Wirtschaftswissenschaftliches Zentrum 2
Universität Basel

Statistik
Dr. Thomas Zehrt

Mittelwerte, Streumass und Konzentrationsmasse

Motivation

In diesem Abschnitt werden zunächst Parameter beschrieben, mit deren Hilfe man die Lage einer (grossen) statistischen Masse auf der Merkmalsachse durch einen einzigen Wert beschreiben kann. Solche Parameter werden als Mittelwert**e** bezeichnet und wie die Pluralform dieses Wortes andeutet, muss man in der Praxis sehr genau überlegen, welcher von ihnen gemeint ist. So sind z.B. die folgenden beiden Aussagen sachlich korrekt und je nachdem, ob man die Intersessen der Ärztelobby oder die der Krankenkassen vertritt könnte man die passende Aussage wählen:

Im Jahr 2005 verdiente ein deutscher Zahnarzt im Mittel

1. 98'150 Euro (hier ist der so genannte Zentralwert oder Median gemeint)
2. 111'105 Euro (hier ist das so genannte arithmetische Mittel gemeint)

Mittelwerte sind nicht alles, insbesondere sind sie blind gegenüber Streuung und in der Praxis ist man meist nicht nur am Mittelwert der Daten interessiert, sondern man möchte auch wissen, wie nahe sich die Daten (im Durchschnitt) beim Mittelwert befinden.

Statistikerwitz (Mittelwerte sind nicht alles): Ein Mann mit seinem Kopf im Tiefkühlschrank und Füssen im Ofen hat es im Durchschnitt recht angenehm.

Benötigtes Schulwissen

- Bruchrechnung und Prozentrechnung
- Summenzeichen und Produktzeichen $\prod_{j=1}^{m} q_j = q_1 \cdot q_2 \ldots q_m$, also z.B. falls $m = 4$ und $q_1 = 1/2$, $q_2 = 1/5$, $q_3 = 3/2$, $q_4 = 1/4$ so ist $\prod_{j=1}^{m} q_j = \frac{1}{2} \cdot \frac{1}{5} \cdot \frac{3}{2} \cdot \frac{1}{4} = \frac{3}{80}$.

1 Mittelwerte

Wir gehen stets von der Urliste aus. Sei also eine statistische Masse mit n Elementen, die wir uns mit $1, 2, \ldots, n$ durchnummeriert denken, und ein (quantitatives) Merkmal X gegeben. Für alle $i = 1, 2, \ldots, n$ bezeichne x_i die Ausprägung des Merkmals X beim Element i.

1.1 Das arithmetische Mittel

Definition 1.1 *Das arithmetische Mittel $\bar{x}$ ist definiert als*

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

Eigenschaften des arithmetischen Mittels:

1. Es gilt stets

$$\sum_{i=1}^{n} (x_i - \bar{x}) \;=\; \sum_{i=1}^{n} x_i - n\bar{x} \;=\; \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i \;=\; 0.$$

   Anders ausgedrückt: Die Summe aller Abweichungen links von $\bar{x}$ ist gleich der Summe aller Abweichungen rechts davon:

$$\sum_{x_i < \bar{x}} (\bar{x} - x_i) \;=\; \sum_{x_i > \bar{x}} (x_i - \bar{x}).$$

   Im physikalischen Sinne ist das gerade die Gleichgewichtsbedingung: Denkt man sich die x-Achse als langen masselosen Stab und darauf an den Positionen i die Masse x_i angebracht, so befindet sich der Stab genau dann im Gleichgewicht, wenn er im Punkt $\bar{x}$ gehalten wird.

2. Das arithmetische Mittel ist im folgenden Sinne ein optimaler Representant der gesamten Datenmenge $\{x_1, x_2, \ldots, x_n\}$: Sei λ eine beliebige reelle Zahl. Wir führen als ,,Strafmass" für die Abweichung eines Wertes x_i von λ die Grösse $(x_i - \lambda)^2$ ein. Die durch das Quadrat stets nichtnegative ,,Gesamtstrafe" ist dann

$$f(\lambda) \;=\; \sum_{i=1}^{n} (x_i - \lambda)^2 \;=\; \sum_{i=1}^{n} x_i^2 - 2\lambda \sum_{i=1}^{n} x_i + n \cdot \lambda^2 \; .$$

   Durch Ableiten der Funktion und Nullsetzen der ersten Ableitung ergibt sich dann, dass $f(\lambda)$ genau dann minimal ist, wenn $\lambda = \bar{x}$ gilt.

Aufgabe 1.1 *Bestimmen Sie das arithmetische Mittel der* 20 *Daten*

6 8 7 5 8 6 6 6 6 5 8 8 6 9 6 8 7 9 5 11

Lösung:

Der direkte Weg ist sehr aufwendig, da wir die Summe aller Daten berechnen und das Ergebnis durch 20 *teilen müssen. Schneller geht es aber, wenn wir die Häufigkeitsauswertung verwenden*

$$\begin{aligned} \bar{x} &= \frac{1}{20}\left(6+6+7+\ldots+6+5+11\right) \\ &= \frac{1}{20}\left(3\cdot 5+7\cdot 6+2\cdot 7+5\cdot 8+2\cdot 9+1\cdot 11\right) \\ &= \end{aligned}$$

1.2 Das gewichtete arithmetische Mittel

Angenommen, Sie hätten den durchschnittlichen Anstieg der Kosten für Ihr Auto zu bestimmen. Diese bestehen aus den Benzinkosten, die um 50 % und den Kosten für Motoröl, die um 10 % gestiegen sind. Das gewöhnliche arithmetische Mittel liefert uns hier einen durchschnittlichen Anstieg um $(50\% + 10\%)/2 = 30\%$. Das ist aber nicht realistisch, da wir sehr viel mehr Geld für Benzin als für Öl ausgeben. Bei einem Ausgabenanteil von sagen wir 4/5 für Benzin und 1/5 für Öl bietet es sich statt dessen an, die verschiedenen Eingänge im arithmetischen Mittel zu gewichten:

$$\text{Gewichtetes arithmetisches Mittel} \;=\; \frac{4}{5}\,50\% + \frac{1}{5}\,10\% = 42\%.$$

