

6. Maßzahlen I: Lageparameter

Ein skeptischer Kunde fragt den Imbissbuden-Pächter, ob in seinem Wildragout auch Pferdefleisch dabei ist: "Ja" "und wieviel Pferdefleisch?" "Halb und halb", sagte daraufhin der Gastronom, "Ein Kaninchen und ein Pferd"

Überblick und Motivation für Maßzahlen

Beim Übergang von der Urliste zur Häufigkeitsverteilung bzw. der Verteilungsfunktion wurden bereits der Informationsgehalt zugunsten der Übersichtlichkeit reduziert ("die Daten aggregiert").

Für viele Fragestellungen enthält jedoch auch die Verteilungsfunktion noch zu viele unnötige Informationen. Speziell, wenn es um Vergleiche oder die kompakte Darstellung eines Sachverhalts geht, ist man an wenigen charakteristischen *Zahlen*, den sog. **Maßzahlen** interessiert, die die Verteilung möglichst gut beschreiben ("weniger ist manchmal mehr"). Man unterscheidet

- **Lageparameter:** z.B. Mittelwerte, Quantile.
- **Streuungsparameter:** z.B. Varianz, Standardabweichung, Spannweite etc.
- Maßzahlen für die Form der Verteilung
- **Konzentrationsmaße**

6.1 Arithmetisches Mittel

Das **arithmetische Mittel** (oder einfach “Mittel” oder “Mittelwert”) ist der am meisten verwendete Lageparameter.

- Anwendung immer dann, wenn kardinalskalierte Daten vorliegen und keine der in Kap. 6.2 bis 6.5 aufgeführten Gründe für die Wahl eines anderen Lageparameters sprechen.
- Je nachdem, ob die Daten als Urliste, als Häufigkeitstabelle für die Merkmalsausprägungen (“gepoolte Daten”) oder klassiert vorliegen, muss man den Mittelwert anders berechnen.

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

arithmetisches Mittel aus der Urliste

$$\bar{x} = \frac{1}{n} \sum_{j=1}^m h_j x_j = \sum_{j=1}^m f_j x_j$$

aus der Häufigkeitstabelle (“gewichtetes arithm. Mittel”)

$$\bar{x} \approx \frac{1}{n} \sum_{k=1}^K h_k x_k^* = \sum_{k=1}^K f_k x_k^*$$

aus klassierten Daten
(x_k^* =Klassenmitte)

Beispiel: Gegeben ist eine geordnete Urliste mit den Werten

2 4 5 5 6 6 6 7 8 8 9 9 10 13.

Ermitteln Sie die Häufigkeitstabelle, klassieren Sie die Daten (Klassenbreite 2, erste Klasse von 0.5 bis 2.5) und bilden Sie für alle drei Datenkategorien den Mittelwert.

6.1 (b) Bemerkungen zur Klassierung

Aus dem vorigen Beispiel sieht man: Klassiert man die Daten erst, bevor man den Mittelwert berechnet, ist das Ergebnis i.A. leicht verfälscht. Daraus ergibt sich

- Wenn die Originaldaten in Form einer Urliste oder Häufigkeitstabelle vorhanden sind, wird der Mittelwert aus diesen Daten berechnet,
- Bei der Klasseneinteilung sollte man die Klassengrenzen so legen, dass im Falle von gleichverteilten Merkmalsausprägungen (diese Annahme steckt ja implizit hinter der Klassierung!) die Formel $\bar{x} = 1/n \sum h_k x^*_k$ exakt gilt.

Zusammen mit der Forderung nach lückenlosen Klassenintervallen ergibt das bei ganzzahligen Merkmalsausprägungen (Absolutskala, wie Stück- oder Personenzahlen) Intervallgrenzen, welche exakt zwischen den möglichen Werten liegen.

