$

Aussagekräftiger als die *absoluten* Häufigkeiten H_i sind die *relativen* Häufigkeiten (relative frequencies) h_i; dazu teilen wir die H_i jeweils durch die Anzahl der Untersuchungsobjekte n:

$$h_i := \frac{H_i}{n} \qquad \text{(Theorie)}$$

bzw

$$h_i := \frac{H_i}{n} * 100\%. \qquad \text{(Praxis)}$$

In Storrer p 11 finden wir in einer Tabelle die Prüfungsnoten derart kompakt zusammengefasst.

Bilder sagen mehr als 1000 Worte und wohl auch mehr als Zahlen; wir wenden uns kurz p 12 in Storrer zu; das *Stabdiagramm* (bar chart) gibt auf den ersten Blick wohl den schnellsten und besten Überblick.

STETIG (continuous): In 2.1 haben wir kurz diskrete und stetige Daten unterschieden. Die Prüfungsnoten sind klar diskret (von Lehrer/innen und Dozent/innen so gemacht). Wenn man 50 Küken wiegt, so ist das Gewicht (abgesehen von physikalisch / philosophischen Fragen) sicher stetig. Wir werden aber meist (unbewusst) Klassen bilden: wir fassen alle Küken mit Gewicht zwischen 103.5 und 104.5 zu 104 g zusammen. Wieder können wir sowohl Urliste wie auch Strichliste und (absolute/relative) Häufigkeiten in einer Tabelle zusammenfassen (Storrer p 12-13). Anstelle des Stabdiagramms (diskret - Stäbe berühren sich nicht) haben wir jetzt ein sogenanntes *Histogramm oder Blockdiagramm*. Die Balken liegen hier aneinander - dies soll illustrieren, dass eigentlich jeder Wert vorkommen kann. Selbsterklärende Begriffe sind dann auch *Klassenbreite (bin width) und Klassenmitte* - Storrer p 14 oben.

Zur Frage der *Klassenbreite*: Frage ans Publikum: Wie gross sollte man diese wählen?

Man sollte nach Möglichkeit darauf achten, dass der Flächeninhalt total 1 ist (vor allem, wenn man thematisch auf eine Dichtefunktion (vgl Kapitel 4) hinarbeiten will). Der Klassenwechsel wird in dieser Vorlesung nicht wichtig sein. Unbestritten sind die Bemerkungen auf Storrer p 15 unten, wenn man in einem Diagramm mehrere Klassenbreiten hat.

(Storrer p 16: Idealisierung I)

Eine weitere wichtige Darstellung ist die Summenhäufigkeitsverteilung (empirical cumulative distribution function). DISKRET: Wir betrachten erneut die Noten der Prüfung. Frage sei:

"Wie viele Personen haben eine ungenügende Note?"

Das ist die Frage nach der Anzahl (oder Anteil) Personen mit Note

$$\leq 3.5.$$

Dies ist leider nicht direkt aus der Tabelle oder dem Stabdiagramm ablesbar. Man muss die Häufigkeit der Noten bis ≤ 3.5 addieren. Dies macht man vorteilhaft gleich für alle möglichen Fragen dieser Art und kommt dann zur *Summenhäufigkeit oder kumulativen Häufigkeit*. Wir gelangen so zu den Darstellungen in Storrer p 17 und 18 (graphische Darstellung desselben).

STETIG: Im Kükenbeispiel von Storrer p 19, 20 und 21 tritt hier ein kleines Problem auf: weil das Merkmal stetig ist und wir Klassen gebildet haben, kommen sämtliche Gewichte, welche eigentlich im Intervall

$$(86.5, 87.5]$$

wären, auf einem Punkt 87 zu liegen. Wenn man dann die graphische Darstellung der Summenhäufigkeit bildet, kann man sich (zu Recht) darüber streiten, wo der Sprung stattfinden sollte (Klassenmitte oder Ende). Weiter hat eine grosse Klassenbreite hier einen ganz klaren Nachteil; der Verlust an Information kommt klarer zum Vorschein als bei den Histogrammen.

(Storrer p 21, 22: Idealisierung II)

Lesen Sie jetzt (nach der Vorlesung) Kapitel 30 von Storrer durch.

2.3 Statistische Masszahlen

"Die Armen werden immer ärmer,
die Reichen immer reicher!"

Wer von Ihnen glaubt, dass diese Aussage richtig ist?

Wir haben auf der letzten Seite bei einer bekannten und trivialen Aussage gesehen, dass Daten sehr komplex sein können. Wünsche darf man haben: wir wollen komplexe Datensätze mit möglichst wenigen Zahlen möglichst umfassend beschreiben.
Für's erste dienen dazu **Lagemasse** (measures of location; Durchschnitt (average, mean), Median (median), Modus (mode)) sowie **Streuungsmasse** (measures of scale; Variationsbreite, Interdezilbereich, Varianz (Variance), Standardabweichung (Standard deviation)).
Betrachten wir erstmal das arithmetische Mittel (Durchschnitt) einer Folge von $n = 7$ Zahlen (zB Noten):

$$3.5, 4, 4, 6, 2.5, 6, 5.$$

Der Durchschnitt ist hier:
Allgemein definieren wir für eine Folge von n reellen Zahlen

$$x_1, x_2, \ldots, x_n$$

das arithmetische Mittel (oder auch Durchschnitt) als

$$\overline{x} := \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

In Storrer p 28 finden Sie noch den Fall der Berechnung eines Durchschnitts, wenn die Daten in Klassen eingeteilt sind.
In der einführenden Diskussion haben wir gesehen, dass bei Löhnen nicht einfach über den Durchschnitt gesprochen werden sollte. Mindestens der Median $\tilde{x}$ sollte auch angeschaut werden. Was ist der Median in obigem Datensatz von Noten

$$3.5, 4, 4, 6, 2.5, 6, 5?$$

Problem wenn n gerade Zahl?

