3,120 Methoden: Statistik

Übungsblatt 1: Lage- und Streuparameter

Mathis Mörke
Michael Schürle
und Alois Weigand

Universität St.Gallen (HSG)

Herbstsemester 2024

Refresher: Grundkonzepte

- **Merkmalsträger:** befragte Personen/ Objekte an denen Messungen vorgenommen werden
- **Beobachtungsmerkmale:** Grössen, auf die sich Forschungsfrage/Messungen beziehen
- **Merkmalsausprägung:** Wert, den man durch Beobachtung eines Merkmals an einer Beobachtungseinheit erhält

Refresher: Grundkonzepte

- **Merkmalsarten**
  - **nominal messbar**: gleich oder anders
    $\Rightarrow$ Nationalität, Beruf, Unternehmensrechtsform
  - **ordinal messbar**: unterscheidbar und sinnvolle Rangordnung
    $\Rightarrow$ sozialer Status, Schulnoten, Credit Ratings
  - **kardinal messbar**: quantitativer Unterschied (numerisch)
    $\Rightarrow$ BIP; Investitionen, Inflation, Kosten, Umsatz (oft in Masseinheiten)

Refresher: Grundkonzepte

- **Grundidee**
  - Datenreihe mit n Beobachtungen: $x_1, x_2, x_3, ..., x_n$
    - Beobachtungsreihe mit verschiedenen Merkmalsausprägungen
  - **Grundgesamtheit:** Menge aller möglichen Werte eines Merkmals (oft Anzahl N statt n)
  - **Stichprobe:** Teilmenge aus der Grundgesamtheit (mit Stichprobenumfang n)

    Vollerhebung der Grundgesamtheit
    ... zu teuer
    ... zu zeitaufwändig,...

    $\Rightarrow$ dazu später mehr

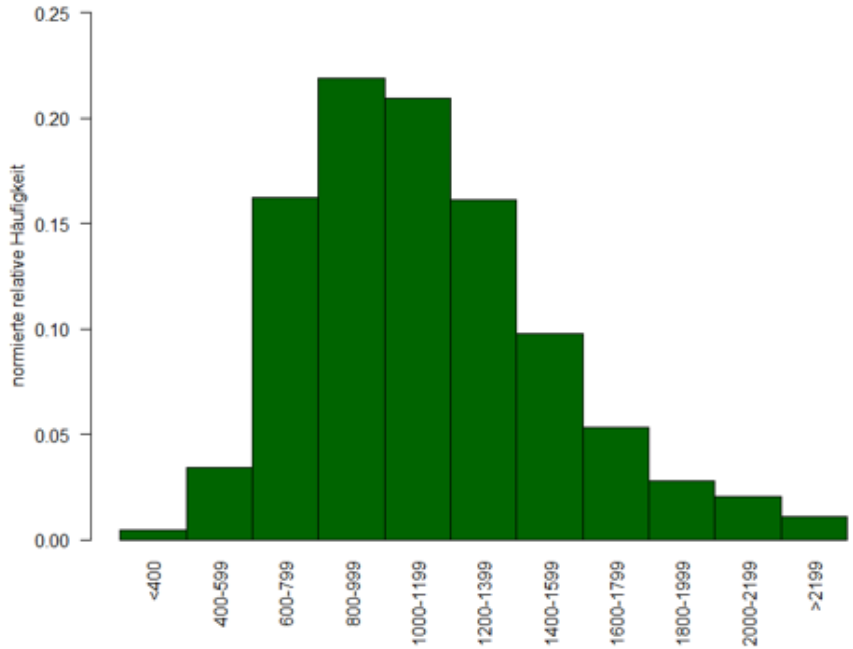
Refresher: Grundkonzepte

- **Grundidee**
  - **Beschreibende Statistik:** Was können wir über die Merkmalswerte lernen?
    - Wie häufig kommen verschiedene Merkmalsausprägungen vor?
      $\Rightarrow$ **Häufigkeitsverteilungen**
    - Was ist die durchschnittliche Grössenordnung der Variablenwerte?
      $\Rightarrow$ **Lageparameter**
    - Wie eng liegen die einzelnen Werte beieinander?
      $\Rightarrow$ **Streuparameter**
  - Lage- und Streuparameter geben eine gute Vorstellung von der Häufigkeitsverteilung, ohne diese genau zu kennen.