Dies wird deutlich, wenn man die Urliste

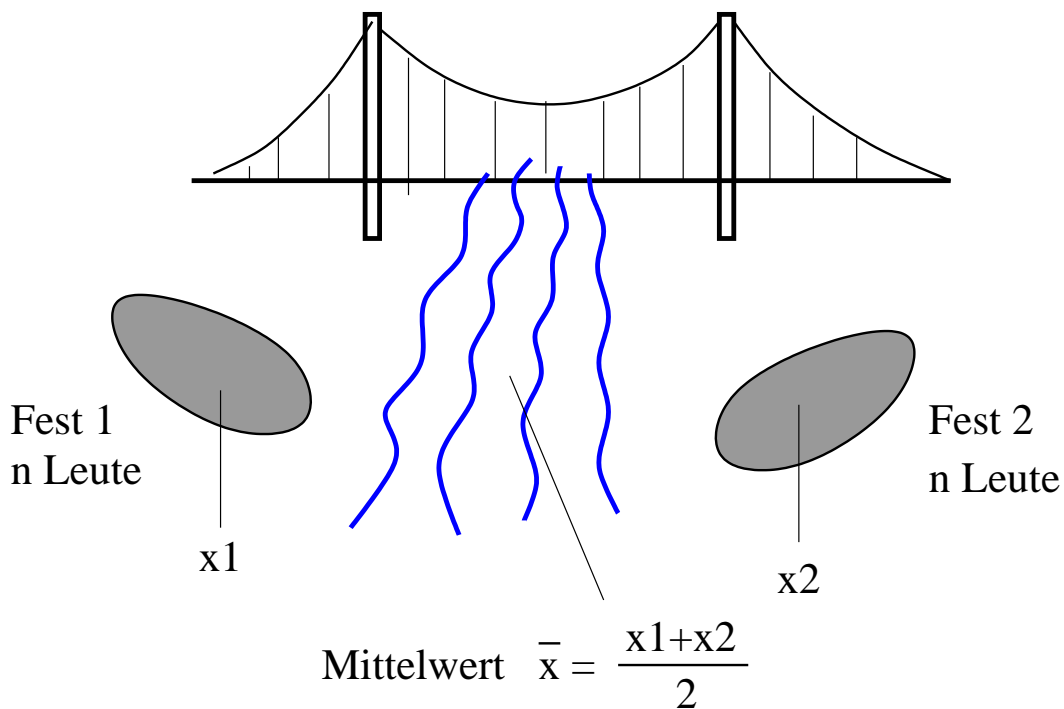
1 2 3 4

in zwei Klassen mit $x_1^u = 0.5$, $x_1^o = x_2^u = 2.5$ und $x_2^o = 4.5$ einteilt und dann den Mittelwert bildet.

6.1 (c) Eigenschaften des arithmetischen Mittels I

- Geeignet zur Beschreibung von eingipfligen und nicht zu unsymmetrischen Verteilungen.

Gegenbeispiel 1: Es wird auf beiden Seiten der Elbe gefeiert. Jeweils n Leute befinden sich an den Stellen x_1 und x_2 . Im Mittel sind die Leute also in der Elbe.



Gegenbeispiel 2: Am Samstag den 12.4. kamen 15 000 zum Fußballspiel, am Samstag den 19.4. kamen 20 000. Die Leute kamen im Mittel am 16.4., also am Mittwoch.

- Es ist empfindlich gegenüber Ausreißern.

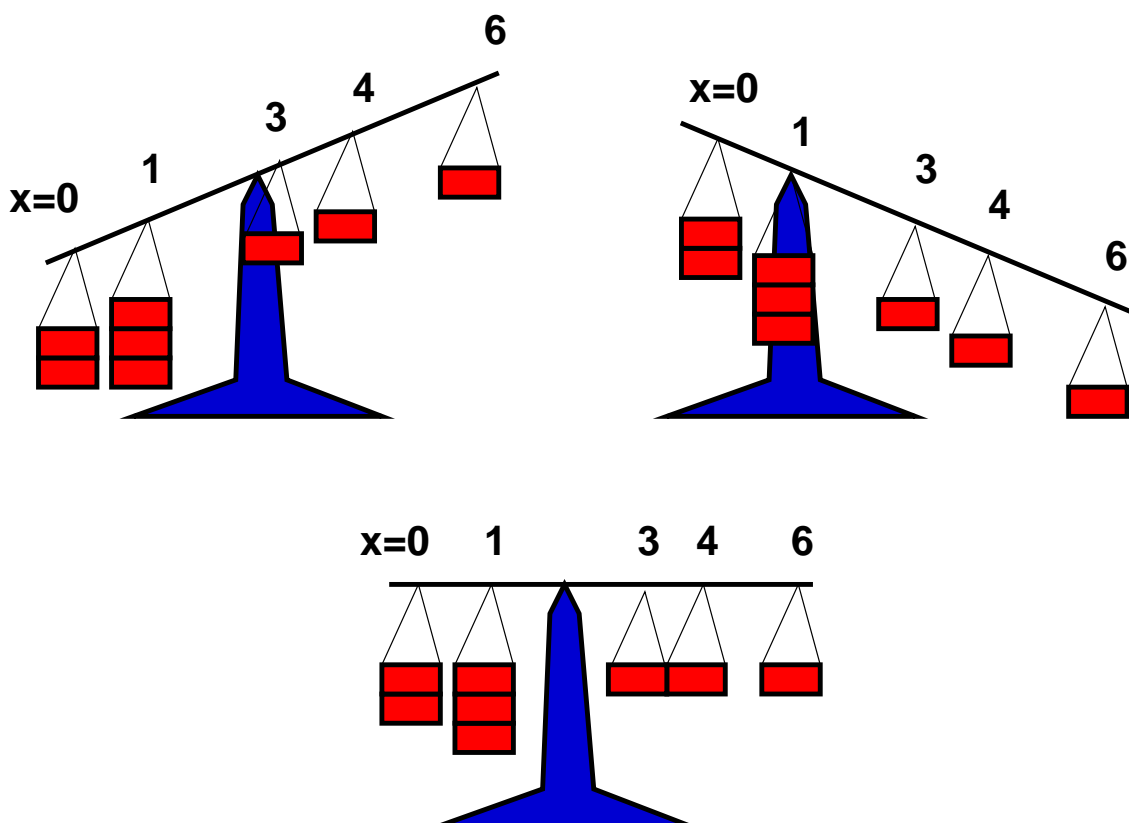
Beispiel: Zu einer Veranstaltung kommen 19 zu Fuß (Geschwindigkeit $v = 5$ km/h) und einer mit dem Auto ($v = 105$ km/h). Bildung des "Geschwindigkeitsmittels" $\bar{v} = 10$ km/h ist nicht sinnvoll.

