2. Eindimensionale (univariate) Datenanalyse

Andreas Eberle
Institut für angewandte Mathematik

Oktober 2008

Kennzahlen, Statistiken

In der Regel interessieren uns nicht so sehr die beobachteten Einzeldaten selbst. Stattdessen betrachten wir verschiedene Kennzahlen und grafische Darstellungen, die einen Datensatz möglichst effizient beschreiben. Solche Kenngrößen nennt man Statistiken. Mathematisch präzise definieren wir:

Definition

Eine *Statistik* ist eine Funktion $S(x_1, x_2, \ldots, x_n)$, die von den beobachteten Merkmalsausprägungen $x_1, x_2, \ldots, x_n$ abhängt.

Beispiele:

- Relative Häufigkeit einer bestimmten Merkmalsausprägung im Datensatz $x_1, x_2, \ldots, x_n$.
- Median der Daten (bei Ordinalskala oder metrischer Skala).
- Mittelwert und Standardabweichung (bei metrischer Skala).

2.1. Häufigkeiten und empirische Verteilungen

Absolute und relative Häufigkeiten

- Die *(absolute) Häufigkeit* einer Merkmalsausprägung a ist

$$n(a) = \text{Anzahl der stat. Einheiten } \omega_i \text{ mit } x_i = a.$$

- Die *relative Häufigkeit* der Merkmalsausprägung a ist

$$h(a) = \frac{n(a)}{n} = \frac{\text{Häufigkeit von } a}{\text{Gesamtgröße der Stichprobe}}.$$

Häufigkeiten und empirische Verteilungen

Empirische Verteilung

Die relativen Häufigkeiten $h_k = h(a_k)$ aller vorkommenden Merkmalsausprägungen a_k $(k = 1, 2, \ldots, m)$ bilden die Gewichte einer *Wahrscheinlichkeitsverteilung*, d.h. die relativen Häufigkeiten sind alle nichtnegativ und summieren sich zu 1:

$$\sum_k h_k = \sum_k \frac{n_k}{n} = \frac{1}{n} \sum_k n_k = 1$$

Definition

Die durch die Gewichte $h_1, h_2, \ldots, h_m$ der möglichen Merkmalsausprägungen festgelegte Wahrscheinlichkeitsverteilung heißt *empirische Verteilung* der Stichprobe.

Häufigkeiten und empirische Verteilungen

Stabdiagramm

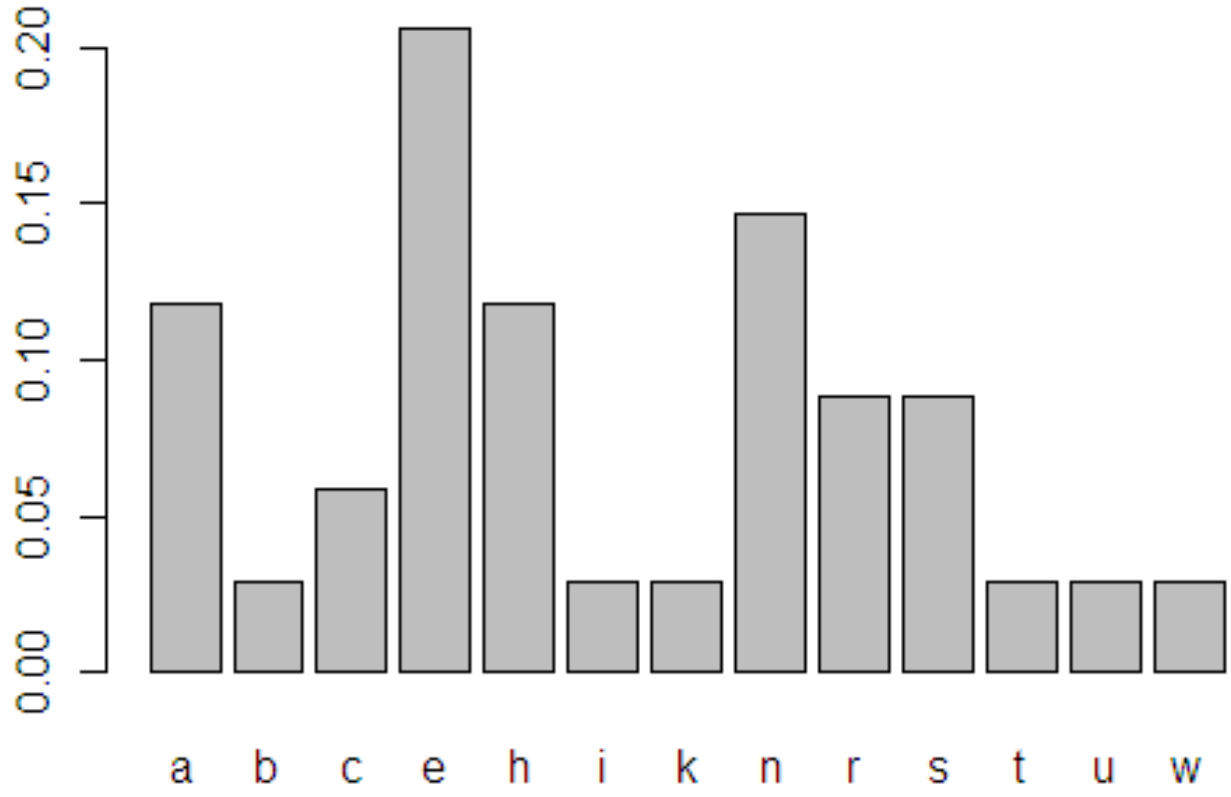
Graphisch stellt man die Häufigkeiten der Merkmalsausprägungen bzw. die empirische Verteilung als *Stabdiagramm* dar.

Beispiel.

Die absoluten und relativen Häufigkeiten der Buchstaben in dem Wort "Eisenbahnschrankenwaerterhaeuschen" sind:

a	*b*	*c*	*e*	*h*	*i*	*k*	*n*	*r*	*s*	*t*	*u*	*w*
4	*1*	*2*	*7*	*4*	*1*	*1*	*5*	*3*	*3*	*1*	*1*	*1*
0.12	*0.03*	*0.06*	*0.21*	*0.12*	*0.03*	*0.03*	*0.15*	*0.09*	*0.09*	*0.03*	*0.03*	*0.03*

Hieraus ergibt sich das Stabdiagramm der empirischen Verteilung:


0.00
0.05
0.10
0.15
0.20
a
b
c
e
h
i
k
n
r
s
t
u
w

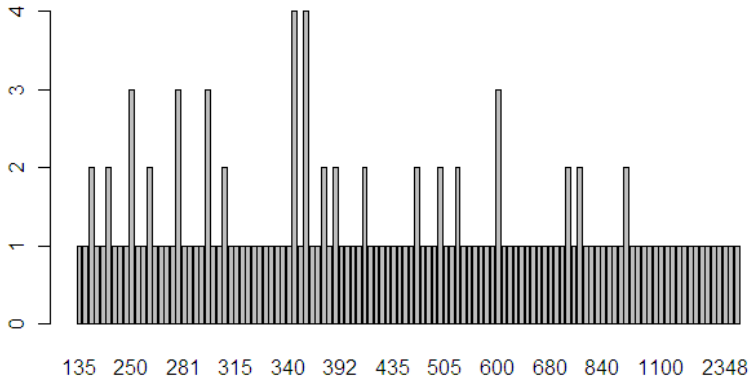
Häufigkeiten und empirische Verteilungen

Klasseneinteilung

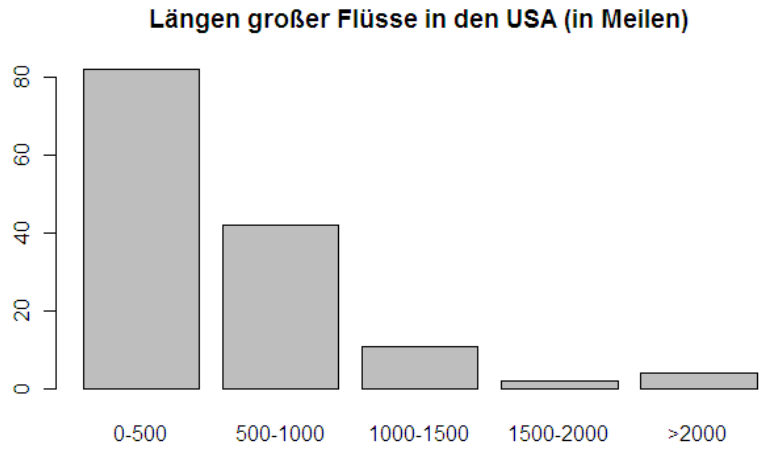
- Für metrisch skalierte Daten, oder allgemein bei einer großen Anzahl möglicher Merkmalsausprägungen, ist eine direkte Darstellung durch eine Häufigkeitstabelle oder ein Stabdiagramm oft nicht sinnvoll, da die Merkmalsausprägungen der Daten (fast) alle verschieden sind.
- Stattdessen kann man die Merkmalsausprägungen zu einer begrenzten Anzahl von *Klassen* zusammenfassen, und die Häufigkeiten der Klassen in einer Tabelle oder einem Stabdiagramm darstellen.
- Dabei ist aber Vorsicht geboten, da die Klasseneinteilung nicht eindeutig festgelegt ist, und Stabdiagramme mit unterschiedlichen Klasseneinteilungen ganz verschiedene subjektive Eindrücke bewirken können. *(How to lie with statistics)*

Beispiel. (Längen größerer Flüsse in den USA)