Definition 1.2 *Ein Vektor* $\mathbf{p} = (p_1, \ldots, p_n)$*, dessen Komponenten die Eigenschaften*

- $0 \le p_1, \ldots, p_n \le 1$ *(alle Komponenten liegen zwischen* 0 *und* 1*)*
- $\sum_{i=1}^{n} p_i = 1$ *(die Summe aller Komponenten ist* 1*).*

haben, wird als Wahrscheinlichkeitsvektor bezeichnet. Dann ist das bzgl. $\mathbf{p}$ *gewichtete arithmetische Mittel* $\bar{x}_{\mathbf{p}}$ *definiert als*

$$\boxed{\bar{x}_{\mathbf{p}} \;=\; \sum_{i=1}^{n} p_i\, x_i.}$$

Ist $p_i = 1/n$ für alle $i = 1, \ldots, n$, so ist $\mathbf{p} = (\frac{1}{n}, \ldots, \frac{1}{n})$ ein Wahrscheinlichkeitsvektor und $\bar{x}_{\mathbf{p}} = \bar{x}$ das gewöhnliche arithmetische Mittel. Jede Merkmalsausprägung wird also mit dem gleichen Gewicht versehen.

Aufgabe 1.2 *Sei* $\mathbf{p} = \left(\frac{1}{8}, \frac{1}{8}, \frac{1}{4}, \frac{3}{8}, \frac{1}{8}\right)$ *ein Wahrscheinlichkeitsvektor. Berechnen Sie das bzgl.* $\mathbf{p}$ *gewichtete aritmetische Mittel der Daten* $4, 5, 6, 2, 6$*. Vergleichen Sie diesen Wert mit dem gewöhnlichen arithmetischen Mittel dieser Daten.*

Lösung:

1.3 Der Zentralwert oder Median

Definition 1.3 *Der Zentralwert oder Median Z ist der mittlere Wert der nach Grösse geordneten Reihe der Merkmalsausprägungen.*

Eigenschaften des Zentralwertes:

1. Z ist unempfindlich gegenüber **Ausreissern**, das sind Messwerte, die weit weg von den übrigen liegen und manchmal durch Mess- oder Schreibfehler entstehen.
2. Auch der Zentralwert hat eine Optimalitätseigenschaft: Wir legen diesmal als Gesamtstrafe für die Abweichungen fest

$$f(\lambda) \quad = \quad \sum_{i=1}^{n} |x_i - \lambda|.$$

$f(\lambda)$ ist genau dann minimal ist, wenn $\lambda = Z$ gilt.

1.4 Quantile

Seien $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ die der Grösse nach geordneten Beobachtungswerte und sei p eine reelle Zahl mit $0 \leq p \leq 1$. Dann ist das p-Quantil $\tilde{x}_p$ dieser Daten definiert als:

$$\tilde{x}_p := \begin{cases} x_{(k)} & \text{falls } n \cdot p \notin \mathbb{Z} \text{ ist } k \text{ die kleinste ganze Zahl mit } k > n \cdot p \\ \frac{1}{2}(x_{(n \cdot p)} + x_{(n \cdot p+1)}) & \text{falls } n \cdot p \in \mathbb{Z} \end{cases}$$

Definition 1.4 *$\tilde{x}_{0.25}$ heisst unteres Quantil, $\tilde{x}_{0.5} = Z$ ist der wohlbekannte Zentralwert und $\tilde{x}_{0.75}$ heisst oberes Quantil.*

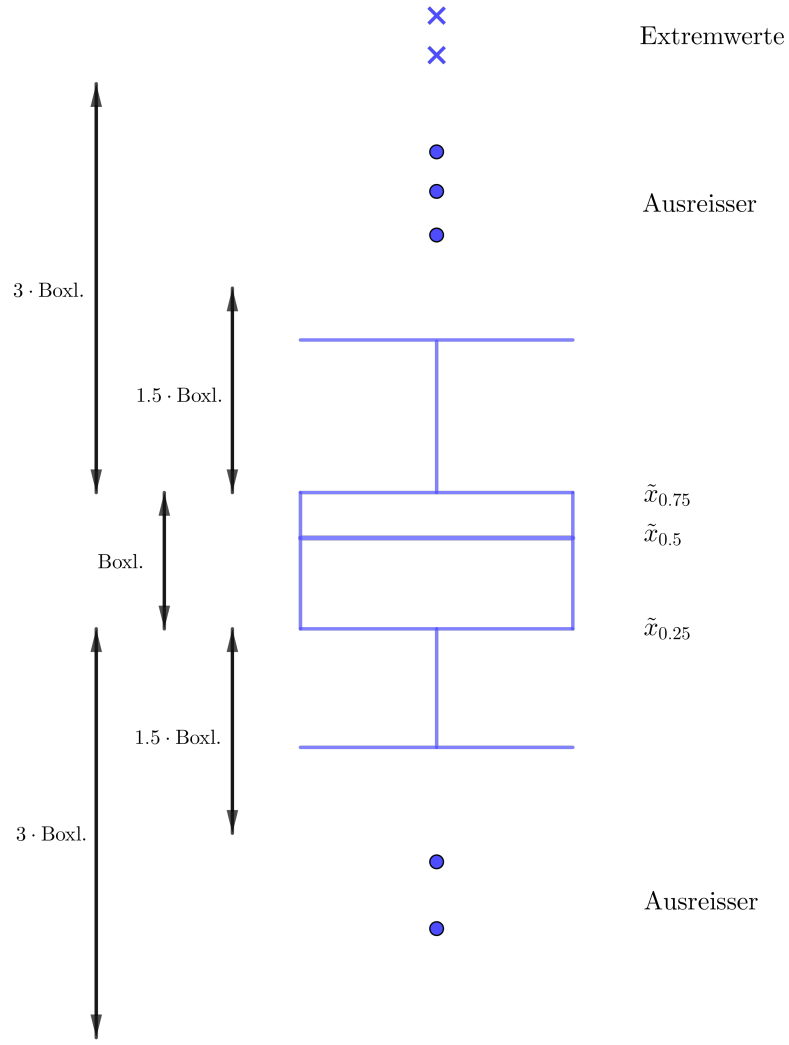
Aufgabe 1.3 *Bestimmen Sie $Z = \tilde{x}_{1/2}$ und $\tilde{x}_{1/3}$ der 20 Daten*

