Übung 1 – Theorie

Querschnittsdaten mit Observationen $i \in \{1, \dots, N\}$ [Teils gruppiert in Gruppen $k \in \{1, \dots, K\}$]

↓

Wir wollen diese Daten verstehen (und später Effekte schätzen und testen)

↓

Zuerst beschreiben wir die Daten (deskriptive Statistik; wie ist Variable X verteilt?)

↓

(i) Wo liegt das "Zentrum" dieser Variablen X und (ii) wie stark ist die Streuung?

(i) • Mittelwert $\bar{X} = \frac{1}{N} \sum_{i=1}^{N} x_i$ $\left[\bar{X} = \sum_{k=1}^{K} f_k \bar{x}_k \right]$

• Median (X) = "50% der Observationen sind grösser als der Median und 50% der Observationen sind kleiner als der Median."

(ii) • Varianz $Var(x) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2$

= " Die durchschnittliche quadratische Abweichung vom Mittelwert $\bar{x}$."

• Standardabweichung $SA(x) = \sqrt{Var(x)}$

Statistik
Übungsaufgaben 1

Deskriptive Statistik - Klassifizierung - Kennzahlen

1. Klassifizieren Sie folgende Daten

   a) Tägliche Aktienrenditen von 100 Firmen zwischen 1995 und 2020

   b) Donald Trump Tweets während seiner Präsidentschaft

   c) Körperlänge der beobachteten Walhaie an einem spezifischen Tag in Westaustralien

   d) Wöchentlicher Fleischkonsum in kg an der UZH Mensa zwischen 1914 und 2020

   e) Einfluss von LSD auf die Netzstruktur von Spinnen

2. Betrachten Sie folgende Rohdaten

$$2, 5, 5, 4, 4, 3, 6, 7, 6, 5, 3, 4$$

   a) Erstellen Sie eine Häufigkeitstabelle.

   b) Berechnen Sie Mittelwert und Median.

   c) Wie ändert sich der Mittelwert und der Median, wenn Sie im obigen Datensatz die 7 in eine 43 verwandeln?

   d) Erklären Sie, wie die Beziehung zwischen Mittelwert und Median Informationen über die Symmetrie einer Verteilung liefert.

B.1.11.2. Calculate the variance for samples where

   a) $N = 10$, $\sum_{i=1}^{N} X_i^2 = 84$, and $\sum_{i=1}^{N} X_i = 20$.

   b) $N = 40$, $\sum_{i=1}^{N} X_i^2 = 380$, and $\sum_{i=1}^{N} X_i = 100$.

(Hint: Think about how you can rewrite the normal variance formula.)

3. Für N Beobachtungen sei $\bar{x}_g$ der Mittelwert in Gruppe g und $f_g = \frac{N_g}{N}$ die relative Häufigkeit der Gruppe g. Zeigen Sie dass folgendes gilt:

$$\bar{x} = \sum_{g=1}^{G} f_g \, \bar{x}_g.$$

4. Warum ist die mittlere Sterbeziffer für Krebs zwischen 1970 und 1999 angestiegen?

	1970		1999	
Alter	Anteil	Sterbeziffer	Anteil	Sterbeziffer
1-14	0.24	7	0.18	2
15-44	0.46	22	0.43	15
45-64	0.20	329	0.25	229
65-84	0.09	1395	0.12	1255
85+	0.01	2830	0.02	3002
Gesamt		231.45		274.7

5. Berechnen Sie die Standardabweichung für die Daten in Aufgabe 4 (jeweils für 1970 und 1999).

1. Klassifizieren Sie folgende Daten

   a) Tägliche Aktienrenditen von 100 Firmen zwischen 1995 und 2020

   b) Donald Trump Tweets während seiner Präsidentschaft

   c) Körperlänge der beobachteten Walhaie an einem spezifischen Tag in Westaustralien

   d) Wöchentlicher Fleischkonsum in kg an der UZH Mensa zwischen 1914 und 2020

   e) Einfluss von LSD auf die Netzstruktur von Spinnen

Einschub: Struktur der Daten

Querschnittsdaten
(zu einem Zeitpunkt)