Refresher: Grundkonzepte

- **Grundidee**

  $\Rightarrow$ **Häufigkeitsverteilungen**
  $\Rightarrow$ **Lageparameter**
  $\Rightarrow$ **Streuparameter**



Refresher: Lageparameter

- Arithmetisches Mittel
  - $\bar{x} = \frac{1}{n} \sum_i^n x_i$ oder $\bar{x} = \frac{1}{n}(x_1 + x_2 + x_3 + ... + x_n)$
- Median
  - mittlerer Wert (einer der Grösse nach sortierten Datenreihe)
  - zerlegt geordnete Reihe in zwei gleiche Teile
  - 50% der Beobachtungen sind grösser und 50% der Beobachtungen sind kleiner
- Modus
  - häufigster Wert

Refresher: Lageparameter

- Median
  - Merkmalswerte (Ausprägungen der Beobachtung)
    - **nach Grösse geordnet**
    - **mindestens ordinalskaliert**: Merkmalsausprägungen können in eine sinnvolle Rangordung gebracht werden.
  - n **ungerade**: $x_{Median} = x[(n+1)/2]$
    $\Rightarrow$ Wert an Position $(n+1)/2$
  - n **gerade**: $x_{Median} = \frac{1}{2}(x[n/2] + x[n/2+1])$
    $\Rightarrow$ Durchschnitt der Werte an Positionen $n/2$ und $n/2+1$
  - Berechnung nach Lehrbuch Weiers, *Introductory Business Statistics*

Refresher: Lageparameter

- Median
  - **Beispiel 1**: 1, 2, 3, 3, 4, 5, 15
  - $n = 7$ **ungerade**: $x_{Median} = x[(n+1)/2]$

    $\Rightarrow$ Wert an der Stelle (geordnet) $(n+1)/2 = (7+1)/2 = 4$
    $\Rightarrow$ Beobachtung mit Ausprägungswert 3

  - **Beispiel 2**: 1, 2, 3, 3, 4, 5, 15, 178
  - $n = 8$ **gerade**: $x_{Median} = \frac{1}{2}(x[n/2] + x[n/2+1])$

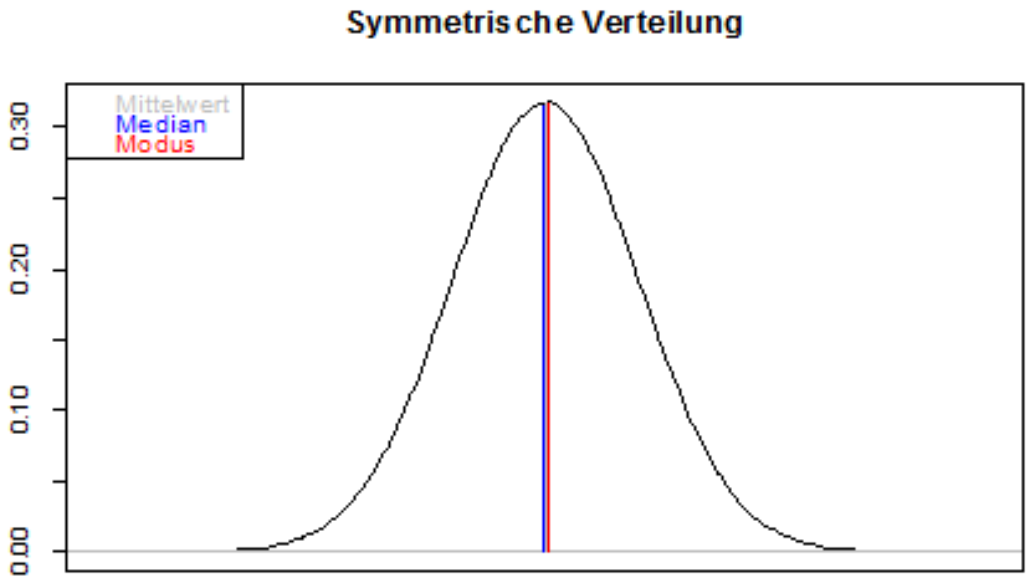
    $\Rightarrow$ Durchschnitt aus den Werten an der Stelle $n/2$ und $n/2+1$
    $\Rightarrow$ an der Stelle $n/2 = 8/2 = 4$: 3
    $\Rightarrow$ an der Stelle $n/2+1 = 8/2+1 = 4+1 = 5$: 4
    $\Rightarrow$ Durchschnitt beider Werte: (3+4)/2=7/2=3.5

Refresher: Lageparameter

- Mittelwert, Median, Modus geben an, wo sich das Zentrum einer Verteilung befindet
- arithmetischer Mittelwert problematisch bei **schiefen Verteilungen**: nicht robust gegen **Ausreisser**
- Median und Modus sind robust gegen Ausreissern
  - aber: nicht alle Beobachtungen gehen in Berechnung ein