6 8 7 5 8 6 6 6 6 5 8 8 6 9 6 8 7 9 5 11

Lösung:

Box-Plots Der Box-Plot eines Datensatzes stellt die Lage des Zentralwertes, des oberen und unteren Quantils, der Extremwerte und der Ausreisser graphisch dar.

- **innerhalb der Box**
  untere Boxgrenze $\tilde{x}_{0.25}$
  obere Boxgrenze $\tilde{x}_{0.75}$
  Linie in der Box $\tilde{x}_{0.5}$
- **ausserhalb der Box**
  - **Extremwerte**: mehr als 3 Boxlängen vom unteren bzw. oberen Boxrand entfernt, wiedergegeben durch „×“
  - **Ausreisser**: zwischen 1.5 und 3 Boxlängen vom oberen bzw. unteren Boxrand entfernt, wiedergegeben durch „•“
  - Der kleinste und der grösste Wert, der jeweils nicht als Ausreisser eingestuft wird, ist durch eine horizontale Strecke darzustellen.

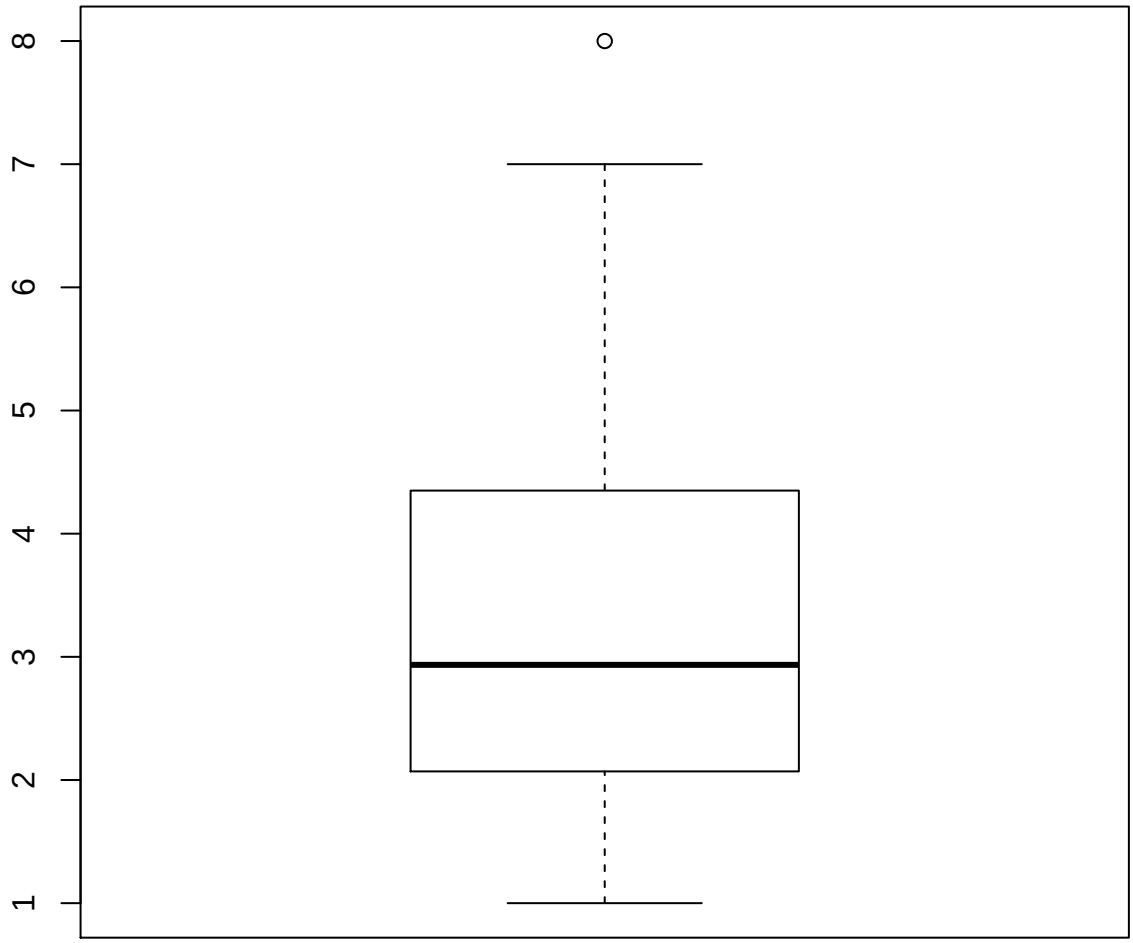


Beispiel 1.1

Für die Daten

$$2.93, 2.58, 2.85, 4.26, 2.94, 4.33, 1.71, 4.42, 3.59, 4.35, 2.07, 1.16, 2.36, 1.16, 4.72, 8, 7, 1$$

sieht der Boxplot wie folgt aus.