	Körpergrösse	Ausbildung
Ind. 1	1.80	Bachelor
Ind. 2	1.60	Master
⋮	⋮	⋮
Ind. N	1.75	Lehre

$\underbrace{}$ N Observationen

Paneldaten
(zu mehreren Zeitpunkten)

	Körpergrösse	Ausbildung
Ind. 1 zu Zeitpunkt 1	1.80	Bachelor
Ind 2, t = 1	1.60	Master
⋮	⋮	⋮
Ind. N, t = 1		
⋮	⋮	⋮
Ind 1., t = T		
⋮	⋮	⋮
Ind. N, t = T		

$\underbrace{}$ N x T Observationen

Zeitreihen

	Körpergrösse Ind. 1
t = 1	1.80
t = 2	1.81
⋮	⋮
t = T	1.85

$\underbrace{}$ T Observationen

Strukter, Verfügbarkeit, Skala, Herkunft

(a) Paneldaten, meistens öffentlich, metrisch stetig, hergeleitet aus Preisdaten

(b) Zeitreihe, öffentlich, Textdaten, Primärdaten

(c) Querschnittsdaten, können öffentlich sein, metrisch stetig, Primärdaten

(d) Zeitreihe, nicht öffentlich, metrisch stetig, administrative Daten

(e) Experiment

2. Betrachten Sie folgende Rohdaten

$$2, 5, 5, 4, 4, 3, 6, 7, 6, 5, 3, 4$$

a) Erstellen Sie eine Häufigkeitstabelle.

b) Berechnen Sie Mittelwert und Median.

c) Wie ändert sich der Mittelwert und der Median, wenn Sie im obigen Datensatz die 7 in eine 43 verwandeln?

d) Erklären Sie, wie die Beziehung zwischen Mittelwert und Median Informationen über die Symmetrie einer Verteilung liefert.

a)

	h_k	f_k
2	1	1/12
3	2	2/12
4	3	3/12
5	3	3/12
6	2	2/12
7	1	1/12

b) $$\bar{x} = \frac{1}{12} \sum_{i=1}^{12} x_i$$

$$= \frac{1}{12}(2 + 3 + 3 + \ldots + 7)$$

$$= \underline{4.5}$$

Median $\quad Med(x)$

(i) Sortiere $\{x_i\}_{i=1}^{N}$ von klein nach gross

$\{2, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 7\}$

(ii) Berechne den Median als

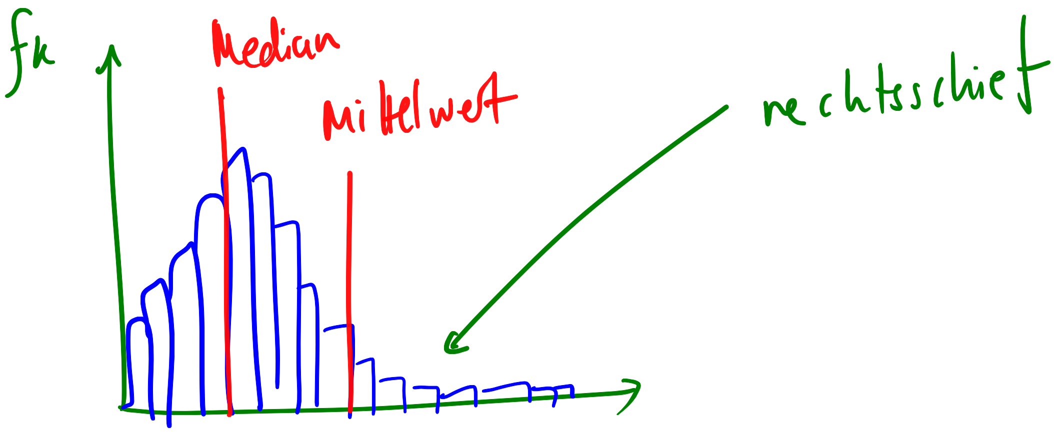
$$Med(x) = \begin{cases} x_{(N+1)/2} & \text{, falls } N \text{ ungerade} \\ \dfrac{x_{(N/2)} + x_{(N/2)+1}}{2} & \text{, falls } N \text{ gerade} \end{cases}$$

→ also: $\text{Med}(x) = \frac{4+5}{2} = \underline{\underline{4.5}}$

(c) Mittelwert $\bar{X}_{NEU} = \underline{\underline{7.5}}$

Median unverändert

(d) Mittelwert + Median einer symmetrischen Verteilung sind gleich.