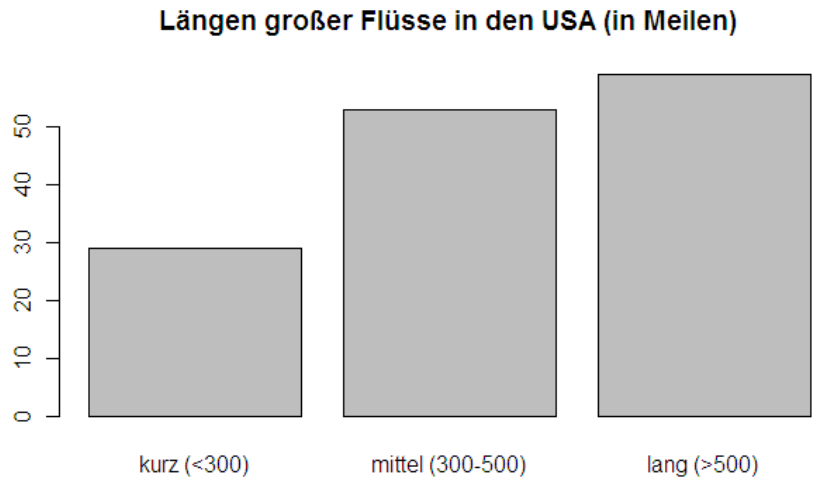
- *Das Stabdiagramm der Längen aller größeren Flüsse in den USA ist wenig aufschlußreich:*

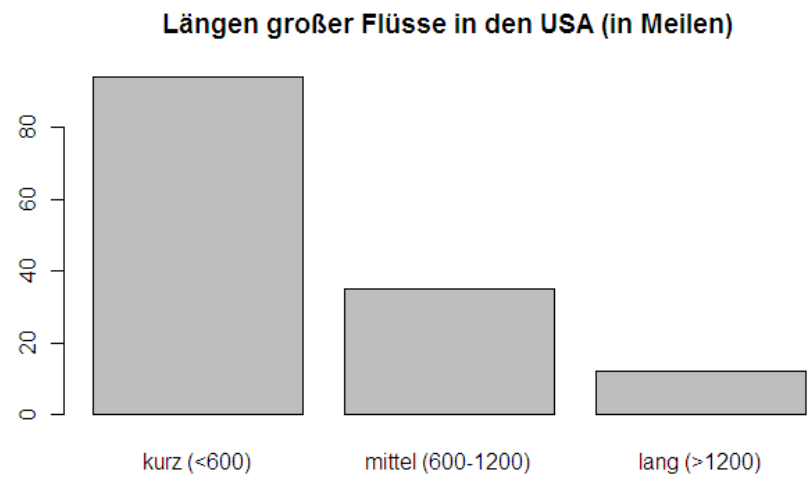


- *Eine Unterteilung in fünf Klassen liefert das folgende deutlich nützlichere Stabdiagramm:*



Zwei mögliche Einteilungen in kurze, mittlere und lange Flüsse, und die entsprechenden Stabdiagramme:

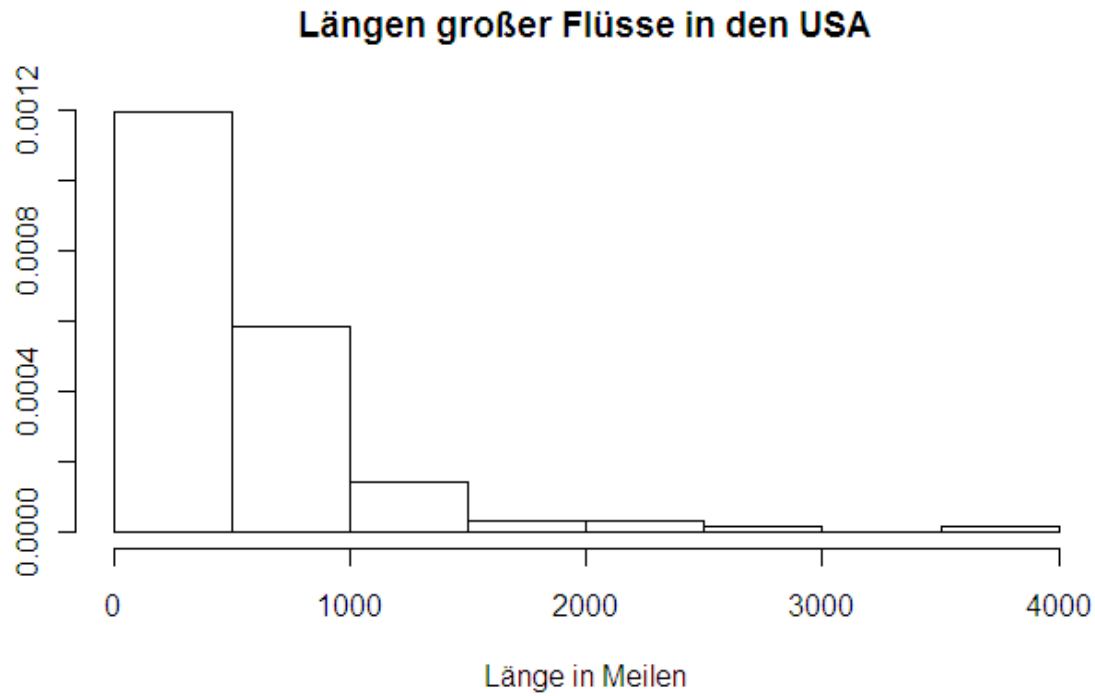




Häufigkeiten und empirische Verteilungen

Histogramme

Um Verzerrungen durch die gewählte Klasseneinteilung zu reduzieren, können wir **metrisch skalierte Daten** durch ein **Histogramm** darstellen. Dazu unterteilen wir die Merkmalsausprägungen wieder in mehrere Klassen (welche oft alle gleich groß sind), und tragen über jede Klasse ein Rechteck auf, dessen Flächeninhalt gleich der relativen Häufigkeit der Klasse ist:



Häufigkeiten und empirische Verteilungen

Empirische Dichte

Wir bezeichnen die Klassengrenzen mit

$$a_0^* < a_1^* < a_2^* < \ldots < a_l^*.$$

Der Funktionswert, der im Histogramm über der Klasse $K_j = (a_{j-1}^*, a_j^*]$ aufgetragen wird, ist

$$f(K_j) = \frac{h(K_j)}{a_j^* - a_{j-1}^*} = \frac{\text{relative Häufigkeit von } K_j}{\text{Größe von } K_j}$$

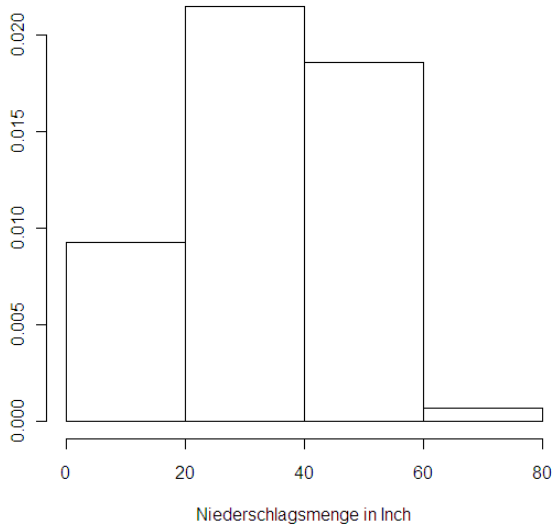
Definition

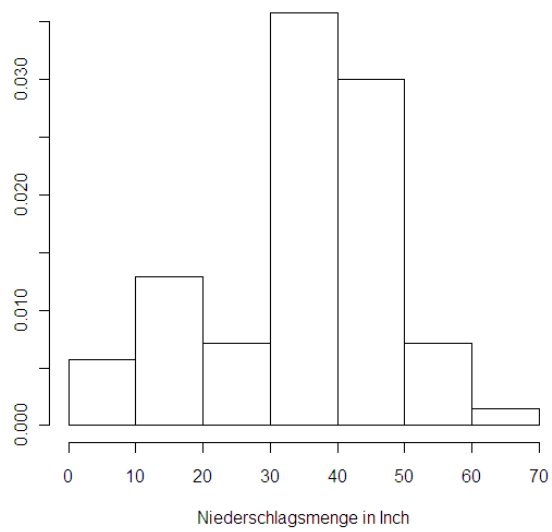
$f(K_j)$ heißt **empirische Dichte** der Klasse K_j.

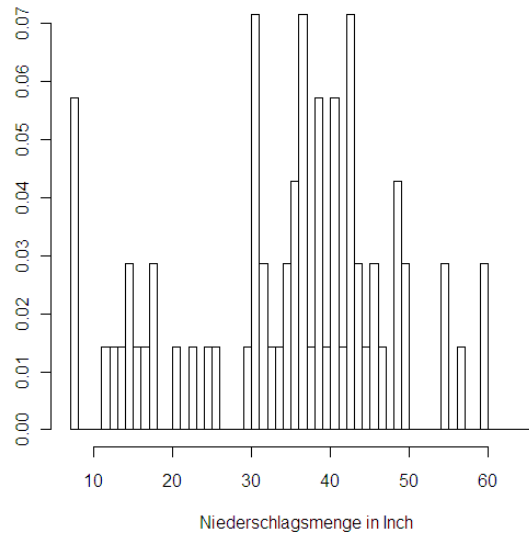
Auch bei Histogrammen hängt die Darstellung von der gewählten Klasseneinteilung ab !

Beispiel.