1.5 Das geometrische Mittel

Angenommen wir kaufen eine Aktie für 1000 CHF. Nach einem Jahr steigt der Kurs auf 1200 CHF, nach zwei Jahren auf 1500 CHF und im dritten Jahr fällt der Kurs wieder auf 1000 CHF. Zinsen und Dividenden sind nicht angefallen und wir suchen die mittlere jährliche Rendite der Aktie. Nun, nichts einfacher als das! Im ersten Jahr (1000 → 1200 CHF) haben wir 20 %, im zweiten Jahr (1200 → 1500 CHF) 25 % und im dritten Jahr (1500 → 1000 CHF) -33.33 % und das macht im Durchschnitt (arithmetisches Mittel)

$$\frac{20\% + 25\% - 33.33\%}{3} \quad = \quad +3.89\%.$$

Wir staunen also nicht schlecht, denn nach drei Jahren sind 1000 CHF wieder 1000 CHF aber die durchschnittliche Rendite ist + 3.89 % ! Hier scheint etwas faul zu sein.

Definition 1.5 *Seien alle Beobachtungswerte x_i positiv. Dann ist das geometrische Mittel G gleich der n-ten Wurzel aus dem Produkt aller x_i:*

$$G \;:=\; \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} \;=\; \sqrt[n]{\prod_{i=1}^{n} x_i}\;.$$

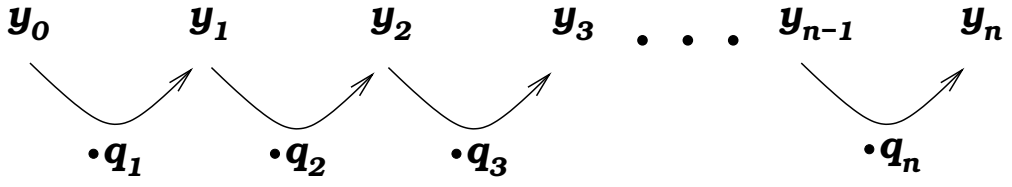
Bei der Berechnung von durchschnittlichen Wachstumsfaktoren ist **stets** das geometrische Mittel anzuwenden.

Definition 1.6 *Sei $y_t > 0$ der Wert einer Grösse in der Periode t (z.B. Umsatz im Monat, Einkommen, Ausgaben,...) , $t = 0, 1, 2, ...$ Dann ist der Wachstumsfaktor q_t von Periode $t-1$ auf Periode t gleich*

$$q_t \;:=\; \frac{\textit{Wert in Periode } t}{\textit{Wert in Periode } t-1} \;=\; \frac{y_t}{y_{t-1}}$$

und die Wachstumsrate r_t von der Periode $t-1$ auf Periode t gleich

$$r_t \;:=\; \frac{\textit{Wertänderung von } t-1 \textit{ auf } t}{\textit{Wert in Periode } t-1} \;=\; \frac{y_t - y_{t-1}}{y_{t-1}} \;=\; q_t - 1.$$



Es gilt somit

$$\begin{aligned} y_n &= y_{n-1} \cdot q_n = y_{n-2} \cdot q_{n-1} \cdot q_n = \dots \\ &= y_0 \cdot q_1 \cdot q_2 \cdots q_n. \end{aligned}$$

Wir suchen nun den durchschnittlichen Wachstumsfaktor $\hat{q}$ von Periode 0 auf Periode n. Für diesen muss nun gelten: Wenn y_0 insgesamt n-mal mit $\hat{q}$ multipliziert wird, kommt y_n heraus, also $y_0 \cdot \hat{q}^n = y_n$. Damit erhalten wir

$$\hat{q}^n = \frac{y_n}{y_0} = \frac{y_1}{y_0} \cdot \frac{y_2}{y_1} \cdot \frac{y_3}{y_2} \cdot \dots \cdot \frac{y_n}{y_{n-1}} = q_1 q_2 q_3 \cdots q_n$$

oder

$$\boxed{\hat{q} = \sqrt[n]{q_1 \cdot q_2 \cdot q_3 \cdots q_n} = \sqrt[n]{\frac{y_n}{y_0}}}$$

und das ist gerade das geometrische Mittel der einzelnen Faktoren q_i. Zur Berechnung ist aber die zweite Formel viel besser geeignet.

Für unser einführendes Beispiel bedeutet das nun zunächst: $y_0 = 1000$, $y_1 = 1200$, $y_2 = 1500$ und $y_3 = 1000$ sowie $q_1 = 1200/1000 = 6/5$, $q_2 = 1500/1200 = 5/4$ und $q_3 = 1000/1500 = 2/3$. Das bedeutet nun

$$\hat{q} = \sqrt[3]{q_1 q_2 q_3} = \sqrt[3]{\frac{6}{5} \cdot \frac{5}{4} \cdot \frac{2}{3}} = \sqrt[3]{\frac{60}{60}} = 1 = \sqrt[3]{\frac{1000}{1000}},$$

und das ist ein sinnvolles Resultat.

Aufgabe 1.4 *Der S. New Power Funds-B zeigte vom 29.06.- 03.07.2008 die folgende Wertentwicklung:*

Datum	*29.06.*	*30.06.*	*01.07.*	*02.07.*	*03.07.*
Kurs	55.50	55.58	56.16	55.16	54.84

Bestimmen Sie die durchschnittliche Wachstumsrate für diesen Zeitraum.