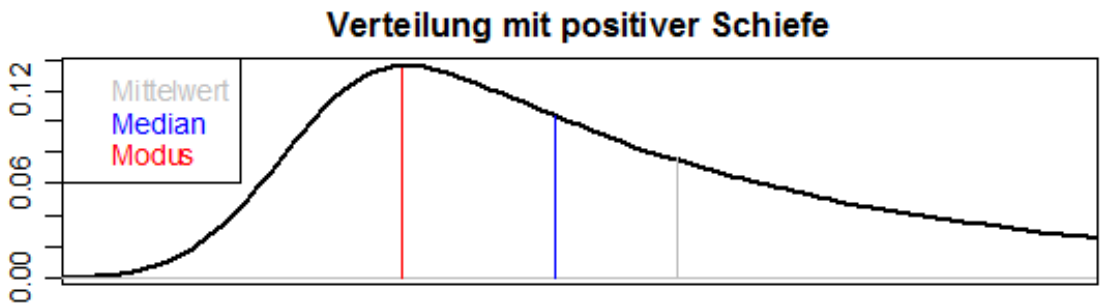
Refresher: Lageparameter

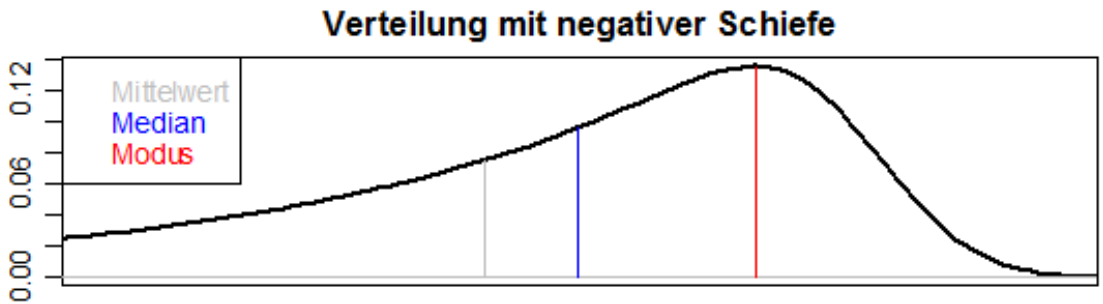
- Lageparameter und Verteilungen
  - symmetrische Verteilung
  - $\Rightarrow$ Mittelwert, Median, Modus nehmen gleichen Wert an



Refresher: Lageparameter

- Lageparameter und Verteilungen
  - schiefe Verteilungen
    - positive Schiefe (rechtschief)
      $\Rightarrow$ Mittelwert $>$ Median $>$ Modus
    - negative Schiefe (linksschief)
      $\Rightarrow$ Mittelwert $<$ Median $<$ Modus





Aufgabe 1

Sie betreiben einen Online-Blog auf dem Sie für Werbezwecke eine von einem Unternehmen gebuchte Anzeige als Text-Link positioniert haben. Da Sie durch die Werbeanzeige mit jedem Klick CHF 0.10 verdienen, interessieren Sie sich für die Anzahl der Besucher Ihres Blogs. Die Anzahl der Besucher in den letzten 7 Tagen beträgt: $17, 18, 19, 19, 21, 22, 25$.
Berechnen Sie für diese **Stichprobe Mittelwert, Median, Modus, die Spannweite, mittlere absolute Abweichung, die Standardabweichung, sowie den Variationskoeffizienten**.

Aufgabe 1

- **Mittelwert**
  - $\bar{x} = \frac{1}{n}\sum_i^n x_i$ oder $\bar{x} = \frac{1}{n}(x_1 + x_2 + x_3 + ... + x_n)$
  - für $n = 7$ Beobachtungen:
  - $\overline{x} = \frac{1}{7}(17 + 18 + 19 + 19 + 21 + 22 + 25) = \frac{141}{7} = 20.14$
- **Median**
  - sortiere Werte nach Grösse: 17, 18, 19, 19, 21, 22, 25
  - n ungerade ($n = 7$)
  - $x_{Median} = x[\frac{n+1}{2}] \Rightarrow$ Wert an Position $\frac{n+1}{2}$ der Reihe
  - $x_{Median} = x[\frac{n+1}{2}] = x[\frac{7+1}{2}] = x_4 = 19$
- **Modus**
  - häufigster Wert: $x_{Modus} = 19$ (zweimal vorhanden)

Aufgabe 1

- **Spannweite**
  - Differenz zwischen grösster und kleinster Merkmalsausprägung $\Rightarrow$ übrigen Werte unberücksichtigt
  - $Range = x_{max} - x_{min} = 25 - 17 = 8$