Die folgenden Histogramme geben den Anteil US-amerikanischer Städte mit verschiedenen Jahresniederschlagsmengen wieder:







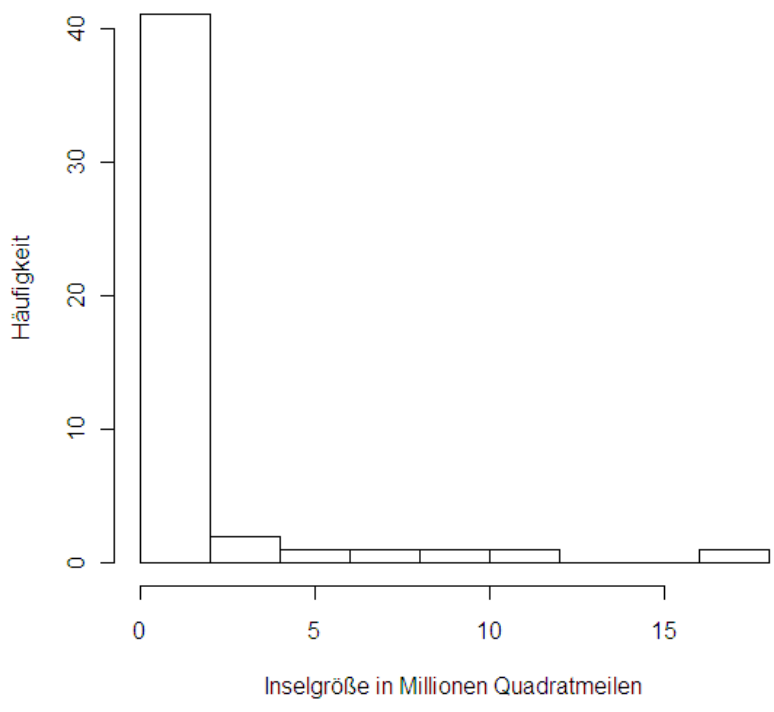
Eine Faustregel zur Festlegung der Klassenanzahl l ist die **Formel von Sturges**:

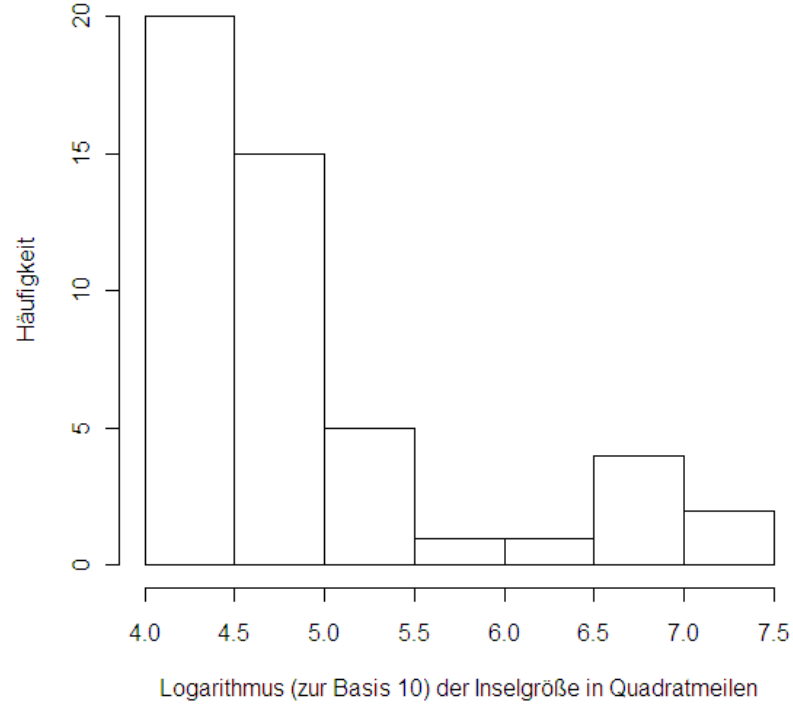
$$l = \lfloor \log_2 n \rfloor + 1$$

Hierbei bezeichnet $\lfloor x \rfloor$ den ganzzahligen Anteil von x, also z.B. $\lfloor 6.98 \rfloor = 6$.

- Im Beispiel wurden 70 Städte betrachtet, daraus ergibt sich $l = 7$. Dies ist gerade die Klassenzahl im mittleren Histogramm.
- Die Formel ist wirklich nur eine *Faustregel* ! Oft ist eine andere Klasseneinteilung sinnvoller.
- Manchmal fällt die relative Häufigkeit bei großen Merkmalsausprägungen sehr rasch, z.B. exponentiell ab. Dann ist es eventuell sinnvoll, die Klassengrößen entsprechend exponentiell anwachsen zu lassen. **VORSICHT - Mögliche Fehlerquelle !**
- *Beispiel im Internet: http://demonstrations.wolfram.com/GroupingCountryData*

- Manchmal ist es auch sinnvoll, anstelle des Histogramms einer Merkmalsausprägung das Histogramm für den Logarithmus der Merkmalsausprägung zu erstellen.
- Die folgenden Histogramme zeigen beispielsweise die Häufigkeit von größeren Inseln auf der Erde in Abhängigkeit von der Größe (Flächeninhalt) der Insel, sowie die Häufigkeitsverteilung der logarithmierten Inselgröße:





WARNUNG: Das zweite Histogramm ist KEIN Histogramm für die Inselgröße selbst, sondern nur ein Histogramm für die logarithmierte Inselgröße. Würden wir auf der x-Achse statt der Logarithmen die Größen selbst auftragen, dann wären die Flächeninhalte nicht proportional zu den Häufigkeiten.

2.2. Empirische Verteilungsfunktion und Quantile

Kumulierte Häufigkeiten

Ordinalskalierte Merkmale können wir in einer natürlichen Reihenfolge anordnen:

$$a_1 < a_2 < \cdots < a_m$$

Wir können dann fragen, wieviele (absolute Häufigkeit) bzw. welcher Anteil (relative Häufigkeit) der beobachteten Daten vor einem bestimmten Merkmal a_k liegt.

- Die k-te *kumulierte absolute Häufigkeit* ist

$$N_k = n_1 + n_2 + \cdots + n_k = \text{Anzahl der } \omega_i \text{ mit } x_i \leq a_k.$$

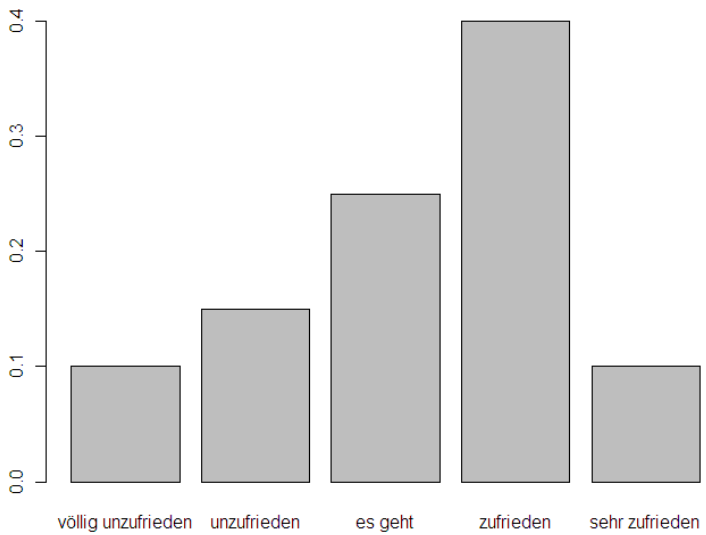
- Die k-te *kumulierte relative Häufigkeit* ist

$$H_k = h_1 + h_2 + \cdots + h_k = \frac{N_k}{n}$$

Empirische Verteilungsfunktion und Quantile

Kumulierte Häufigkeiten - Beispiel

Bei einer (rein fiktiven) Umfrage nach der Zufriedenheit mit den Studienbedingungen an der Uni Bonn ergibt sich folgendes Bild:



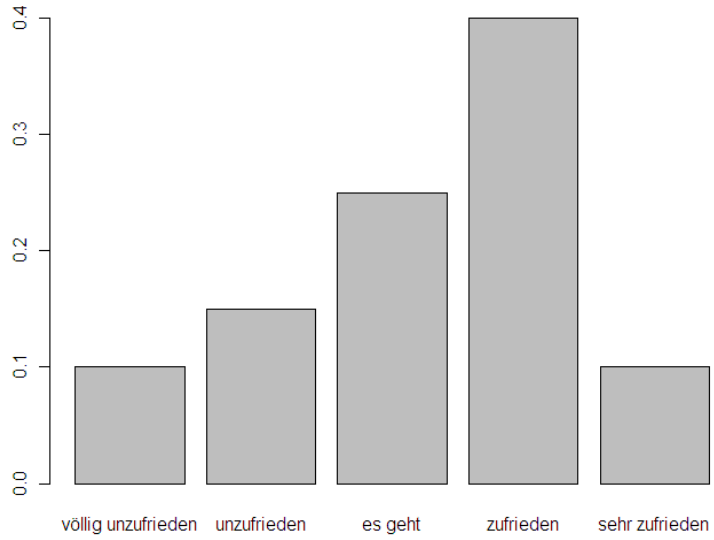
- Welcher Anteil der Studenten ist nicht wirklich zufrieden ?
-

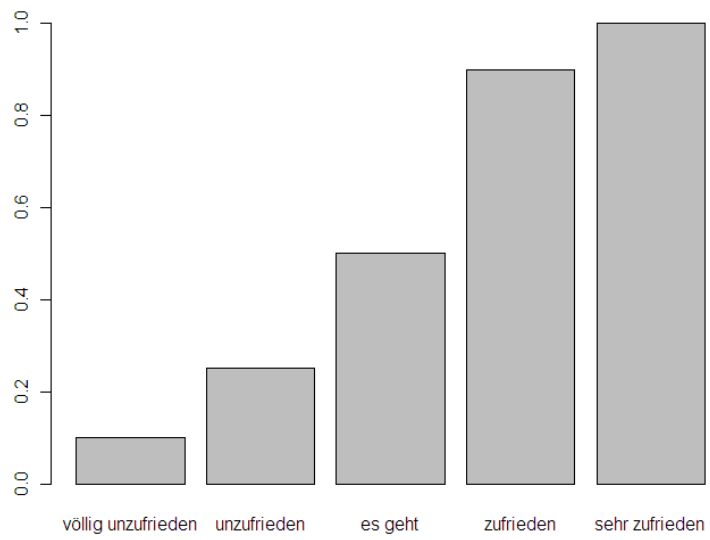
$$\begin{aligned} H_{\text{es geht}} &= 0.1 + 0.15 + 0.25 \\ &= 50\% \end{aligned}$$

- Welcher Anteil ist nicht wirklich unzufrieden ?
-

$$\begin{aligned} & 1 - H_{\text{unzufrieden}} \\ = \; & 1 - (0.1 + 0.15) \\ = \; & 75\% \end{aligned}$$

Die Stabdiagramme der relativen Häufigkeiten h_k und der kumulierten relativen Häufigkeiten H_k:





Empirische Verteilungsfunktion und Quantile

Empirische Verteilungsfunktion (ecdf=empirical cumulative distribution function)

Bei metrisch skalierten Merkmalen können wir allgemeiner nach dem Anteil der Daten fragen, der eine bestimmte reellen Zahl x nicht überschreitet.