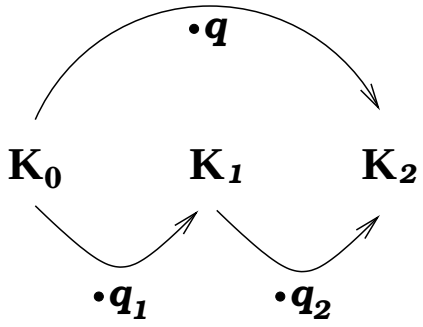
Lösung:

Einfache und logarithmische Rendite Wachstumsraten und Wachstumsfaktoren spielen in der Finanzmathematik eine wichtige Rolle. Wir wollen zunächst die eng verwandten Begriffe der einfachen und logarithmischen Rendite einführen und beziehen uns dabei auf die Anfangs- und Endkurse einer Aktie.

$$\text{einfache Rendite} \;=\; \frac{\text{Endkurs - Anfangskurs}}{\text{Anfangskurs}} \;=\; r \;=\; q-1$$

$$\text{logarithmische Rendite} \;=\; \ln\left(\frac{\text{Endkurs}}{\text{Anfangskurs}}\right) \;=\; \ln(q)$$

Die Bedeutung der einfachen Rendite ist klar, trotzdem wird in der Praxis meist mit der logarithmischen Rendite gerechnet. Ein Grund dafür ist, dass diese additiv ist. Dazu betrachten wir das folgende Beispiel einer Kapitalentwicklung über zwei Tage hin:



	$K_0 \to K_1$	$K_1 \to K_2$	$K_0 \to K_2$
einf. Rendite	$q_1 - 1$	$q_2 - 1$	$q - 1 \;=\; q_1 q_2 - 1$
log. Rendite	$\ln(q_1)$	$\ln(q_2)$	$\ln(q) \;=\; \ln(q_1 q_2) \;=\; \ln(q_1) + \ln(q_2)$

Man kann erkennen, dass sich die beiden logarithmischen Teilrenditen einfach zur Gesamtrendite addieren lassen. Das ist eine schöne Eigenschaft, die die einfache Rendite nicht hat!

2 Varianz und Standardabweichung

Beispiel 2.1 *Stellen wir uns zwei Studierende A und B vor, die jeweils acht Einzelleistungen erbracht haben, aus denen eine Gesamtnote errechnet wird. A hat achtmal die Note* 3 *und B viermal die Note* 5 *und viermal die Note* 1*. Sowohl arithmetisches Mittel als auch Zentralwert sind beide gleich* 3*, obwohl sich A und B in der Konstanz ihrer Leistungen völlig unterscheiden.*

Offenbar sind die Mittelwerte blind gegenüber der **Streuung** der Einzeldaten. Um die Mittelwerte sinnvoll zu ergänzen, benötigen wir Masszahlen, die etwas über die Abweichung der Einzeldaten vom Mittelwert aussagen, eben die **Streuungsparameter**. Sei also eine statistische Masse mit n Elementen, die wir uns mit $1, 2, \ldots, n$ durchnummeriert denken, und ein (quantitatives) Merkmal X gegeben. Für alle $i = 1, 2, \ldots, n$ bezeichne x_i die Ausprägung des Merkmals X beim Element i.

Definition 2.1 *Die (empirische) Varianz $var(x)$ der Daten $x_1, x_2, \ldots, x_n$ ist definiert durch*

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

Die Standardabweichung $\sigma(x) = \sqrt{var(x)}$ ist die positive Quadratwurzel aus der Varianz.

Bemerkung 2.1 *Viele Autoren bezeichnen die Grösse*

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

als empirische Varianz. Im Skript ,,Schätzen“ werden Sie sehen, dass beide Werte nützlich sind und gebraucht werden, um so genannte erwartungstreue Schätzer zu konstruieren. Auch wenn Sie die Varianz mit Hilfe der implementierten Funktion an Ihrem Taschenrechner bestimmen ist es zunächst nicht klar, ob durch n oder $n-1$ dividiert wird.

Wird in der Prüfung nach der empirischen Varianz (oder Kovarianz) gefragt, werde ich beide Varianten als richtig bewerten!

Beispiel 2.2 *Wir wollen die Varianz der Daten* $6, 8, 7, 5, 8, 6, 6, 6, 6, 5, 8, 8, 6, 9, 6, 8, 7, 9, 5, 11$ *bestimmen. Wir wissen bereits, dass $\bar{x} = 7$ gilt. Damit gilt*

$$\begin{aligned}
& var(x) \\
= \;& \frac{1}{20} \sum_{i=1}^{20} (x_i - 7)^2 \\
= \;& \frac{1}{20} \left[\, 3 \cdot (5-7)^2 + 7 \cdot (6-7)^2 + 2 \cdot (7-7)^2 + 5 \cdot (8-7)^2 + 2 \cdot (9-7)^2 + 1 \cdot (11-7)^2 \, \right] \\
= \;& 2.526315789 \;\; (\textit{mit TI-30XIIS})
\end{aligned}$$

Ergänzen Sie die fehlenden Rechenschritte:

$$
\begin{aligned}
var(x) &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \\
&= \frac{1}{n} \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2
\end{aligned}
$$

3 *Schiefe und Exzess(zur Information)*

Neben den Mittelwerten und dem Streumass werden zur Charakterisierung einer Stichprobe $x_1, x_2, \ldots, x_n$ weitere statistische Masszahlen verwendet.

Definition 3.1 *Es seien die Daten $x_1, x_2, \ldots, x_n$ gegeben. Die (empirische) Schiefe g_1 ist*

$$g_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^3}} .$$

Der (empirische) Exzess (Wölbung) g_2 ist

$$g_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} - 3 .$$

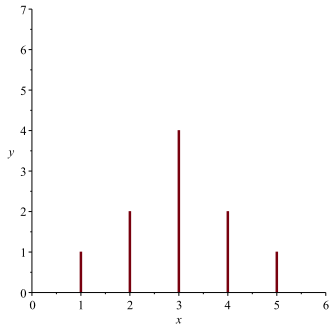
Die Schiefe ist ein Mass für die Symmetrie der Häufigkeitsverteilung, insbesondere ist $g_1 = 0$, wenn die Verteilung symmetrisch zum Mittelwert $\bar{x}$ liegt. Der Exzess kann als ein Mass für die Steilheit der Häufigkeitsverteilung angesehen werden. Hier einige Beispiel.

