Übung 4

-

Häufigkeitsverteilungen und ihre Parameter

Musterlösung

Aktuelle Version: 19. Juli 2022

Hinweise:

- Übungen sind mit Vorteil alleine zu lösen.
- Benutzen Sie die Musterlösungen nur zur Korrektur.
- Die Übungen sind wichtige Vorbereitungen für die Prüfung. Lösen sie die Übungen sorgfältig und stellen Sie die Lösungswege übersichtlich dar.
- (Ergänzte) Vorlesungsunterlagen und Fachbücher helfen beim Lösen von Übungen und bringen gleichzeitig eine erweiterte Ansicht auf die Problemstellung.
- Wenn Sie die Übungen nicht verstehen, fragen Sie!

Übung 1. *Fragen*

1. Beschreiben Sie die Aufgaben, die die Parameter von Häufigkeitsverteilungen zu erfüllen haben!

   Die Dichte- oder Häufigkeitsverteilung gibt sicher den besten Eindruck, wie die Daten über einen bestimmten Bereich verteilt sind. Dennoch ist es wünschenswert, Kennwerte zu definieren, die eine kompakte Aussage über die Eigenschaften von Daten zulassen. Hierzu wurden einerseits die Lageparameter und weiter die Streuungsparameter definiert.

2. Welche Auffassungen von Mitte liegen Modus, Median und arithmetischem Mittel zugrunde? Beschreiben Sie die Vor- und Nachteile dieser Mittelwerte!

   Dem *Modus* liegt die Auffassung zugrunde, dass das Zentrum der Verteilung beim häufigsten Wert liegt. Der Modus hat den Vorteil, dass er einfach für allen Skalenniveaus bestimmbar und insensitiv gegenüber Ausreissern ist. Nachteil ist, dass er nicht immer eindeutig ist.

   Dem *Median* liegt die Auffassung zugrunde, dass das Zentrum der Verteilung in der mittleren Position der Rangordnung der Werte liegt. Das Zentrum der Verteilung teilt also den sortierten Datensatz in zwei gleich grosse Teile. Der Median hat den Vorteil, dass er bereits für ordinalskalierte Daten bestimmt werden kann und insensitiv gegenüber Ausreissern ist.

   Dem *arithmetischen Mittel* liegt die Auffassung zugrunde, dass das Zentrum der Verteilung dem Durchschnitt der Werte entspricht. Nachteil ist insbesondere, dass er sensitiv gegenüber Ausreissern ist.

3. Erklären Sie den Unterschied zwischen arithmetischem und geometrischem Mittel!

   Das arithmetische Mittel berechnet den Durchschnitt der Werte, das geometrische Mittel die durchschnittliche *Veränderung* der Werte.

4. Beschreiben Sie die beiden zum praktischen Einsatz kommenden Konzepte zur Ermittlung der Streuung!

   Zur Beschreibung der Streuung haben wir zwei Möglichkeiten kennengelernt. Einmal der mittlere absolute Abstand zum Erwartungswert und einmal der mittlere quadratische Abstand zum Erwartungswert.

   Der *mittlere absolute Abstand* ist einheitenbehaftet und erlaubt direkt eine Aussage über die Spannweite eines Mittelwertes und ist nach Möglichkeit in der Beschreibung der erfassten Datenmenge zu benutzen.

   Der *mittlere quadratische Abstand* findet besondere Bedeutung in der schliessenden Statistik und wird uns noch im Zusammenhang mit den Verteilungsfunktionen beschäftigen.

5. Wodurch unterscheidet sich der Variationskoeffizient von den anderen Streuparametern?

   Der Varationskoeffizient misst im Gegensatz zu den anderen Streuparametern nicht die absolute Streuung, sondern die *relative* Streuung bezogen auf die zentrale Tendenz, genauer gesagt die Standardabweichung bezogen auf den arithmetischen Mittelwert.