Definition

Die empirische Verteilungsfunktion der Daten $x_1, x_2, \ldots, x_n$ ist die durch

$$\begin{aligned} F_n(x) &= \text{relative Häufigkeit des Intervalls } (-\infty, x] \\ &= \frac{\text{Anzahl der } \omega_i \text{ mit } x_i \leq x}{n} \end{aligned}$$

definierte Funktion $F_n : (-\infty, \infty) \to [0, 1]$.

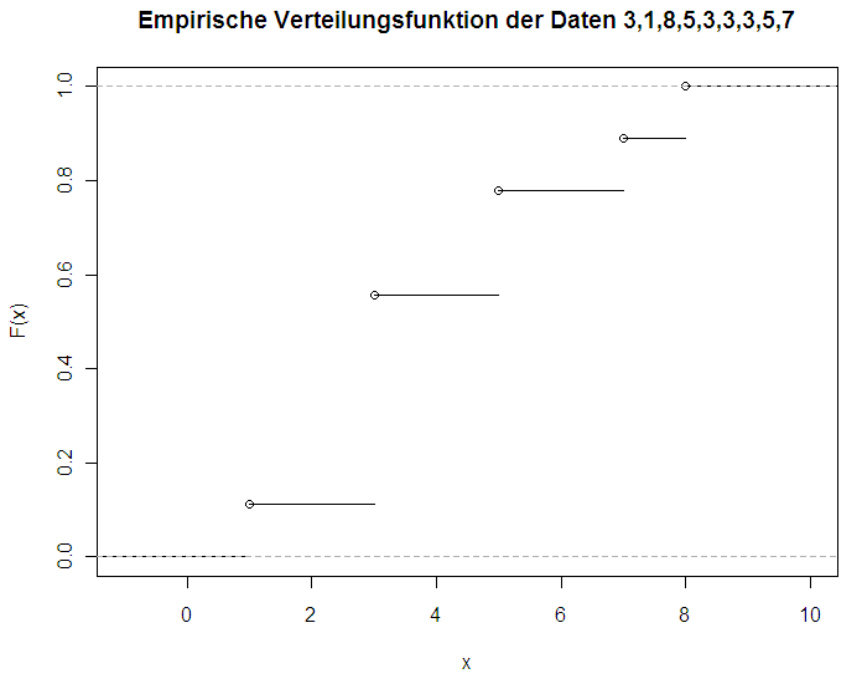
Empirische Verteilungsfunktion und Quantile

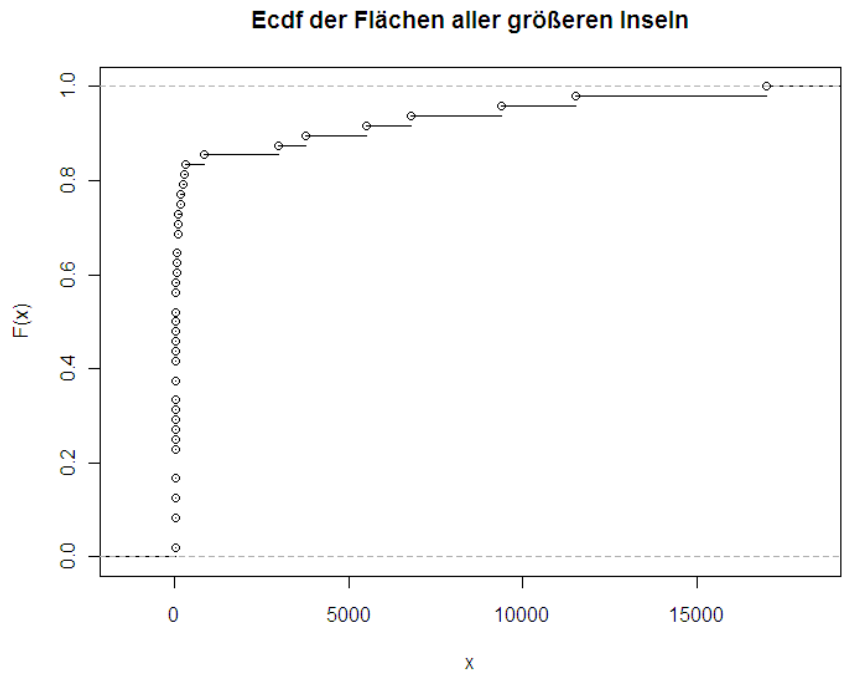
Empirische Verteilungsfunktion

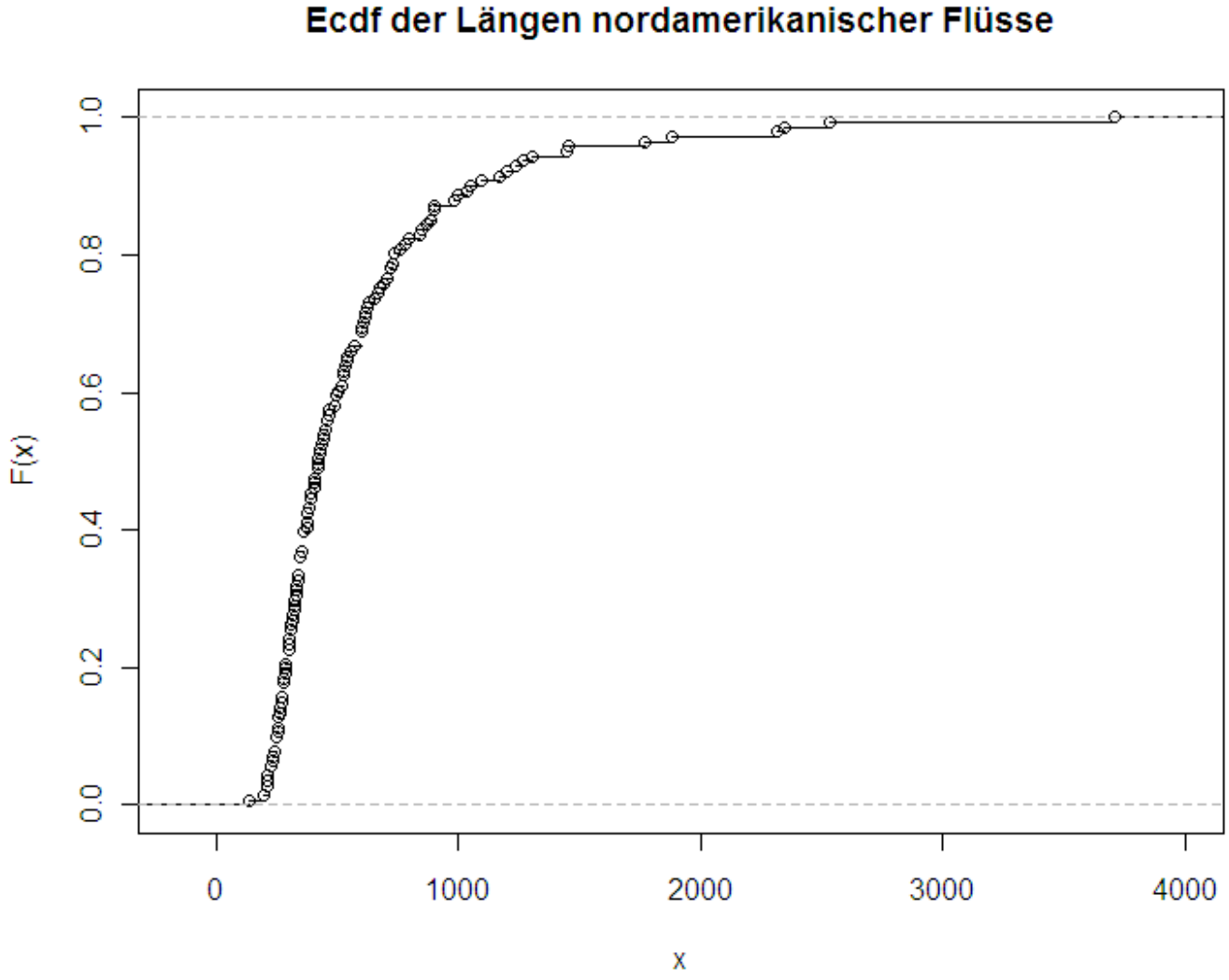
$$F_n(x) = \frac{\text{Anzahl der } \omega_i \text{ mit } x_i \leq x}{n}$$

Die emp. Verteilungsfunktion ist eine Treppenfunktion, die bei jedem Datenwert um $\frac{1}{n} \times$ *Häufigkeit des Datenwerts* nach oben springt.

