Kapitel 7
Einführung in die beschreibende Statistik

Lernzielangaben Kapitel 7

- Das Aufstellen von Häufigkeitstabellen sollte verstanden sein.
- Wichtige Lage- und Streuemaße sollten verinnerlicht sein.
- Grundlegende grafische Darstellungen von Datensätzen sind bekannt.
- Maße zur Beschreibung bivariater linearer Abhängigkeiten können angewendet werden.
- Grundzüge der linearen Regression sind bekannt.

Kapitel 7.1
Grundbegriffe und Häufigkeitstabellen

Statistische Grundbegriffe

- **Statistische Einheit**: Objekte, an denen interessierende Größen erfasst werden.
- **Grundgesamtheit**: Menge aller für die Fragestellung relevanten statistischen Einheiten.
- **Teilgesamtheit**: Teilmenge der Grundgesamtheit.
- **Stichprobe**: Tatsächlich untersuchte Teilmenge der Grundgesamtheit bzw. Teilgesamtheit.
- **Merkmal**: Interessierende Größe bzw. Variable.
- **Merkmalsausprägung**: Konkreter Wert des Merkmals für eine bestimmte statistische Einheit.

Stetige und diskrete Merkmale

Die Merkmale, welche von den statistischen Einheiten erhoben werden sollen, werden in zwei verschiedene **Merkmalsarten** eingeteilt.

Kann ein Merkmal endlich viele oder abzählbar unendlich viele Ausprägungen annehmen, so wird dieses Merkmal als **diskret** bezeichnet.

Ist dagegen die mögliche Anzahl der Merkmalsausprägungen überabzählbar unendlich, so wird dieses Merkmal als **stetig** bezeichnet.

Skalenniveaus

- **Nominalskala**: Ein Merkmal ist nominalskaliert, wenn lediglich die Gleichheit oder Andersartigkeit verschiedener Ausprägungen festgestellt werden kann.
- **Ordinalskala**: Ein Merkmal ist ordinalskaliert, wenn die möglichen Merkmalsausprägungen unterscheidbar sind und zusätzlich in eine natürliche oder sinnvoll festzulegende Rangordnung gebracht werden können.
- **Intervallskala**: Ein Merkmal ist intervallskaliert, wenn zusätzlich zur Rangfolge die Differenz zweier beliebiger Merkmalsausprägungen sinnvoll interpretiert werden kann.
- **Verhältnisskala**: Ein Merkmal ist verhältnisskaliert, wenn zusätzlich zur Differenz auch der Quotient zweier beliebiger Merkmalsausprägungen sinnvoll interpretiert werden kann.

Absolute und relative Häufigkeiten (1)

- Um über die vorliegenden (bzw. gewonnenen) Daten einen besseren Überblick zu erhalten, werden gleichwertige Merkmalsausprägungen oftmals zusammengefasst.
- Die einzelnen Merkmalsausprägungen seien mit x_i bezeichnet.
- Der Index i „durchläuft“ die Werte $1, \ldots, n$, sodass insgesamt n (nicht notwendigerweise verschiedene) Merkmalsausprägungen existieren.
- Dabei werden für die einzelnen Merkmalsausprägungen sogenannte **absolute Häufigkeiten** (n_k) bzw. **relative Häufigkeiten** ($h_k = n_k/n$) berechnet ($k = 1, \ldots, K$).
- Es gilt $\sum_{k=1}^{K} n_k = n$ bzw. $\sum_{k=1}^{K} h_k = 1$.
- Für die einzelnen absoluten Häufigkeiten n_k gilt: $n_k \in \{0, 1, 2, \ldots, n\}$.
- Für die einzelnen relativen Häufigkeiten h_k gilt: $h_k \in [0, 1]$.