6.1(d) Eigenschaften des arithmetischen Mittels II

Ein Statistiker ist jemand, der mit dem Kopf im Gefrierschrank und mit den Beinen im Backofen steckt und sagt: "Also meine Durchschnittstemperatur ist eigentlich optimal!"

Statistik, FH München

Betrachtet man die einzelnen zu mittelnden Größen als "Gewichte" vom Gewicht h_i , die an die Stellen x_i einer Waage mit verschiebbaren Lastarm gehängt werden, so gibt das arithmetische Mittel den Punkt an, bei dem die Waage im Gleichgewicht ist!



Das Bild entspricht etwa folgender Häufigkeitstabelle:

| | | | | | |
|-------|---|---|---|---|---|
| x_i | 0 | 1 | 3 | 4 | 6 |
| h_i | 2 | 3 | 1 | 1 | 1 |

6.1(e) Eigenschaften des arithmetischen Mittels III

- Hängt das Merkmal Y linear vom Merkmal X ab (so dass für die Ausprägungen gilt: $y_i = 1/x_i$), dann gilt Selbiges für das arithmetischen Mittel \bar{y} in Abhängigkeit von \bar{x} .

$$y_i = a + bx_i \Rightarrow \bar{y} = a + b\bar{x}$$

Dies sieht man leicht wie folgt:

$$\begin{aligned}\bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i \\ &= \frac{1}{n} \sum_{i=1}^n (a + bx_i) \\ &= \frac{1}{n} \sum_{i=1}^n a + \frac{b}{n} \sum_{i=1}^n x_i \\ &= a + b \left(\frac{1}{n} \sum_{i=1}^n x_i \right) \\ &= \underline{\underline{a + b\bar{x}}}\end{aligned}$$

Beispiel: Mietwagenpreis $Y = a + bX$, a : Fixkosten; b : Kilometerpreis; X : Gefahrene Kilometer; \bar{x} : im Mittel gefahrene Kilometer; \bar{y} =mittlerer Mietwagenpreis.

- Der Mittelwert minimiert die Quadratsumme der Abweichungen:

$$S = \sum_{i=1}^n (x_i - c)^2 = \min \text{ falls } c = \bar{x}$$

Beweis: S nach c ableiten und nullsetzen.

6.2 Geometrisches Mittel

$$\bar{x}_G = \left(\prod_{i=1}^n x_i \right)^{\frac{1}{n}} = \sqrt[n]{\prod_{i=1}^n x_i}$$

geometrisches Mittel aus der Urliste (n ist Zahl der Daten)

$$\bar{x}_G = \left(\prod_{j=1}^m x_j^{h_j} \right)^{\frac{1}{n}}$$

aus der Häufigkeitstabelle (m ist Zahl der verschiedenen Ausprägungen)

- Alle Werte müssen positiv sein
- Werden Werte y_i durch Logarithmieren der ursprünglichen Ausprägungen x_i gebildet, so ist das *arithmetische* Mittel der y_i gleich dem Logarithmus des *geometrischen* Mittels der x_i .

$$y_i = \log(x_i) \Rightarrow \bar{y} = \log(\bar{x}_G)$$

denn:

$$\begin{aligned} \log(\bar{x}_G) &= \log \left[\left(\prod_{i=1}^n x_i \right)^{\frac{1}{n}} \right] \\ &= \frac{1}{n} \log \left(\prod_{i=1}^n x_i \right) \\ &= \frac{1}{n} \left(\sum_{i=1}^n \log x_i \right) \\ &= \frac{1}{n} \sum_{i=1}^n y_i = \underline{\underline{\bar{y}}} \end{aligned}$$

6.2 (b) Anwendungsbeispiele

Das geometrische Mittel ist sachlich sinnvoll

- Bei Wachstumsvorgängen zur Beschreibung der Wachstumsfaktoren (nicht der Wachstumsraten)!

Beispiel: Die Zahl der Erdgasautos verdoppelte sich jeweils pro Jahr in den Jahren 1997-1999 und wuchs um je 50% pro Jahr in 2000 und 2001. Wie groß ist die mittlere Wachstumsrate von 1997-2001? Vergleichen Sie mit dem arithmetischen Mittel!

- Bei einigen Fragestellungen, wenn um mehrere Größenordnungen unterschiedliche Werte verglichen werden.

Beispiel: Über eine Woche gemittelte Lautstärke \bar{y} beim Musikhören z.B.

- So-Mo: Zimmerlautstärke (65 Dezibel);
- Sa: Rockkonzert (100 Dezibel)

Nun ist aber Y (Lautstärke in Dezibel) ein logarithmischer Maßstab der am Ohr ankommenden physikalischen Schallintensität $X \propto 10^{\frac{y}{10}}$. Subjektiv wird das arithmetische Mittel von Y gebildet, was dem geometrischen Mittel der um etwa den Faktor 3000 schwankenden Intensitäten entspricht.