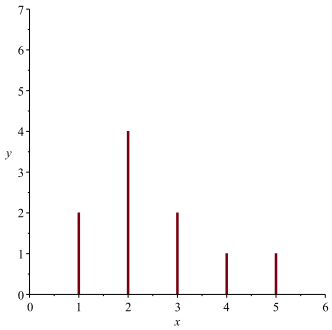
Aufgabe 3.1

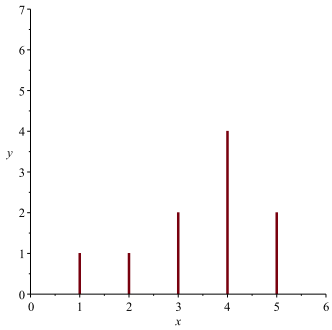
- *Wir untersuchen die folgenden drei Häufigkeitsverteilungen mit den stets gleichen 5 Merkmalsausprägungen $a_j = 1, 2, 3, 4, 5$ und den jeweils zugehörigen Häufigkeiten h_j:*

a_j	1	2	3	4	5
h_j *1. Probe*	1	2	4	2	1
h_j *2. Probe*	2	4	2	1	1
h_j *3. Probe*	1	1	2	4	2

Berechnen Sie für alle drei Proben die Zahl g_1.





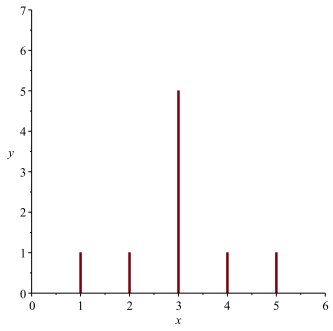


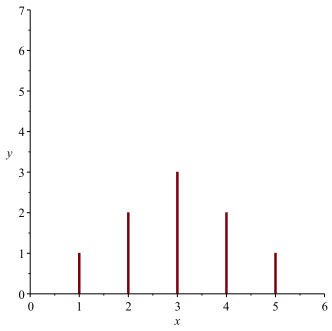
symmetrisch	*rechtsschief*	*linksschief*
$\bar{x} =$	$\bar{x} =$	$\bar{x} =$
$g_1 =$	$g_1 =$	$g_1 =$

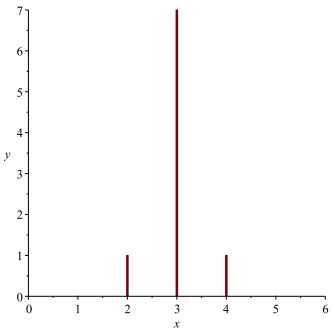
- *Wir untersuchen die folgenden drei Häufigkeitsverteilungen mit den stets gleichen 5 Merkmalsausprägungen $a_j = 1, 2, 3, 4, 5$ und den jeweils zugehörigen Häufigkeiten h_j:*

a_j	1	2	3	4	5
h_j *1. Probe*	1	1	5	1	1
h_j *2. Probe*	1	2	3	2	1
h_j *3. Probe*	0	1	7	1	0

Berechnen Sie für alle drei Proben die Zahl g_2.







mesokurtisch	*platykurtisch*	*leptokurtisch*
$\bar{x} =$	$\bar{x} =$	$\bar{x} =$
$g_2 = 0$	$g_2 = -0.75$	$g_2 = 1.5$

4 Konzentrationsmasse: Lorenzkurve und Ginikoeffizient

Hier soll das allgemeine Problem behandelt werden, wie sich die Merkmalssumme $\sum_i^n x_i$ auf die einzelnen Ausprägungen verteilt. Wir wollen dabei ein Mass entwickeln, das es uns ermöglicht Verteilungen im Hinblick auf ihre ‚Verteilungsgerechtigkeit, zu vergleichen.

Beispiel 4.1 *In einem Dorf gibt es drei Bauern, der erste mit einer Kuh, der zweite mit zwei Kühen und der dritte mit sieben Kühen.*

Wie gross ist die Ungleichheit beim Kuhbesitz?

Keine der bisher behandelten Grössen ist zur Beantwortung dieser Frage geeignet. Die Standardabweichung ist zum Beispiel 2.62*. Bekäme nun jeder Bauer* 30 *Kühe geschenkt (die Bauern hätten dann* 31*,* 32 *bzw.* 37 *Kühe), hätte die Ungleichheit intuitiv abgenommen, die Standardabweichung hätte sich aber nicht geändert.*

Die beobachteten Merkmalsausprägungen $x_1, x_2, \ldots, x_n$ eines (metrisch skalierten und häufbaren) Merkmals, das nur nicht-negative Ausprägungen besitzt, werden der Grösse nach geordnet: $0 \leq x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. Wir berechnen:

$$u_i = \frac{i}{n} \quad \text{für } i = 0, 1, \ldots, n \quad \text{sowie} \quad v_0 = 0$$

$$v_i = \frac{\sum_{j=1}^{i} x_{(j)}}{\sum_{j=1}^{n} x_{(j)}} \quad \text{für } i = 1, \ldots, n$$

Für $i = 1, \ldots, n$ ist v_i somit der (relative) Anteil am Gesamtreichtum, der auf die i kleinsten Merkmalsträger entfällt.