- **Mittlere Absolute Abweichung**
  - für $\overline{x} = 20.14$
  - $MAD_S = \frac{1}{n-1} \sum_{i=1}^{n} |x_i - \overline{x}| = \frac{1}{6} \cdot 15.14 = 2.52$

x_i	$x_i - \overline{x}$	$\lvert x_i - \overline{x} \rvert$
17	-3.14	3.14
18	-2.14	2.14
19	-1.14	1.14
19	-1.14	1.14
21	0.86	0.86
22	1.86	1.86
25	4.86	4.86
$\sum$		15.14

Aufgabe 1

- **Varianz**
  - Streuung oder Abweichung vom Mittelwert
  - $s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{1}{6} \cdot 44.8572 = 7.48$
  - durchschnittliche quadrierte Abweichungen vom Mittelwert
  - **bei Stichproben:** durch (n-1) teilen !

x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
17	-3.14	9.8596
18	-2.14	4.5796
19	-1.14	1.2996
19	-1.14	1.2996
21	0.86	0.7396
22	1.86	3.4596
25	4.86	23.6196
$\sum$		44.8572

Aufgabe 1

- **Standardabweichung:**

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{7.48} = 2.73$$

- **Variationskoeffizient**
  - $VC_x = \frac{s_x}{|\overline{x}|} = \frac{2.73}{20.14} = 0.136$
  - relatives Streungsmass
  - Standardabweichung relativ zum Mittelwert
  - Vergleich von Streuungen von Beobachtungsreihen mit unterschiedlichem Mittelwert
  - macht keinen Sinn, wenn Mittelwert nahe null ist

Aufgabe 2

Eine Stadt in der Schweiz hat 2567 (in Tausend) Haushalten. Darunter sind 457 (Tausend) Haushalte mit jeweils einer Person, 628 (Tausend) Haushalte mit zwei Personen, 612 (Tausend) Haushalte mit drei Personen, 526 (Tausend) Haushalte mit vier Personen, sowie 344 (Tausend) Haushalte mit fünf oder mehr Personen.

1. Berechnen Sie die **Häufigkeitsverteilung** sowohl **absolut** als auch **relativ** gesehen **sowie die kumulierte** Häufigkeitsverteilung.
2. In wie vielen Haushalten leben **drei oder weniger Personen** (sowohl absolut als auch in Prozentangaben)?

Aufgabe 2

- Häufigkeitsverteilung
  - absolute Häufigkeiten: n_i
  - relative Häufigkeiten: $f(x_i) = n_i/n$
  - kumulierte Häufigkeiten: $F(x_i) = \sum f(x_i)$

Haushalte mit ... Personen: x_i	n_i	$f(x_i) = \frac{n_i}{n}$	$\sum f(x_i) = F(x_i)$
1	457	0.178	0.178
2	628	0.245	0.423
3	612	0.238	0.661
4	526	0.205	0.866
≥ 5	344	0.134	1.000
$\sum$	2567	1.000	

Aufgabe 2

- Häufigkeitsverteilung

Haushalte mit ... Personen: x_i	n_i	$f(x_i) = \frac{n_i}{n}$	$\sum f(x_i) = F(x_i)$
1	457	0.178	0.178
2	628	0.245	0.423
3	612	0.238	0.661
4	526	0.205	0.866
≥ 5	344	0.134	1.000
$\sum$	2567	1.000	

- In wie vielen Haushalten drei oder weniger Personen?
  - $F(x_i = 3) = 66.1\%$ (3 Personen oder weniger)
    $2567 \times 66.1\% = 1696.79 = 1697$
  - oder: $457 + 628 + 612 = 1697$ $(n_1 + n_2 + n_3)$

Aufgabe 3

Die Bevölkerung einer aufstrebenden Gemeinde ist von 2013 bis 2017 wie folgt gewachsen:

Jahre	2013	2014	2015	2016	2017
Bevölkerung	20000	32000	41600	44387	48826

Berechnen Sie die **durchschnittliche Zuwachsrate** der vier Jahre.

Aufgabe 3

- Idee: einzelne Wachstumsraten
  - 2014: $\frac{(32000-20000)}{20000} = 0.6$
  - 2015: $\frac{(41600-32000)}{32000} = 0.3$
  - 2016: $\frac{(44387-41600)}{41600} = 0.067$
  - 2017: $\frac{(48826-44387)}{44387} = 0.10$

  $\Rightarrow (0.6+0.3+0.067+0.10)/4=0.26675$

- **Vorsicht: Das arithmetische Mittel (von 26.675%) als durchschnittliche Wachstumsrate ist falsch!**

Aufgabe 3

- Wachstumsraten
  - **Vorsicht: Das arithmetische Mittel (von 26.675%) als durchschnittliche Wachstumsrate ist falsch!**
  - Wendet man die konstante durchschnittliche Zuwachsrate von 26.675% auf alle vier Jahre an, dann ergibt sich eine andere Bevölkerung am 31.12.2016 als tatsächlich vorhanden.