Absolute und relative Häufigkeiten (2)

- An einer Statistikklausur haben $n = 27$ Studierende teilgenommen. Die einzelnen Noten $(x_i,\ i = 1, \ldots, 27)$ waren:

$$
\begin{aligned}
&x_1 = 2.3,\ x_2 = 1.3,\ x_3 = 3.0,\ x_4 = 5.0,\ x_5 = 2.3,\ x_6 = 4.0,\\
&x_7 = 5.0,\ x_8 = 2.0,\ x_9 = 4.0,\ x_{10} = 3.3,\ x_{11} = 5.0,\ x_{12} = 5.0,\\
&x_{13} = 3.0,\ x_{14} = 1.3,\ x_{15} = 5.0,\ x_{16} = 3.7,\ x_{17} = 3.3,\ x_{18} = 2.3,\\
&x_{19} = 3.3,\ x_{20} = 4.0,\ x_{21} = 5.0,\ x_{22} = 3.3,\ x_{23} = 5.0,\ x_{24} = 5.0,\\
&x_{25} = 4.0,\ x_{26} = 5.0,\ x_{27} = 1.7.
\end{aligned}
$$

- Diese Noten stellen dabei lediglich „Kodierungen“ für „sehr gut“, „gut“, „befriedigend“, usw. dar.
- Berechnet man für diese Liste die absoluten und relativen Häufigkeiten für jede vergebene Note, so erhält man folgende Tabelle:

Absolute und relative Häufigkeiten (3)

k	x_k	n_k	h_k
1	etwas schlechter als sehr gut (1.3)	2	0.0741
2	etwas besser als gut (1.7)	1	0.0370
3	gut (2.0)	1	0.0370
4	etwas schlechter als gut (2.3)	3	0.1111
5	befriedigend (3.0)	2	0.0741
6	etwas schlechter als befriedigend (3.3)	4	0.1481
7	etwas besser als ausreichend (3.7)	1	0.0370
8	ausreichend (4.0)	4	0.1481
9	ungenügend (5.0)	9	0.3333
		$\sum_{k=1}^{K} n_k = 27$	$\sum_{k=1}^{K} h_k = 1$

Kapitel 7.2
Grafische Darstellungen univariater Datensätze

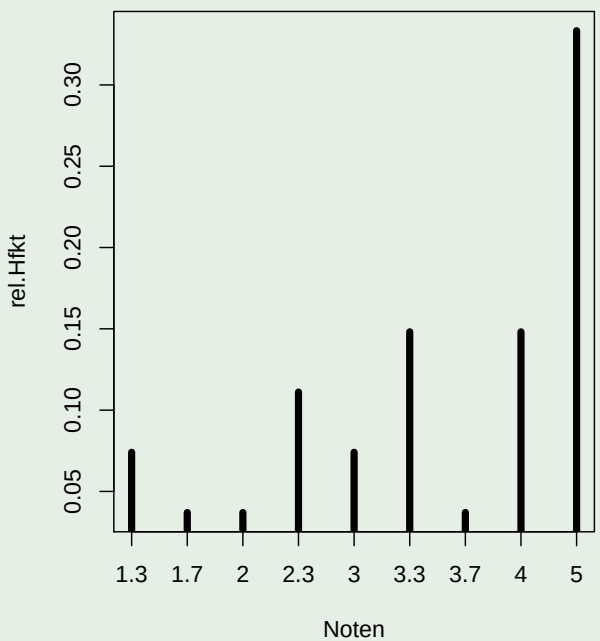
Stab- und Kreisdiagramm (1)

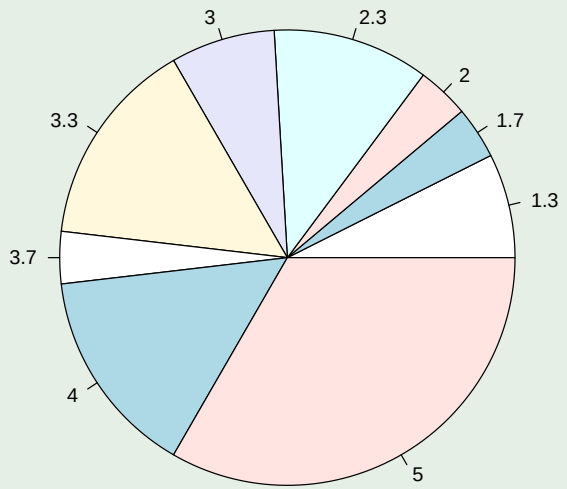
Um univariate Datensätze diskreter Merkmale grafisch darzustellen gibt es verschiedene Möglichkeiten. Die bekanntesten sind sicherlich **Stabdiagramme** und **Kreisdiagramme**.