B.1.11.2. Calculate the variance for samples where

a) $N = 10$, $\sum_{i=1}^{N} X_i^2 = 84$, and $\sum_{i=1}^{N} X_i = 20$.

b) $N = 40$, $\sum_{i=1}^{N} X_i^2 = 380$, and $\sum_{i=1}^{N} X_i = 100$.

(Hint: Think about how you can rewrite the normal variance formula.)

$$Var(X) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

$$= \frac{1}{N} \sum_{i=1}^{N} x_i^2 - 2x_i\bar{x} + \bar{x}^2$$

$$= \frac{1}{N} \left(\sum_{i=1}^{N} x_i^2 - 2\bar{x} \underbrace{\sum_{i=1}^{N} x_i}_{= N\bar{x} = \cancel{N} \cdot \frac{1}{\cancel{N}} \sum_{i=1}^{N} x_i} + \sum_{i=1}^{N} \bar{x}^2 \right)$$

$$= \frac{1}{N} \left(\sum_{i=1}^{N} x_i^2 - \underbrace{2\bar{x}^2 N + N\bar{x}^2}_{-N\bar{x}^2} \right)$$

$$Var(X) = \frac{1}{N} \sum_{i=1}^{N} x_i^2 - \bar{x}^2$$

$$\left(\frac{1}{N} \sum_{i=1}^{N} x_i \right)^2$$

a) $Var(X) = \frac{1}{10} \cdot 84 - \left(\frac{1}{10} \cdot 20\right)^2 = \underline{\underline{4.4}}$

b) $Var(X) = \frac{1}{40} \cdot 380 - \left(\frac{1}{40} \cdot 100\right)^2 = \underline{\underline{3.25}}$

3. Für N Beobachtungen sei $\bar{x}_g$ der Mittelwert in Gruppe g und $f_g = \frac{N_g}{N}$ die relative Häufigkeit der Gruppe g. Zeigen Sie dass folgendes gilt:

$$\bar{x} = \sum_{g=1}^{G} f_g \, \bar{x}_g.$$

$$\bar{x} = \sum_{g=1}^{G} f_g \bar{x}_g = \sum_{g=1}^{G} \frac{\cancel{N_g}}{N} \frac{1}{\cancel{N_g}} \sum_{i=1}^{N_g} x_i$$

$$= \frac{1}{N} \underbrace{\sum_{g=1}^{G} \sum_{i=1}^{N_g}}_{\text{alle Observationen}} x_i$$

4. Warum ist die mittlere Sterbeziffer für Krebs zwischen 1970 und 1999 angestiegen?

f_g $\bar{x}_g$

	1970		1999	
Alter	Anteil	Sterbeziffer	Anteil	Sterbeziffer
1-14	0.24	7	0.18	2
15-44	0.46	22	0.43	15
45-64	0.20	329	0.25	229
65-84	0.09	1395	0.12	1255
85+	0.01	2830	0.02	3002
Gesamt		231.45		274.7

$\bar{x}$

$$\bar{X} = \sum_{g=1}^{G} f_g \bar{x}_g = (0.24 \cdot 7 + 0.46 \cdot 22 + \ldots + 0.01 \cdot 2830)$$

$$= 231.45$$

5. Berechnen Sie die Standardabweichung für die Daten in Aufgabe 4 (jeweils für 1970 und 1999).

$$SA(X) = \sqrt{Var(X)}$$

$$SA(X) = \sqrt{\sum_{g=1}^{G} f_g (\bar{x}_g - \bar{x})^2}$$

$$1970: \quad SA(X) = \sqrt{0.24\left(7 - 231.45\right)^2 + \ldots + 0.01\left(2830 - 231.45\right)^2}$$

$$= \underline{\underline{472.81}}$$

$$1999: \quad SA(X) = \underline{\underline{554.07}}$$