Jahr	Bevölkerung am 31.12.	Zuwachs absolut	Zuwachs relativ
2013	20000	-	-
2014	25335 (statt 32000)	5335	26.675%
2015	32093 (statt 41600)	6758	26.675%
2016	40654 (statt 44387)	8561	26.675%
2017	51498 (statt 48826)	10844	26.675%

  - Die als **arithmetisches Mittel** errechnete *durchschnittliche* Wachstumsrate kann also **nicht korrekt** sein.

Aufgabe 3

- Wachstumsraten
  - **arithmetisches Mittel nicht richtig!**
  - Sind Merkmalswerte relative Änderungen (Zuwachsraten), dann wird das **geometrische Mittel** verwendet
  - Gesamtänderung wird durch das Produkt beschrieben, da Wachstumsraten der einzelnen Jahre voneinander abhängen

  - **geometrische Mittel**: $G_{\bar{x}} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$
    $\rightarrow$ einzelne Merkmalswerte multipliziert und n-te Wurzel aus dem Produkt

  - **Achtung**: geometrisches Mittel aus den Zuwachsfaktoren, Zinsfaktoren oder Wachstumsfaktoren berechnet
    $\rightarrow$ Wachstumsrate r_i; Wachstumsfaktor $(1 + r_i) = x_i$

Aufgabe 3

- Bilde den Wachstumsfaktor
  - 2014: $\frac{32000}{20000} = 1.6$
  - 2015: $\frac{41600}{32000} = 1.3$
  - 2016: $\frac{44387}{41600} = 1.067$
  - 2017: $\frac{48826}{44387} = 1.1$

Aufgabe 3

- Wachstumsraten
  - **Daraus folgt:** $G_{\bar{x}} = \sqrt[4]{1.6 \times 1.3 \times 1.067 \times 1.1} = 1.2499$
    $\rightarrow$ gerundet eine durchschnittliche Wachstumsrate von 25% (für die 4 Jahre)

Jahr	Bevölkerung am 31.12.	Zuwachs absolut	Zuwachs relativ
2013	20000	-	-
2014	25000	5000	25.0%
2015	31250	6250	25.0%
2016	39063	7813	25.0%
2017	48829	9766	25.0%

  - geringfügige Differenz zwischen 48829 und 48826 (Rundungsfehler)

Aufgabe 4

Die durchschnittlichen wöchentlichen Ausgaben eines HSG-Studenten betragen zu Beginn des 1. Semester 250 CHF. Mittlerweile am Ende des 3. Semester angelangt, sind auch seine Ausgaben parallel mit seinen ökonomischen Kenntnissen in die Höhe geschnellt. Die wöchentlichen Ausgaben liegen nun bei etwa 600 CHF. Wie hoch war seine **durchschnittliche Ausgabensteigerung pro Semester**?

Aufgabe 4

- durchschnittliche Ausgabensteigerung pro Semester gesucht
  - geometrisches Mittel

  - $G = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$ oder $G = \sqrt[n]{K_n/K_0}$
  - K_n ist Endbestand in Periode n
  - K_0 ist Anfangsbestand in Periode 1

  - Wachstumsfaktoren:
    - $x_1 = (1 + r_1) = K_1/K_0$,
    - $x_2 = (1 + r_2) = K_2/K_1$,
    - $x_3 = (1 + r_3) = K_3/K_2$, ...

  $\Rightarrow G = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} = \sqrt[n]{\frac{K_1}{K_0} \cdot \frac{K_2}{K_1} \cdot \ldots \frac{K_n}{K_{n-1}}}$ oder
  $G = \sqrt[n]{K_n/K_0}$ (durch Kürzen)

Aufgabe 4

- durchschnittliche Ausgabensteigerung pro Semester gesucht
  - geometrisches Mittel
  - $G = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$ oder $G = \sqrt[n]{K_n / K_0}$
  - $G = \sqrt[3]{600/250} = \sqrt[3]{2.4} = 1.34$ (Wachstumsfaktor)
    $\Rightarrow$ durchschnittliche Steigerung von 34% (Wachstumsrate)pro Semester

Aufgabe 5

Gegeben ist die 2017 veröffentlichte Arbeitslosenstatistik eines Landes mit zwei strukturell unterschiedlichen Arbeitsmarktregionen.