6.2 (c) Aufgaben

1. Ein 7-jähriger Bundesschatzbrief liefert am Ende des 1. Jahrs 1 Prozent Zinsen, im 2. Jahr 2% usw. bis am Ende des 7. jahres 7% Zinsen bezahlt werden. Um wieviel Prozent wächst der Geldbetrag im Mittel in einem Jahr?
2. Ein Aktienfonds hat in den letzten Jahren folgende prozentuale Gewinne/Verluste gemacht:

| Jahr | 2001 | 2002 | 2003 | 2004 | 2005 |
|--------|------|------|------|------|------|
| G/V(%) | -50 | -25 | 0 | 25 | 50 |

Hat der Anleger am Ende sein Geld wieder?

6.3 Harmonisches Mittel

$$\bar{x}_H = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}} \quad \text{bzw.} \quad \frac{1}{\sum_{j=1}^m \frac{f_j}{x_j}} \quad \text{bzw.} \quad \frac{1}{\sum_{k=1}^K \frac{f_k}{x_k^*}}$$

- Das **harmonische Mittel** entspricht dem Kehrwert der arithmetischen Mittel der Kehrwerte
- Werden Werte $y_i = 1/x_i$ durch den Kehrwert der ursprünglichen Ausprägungen x_i gebildet, so ist das *arithmetische* Mittel der y_i gleich dem Kehrwert des *harmonischen* Mittels der x_i :

$$y_i = \frac{1}{x_i} \quad \Rightarrow \quad \bar{y} = \frac{1}{\bar{x}_H}$$

- Man kann zeigen: \bar{x}_H ist immer kleiner oder gleich \bar{x} .
- Die Entscheidung, wann das harmonische Mittel sinnvoll ist, setzt Kenntnis des Sachverhalts voraus! Kandidaten sind wegen der obigen Beziehung v.a. Verhältniszahlen, bei denen der *Zähler* zur Gewichtung verwendet wird, z.B.
 - Ermittlung von Durchschnittsgeschwindigkeiten
 - Mehrfacher Einkauf ein- und desselben Gutes mit schwankendem Stückpreis $X(t)$ für jeweils einen festen Geldbetrag

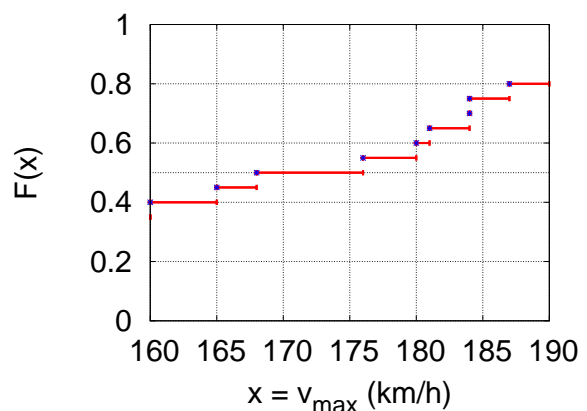
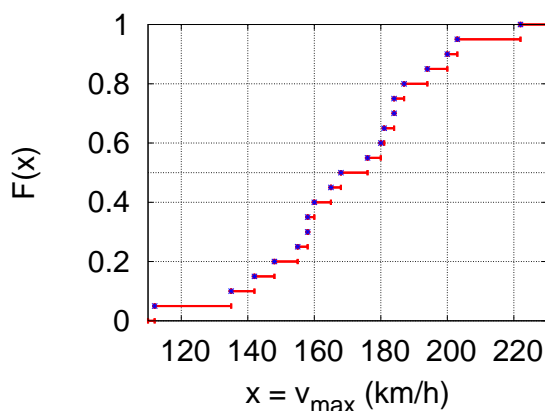
6.3 (b) Anwendungsbeispiele

1. *Mittlere Geschwindigkeit bei Autofahrt:* Fahrt über 180 km, davon 20 km mit 10 km/h (Stau), 20 km mit Tempo 80 und 30 km mit Tempo 120 (Tempolimits) sowie den Rest von 110 km mit 180 km/h (freie Fahrt). Wie groß ist die Durchschnittsgeschwindigkeit? Berechnen Sie diese “direkt” und zeigen Sie dann, dass x_H das “Mittel der Wahl” ist!
2. *Cost-Average Effekt:* Mittlerer Einstandskurs bei Ansparplänen und allgemein bei regelmäßigem Kaufen für einen festen Betrag, siehe [Exkurs zu dieser Vorlesung](#) .