- **Stabdiagramm**: Für K verschiedenwertige Merkmalsausprägungen werden die Tupel (x_k, n_k) bzw. (x_k, h_k) $(k = 1, \ldots, K)$ in ein Koordinatensystem eingetragen. Anschließend wird von jedem Koordinatenpunkt das Lot zur Abszisse eingezeichnet.
- **Kreisdiagramm**: Die jeweiligen Winkel bzw. einzelnen Flächen der K Kreissektoren werden proportional zu den relativen Häufigkeiten h_k $(k = 1, \ldots, K)$ gesetzt. D.h. für den Winkel eines Kreissegments erhält man: $\alpha_k = h_k \cdot 360°$.

Stab- und Kreisdiagramm (2)

Für die Noten der Statistikklausur erhält man ein Stab- bzw. Kreisdiagramm wie folgt:





Stab- und Kreisdiagramm (3)

Das Stabdiagramm kann hierbei sofort mit den bereits berechneten Werten gezeichnet werden. Für die einzelnen Winkel des Kreisdiagramms erhält man:

k	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
n_k	2	1	1	3	2	4	1	4	9
h_k	0.0741	0.0370	0.0370	0.1111	0.0741	0.1481	0.0370	0.1481	0.3333
x_k	1.3	1.7	2.0	2.3	3.0	3.3	3.7	4.0	5.0
α_k	26.7°	13.3°	13.3°	40.0°	26.7°	53.3°	13.3°	53.3°	120.0°

Kumulierte Häufigkeiten, empirische Verteilungsfunktion (1)

- Im Hinblick auf verschiedene Fragestellungen interessiert man sich oftmals für die Anzahl oder den Anteil der Beobachtungen, die einen bestimmten vorgegebenen Wert nicht überschreiten.
- Die kumulierten Häufigkeiten werden mit $H(x_k)$ $(k = 1, \ldots, K)$ bezeichnet.
- Die kumulierten relativen Häufigkeiten (empirische Verteilungsfunktionswerte) werden mit $\widehat{F}(x_k)$ $(k = 1, \ldots, K)$ bezeichnet.

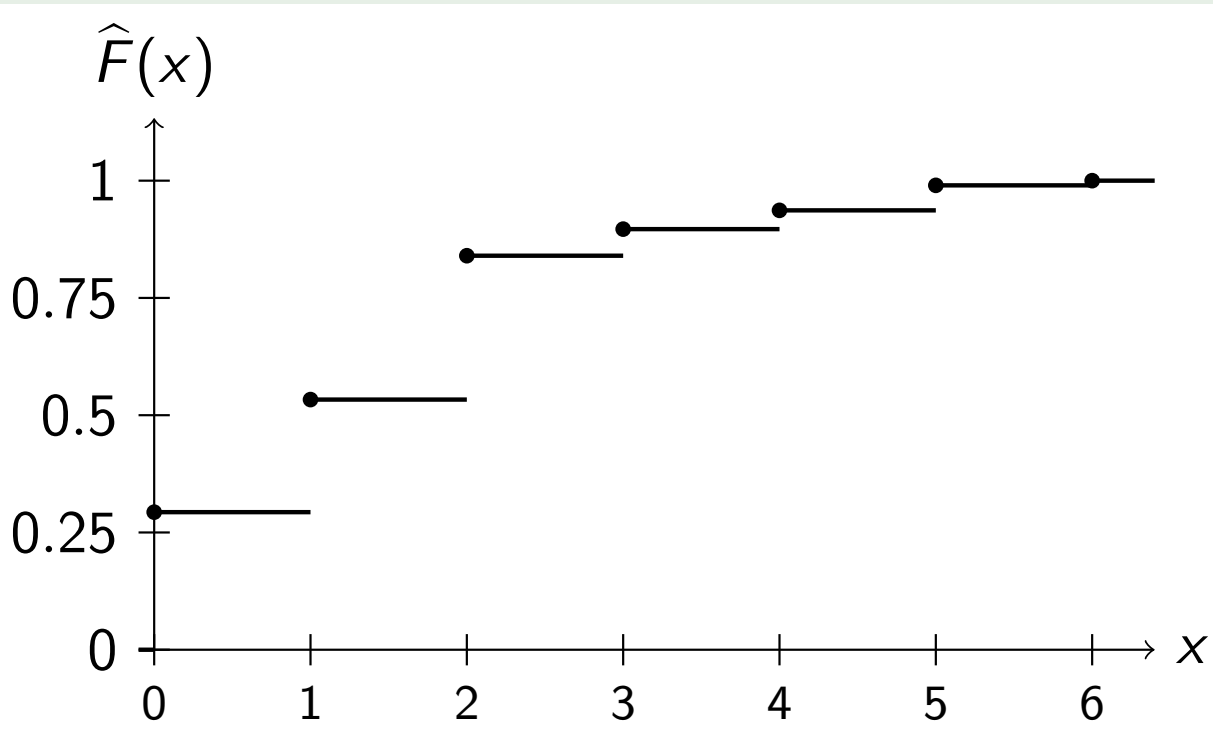
Kumulierte Häufigkeiten, empirische Verteilungsfunktion (2)