Beispiel.






Ecdf der Längen nordamerikanischer Flüsse
F(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
1000
2000
3000
4000
x

Empirische Verteilungsfunktion und Quantile

Quantile

Oft sucht man den kleinsten Wert x, der von einem bestimmten Anteil der Beobachtungsdaten nicht überschritten wird. Beispielsweise ist der *Median* ein Wert, der die Daten in zwei gleich große Teile teilt - also von 50 % der Daten nicht überschritten, und von 50 % der Daten nicht unterschritten wird.

- Allgemeiner definiert man für $0 \leq p \leq 1$:

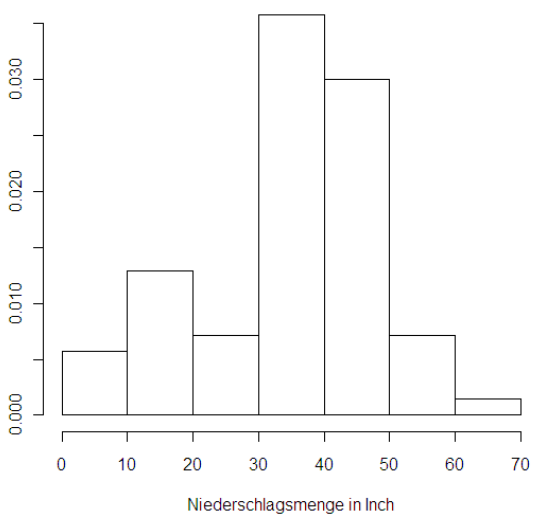
Definition

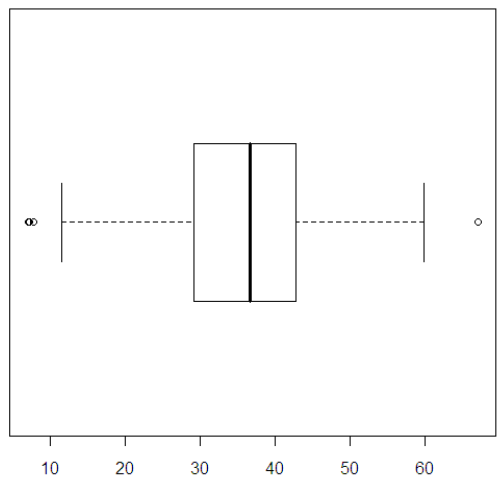
Ein p-Quantil x_p ist ein Wert, der von $100 \cdot p\%$ der Daten nicht überschritten, und von $100 \cdot (1-p)\%$ der Daten nicht unterschritten wird.

- Links von einem p-Quantil liegen also höchstens $100 \cdot p\%$ der Daten, rechts davon höchstens $100 \cdot (1-p)\%$ der Daten.
- Ist das Quantil selbst im Datensatz enthalten, dann kann es sein, daß an beiden Seiten echt weniger als $100 \cdot p\%$ bzw. $100 \cdot (1-p)\%$ der Daten liegen.

- Das mittlere Quantil $x_{0.5}$ heißt **Median**. Die Quantile $x_{0.25}$ und $x_{0.75}$ heißen **unteres** bzw. **oberes Quartil**.
- In einem **Boxplot** trägt man ein Rechteck auf, dessen untere und obere Grenze beim unteren und oberen Quartil liegen, und markiert den Median durch einen Querstrich. **Ausreißer**, d.h. Daten, die um mehr als das 1.5 fache der Kastenbreite von den Quartilen entfernt sind, werden markiert, und das Minimum und Maximum aller übrigen Daten wird durch einen Strich gekennzeichnet.

Beispiel. (Jahresniederschläge in US-amerikanischen Städten)



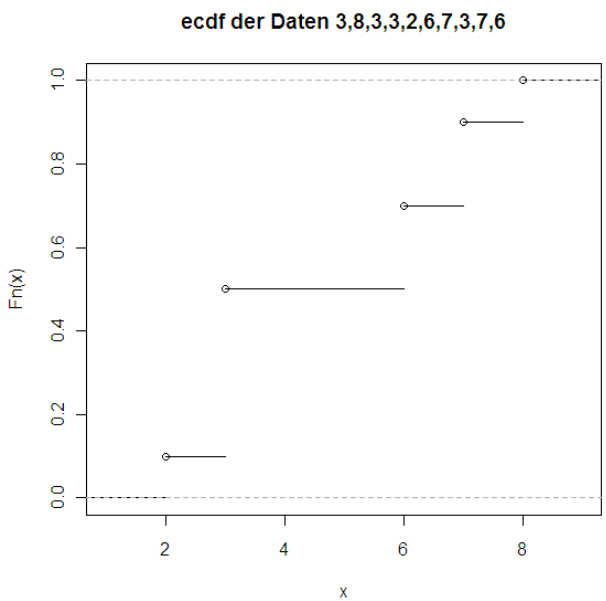


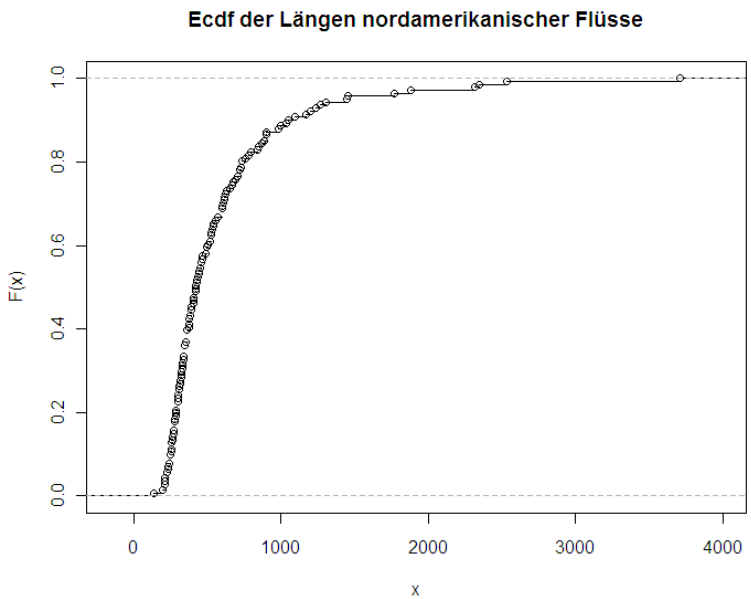
Empirische Verteilungsfunktion und Quantile

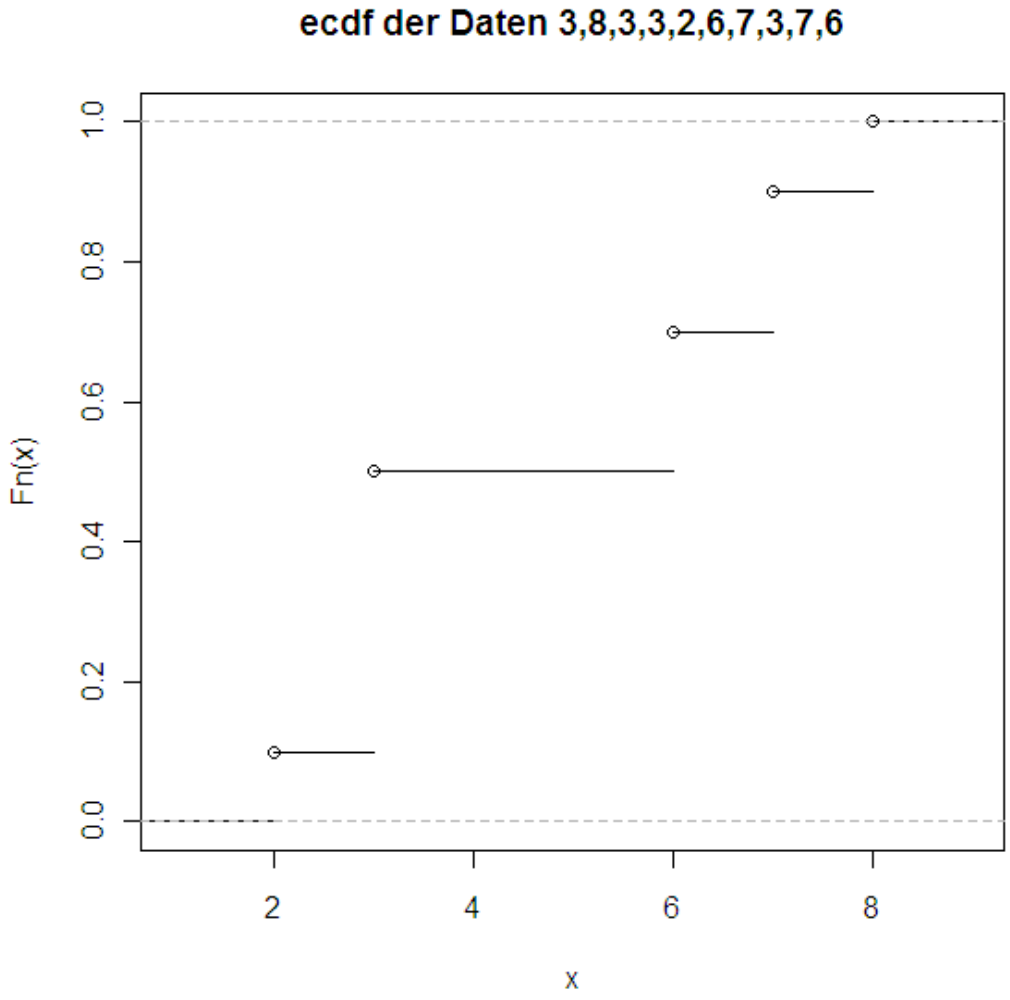
Ein Wert x ist genau dann ein p-Quantil, wenn die empirische Verteilungsfunktion links von x unterhalb von p, und rechts von x oberhalb von p liegt. Also können wir das p-Quantil aus dem Graphen der empirischen Verteilungsfunktion ablesen:

- *Ein p-Quantil x_p ist ein Wert, bei dem die empirische Verteilungsfunktion den Level p überquert.*

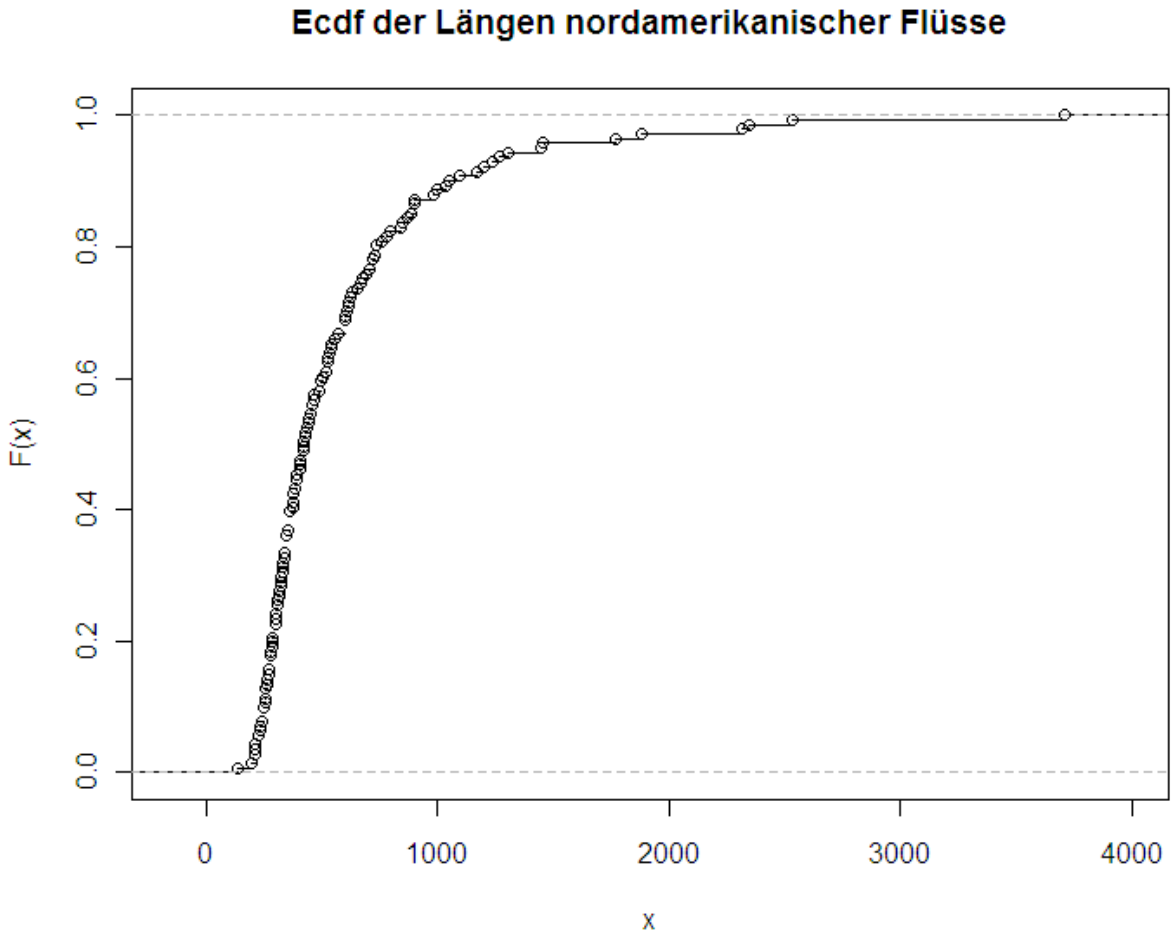
Beispiel. (Bestimmen Sie die Quantile $x_{0.25}$, $x_{0.5}$ und $x_{0.75}$!)







- $x_{0.25} = 3$, $x_{0.75} = 7$. Mediane sind alle Zahlen im Intervall $[3, 6]$!
- **Konvention:** In so einem Fall wird manchmal der mittlere Wert des Intervalls $[3, 6]$, also 4.5, als Median angegeben.



$x_{0.25} = 310$, $x_{0.5} = 425$, $x_{0.75} = 680$.

Empirische Verteilungsfunktion und Quantile

Wie bestimmt man die Quantile ?

Alternativ zum Ablesen aus der empirischen Verteilungsfunktion kann man die Quantile auch folgendermaßen bestimmen:

1. Ordne die Daten der Größe nach: $x_1 \leq x_2 \leq \cdots \leq x_n$
2. Berechne die Anzahl $k = p \cdot n$ von Daten, die mindestens unterhalb des p-Quantils, oder genau beim p-Quantil liegen sollen.

   - Ist k keine ganze Zahl, dann ist

     $$x_{(p)} = x_{\lceil k \rceil}$$

     das (eindeutige) p-Quantil.
     (Wir runden k zu $\lceil k \rceil$ auf. Unterhalb von $x_{\lceil k \rceil}$ oder gleichauf liegen dann $\lceil k \rceil$, also mehr als k Daten, und oberhalb oder gleichauf liegen mehr als $n - k$ Daten).
   - Ist k eine ganze Zahl, dann sind alle Werte, die zwischen x_k und x_{k+1} liegen, p-Quantile.

Beispiel. (Datensatz 3,8,3,3,2,6,7,3,7,6)

- *Ordnen:* $x_1 = 2$, $x_2 = 3$, $x_3 = 3$, $x_4 = 3$, $x_5 = 3$, $x_6 = 6$, $x_7 = 6$, $x_8 = 7$, $x_9 = 7$, $x_{10} = 8$, $n = 10$.
- **Median:** $k = 0.5 \cdot 10 = 5$ *Daten sollten unterhalb bzw. beim Median* $x_{(0.5)}$ *liegen - entsprechend sollten* 5 *Daten oberhalb bzw. beim Median liegen. Dies ist aber immer der Fall, wenn* $x_{(0.5)}$ *zwischen* x_5 *und* x_6 *liegt. Also ist jeder Wert zwischen* $x_5 = 3$ *und* $x_6 = 6$ *ein Median !*
- **Unteres Quartil:** *Mindestens* $k = 0.25 \cdot 10 = 2.5$ *Daten sollten unterhalb bzw. bei* $x_{(0.25)}$ *liegen - entsprechend sollten mindestens* 7.5 *Daten oberhalb bzw. beim Median liegen. Das geht nur für* x_3 *- in diesem Fall liegen 3 Daten unterhalb oder gleichauf, und 8 Daten oberhalb oder gleichauf. Also:*

$$x_{(0.25)} = x_3 = 3$$

- **Oberes Quartil:** $k = 0.75 \cdot 10 = 7.5$, *also* $x_{(0.75)} = x_8 = 7$.

2.3. Mittelwerte

Arithmetisches Mittel

Definition

Das arithmetische Mittel (Durchschnittswert) $\bar{x}$ von metrischen Daten $x_1, \ldots, x_n$ ist

$$\bar{x} = \frac{1}{n}\sum_i x_i = \frac{1}{n}(x_1 + \cdots + x_n)$$

Beispiel. (Monatseinkünfte eines Bauernhofes)

Monat	*Jan*	*Feb*	*März*	*April*	*Mai*	*Juni*	*Juli*	*Aug.*	*Sept.*	....
Einkünfte	*500*	*450*	*520*	*600*	*1 400*	*1 900*	*2 000*	*1 500*	*2 300*	....

Wieviel Euro darf der Besitzer pro Monat ausgeben, ohne sich zu verschulden?

$$\begin{aligned}\bar{x} &= \frac{1}{12}(500 + 450 + 520 + 600 + 1400 + 1900 + 2000 + 1500 + \cdots) \\ &= 1201,66\end{aligned}$$

Mittelwerte

Berechnung des arithmetischen Mittels aus der empirischen Verteilung

Sind n_1, n_2,...,n_m die Häufigkeiten der vorkommenden Merkmalsausprägungen a_1, a_2,...,a_m, und h_1, h_2,...,h_m die entsprechenden relativen Häufigkeiten, dann gilt

$$\overline{x} = \frac{1}{n}\sum_i x_i = \frac{1}{n} \cdot (n_1 a_1 + n_2 a_2 + \cdots + n_m a_m).$$

Theorem

$$\overline{x} = \frac{1}{n}\sum_k n_k a_k = \sum_k h_k a_k$$

- $\overline{x}$ ist also gleich dem **gewichteten arithmetischen Mittel** $\sum h_k a_k$ der beobachteten Merkmalsausprägungen a_k, wobei die Gewichte h_k die relativen Häufigkeiten der Merkmalsausprägungen sind.