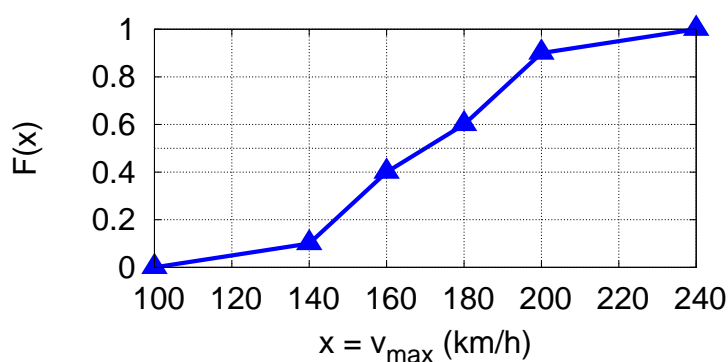
6.4 Median und Quantile

Der **Median (Zentralwert)** $x_{0.5}$ gibt den Wert der Merkmalsausprägung X an, bei dem 50% der Merkmalsausprägungen unterhalb und 50% oberhalb liegen. Im Gegensatz zum arithmetischen Mittel ist der Median auch für *ordinalskalierte* Daten x_i anwendbar. Konkret:

Der Median ist durch $F(x_{0.5}) = 0.5$ definiert. Gilt $F(x) = 0.5$ für ein ganzes Intervall $x_u \leq x \leq x_o$, so ist $x_{0.5} = \frac{1}{2}(x_u + x_o)$ gleich der Intervallmitte.



(i) Bestimmung des Medians aus der geordneten Urliste



(ii) Bestimmung des Medians bei klassierten Daten

6.4 (b) Berechnungsvorschrift für den Median

Liegt eine geordnete Urliste vor (eine Häufigkeitstabelle kann man zu einer Urliste expandieren), so folgt aus der Definition die Vorschrift

$$x_{0.5} = \begin{cases} x_{\lfloor \frac{n+1}{2} \rfloor} & (n \text{ ungerade}) \\ \frac{1}{2} \left(x_{\lfloor \frac{n}{2} \rfloor} + x_{\lfloor \frac{n}{2} \rfloor + 1} \right) & (n \text{ gerade}) \end{cases}$$

Aufgabe: Zeigen Sie dies!

Für klassierte Daten und unter der Annahme einer Gleichverteilung innerhalb der Klassen ergibt sich folgende "Feinberechnung" des Medians:

$$x_{0.5} = x_{k'}^u + \frac{0.5 - F_{k'-1}}{f_{k'}} \Delta x_{k'}$$

mit k' so dass $F_{k'-1} < 0.5$, aber $F_{k'} \geq 0.5$

Aufgabe: Zeigen Sie dies!

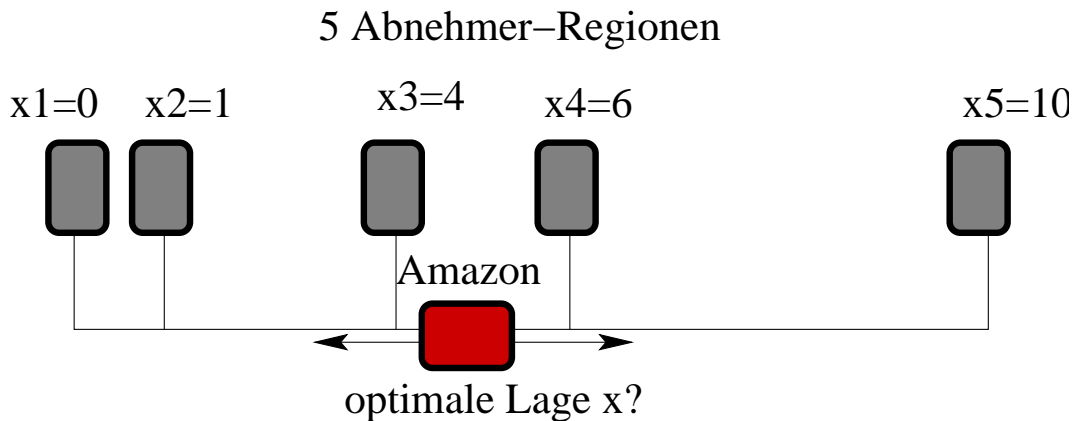
6.4 (c) Eigenschaften des Medians

- Er ist im Gegensatz zum arithmetischen Mittel unempfindlich gegenüber Fehlern
- Er nutzt im Gegensatz zum arithmetischen Mittel nicht die ganze Information der Häufigkeitsverteilung aus
- Der Median minimiert die Betragssumme der Abweichungen:

$$S = \sum_{i=1}^n |x_i - c| = \min \text{ falls } c = x_{0.5}$$

Beweis: F nach c ableiten (die Ableitung von $|x|$ ist $= 1$ für $x > 0$ und gleich -1 für $x < 0$) und nullsetzen.

Anwendung: Planung des Auslieferungslagers von Amazon



Wo sollte das Auslieferungslager von Amazon optimal stehen, damit die Wege minimiert werden?