6. Welche Parameter können mit welchen Skalenniveaus berechnet werden?

Lageparameter	Nominal	Ordinal	Intervall	Absolut
Modus	✓	✓	✓	✓
Median		✓	✓	✓
Perzentil		✓	✓	✓
Arithmetisches Mittel			✓	✓
Harmonisches Mittel				✓
Geometrisches Mittel				✓

Streumass	Nominal	Ordinal	Intervall	Absolut
Spannweite			✓	✓
Zentraler Quartilsabstand			✓	✓
Mittlere absolute Abweichung			✓	✓
Varianz			✓	✓
Standardabweichung			✓	✓
Variationskoeffizient				✓

Übung 2. *Lage- und Streuparameter*
Aus einer Messreihe ergibt sich die folgende Urliste:

$$x_i = (3.2, 3.1, 3.4, 3.6, 3.4, 3.1, 3.3, 1.9, 2.0)$$

1. Bestimmen Sie die folgenden Werte: Arithmetisches Mittel, erstes Quartil, drittes Quartil, geometrisches Mittel, harmonisches Mittel, Maximum, Median, Minimum, mittlere absolute Abweichung, Modus, Spannweite, Standardabweichung, Stichprobenvarianz, Varianz, Variationskoeffizient, zentraler Quartilsabstand. Ordnen sie die Werte den Lage- und Streuparametern zu.

   Lageparameter

   (a) Modus $Mo = 3.1, 3.4$
   (b) Median $Me = 3.2$
   (c) Erstes Quartil $Q1 = \frac{2.0+3.1}{2} = 2.55$
   (d) Drittes Quartil $Q3 = \frac{3.4+3.4}{2} = 3.4$
   (e) Arithmetisches Mittel $\bar{x} = \frac{1}{n}\sum x_i = 3$
   (f) Harmonisches Mittel $\overline{MH} = \frac{n}{\sum \frac{1}{x_i}} = 2.86$
   (g) Geometrisches Mittel $MG = \sqrt[n]{\prod x_i} = 2.93$

   Streuparameter

   (a) Minimum $x_{min} = 1.9$
   (b) Maximum $x_{max} = 3.6$

(c) Spannweite $R = x_{max} - x_{min} = 1.7$

(d) Zentraler Quartilsabstand $ZQA = Q3 - Q1 = 0.85$

(e) Mittlere absolute Abweichung $\delta = \frac{1}{n}\sum |x_i - \bar{x}| = 0.47$

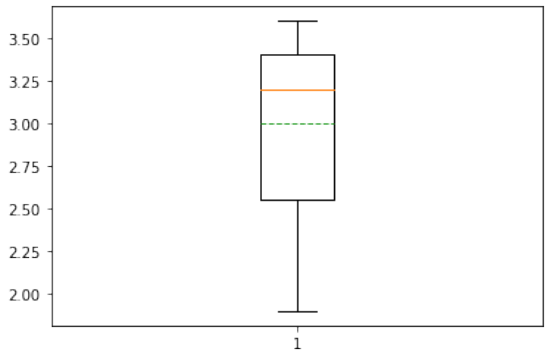
(f) Varianz $\sigma^2 = \frac{1}{n}\sum (x_i - \bar{x})^2 = 0.34$

(g) Standardabweichung $\sigma = 0.58$

(h) Stichprobenvarianz $s^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2 = \frac{n}{n-1}\sigma^2 = 0.38$

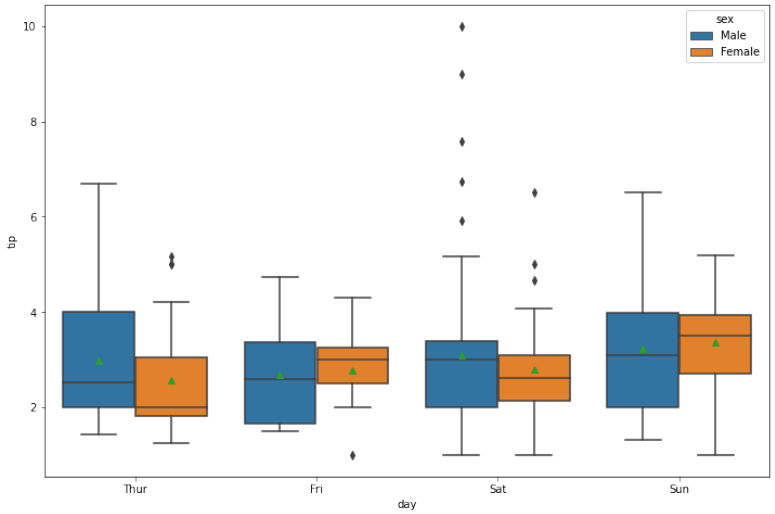
(i) Variationskoeffizient $VK = 100\frac{\sigma}{\bar{x}} = 19.3\%$