Beispiel. (Anzahl herausragender Entdeckungen 1860-1959)

1860	1861	...	1886	1887	1888	1889	1890	...	1959
5	3	...	12	3	10	9	2	...	0

- *In wievielen Jahren gab es 0,1,2, ... herausragende Entdeckungen ?*
-

a_k	0	1	2	3	4	5	6	7	8	9	10
n_k	9	12	26	20	12	7	6	4	1	1	1
h_k	.09	.12	.26	.2	.12	.07	.06	.04	.01	.01	.01

- *Die mittlere Anzahl herausragender Entdeckungen pro Jahr betrug also zwischen 1860 und 1959:*

$$\overline{x} = \frac{9}{100} \cdot 0 + \frac{12}{100} \cdot 1 + \frac{26}{100} \cdot 2 + \frac{20}{100} \cdot 3 + \cdots + \frac{1}{100} \cdot 12 = 3,1$$

- *Dagegen betrug die mittlere Anzahl zwischen 1886 und 1890:*

$$\overline{x} = \frac{1}{5} \cdot (12 + 3 + 10 + 9 + 2) = 7,2$$

Die Formel

$$\overline{x} = \frac{1}{n}\sum_k n_k a_k = \sum_k h_k a_k$$

kann man auch zur näherungsweisen Berechnung des Mittelwerts von Daten, die in Klassen eingeteilt sind, verwenden. Für a_k setzt man dann den Mittelpunkt der k-ten Klasse ein.

Beispiel. (Baumwollerträge im Westen der USA)

Klasse	*Mittelpunkt*	*H'keit*
215–235	*225*	*4*
235–255	*245*	*6*
255–275	*265*	*13*
275-295	*285*	*21*
295-315	*305*	*15*
315–335	*325*	*7*
335–355	*345*	*5*
355–375	*365*	*4*
Summe		*75*

Der mittlere jährliche Flächenertrag pro Farm ist also ungefähr

$$\begin{aligned}\overline{x} &\approx \frac{1}{75}(4\cdot 225 + 6\cdot 245 + \\ &\qquad + \cdots + 4\cdot 365) \\ &\approx 291\end{aligned}$$

Mittelwerte

Vergleich von arithmetischem Mittel und Median

Das arithmetische Mittel und der Median sind **Maßzahlen für die Lage** der empirischen Verteilung. Eine weitere Lagemaßzahl, die auch für nominalskalierte Daten anwendbar ist, ist der *Modus*. Der Modus ist diejenige Merkmalsausprägung, die am häufigsten im Datensatz vorkommt.

- Das arithmetische Mittel gibt den **Schwerpunkt** von metrischen Daten an. Es ist die am meisten verwendete Lagemaßzahl für metrische Daten, und spielt in der schließenden Statistik eine zentrale Rolle.
- Allerdings hat das arithmetische Mittel auch einen Nachteil: Es ist **nicht stabil unter "Ausreißern"**, also Merkmalsausprägungen, die in der Urliste zwar selten vorkommen, aber sehr hohe oder sehr stark negative Werte annehmen!

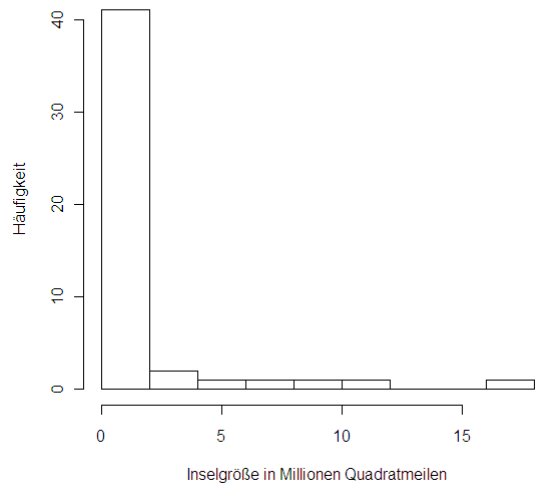
Beispiel.

Wir untersuchen das Durchschnittseinkommen in einer Kleinstadt nahe Seattle. Während wir das tun, zieht Bill Gates in diese Stadt. Das Durchschnittseinkommen steigt sprunghaft. Da wir aber an der sozialen Lage in der Stadt (die sich im Wesentlichen nicht geändert hat) interessiert sind, und weniger an Herrn Gates, vermasselt uns das den Durchschnitt. Reumütig kehren wir zum Median zurück, denn der hat sich ja nur von einem bestimmten Einkommen zum nächsthöheren - also fast gar nicht - verschoben.

- **Median und Modus sind stabil unter Ausreißern, das arithmetische Mittel jedoch nicht !**
- Online-Demonstration zu Mittelwert und Median siehe *http://onlinestatbook.com/stat_sim/descriptive*

Beispiel. (Grössen von Inseln auf der Erde)

- *Der Mittelwert der Größen aller Inseln über 10.000 Quadratmeilen auf der Erde beträgt 1.252.729 Quadratmeilen, der Median hingegen nur 41.000 Quadratmeilen.*
- *Ein Blick auf das Histogramm erklärt, warum der Unterschied so groß ist:*



- *In diesem Fall wäre es (abhängig von der Fragestellung) eventuell sinnvoll, statt des arithmetischen das geometrische Mittel zu betrachten - dieses beträgt nur 85.314.*

Mittelwerte

Geometrisches Mittel

Definition

Das **geometrische Mittel** von positiven Zahlen $x_1, x_2, \ldots, x_n$ ist

$$\overline{x}_g = \sqrt[n]{x_1 x_2 \cdots x_n}$$

- Der Logarithmus des geometrischen Mittels ist das arithmetische Mittel der logarithmierten Daten:

$$\log \overline{x}_g = \log\left((x_1 x_2 \cdots x_n)^{1/n}\right) = \frac{1}{n}\sum_i \log x_i$$

- Das Wechseln vom arithmetischen zum geometrischen Mittel entspricht also dem Übergehen zu einer **logarithmischen Skala** !
  (Wir hatten schon gesehen, daß dies bei Inselgrößen sinnvoll sein kann)
- Allgemein können das geometrische Mittel bzw. eine logarithmische Skala sinnvoll sein, wenn es auf die Verhältnisse der Daten zueinander ankommt, z.B. Populationswachstum, Verzinsung,....

Mittelwerte

Lineare Skalentransformationen

Sie machen eine Präsentation, bei der Sie den durchschnittlich erzielten Preis $\bar{x}$ für Quitscheentchen ihrem Chef vorstellen. Die Produktion und der Vertrieb jedes Quietscheentchens sei immer gleich teuer und koste einen Preis P. Der Chef sagt, ihn interessiere nur der durchschnittliche Gewinn/Quietscheentchen. Wenn x_i die erzielten Preise sind, und n die Zahl der verkauften Entchen, dann ist der Gewinn für ein bestimmtes Entchen also gerade $g_i = x_i - P$. Jetzt schnell aufsummieren und teilen? Nein! Sie antworten $\bar{x} - P$! Denn:

$$\bar{g} = \frac{1}{n}\sum_i g_i = \frac{1}{n}\sum_i (x_i - P) = \frac{1}{n}\left(\sum_i x_i - nP\right) = \frac{1}{n}\sum_i x_i - \frac{1}{n}nP = \bar{x} - P$$

Mittelwerte

Lineare Skalentransformationen

Als nächstes will Ihr Chef den Gewinn in Dollar wissen, Ihr Preis war aber in Euro. Zufällig kennen Sie den Umrechnungskurs Euro nach Dollar, nämlich K=1.4. Der Gewinn für ein bestimmtes Entchen ist also in Dollar $g_i^{Dollar} = Kg_i$. Jetzt schnell multiplizieren, aufsummieren und durch die Anzahl teilen? Nö, Sie sagen $\bar{g}^{Dollar} = K\bar{g}$! Denn:

$$\bar{g}^{Dollar} = \frac{1}{n}\sum_i Kg_i = K\frac{1}{n}\sum_i g_i = K\bar{g}$$

- Bei **linearen** Skalentransformationen, also Verschiebungen des Nullpunkts und Umrechnung in eine andere Einheit, transformiert sich das arithmetische Mittel **linear**, d.h.

$$z_i = Kx_i + a \Rightarrow \bar{z} = K\bar{x} + a$$

- Für Modus und Median gilt dasselbe! Für den Median haben wir z.B.

$$z_i = Kx_i + a \Rightarrow \tilde{z}_{0.5} = K\tilde{x}_{0.5} + a$$

WARNUNG: Der Median und der Modus sind auch stabil unter monotonen **nichtlinearen** Skalentransformationen, **dies gilt aber nicht für das arithmetische Mittel** ! Wenn Sie die Koordianten nichtlinear tranformieren, müssen Sie das arithmetische Mittel also ganz neu berechnen.

Beispiel.

Das arithmetische Mittel der Logarithmen $\log x_i$ *ist nicht gleich dem Logarithmus des arithmetischen Mittels* $\overline{x}$ *(sondern gleich dem Logarithmus des geometrischen Mittels* $\overline{x}_g$*, siehe oben).*

$$\overline{\log x} \neq \log \overline{x}$$

2.4. Streuungsmaße

Spannweite und MAD

Streuungsmaße sind Statistiken, die messen, ob die Daten "eher nah um einen Mittelpunkt herumliegen" oder eher "weit davon abweichen". Als Mittelpunkte dienen hier die Lagemaße Modus, Median und arithmetisches Mittel.