6.4 (d) Quantile

Das **q -Quantil** x_q gibt den Wert der Merkmalsausprägung X an, bei dem ein Anteil q der Merkmalsausprägungen unterhalb von x liegen.

q -Quantile sind also eine Verallgemeinerung des Medians. Bei klassierte Daten gilt damit:

$$x_q = x_{k'}^u + \frac{q - F_{k'-1}}{f_{k'}} \Delta x_{k'}$$

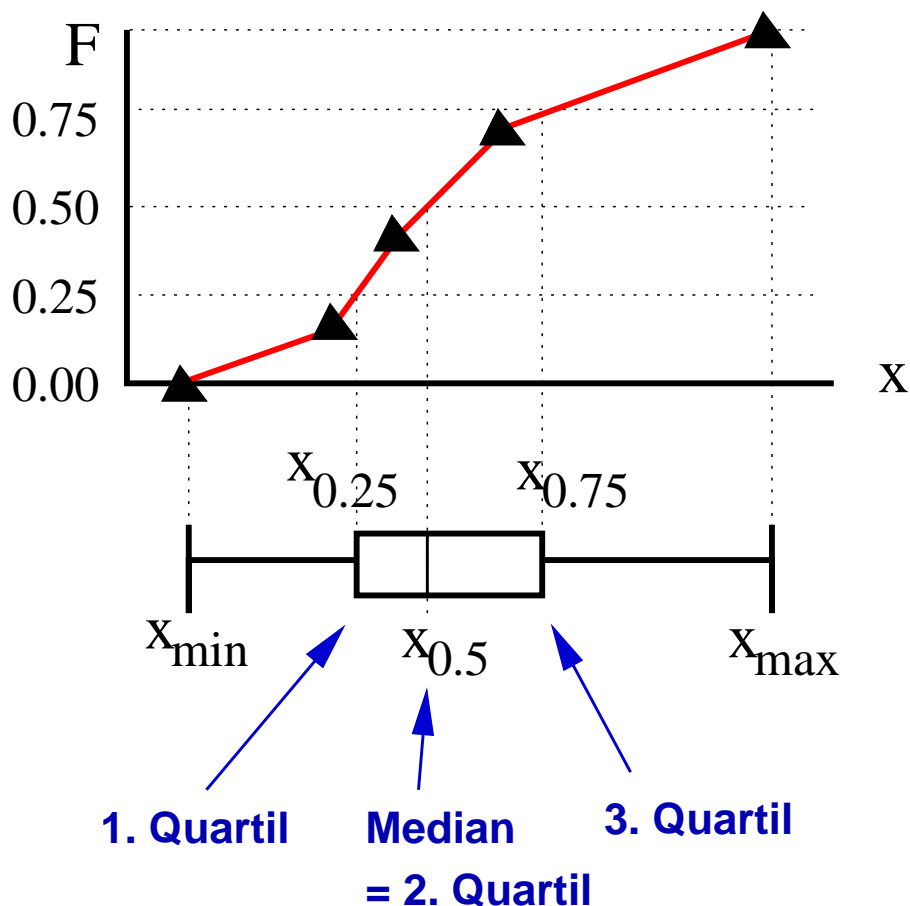
mit k' so dass $F_{k'-1} < q$, aber $F_{k'} \geq q$

Insbesondere spricht man von

- **Quartilen:** $x_{0.25}$ = erstes Quartil, Median = zweites Quartil, $x_{0.75}$ = drittes Quartil.
- **Dezilen:** $x_{0.10}$ = erstes Dezil, $x_{0.20}$ = zweites Dezil, etc.
- **Perzentilen:** $x_{0.01}$ = erstes Perzentil, etc.

6.4 (e) Boxplot

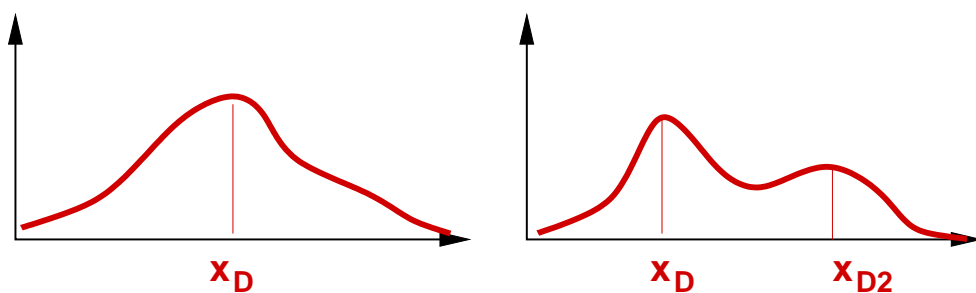
Eine kompakte Darstellung der Situation einschließlich der Streuung wird durch den **Boxplot** oder auch **Box & Whisker-Plot** ermöglicht. Die Lage und Breite der "Box" ist dabei durch das erste und dritte Quartil gegeben, die "Whiskers" (Barthaare) erstrecken sich bis zu den Extremwerten.



6.5 Modus

Der **Modus** bzw. **Modalwert** x_D ist die Merkmalsausprägung, die am häufigsten vorkommt.

- Der Modus ist für *beliebig skalierte* Daten definiert. Für nominalskalierte Daten ist er der einzige Lageparameter.
- Im Ggs zu den anderen Lageparametern kann der Modus mehrdeutig sein bzw. es kann mehrere “Modi” geben. Gibt es z.B. zwei Modi (doppelgipflige Verteilung), spricht man von einer **bimodalen** Verteilung (warum ist z.B. die Geschwindigkeitsverteilung auf Autobahnen oft nahezu bimodal?)