2. Zeichnen Sie einen Boxplot!



Übung 3. *Interpretation von Boxplots*

Der folgende Grafik zeigt die Verteilung von gegebenem Trinkgeld nach dem Essen nach Wochentag und Geschlecht.



Interpretieren Sie die Boxplots!

Spannweite Je nach Tag und Geschlecht variiert die Bandbreite an gegebenem Trinkgeld stark. Freitags geben zum Beispiel alle etwa ungefähr gleich viel Trinkgeld, am Samstag variiert es stärker.

Ausreisser Am Samstag gibt es starke Ausreisser, an anderen Tag nicht.

Mitte Der Median liegt am Donnerstag deutlich tiefer als an den Wocheendtagen und relative nahe an Q1, d.h. viele gegeben an diesem Tag wenig Trinkgeld. Der Median liegt am Samstag (Männer) resp. Sonntag (Frauen) deutlich höher als sonst und näher an Q3, d.h. viele gegen an diesem Tag viel Trinkgeld. Durch die Extremwerte wird jedoch durchschnittlich (arithmetisches Mittel) immer etwa gleich viel Trinkgeld gegeben.

Verteilung Die meisten Verteilungen sind rechtsschief: Der Modus ist kleiner als der Mittelwert und/oder die oberen Antennen/Aussreisser sind länger. Einige Verteilungen sind beinahe symmetrisch (Frauen/Freitag, Frauen/Sonntag).

Übung 4. *Mittelwerte*

Eine moderne Abfüllanlage füllt 50.000 Flaschen pro Stunde ab, eine ältere Anlage nur 30.000 Flaschen pro Stunde. Wie viele Flaschen werden durchschnittlich pro Stunde abgefüllt, wenn auf der modernen Anlage 300.000 Flaschen und auf der älteren 150.000 Flaschen abgefüllt werden?

Wir bilden das harmonische Mittel mit $h_1 = 300000, x_1 = 50000$ und $h_2 = 150000, x_2 = 30000$:

$$MH = \frac{\sum h_i}{\sum \frac{h_i}{x_i}} = \frac{h_1 + h_2}{\frac{h_1}{x_1} + \frac{h_2}{x_2}} = 40909$$

Übung 5. *klassifizierte Häufigkeiten*

Bei der Asseveratio AG wurden im letzten September 400 Lebensversicherungsverträge abgeschlossen. Nachstehend ist die klassifizierte Häufigkeitsverteilung für die Versicherungssummen in Tausend CHF angegeben.

Versicherungssumme von	bis	h_i
4	10	20
10	20	160
20	30	80
30	40	40
40	80	88
80	120	12

a) Berechnen Sie die durchschnittliche Versicherungssumme!

i	Versicherungssumme von	bis	h_i	x_i	$x_i h_i$
1	4	10	20	7	140
2	10	20	160	15	2400
3	20	30	80	25	2000
4	30	40	40	35	1400
5	40	80	88	60	5280
6	80	120	12	100	1200
			400		12420

$$\bar{x} = \frac{1}{n} \sum x_i h_i = \frac{12420}{400} = 31050$$

b) Berechnen und interpretieren Sie den Modus, den Median und das 1. Quartil.