- Die **Spannweite** w von metrischen Daten ist der Abstand zwischen dem kleinsten und größten Beobachtungswert:

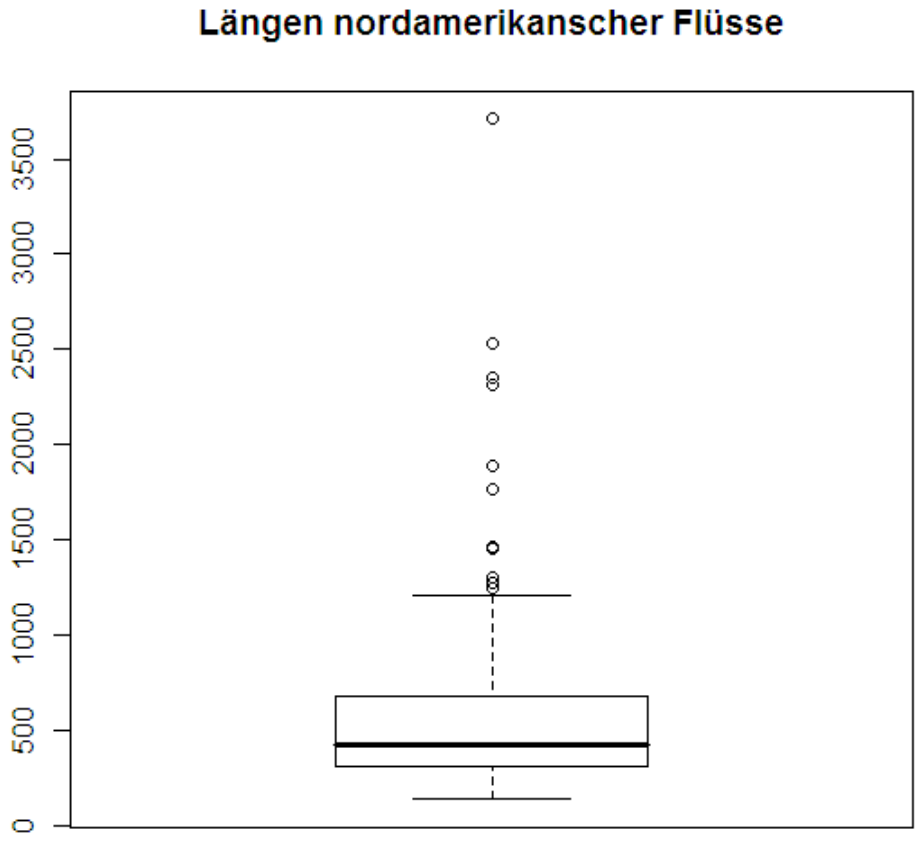
$$w = \max x_i - \min x_i$$

- Die Streuung der Daten um den Median kann man durch Angabe des **unteren** und **oberen Quartils** quantifizieren. Diese Streuungsmaße werden im **Boxplot** grafisch dargestellt.
- Alternativ betrachtet man den **Median der absoluten Abweichung vom Median** (MAD=median absolute deviation):

$$MAD = \text{med} \left\{ \left| x_i - x_{(0.5)} \right| \right\}$$

Streuungsmaße

Ein Boxplot



Dargestellt sind Median, unteres und oberes Quartil, Ausreißer, sowie die von Ausreißern bereinigte Spannweite.

Streuungsmaße

Empirische Varianz

Definition. Die **(empirische) Varianz** der metrischen Daten $x_1, \ldots, x_n$ ist

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

- Die empirische Varianz ist also die **mittlere quadratische Abweichung** der Beobachtungswerte vom arithmetischen Mittel. Die positive Quadratwurzel σ aus der Varianz heißt **Standardabweichung**.
- Für Stichproben aus einer Grundgesamtheit betrachtet man auch häufig die renormierte Stichprobenvarianz

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

- Den Grund für den Faktor $\frac{1}{n-1}$ anstelle von $\frac{1}{n}$ werden wir später sehen.

Berechnungsformel für die Varianz:

$$\sigma^2 = \overline{x^2} - \bar{x}^2$$

Beweis:

$$\begin{aligned}\sigma^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2 &= \frac{1}{n}\sum_i (x_i^2 - 2\bar{x}x_i + \bar{x}^2) \\ &= \frac{1}{n}\sum_i x_i^2 - 2\bar{x}\bar{x} + \frac{n}{n}\bar{x}^2 = \overline{x^2} - \bar{x}^2\end{aligned}$$

Beispiel. (Kinderzahl von Familien)

- *Zahl der Kinder:* 3,0,2,2,1,3,1,1
- *Arithmetisches Mittel:* $\bar{x} = \frac{13}{8} = 1.625$
- *Varianz:* $\overline{x^2} = \frac{29}{8} = 3.625$

$$\Longrightarrow \sigma^2 = \overline{x^2} - \bar{x}^2 = 3.625 - (1.625)^2 = 0.984$$

- *Standardabweichung:* $\sigma = \sqrt{0.984} = 0.992$

Streuungsmaße

Berechnung der Varianz aus der empirischen Verteilung

Theorem

Sind n_1, n_2,...,n_m die Häufigkeiten der vorkommenden Merkmalsausprägungen a_1, a_2,...,a_m, und h_1, h_2,...,h_m die entsprechenden relativen Häufigkeiten, dann gilt

$$\sigma^2 = \frac{1}{n}\sum_k n_k(a_k - \bar{x})^2 = \sum_k h_k(a_k - \bar{x})^2$$

- σ^2 ist also gleich dem **gewichteten arithmetischen Mittel** $\sum h_k(a_k - \overline{x})^2$ der quadratischen Abweichungen der beobachteten Merkmalsausprägungen a_k von $\overline{x}$, wobei die Gewichte h_k die relativen Häufigkeiten der Merkmalsausprägungen sind.
- Die Formel kann auch zur näherungsweisen Berechnung der Varianz von Daten, die in Klassen eingeteilt sind, verwendet werden. a_k ist dann der Klassenmittelpunkt der k-ten Klasse.

Beispiel. (Baumwollerträge im Westen der USA)

Klasse	*Mittelpunkt*	*H'keit*
215–235	*225*	*4*
235–255	*245*	*6*
255–275	*265*	*13*
275-295	*285*	*21*
295-315	*305*	*15*
315–335	*325*	*7*
335–355	*345*	*5*
355–375	*365*	*4*
Summe		*75*

$$\begin{aligned}\overline{x} &\approx \frac{1}{75}(4\cdot 225 + 6\cdot 245 + \\ &\qquad + \cdots + 4\cdot 365) \\ &\approx 291\end{aligned}$$

$$\begin{aligned}\sigma^2 &\approx \frac{1}{75}(4\cdot(225-291)^2 \\ &\qquad +6\cdot(245-291)^2 \\ &\qquad +\cdots+4\cdot(365-291)^2) \\ &\approx 1162\end{aligned}$$

$$\sigma \approx \sqrt{1162} \approx 34$$

Streuungsmaße

MSE, Arithmetisches Mittel als bester Prognosewert

Angenommen, wir wollen die Beobachtungswerte einer Meßreihe möglichst gut prognostizieren. Der **mittlere quadratische Fehler (MSE=mean square error)** eines Prognosewerts c ist

$$MSE = \frac{1}{n}\sum_i (x_i - c)^2$$

Die Varianz ist die mittlere quadratische Abweichung von $\overline{x}$.

Theorem

$$MSE = \sigma^2 + (c - \overline{x})^2$$

Insbesondere ist der mittlere quadratische Prognosefehler minimal, wenn wir $c = \overline{x}$ wählen. Das arithmetische Mittel ist also der **beste Prognosewert** *bzgl. des mittleren quadratischen Fehlers.*

- *Beweis:*

$$\begin{aligned} MSE &= \frac{1}{n}\sum_i (x_i - c)^2 = \frac{1}{n}\sum_i (x_i - \overline{x} + \overline{x} - c)^2 \\ &= \frac{1}{n}\sum_i (x_i - \overline{x})^2 + \frac{2}{n}\sum_i (x_i - \overline{x})(\overline{x} - c) + \frac{1}{n}\sum_i (\overline{x} - c)^2 \\ &= \sigma^2 + 0 \cdot (\overline{x} - c) + \frac{n}{n} \cdot (\overline{x} - c)^2 = \sigma^2 + (\overline{x} - c)^2 \end{aligned}$$

- *Bemerkung:* Der Median ist der beste Prognosewert bzgl. des mittleren absoluten Fehlers

$$MAE = \frac{1}{n}\sum_i |x_i - c|$$

Streuungsmaße

Variationskoeffizient, relativer Prognosefehler

Oft interessiert, wie stark die Daten im Verhältnis zum Prognosewert streuen.

Definition

1. Der **Variationskoeffizient** v von positiven metrischen Daten x_i ist das Verhältnis

$$v = \sigma / \overline{x}$$

   von Standardabweichung und Mittelwert.

2. Entsprechend definiert man den **relativen Fehler** eines Prognosewerts c als

$$\varepsilon = \sqrt{MSE} / \overline{x}$$

Beispiel.

1. *Baumwollerträge:* $v = \sigma / \overline{x} = 34/291 = 11.7\%$
   Die beobachteten Werte streuen weniger stark um den Mittelwert.

2. *Kinderzahl:* $v = \sigma / \overline{x} = 0.992/1.625 = 61.0\,\%$
   Die beobachteten Werte streuen mäßig stark.
3. *Flusslängen:* $v = \sigma / \overline{x} = 493.8/591.2 = 83.5\,\%$
   Die beobachteten Werte streuen stark.

Streuungsmaße

Entropie

Randnotiz: Ein Maß für die Streuung nominalskalierter Daten ist die **Entropie**

$$H = -\sum_k h_k \log h_k$$

- Die Entropie beschreibt den mittleren Informationsgehalt der Daten.
- Tritt nur eine einzige Merkmalsausprägung auf, dann ist die Entropie minimal - nämlich gleich 0.
- Tritt jede mögliche Merkmalsausprägung gleich oft auf, dann ist die Entropie maximal - nämlich gleich $\log m$, wobei m die Anzahl der Merkmalsausprägungen ist.

Übersicht

Lage- und Streumaße

	Lagemaße	**Streumaße**
Nominalskala	Modus	Entropie
Ordinalskala	Median	Quartile, MAD
Metrische Skala	arithmetisches Mittel	Varianz, Standardabweichung
Verhältnisskala	geometrisches Mittel	