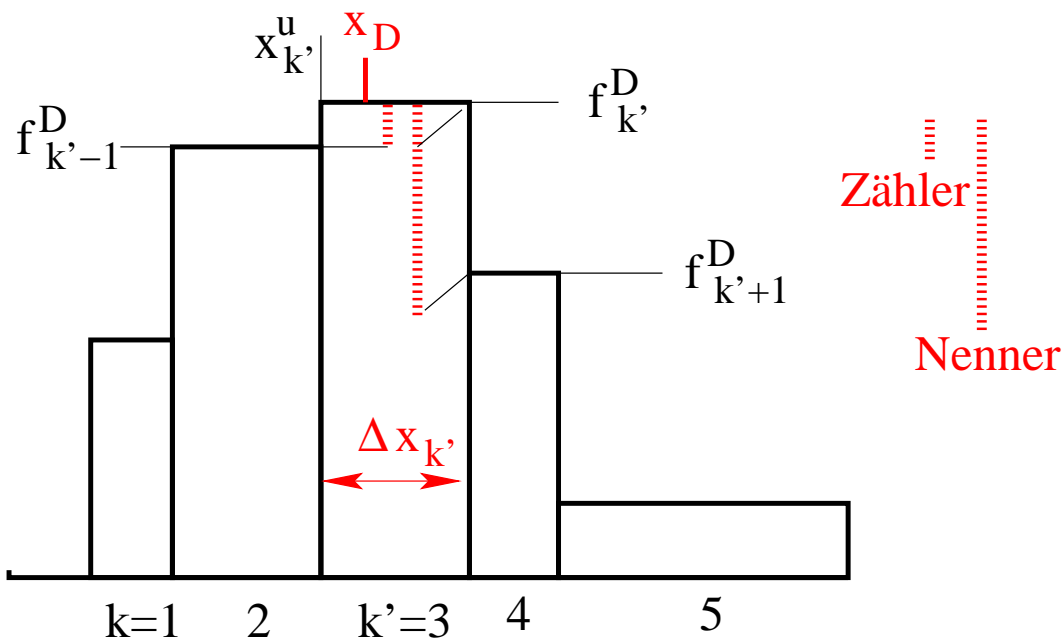
- Bei klassierten kardinalskalierten Daten liegt der Modus, dann auch **Dichtemittel** genannt, innerhalb der Klasse i mit der höchsten Häufigkeitsdichte f_i^* .

6.5(b) Modus bei klassierten Daten

Wie bei den Quantilen gibt es für klassierte Daten eine "Feinberechnung", bei der Parabeln durch jeweils drei Punkte der Punktmenge $\{(x_k^*, f_k^D)\}$ gelegt werden, wobei $f_k^D > f_{k-1}^D$ und $f_k^D > f_{k+1}^D$:

$$x_D = x_{k'}^u + \frac{f_{k'}^D - f_{k'-1}^D}{(f_{k'}^D - f_{k'-1}^D) + (f_{k'}^D - f_{k'+1}^D)} \Delta x_{k'}$$

mit k' der Klasse mit der höchsten Häufigkeitsdichte f^* .



Da diese Formel in der Herleitung endliche (durch parabolische Splines approximierte) Dichten in beiden Nachbarklassen voraussetzt, darf sie nicht angewandt werden, falls das Dichtemaximum innerhalb der Randklassen liegt.

6.6 (a) Diskussion der verschiedenen Mittelwerte

- Das arithmetische Mittel \bar{x} ist der “Standard-Mittelwert”
- Das harmonische Mittel x_H wird bei Mittelwerten von *Verhältniszahlen* verwendet, wenn die im *Zähler* stehende Größe zur Gewichtung herangezogen werden soll: z.B. Geschwindigkeit=Strecke/Zeit, Leistung=Arbeit/Zeit, Kfz-Dichte=Zahl der Kfz/Zahl der Einwohner etc
- Das geometrische Mittel x_G wird zur Mittelung von Wachstumsprozessen verwendet, bei der es auf die *prozentuale Änderung* ankommt.
- Der Median ist das “Mittel” der Wahl, wenn Extremwerte und/oder offene Randklassen keine Rolle spielen sollen, sowie bei ordinalskalierten Daten
- Der Modus ist sinnvoll bei mehrgipfligen (multimodalen) Verteilungen, bei stark asymmetrischen Verteilungen sowie bei nominalskalierten Daten
- Grundsätzlich gilt:

– Immer:

$$x_H \leq x_G \leq \bar{x}$$

– Bei symmetrischen unimodalen Verteilungen:

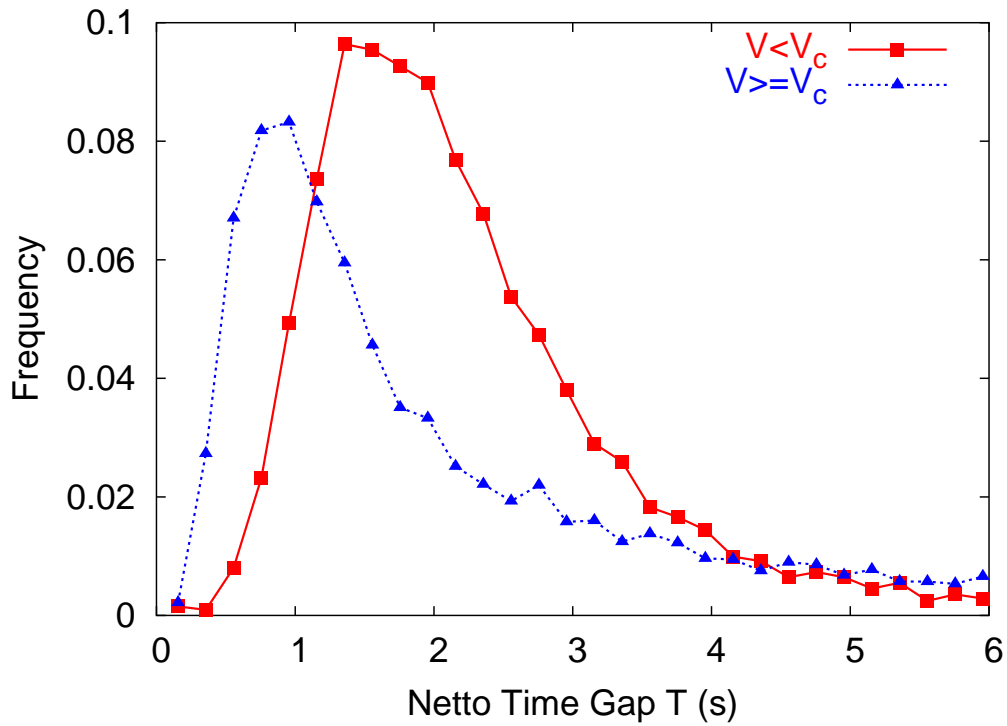
$$\bar{x} = x_{0.5} = x_D$$

– Bei rechtsschiefen (=linkssteilen) Verteilungen:

$$x_D < x_{0.5} < \bar{x},$$

bei rechtssteilen Verteilungen umgekehrt.

6.6 (b) Beispiels-Aufgabe (siehe auch Kap. 7, 8)



Diskutieren Sie anhand der Verteilung der Folgezeiten auf einer deutschen Autobahn verschiedene Lagemaße:

- Zunächst die Form der Verteilung: Linkssteil oder rechtssteil?
- Welcher Mittelwert bzw. allgemeines Lagemaß ist für die Aufgabenstellungen (i) Abschätzung des Verkehrsflusses, (ii) Bestimmung des Anteils der “Abstandssünder”, (iii) Bestimmung des wahrscheinlichsten zeitlichen Abstandes angemessen?
- Wie sieht die Reihenfolge der numerischen Werte der verschiedenen Mittelwerte aus? (zunächst qualitativ ohne Rechnung).
- Klassieren Sie nun die Daten anhand der Grafik (welche Fehlerquellen gibt es dabei?) und berechnen Sie die Mittelwerte

6.6 (c) Weitere Aufgabe: Die Radarfalle

Um den Einsatzort einer Radarfalle zu testen, werden für 1 Tag die Geschwindigkeiten an der geplanten Stelle (zunächst unauffällig, ohne Blitzgerät) gemessen. Das Ergebnis liegt in Form von klassierte Daten vor:

| Geschwindigkeits-Klasse | ≤ 60 | 60-80 | 80-90 | 90-100 | 100-120 | > 120 |
|-------------------------|-----------|-------|-------|--------|---------|---------|
| Zahl der Fahrzeuge | 1000 | 1200 | 900 | 500 | 300 | 100 |

- Wie groß ist das arithmetische Mittel und der Median? Bei welchem Mittelwert muss man dabei zusätzliche Annahmen treffen? Nehmen Sie ggf. eine Minimalgeschwindigkeit von 40 km/h und eine Maximalgeschwindigkeit von 140 km/h an.
- Zeichnen Sie die Verteilungsfunktion und den Boxplot!
- Nach einer Empfehlung sollte man das Tempolimit (wenn keine sonstigen Gründe dagegensprechen) so legen, dass es 85% der Fahrer bei freier Fahrt und ohne Polizei und Kontrolle sowieso einhalten würden. Wie würde gemäß dieser Empfehlung das Tempolimit lauten?
- Das tatsächliche Tempolimit beträgt 90 km/h. (Es sind nur "runde" Werte für das Tempolimit möglich). Die Polizei will nun abschätzen, was sie an einem Tag wie den untersuchten an Einnahmen erwarten kann. Das verwendete Messgerät hat bei 90 km/h eine Toleranz von 3 km/h, die zugunsten des Fahrers abgezogen wird. Eine Überschreitung von 5-15 km/h kostet 20 €, höhere Überschreitungen werden pauschal mit 80 € verrechnet. Welche Einnahmen können erwartet werden?
- Warum ist es nicht sinnvoll, anhand dieser Daten den erwarteten Anteil der Führerscheinentzüge (Überschreitung um über > 35 km/h) auszurechnen?
- Den lauernden Beamten wird es langweilig, so dass sie eine Wette machen, mit welcher Geschwindigkeit das nächste Fahrzeug vorbeikommt. Wer am nächsten dranliegt, hat gewonnen. Ein Polizist hat Statistik noch in Erinnerung. Welchen Wert sagt er?