Der *Modus* ist die häufigste Nennung. Da hier die Klassenbreite unterschiedlich ist, muss erst die Klasse mit der höchsten Dichte gefunden werden. Die Dichte berechnet sich zu $d_i = \frac{h_i}{x_i^o - x_i^u}$.

i	Versicherungssumme von	bis	h_i	d_i
1	4	10	20	3.3
2	10	20	160	16.0
3	20	30	80	8.0
4	30	40	40	4.0
5	40	80	88	2.2
6	80	120	12	0.3
			400	

Die Modusklasse ist die Klasse höchster Dichte, also Klasse 2 (10000-20000) mit $d_2 = 16.0$. Jetzt berechnet sich der Modus zu:

$$Mo = x_2^u + \frac{d_2 - d_1}{(d_2 - d_1) + (d_2 - d_3)}(x_2^o - x_2^u)$$

$$Mo = 10000 + \frac{16000 - 3300}{(16000 - 3300) + (16000 - 8000)}(20000 - 10000) = 16135$$

Die am häufigsten abgeschlossene Vertragssumme ist also $Mo = 16135$.

Der *Median* ist der Wert der geordneten Mitte. Von 400 ist der Mittelwert 200, diese Anzahl kommt in der 3. Klasse zu liegen. Damit ergibt sich der Median zu

i	Versicherungssumme von	bis	h_i	H_i
1	4	10	20	20
2	10	20	160	180
3	20	30	80	260
4	30	40	40	300
5	40	80	88	388
6	80	120	12	400
			400	

$$Me = x_3^u + \frac{\frac{n}{2} - H_2}{h_3}(x_3^o - x_3^u)$$

$$Me = 20000 + \frac{200 - 180}{80}(30000 - 20000) = 22500$$

Die *erste Quantilklasse* ist bei $\frac{400}{4} = 100$ die zweite Klasse. Analog zum Median berechnet sich das erste Quantil zu:

$$Q1 = x_2^u + \frac{\frac{n}{4} - H_1}{h_2}(x_2^o - x_2^u)$$

$$Q1 = 10000 + \frac{100 - 20}{160}(20000 - 10000) = 15000$$

c) Warum ist der Median deutlich kleiner als das arithmetische Mittel?

Weil (nur) einige höherwertige Verträge abgeschlossen werden.

d) Berechnen Sie die Spannweite, den zentralen Quartilsabstand, den zentralen 80%-Dezilabstand und die mittlere absolute Abweichung!

Die *Spannweite* ist die Differenz zwischen der oberen Klassengrenze der obersten Klasse und der unteren der untersten Klassengrenze:

$$R = x_6^o - x_1^u = 120000 - 4000 = 116000$$

Der *zentrale Quartlisabstand* sind die 50% um den Median. Das dritte Quartil liegt bei 300 und entspricht genau der oberen Grenze der 4. Klasse x_4^o. Somit ist:

$$ZQA = Q3 - Q1 = 40000\breve{}15000 = 25000$$

Das heisst, die mittleren 50% der Versicherungsverträge haben eine Spannweite von $ZQA = 25000$.

Der *zentrale 80%-Dezilabstand* sind die 80% um den Median. Das erste Dezil liegt in der 2. Klasse ($\frac{400}{10} = 40$), das letzte in der 5. Klasse ($\frac{9 \cdot 400}{10} = 360$):

$$D1 = x_2^u + \frac{\frac{n}{10} - H_1}{h_2}(x_2^o - x_2^u)$$

$$D1 = 10000 + \frac{40 - 20}{160}(20000 - 10000) = 11250$$

$$D9 = x_5^u + \frac{\frac{9n}{10} - H_4}{h_5}(x_5^o - x_5^u)$$

$$D9 = 40000 + \frac{360 - 300}{88}(80000 - 40000) = 67270$$

$$I_{80} = D9 - D1 = 67270 - 11250 = 56020$$

Das heisst, die mittleren 80% der Versicherungsverträge haben eine Spannweite von $I_{80} = 56020$.

Die *mittlere absolute Abweichung* gibt die durchschnittliche Entfernung der Werte zum arithmetischen Mittelwert an:

i	Versicherungssumme von	bis	h_i	x_i	$\lvert x_i - \bar{x}\rvert h_i$
1	4	10	20	7	481.0
2	10	20	160	15	2568.0
3	20	30	80	25	484.0
4	30	40	40	35	158.0
5	40	80	88	60	2547.6
6	80	120	12	100	827.4
			400		7066.0

$$\delta = \frac{1}{n}\sum |x_i - \bar{x}|h_i = \frac{7066000}{400} = 17665$$

e) Berechnen Sie die Varianz, die Standardabweichung und den Variationskoeffizienten.