	Arbeitslose	Erwerbspersonen	Arbeitslosenquote
Region A	1,851,086	33,781,938	5.48%
Region B	861,968	8,421,990	10.23%

1. Berechnen Sie die **durchschnittliche Arbeitslosenquote** des Landes unter der Annahme, dass Ihnen die **Anzahl der Arbeitslosen unbekannt** ist.

2. Berechnen Sie die **durchschnittliche Arbeitslosenquote** des Landes unter der Annahme, dass Ihnen die **Anzahl der Erwerbspersonen unbekannt** ist.

Aufgabe 5

- Refresher: **Harmonisches Mittel vs arithmetisches Mittel**
  - **Verhältniszahl**: $v = \frac{z}{n}$
    z.B. durchschnittliche Arbeitslosenquote als Verhältnis der durchschnittl. Arbeitslosen zu den Erwerbspersonen im Durschnitt
  - für $v_1 = \frac{z_1}{n_1}$ und $v_2 = \frac{z_2}{n_2}$

    $\Rightarrow z = z_1 + z_2$ und $n = n_1 + n_2$

    $\Rightarrow v = \frac{z_1 + z_2}{n_1 + n_2}$

    (z für Zähler, n für Nenner)

Aufgabe 5

- Refresher: **Harmonisches Mittel** versus arithmetisches Mittel
  - Fall 1: z_1 und z_2 sind bekannt (Nenner n_1 und n_2 unbekannt)
  - $v = \frac{z_1+z_2}{n_1+n_2} = \frac{z_1+z_2}{z_1 \times \frac{n_1}{z_1} + z_2 \times \frac{n_2}{z_2}} = \frac{z_1+z_2}{\frac{z_1}{(\frac{z_1}{n_1})} + \frac{z_2}{(\frac{z_2}{n_2})}} = \frac{z_1+z_2}{\frac{z_1}{v_1} + \frac{z_2}{v_2}}$

    $\Rightarrow v$ ist das gewichtete harmonische Mittel von $v_1 = \frac{z_1}{n_1}$ und $v_2 = \frac{z_2}{n_2}$ mit den Gewichten z_1 und z_2

Aufgabe 5

- Refresher: **Harmonisches Mittel** versus arithmetisches Mittel
  - Merkmalsausprägungen als Quotienten $x_j = \frac{a_j}{b_j}$
  - Nenner b_j unbekannt $\rightarrow b_j = \frac{a_j}{x_j}$
  - $\bar{x}_H = \frac{\sum_{j=1}^{m} a_j}{\sum_{j=1}^{m} b_j} = \frac{\sum_{j=1}^{m} a_j}{\sum_{j=1}^{m} (\frac{a_j}{x_j})}$
  - a_j ist gegeben und b_j nicht: **gewogenes harmonisches Mittel** mit den Gewichten a_j

    $\bar{x}_H = \frac{\sum_{j=1}^{m} a_j}{\sum_{j=1}^{m} \frac{a_j}{x_j}}$

Aufgabe 5

- Refresher: Harmonisches Mittel versus **arithmetisches Mittel**
  - Fall 1: n_1 und n_2 sind bekannt (Zähler z_1 und z_2 unbekannt)
  - $v = \frac{z_1+z_2}{n_1+n_2} = \frac{n_1 \times \frac{z_1}{n_1} + n_2 \times \frac{z_2}{n_2}}{n_1+n_2} = \frac{n_1 \times v_1 + n_2 \times v_2}{n_1+n_2}$

    $\Rightarrow v$ ist das gewichtete arithmetische Mittel von v_1 und v_2 mit den Gewichten n_1 und n_2

Aufgabe 5

- Berechnen Sie die durchschnittliche Arbeitslosenquote des Landes unter der Annahme, dass Ihnen die Anzahl der Arbeitslosen unbekannt ist.

	Arbeitslose	Erwerbspersonen	Arbeitslosenquote
Region A		33,781,938	5.48%
Region B		8,421,990	10.23%

  - Da von den beiden Verhältniszahlen die Nenner bekannt sind, ergibt sich der Durchschnitt als gewichtetes arithmetisches Mittel:

  - $ALQ_{total} = \frac{33{,}781{,}938 \times 5.48 + 8{,}421{,}990 \times 10.23}{33{,}781{,}938 + 8{,}421{,}990} = 6.43\%$

Aufgabe 5

- Berechnen Sie die durchschnittliche Arbeitslosenquote des Landes unter der Annahme, dass Ihnen die Anzahl der Erwerbspersonen unbekannt ist.