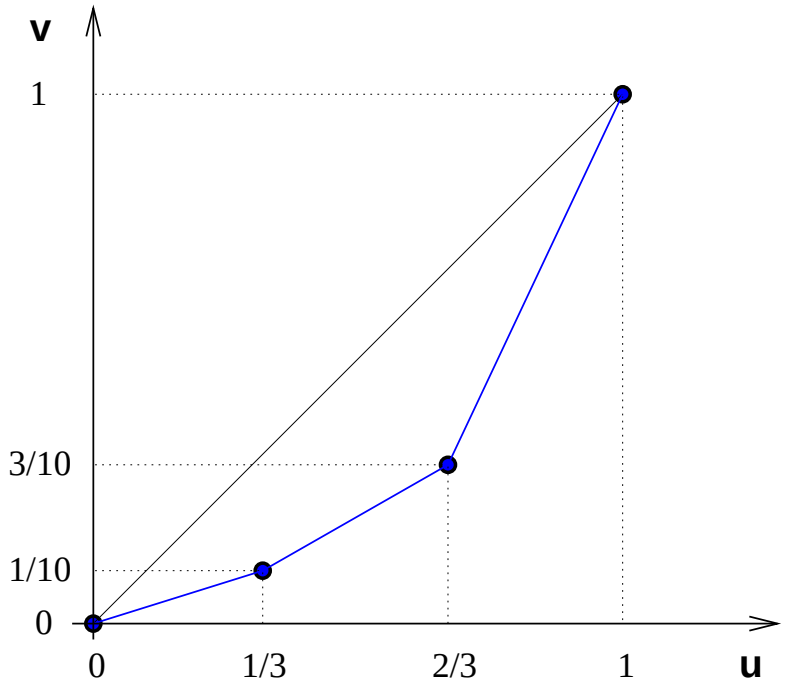
Definition 4.1
Die Lorenzkurve ist der Streckenzug, der durch die Punkte $(u_0, v_0), (u_1, v_1), \ldots, (u_n, v_n)$ verläuft. Der Ginikoeffizient G ist definiert durch $G = 2 \cdot F$, wobei F die Fläche zwischen der Diagonalen und der Lorenzkurve ist.

Satz 1 *Für den Ginikoeffizienten G gilt*

$$G = \frac{2\sum_{i=1}^{n} i \cdot x_{(i)} - (n+1)\sum_{i=1}^{n} x_{(i)}}{n\sum_{i=1}^{n} x_{(i)}} = 1 - \frac{1}{n}\left(2\sum_{i=1}^{n-1} v_i + 1\right)$$

Beispiel 4.2 *Für unser obiges Beispiel gilt:* $x_1 = 1$ *(Kühe),* $x_2 = 2$ *und* $x_3 = 7$ *sowie:*

$$\begin{aligned} u_0 &= \frac{0}{3} & v_0 &= 0 \\ u_1 &= \frac{1}{3} & v_1 &= \frac{1}{1+2+7} = \frac{1}{10} \\ u_2 &= \frac{2}{3} & v_2 &= \frac{1+2}{1+2+7} = \frac{3}{10} \\ u_3 &= \frac{3}{3} & v_3 &= \frac{1+2+7}{1+2+7} = 1 \end{aligned}$$



Für den Ginikoeffizienten gilt mit Hilfe der zweiten Formel:

$$G \;=\; 1 - \frac{1}{3}\left(2 \cdot \left(\frac{1}{10} + \frac{3}{10}\right) + 1\right) \;=\; \frac{2}{5}$$

Berechnen Sie G *für den zweiten Fall* $x_1 = 31$, $x_2 = 32$ *und* $x_3 = 37$*!*

5 Übungsaufgaben

1. Wir untersuchen den Datensatz: 12, 8, 5, 15, 6, 7, 14 und 6. Bestimmen Sie

   (a) das arithmetische Mittel,

   (b) das bezüglich $\mathbf{p} = \left(0, \frac{1}{20}, \frac{2}{20}, \frac{3}{20}, \frac{4}{20}, \frac{5}{20}, \frac{5}{20}, 0\right)$ gewichtete arithmetische Mittel,

   (c) den Zentralwert,

   (d) das geometrische Mittel,

   (e) die Varianz und

   (f) die Standardabweichung.

   Kennen Sie die Statistikfunktionen Ihres (einfachen) Taschenrechners?

2. Eine Gemeinde besteht aus genau zwei Ortsteilen: dem Oberdorf mit genau 600 Einwohnern und dem Unterdorf mit genau 400 Einwohnern. Bei der Wahl zum Bürgermeister stimmen im Oberdorf genau 59% und im Unterdorf genau 75% für den Kandidaten I. Wieviel Prozent aller Einwohner der Gemeinde stimmen dann für den Kandidaten I?

3. Ein Computerhersteller konnte seine Umsätze von 1990 im Jahr 1991 auf 107% gegenüber dem Vorjahr steigern. In der folgenden Jahren 1992-1995 gelang das gegenüber dem Vorjahr jeweils auf 102%, 108%, 111% und 105%. Bestimmen Sie den durchschnittlichen Wachstumsfaktor.

4. (a) Aus einer Messung stammen die drei Entfernungen $x_1 = 1\ km$, $x_2 = 2\ km$ und $x_3 = 4\ km$. Um mit einer anderen Messreihe vergleichen zu können, sollen die Daten in die Masseinheit Meter (kurz: m) umgerechnet werden, was mit der Vorschrift

   $$y_i \ = \ 1000 \cdot x_i \qquad\qquad \text{für } i = 1, 2, 3$$

   geschieht. Berechen Sie $\bar{x}$, $\bar{y}$, $var(x)$ und $var(y)$. Wie hängen die beiden Mittelwerte bzw. die beiden Varianzen zusammen?

   (b) Seien $x_1, \ldots, x_n$ Daten und $a, b \in \mathbb{R}$ reelle Zahlen. Wir betrachten nun die linear transformierten Daten $y_1, \ldots, y_n$ mit $y_i = a \cdot x_i + b$ für $i = 1, \ldots, n$. Beweisen Sie die folgenden Gleichungen:

   $$\boxed{\bar{y} \ = \ a \cdot \bar{x} + b \qquad \text{und} \qquad var(y) \ = \ a^2 \cdot var(x)}$$

5. In einer Grossgemeinde gibt es 10 Facharztpraxen, die sich in kleine, mittlere und grosse Praxen einteilen lassen (innerhalb einer Gruppe werde jeweils das gleiche Einkommen erziehlt). Im Jahr 2002 erhielten alle 10 Praxen zusammen ein Einkommen von 3 Millionen CHF. Allein 40 % davon entfiehlen auf die einzige grosse Facharztpraxis, während die 5 kleinen Praxen nur insgesamt ein Einkommen von 600 000,- CHF erziehlten.

   (a) Zeichnen Sie die Lorenzkurve.