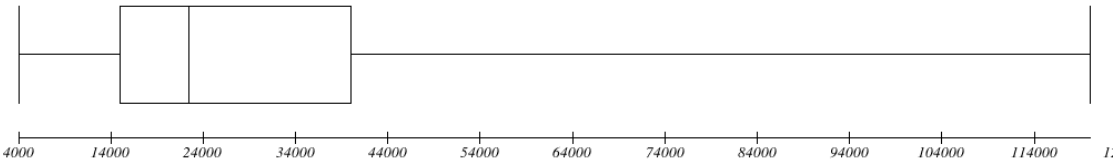
i	Versicherungssumme von	bis	h_i	x_i	$(x_i - \bar{x})^2 h_i$
1	4	10	20	7	11568.1
2	10	20	160	15	41216.4
3	20	30	80	25	2928.2
4	30	40	40	35	624.1
5	40	80	88	60	73753.0
6	80	120	12	100	57049.2
			400		187139.0

$$\sigma^2 = \frac{1}{n}\sum (x_i - \bar{x})^2 h_i = \frac{187139000}{400} = 467800$$

$$\sigma = \sqrt{\sigma^2} = 684$$

$$VK = 100\frac{\sigma}{\bar{x}} = 100\frac{684}{31050} = 2.20\%$$

f) Zeichnen Sie einen Boxplot!



Zusatzaufgaben

Übung 6. *Gewichtetes Mittel*
Vier Schulklassen haben die folgenden Abwesenheitszeiten in Stunden pro Monat:

Klasse	Klassengrösse	Abwesenheitszeit
A	20	4
B	25	7
C	28	2
D	32	11

Wie gross ist durchschnittliche Abwesenheitszeit pro Schüler?

Da der n-fache arithmetische Mittelwert der Summe der einzelnen Werte entspricht

$$n\bar{x} = \sum x_i$$

können wir für die Berechnung des Gesamtdurchschnitts entsprechend die einzelnen Mittelwerte verwenden. Wir gelangen zum gewichteten arithmetischen Mittel (GAM):

$$GAM = \frac{\sum n_j \bar{x}_j}{\sum n_j} = \frac{4 \cdot 20 + 7 \cdot 25 + 2 \cdot 28 + 11 \cdot 32}{20 + 25 + 28 + 32} = 6.31$$

Übung 7. *Berechnung von Varianzen*
Die Varianz kann auf zwei Wegen berechnet werden:

$$\sigma^2 = \frac{1}{n} \sum (x_i - \bar{x})^2 \tag{1}$$

$$\sigma^2 = \frac{1}{n} \sum x_i^2 - \bar{x}^2 \tag{2}$$

1. Welche Gleichung bevorzugen sie für die Implementierung und warum?
   Der Mittelwert berechnet sich:

   $$\bar{x} = \frac{1}{n} \sum x_i$$

   Eine Implementierung der ersten Gleichung wäre:

1 $\bar{x} = 0$, $\sigma^2 = 0$
2 **for** each x_i:
3 $\quad \bar{x}$ += x_i
4 $\bar{x}$ *= $\frac{1}{n}$
5 **for** each x_i:
6 $\quad \sigma^2$ += $(x_i - \bar{x})^2$
7 σ^2 *= $\frac{1}{n}$

Bei der ersten Gleichung muss bereits vor der Berechnung der Varianz einmal über alle x-Werte iteriert werden, bei der zweiten Gleichung hingegen reicht eine Iteration:

1 $\bar{x} = 0$, $\sigma^2 = 0$
2 **for** each x_i:
3 $\quad \bar{x}$ += x_i
4 $\quad \sigma^2$ += x_i^2
5 $\bar{x}$ *= $\frac{1}{n}$
6 $\sigma^2 = \frac{1}{n}\sigma^2 - \bar{x}^2$

Die zweite Gleichung bedeutet daher weniger Rechenaufwand und ist damit zu bevorzugen.