	Arbeitslose	Erwerbspersonen	Arbeitslosenquote
Region A	1,851,086		5.48%
Region B	861,968		10.23%

  - Da von den beiden Verhältniszahlen die Zähler bekannt sind, ergibt sich der Durchschnitt als gewichtetes harmonisches Mittel:

  - $ALQ_{total} = \frac{1{,}851{,}086+861{,}968}{\frac{1{,}851{,}086}{5.48}+\frac{861{,}968}{10.23}} = 6.43\%$

Aufgabe 6

Im Rahmen einer Untersuchung über Wohnverhältnisse von Studenten wurde für fünf Bezirke einer Universitätsstadt die Anzahl der in dem jeweiligen Bezirk wohnenden Studenten ermittelt. Diese Anzahl wurde zur Einwohnerzahl des jeweiligen Bezirks in Beziehung gesetzt und eine Studentendichte als Anzahl der Studenten pro 100 Einwohner bestimmt.

Bezirk	1	2	3	4	5
Anzahl der Studenten	300	1500	200	800	700
Studentendichte	15	30	4	5	10

Berechnen Sie die **durchschnittliche Studentendichte** aus den Studentenzahlen.

Aufgabe 6

- Harmonisches Mittel
  - $\bar{x}_H = \frac{\sum a_j}{\sum b_j} = \frac{\sum a_j}{\sum (\frac{a_j}{x_j})}$
  - Zähler zur Gewichtung
  - nur Studentenzahlen (Zähler) und Studentendichte ist bekannt:

$$H = \frac{300+1500+200+800+700}{300/15+1500/30+200/4+800/5+700/10} = \frac{3500}{350} = 10$$

Aufgabe 7

Der folgende Datensatz zeigt eine repräsentative Auswahl der Monatseinkommen von Saisonarbeitskräften (bereits der Grösse nach geordnet). Ermitteln Sie **Median**, die **Quartile** sowie das **erste Dezil**. Bestimmen Sie den **Interquartilsabstand**.

2200	2200	2269	2286	2302	2308	2335	2344	2353	2375
2467	2502	2507	2534	2562	2572	2598	2600	2606	2614
2669	2678	2691	2706	2708	2712	2770	2791	2798	2826
2827	2858	2867	2873	2876	2905	2969	2978	2989	2996
3000	3007	3020	3200	3229	3252	3506	3506	3582	3609

Zeichnen Sie die (empirische) Verteilungsfunktion und einen Boxplot.

Aufgabe 7

- Median
  - $n = 50$: **gerade**: $x_{Median} = \frac{1}{2}(x[n/2] + x[n/2+1])$

2200	2200	2269	2286	2302	2308	2335	2344	2353	2375
2467	2502	2507	2534	2562	2572	2598	2600	2606	2614
2669	2678	2691	2706	2708	2712	2770	2791	2798	2826
2827	2858	2867	2873	2876	2905	2969	2978	2989	2996
3000	3007	3020	3200	3229	3252	3506	3506	3582	3609

  - $x_{Median} = \frac{1}{2}(x[25] + x[26]) = \frac{1}{2}(2708 + 2712) = 2710$

Aufgabe 7

- **Refresher: Quantile**
  - statistische Reihe kann in vier (Quartile), zehn (Dezentile), 100 (Perzentile)...
  - **allgemein: q-Quantil**
  - q-Quantil (z.B. 0.3): mindestens $100q\%$ der Werte sind kleiner oder gleich dem Wert x und mindestens $100(1-q)\%$ sind grösser
  - unteres Quartil (Q_1) $\Rightarrow$ 25%-Quartil ($q = 0.25$)
  - oberes Quartil (Q_3) $\Rightarrow$ 75%-Quartil ($q = 0.75$)
  - Median (Q_2) $\Rightarrow$ 50%-Quartil ($q = 0.50$)
  - Dezile: $x[0.1], x[0.2], x[0.3], ..., x[0.9]$
  - Perzentile: $x[0.01], x[0.02], x[0.03], ..., x[0.98], x[0.99]$

Aufgabe 7

- Quartile
  - Berechnung folgt dem Lehrbuch Weiers, *Introductory Business Statistics*
  - $Q_1 = x[(n+1)/4] = x[12.75] = 0.25 \cdot x[12] + 0.75 \cdot x[13]$
    $\rightarrow Q_1 = 0.25 \cdot 2502 + 0.75 \cdot 2507 = 2505.75$
  - $Q_2 = x[2(n+1)/4] = x[25.50] = 0.50 \cdot x[25] + 0.50 \cdot x[26]$
    $\rightarrow Q_2 = 0.50 \cdot 2708 + 0.50 \cdot 2712 = 2710$
  - $Q_3 = x[3(n+1)/4] = x[38.25] = 0.75 \cdot x[38] + 0.25 \cdot x[39]$
    $\rightarrow Q_3 = 0.75 \cdot 2978 + 0.25 \cdot 2989 = 2980.75$
  - Anmerkung: $x[i]$ entspricht der Merkmalsausprägung an der i-ten Stelle einer geordneten Reihenfolge