   (b) Berechnen Sie den Gini-Koeffizienten.

Resultate einiger Übungsaufgaben

1. (a) $\frac{73}{8} = 9.125$
   (b) $\frac{48}{5} = 9.6$
   (c) $\frac{15}{2} = 7.5$
   (d) $8.4257 \;=\; (12 \cdot 8 \cdot 5 \cdot 15 \cdot 6 \cdot 7 \cdot 14 \cdot 6)^{1/8}$
   (e) 13.609
   (f) $3.689 \;=\; \sqrt{13.609}$

2. 65.4%

   Hinweis: Für die gesuchte Zahl p sollte folgendes gelten:

$$\underbrace{\frac{59}{100} \cdot 600}_{\substack{\text{Anzahl Wähler}\\ \text{im O-Dorf}}} \quad + \quad \underbrace{\frac{75}{100} \cdot 400}_{\substack{\text{Anzahl Wähler}\\ \text{im U-Dorf}}} \quad = \quad \underbrace{p \cdot 1000}_{\substack{\text{Anzahl Wähler}\\ \text{im Dorf}}}$$

3. $\sqrt[5]{1.07 \cdot 1.02 \cdot 1.08 \cdot 1.11 \cdot 1.05} = 1.0656$

4. (a) Direkte Rechnung ergibt zunächst

$$\begin{aligned} x_1 &= 1\,[km] \;\rightarrow\; y_1 = 1000\,[m] \\ x_2 &= 2\,[km] \;\rightarrow\; y_2 = 2000\,[m] \\ x_3 &= 4\,[km] \;\rightarrow\; y_3 = 4000\,[m] \end{aligned}$$

   und

$$\begin{aligned} \bar{x} &= \frac{1}{3}(1+2+4) \;=\; \frac{7}{3} \\ \bar{y} &= \frac{1}{3}(1000+2000+4000) \;=\; \frac{7000}{3} \end{aligned}$$

$$\begin{aligned} var(x) &= \frac{1}{3}\left(\left(1-\frac{7}{3}\right)^2 + \left(2-\frac{7}{3}\right)^2 + \left(4-\frac{7}{3}\right)^2\right) \\ &= 1.555'555 \\ var(y) &= \frac{1}{3}\left(\left(1000-\frac{7000}{3}\right)^2 + \left(2000-\frac{7000}{3}\right)^2 + \left(4000-\frac{7000}{3}\right)^2\right) \\ &= 1'555'555 \end{aligned}$$

   Und man erkennt die folgenden Zusammenhänge

$$\bar{y} \;=\; \frac{7000}{3} \;=\; 1000 \cdot \frac{7}{3} \;=\; 1000 \cdot \bar{x}$$

   (das arithmetische Mittel transformiert sich genau so, wie die einzelnen Daten) und

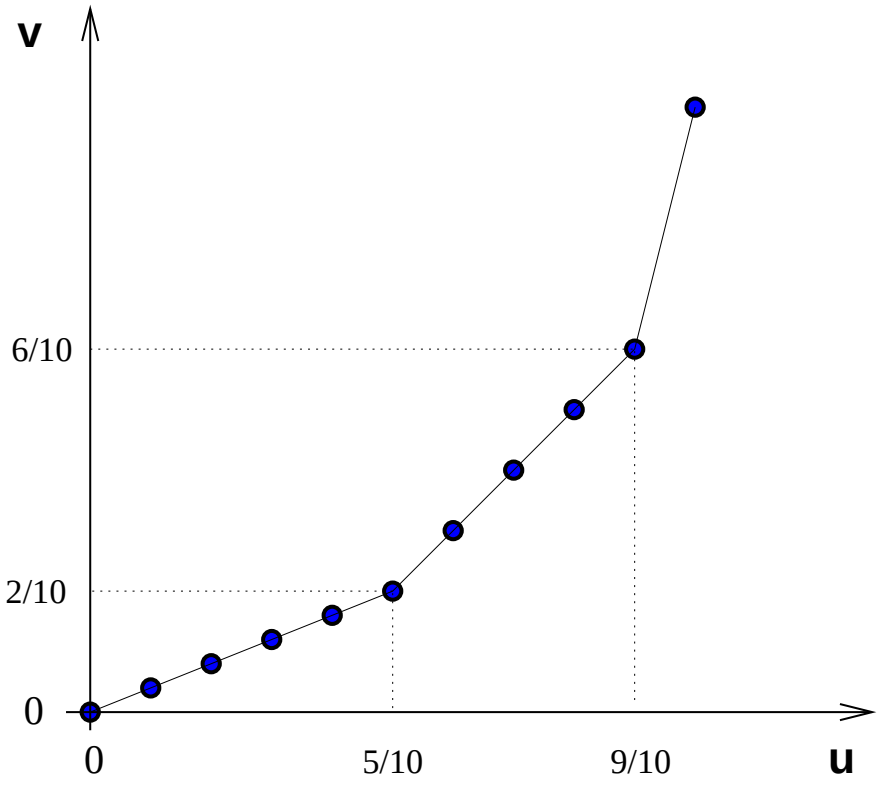
$$var(y) \;=\; 1'555'555 \;=\; 1'000'000 \cdot 1.555'555 \;=\; 1'000'000 \cdot var(x)$$

   (der Übergangsfaktor ist $1'000^2$).

(b)

5.

i	u_i	v_i
0	0	0
1	1/10	4/100
2	2/10	8/100
3	3/10	12/100
4	4/10	16/100
5	5/10	20/100
6	6/10	30/100
7	7/10	40/100
8	8/10	50/100
9	9/10	60/100
10	10/10	100/100



Der Ginikoeffizient ist

$$\begin{aligned} G &= 1 - \frac{1}{10}\left[2 \cdot \frac{1}{100}\left(4+8+12+16+20+30+40+50+60\right)+1\right] \\ &= \ldots \\ &= 0.42 \end{aligned}$$