2. Zeigen sie durch Umformung, dass beide Gleichungen äquivalent sind.

$$\sigma^2 = \frac{1}{n}\sum (x_i - \bar{x})^2$$
$$\sigma^2 = \frac{1}{n}\sum x_i^2 - 2x_i\bar{x} + \bar{x}^2$$
$$\sigma^2 = \frac{1}{n}\sum x_i^2 - \frac{1}{n}\sum 2x_i\bar{x} + \frac{1}{n}\sum \bar{x}^2$$
$$\sigma^2 = \frac{1}{n}\sum x_i^2 - 2\bar{x}\frac{1}{n}\sum x_i + \bar{x}^2\frac{1}{n}\sum$$
$$\sigma^2 = \frac{1}{n}\sum x_i^2 - 2\bar{x}\bar{x} + \bar{x}^2 1$$
$$\sigma^2 = \frac{1}{n}\sum x_i^2 - \bar{x}^2$$

Übung 8. *Fortlaufende Berechnung von Mittelwerten*
Sie messen ein mit einem Mikcrocontroller fortlaufend Messwerte. Da ihr Speicherplatz beschränkt ist, möchten Sie nur jeweils den aktuellen Mittelwert sowie die Anzahl Messungen im Speicher behalten.

Wie können Sie den Mittelwert fortlaufend berechnen?

Der Mittelwert berechnet sich:

$$\bar{x} = \frac{1}{n} \sum x_i$$

Respektive zu jedem Zeitpunkt $i = 1, 2, 3, ...$

$$\begin{aligned}
\bar{x}_1 &= \frac{1}{1}(x_1) \\
\bar{x}_2 &= \frac{1}{2}(x_1 + x_2) \\
\bar{x}_3 &= \frac{1}{3}(x_1 + x_2 + x_3) = \frac{1}{3}(2\bar{x}_2 + x_3) \\
&...
\end{aligned}$$

Der Mittelwert kann demnach fortlaufend berechnet werden mit:

$$\bar{x}_n = \frac{1}{n}((n-1)\bar{x}_{n-1} + x_n) = \bar{x}_{n-1} + \frac{x_n - \bar{x}_{n-1}}{n}$$

Übung 9. *Parameter*
Sie haben die folgende Urliste:

$$\begin{aligned}
x_i = (&56, 77, 72, 54, 65, 1, 1, 85, 85, 45, 9, 61, 55, 1, 84, \\
&81, 89, 83, 93, 24, 20, 87, 29, 73, 7, 8, 10, 82, 10, 18)
\end{aligned}$$

1. Berechnen Sie Modus, Median und arithmetisches Mittel.
   (a) Modus $Mo = 1$
   (b) Median $Me = 55.5$
   (c) Arithmetisches Mittel $\bar{x} = 48.8$
2. Berechnen Sie Minimum, Maximum, Spannweite, Varianz und Standardabweichung.
   (a) Minimum $x_{min} = 1$
   (b) Maximum $x_{max} = 93$
   (c) Spannweite $R = 92$
   (d) Varianz $\sigma^2 = 1076.9$
   (e) Standardabweichung $\sigma = 32.8$

3. Bilden Sie 10 Klassen und berechnen damit arithmetisches Mittel und Varianz.

	x_i	h_i	$x_i h_i$	$(x_i - \bar{x})^2 h_i$
0-10	5	8	40	15254
10-20	15	2	30	2267
20-30	25	2	50	1120
30-40	35	0	0	0
40-50	45	1	45	13
50-60	55	3	165	120
60-70	65	2	130	534
70-80	75	3	225	2080
80-90	85	8	680	10561
90-100	95	1	95	2147
Σ	30		1460	34097

$$\bar{x} = \frac{1}{n} \sum x_i h_i = \frac{1460}{30} = 48.7$$

$$\sigma^2 = \frac{1}{n} \sum (x_i - \bar{x})^2 h_i = \frac{34097}{30} = 1136.6$$

4. Was fallt Ihnen auf, wenn Sie die Resultate vergleichen?

   Durch die Klassenbildung werden die berechneten Werte ungenauer. Es wird mit der Klassenmitte gerechnet, die tatsächlichen Werte sind jedoch nicht so innerhalb der Klasse verteilt, dass x_i der Klassenmitte entspricht. Je grösser die Anzahl Klassen k, desto genauer werden die Werte tendenziell.