Vorteil von $\tilde{x}$ gegenüber $\overline{x}$?

In Storrer p 30 - 31 finden Sie noch den Fall der Berechnung von $\tilde{x}$, wenn die Daten in Klassen eingeteilt sind.

Als letzter Ausdruck folgt der Modus: In der Legislaturperiode 2019-2023 ist die Sitzverteilung im Zürcher Kantonsrat folgendermassen: AL 6, Die Mitte 8, EDU 4, EVP 8, FDP.Die Liberalen 29, GLP 23, Grüne 22, SP 35, SVP 45. Der Modus liegt hier bei der SVP - sie hat zur Zeit am meisten Sitze!

Arithmetisches Mittel, Median und Modus können alle gleich sein (Situation A) oder auch zum Beispiel alle verschieden (Situation B):

Ein Artikel vom Dozenten: https://schweizermonat.ch/durchschnitt-oder-median

Wir kommen zu den Streuungsmassen - 2 davon kann man sehr schnell abhaken: die Variationsbreite ist einfach der grösste Wert minus den kleinsten Wert (bei den Noten also $6 - 2.5 = 3.5$). Das ist einfach die Spannweite der Daten. Man sollte nie nur diese Zahl anschauen (Ausreisser! (outlier)). Wir können bei den der Grösse nach geordneten Daten einfach die kleinsten 10 % und die grössten 10 % der Daten ignorieren und die Spannweite der verbleibenden Daten berechnen - das ist der Interdezilbereich. Man ist dann schon robuster gegenüber Ausreissern.

Nach diesen (sehr einfachen) Massen kommen zwei zentral wichtige Masse: Varianz und Standardabweichung. Wir wollen ein sinnvolles Mass für die Streuung der Daten um das Lagemass (zum Beispiel um das arithmetische Mittel) definieren. Wenn wir Daten

$$x_1, x_2, \ldots, x_n$$

mit arithmetischem Mittel $\overline{x}$ haben, so kann man die Werte

$$(x_1 - \overline{x}), (x_2 - \overline{x}), \ldots, (x_n - \overline{x})$$

berechnen (Differenz zum arithmetischen Mittel). Berechnen Sie hiervon den Durchschnitt:

Nach diesem Fehlschlag fallen folgende Vorschläge ein, um ein sinnvolles Streuungsmass zu definieren:

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}|$$

und

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

Beide Masse werden (zu Recht) verwendet. Wir werden uns in dieser Vorlesung auf den zweiten Ausdruck mit dem Quadrat konzentrieren (es ist mathematisch übrigens einfacher).

Die Summe

$$SS_{xx} := \sum_{i=1}^{n} (x_i - \overline{x})^2 \qquad (\textbf{S}\text{um of } \textbf{S}\text{quares x and x})$$

und Varianten werden in dieser Vorlesung noch oft vorkommen - untersuchen wir sie:

Wir definieren die empirische Varianz s^2 als

$$s^2 := \frac{1}{n-1} SS_{xx}.$$

Wir werden später in dieser Vlsg begründen, warum wir hier durch $(n-1)$ statt durch n teilen. Nach obigen Berechnungen haben wir also die folgenden drei gleichwertigen Formeln zur Auswahl:

$$\begin{aligned} s^2 &= \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 \\ &= \frac{1}{n-1} \big(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2 \big) \\ &= \frac{1}{n-1} \big(\sum_{i=1}^{n} x_i^2 - \frac{1}{n} (\sum_{i=1}^{n} x_i)^2 \big). \end{aligned}$$

Berechnen Sie bei den Noten $3.5, 4, 4, 6, 2.5, 6, 5$ die empirische Varianz (HA): 1.7024

Teilt R (oder Ihr TR) durch n oder $n-1$?

Das letzte Streuungsmass, welches wir definieren, ist die empirische Standardabweichung (in R "sd(a)" für Englisch **s**tandard **d**eviation):

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

Bemerkung zu Varianz und Standardabweichung: Die Varianz ist ein quadratisches Mass; wenn wir davon die Wurzel ziehen, erhalten wir ein Mass, welches mit den Differenzen direkt in Bezug gesetzt werden kann.

Aufgabe 31-8:

Untersuchen Sie $\bar{x}$ und s^2 wenn $n = 1$:

Wichtig:

1. Lesen Sie jetzt das komplette Kapitel im Storrer II selber durch (Kapitel 29-31).
2. Lösen Sie danach mindestens 5 Aufgaben hinten im Kapitel und vergleichen Sie mit den Lösungen am Schluss des Buches. Bei Bedarf lösen Sie mehr Aufgaben.
3. Gehen Sie in die Übungsstunde. Drucken Sie das Übungsblatt dazu *vorher* aus, lesen Sie *vorher* die Aufgaben durch und machen sich erste Gedanken dazu (zum Beispiel, wie man sie lösen könnte).
4. Dann lösen Sie das Übungsblatt: zuerst immer selber probieren, falls nicht geht: Tipp von Mitstudi benutzen, falls immer noch nicht geht: Lösung von Mitstudi anschauen, 1 Stunde warten, versuchen, aus dem Kopf heraus wieder zu lösen, falls immer noch nicht geht: Lösung von Mitstudi abschreiben (und verstehen - also sollte man insbesondere keine Fehler abschreiben!).
5. Lösen Sie die entsprechenden Prüfungsaufgaben im Archiv.