Aufgabe 7

- Interquartilabstand
  - Differenz zwischen dem 75%-Quantil und dem 25%-Quantil
  - $IQA = Q_3 - Q_1 = 2980.75 - 2505.75 = 475$
- Dezile
  - analog zu den Quartilen:
  - $D_1 = x[(n+1)/10] = x[5.1] = 0.90 \cdot x[5] + 0.10 \cdot x[6]$
    $\rightarrow D_1 = 0.90 \cdot 2302 + 0.10 \cdot 2308 = 2302.60$
  - $D_3 = x[(n+1)3/10] = x[15.3] = 0.70 \cdot x[15] + 0.30 \cdot x[16]$
    $\rightarrow D_3 = 0.70 \cdot 2565 + 0.30 \cdot 2572 = 2567.10$

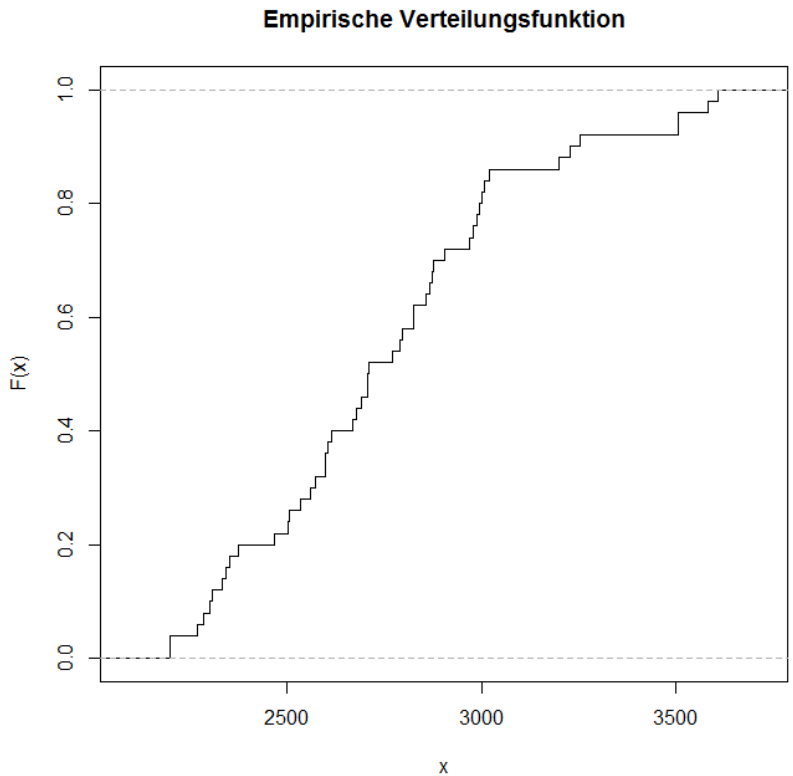
Aufgabe 7

- Empirische Verteilungsfunktion (theoretisch)

x_i	$f(x_i)$	$F(x_i) = \sum f(x_i)$
2200	0.04	0.04
2269	0.02	0.06
2286	0.02	0.08
2303	0.02	0.10
2335	0.02	0.12
$\vdots$	$\vdots$	$\vdots$

Aufgabe 7

- Empirische Verteilungsfunktion
  - q-Quantil: Position (x-Wert) auf der (empirischen) Verteilungfunktion
  - mindestens $100 \times q$ % der Werte sind kleiner oder gleich dem Wert x und mindestens $100 \times (1 - q)$ % sind grösser
  - z.B. Median (0.5- Quantil) bei $x_i = 2710 \rightarrow F(x) = 0.5$ oder 50%



Aufgabe 7

- Boxplots
  - Die Häufigkeitsverteilung kann mit den folgenden Werten beschrieben werden:
    $(x_{min}, x[0.25], x[0.50], x[0.75], x_{max})$
  - Datensatz in 4 gleich grosse Teile zerlegt
  - **Lageparameter**: Median
  - **Streuungsmasse**: *IQA* und die Spannweite $(x_{max} - x_{min})$
  - Länge der Box entspricht dem *IQA*: 50% der Werte um Median

Aufgabe 7

- Boxplots