In einem Unternehmen wurde an $n = 300$ Arbeitstagen die jeweilige Anzahl defekt produzierter Stücke gezählt. Dabei ergab sich die erste und zweite Spalte der untenstehenden Tabelle mit x_k als Anzahl defekter Stücke pro Tag und n_k als Anzahl Tage ($k = 1, \ldots, 7$).

k	x_k	n_k	h_k	$H(x_k)$	$\widehat{F}(x_k)$
1	0	88	0.2933	88	0.2933
2	1	72	0.2400	160	0.5333
3	2	92	0.3067	252	0.8400
4	3	17	0.0567	269	0.8967
5	4	12	0.0400	281	0.9367
6	5	16	0.0533	297	0.9900
7	6	3	0.0100	300	1

Kumulierte Häufigkeiten, empirische Verteilungsfunktion (3)

Die empirische Verteilungsfunktion für die Anzahl defekter Stücke pro Tag ergibt sich wie folgt:



Kapitel 7.3
Lage- und Streuemaße

Maßzahlen zur Beschreibung von Datensätzen

- Durch die grafische Darstellung univariater Datensätze lassen sich einige charakteristische Eigenschaften von Häufigkeitsverteilungen erkennen.
- Es treten jedoch unmittelbar weitere Fragen auf, welche nicht aus den einzelnen Grafiken abgelesen werden können. Dies sind z.B. Fragen folgender Art:
  - Wo liegt das **Zentrum** der Daten?
  - Wie stark **streuen** die Daten um das Zentrum?
  - Liegen die Daten **symmetrisch** um das Zentrum oder liegt eine gewisse **Schiefe** vor?
  - Gibt es **Ausreißer** in den Daten?
- Um diese Fragestellungen beantworten zu können, werden verschiedene Maßzahlen, wie z.B. Maßzahlen zur **Lage** und zur **Streuung**, benötigt.

Arithmetisches Mittel und Varianz

- Das **arithmetische Mittel** $\bar{x}$ beschreibt die mittlere **Lage** eines Datensatzes:

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

- Die **empirische Varianz** s^2 beschreibt die Streuung um das arithmetische Mittel eines Datensatzes:

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

- Die (positive) Wurzel aus der Varianz s^2 wird **Standardabweichung** genannt und mit s bezeichnet.

Der Modus

Der **Modus** ist ein recht einfaches **Lagemaß**, das angibt, welche Merkmalsausprägungen am häufigsten im Datensatz vorhanden sind.

Für gegebene Ausprägungen ist der Modus wie folgt definiert:

$$\bar{x}^{\text{Mod}} := \{x_j \mid h_j \geq h_i,\ i, j = 1, \ldots, n\}$$

Der Modus wird auch oft als **Modalwert** bezeichnet, muss für gegebene Merkmalsausprägungen x_i $(i = 1, \ldots, n)$ **nicht notwendigerweise eindeutig** sein und kommt in jedem Datensatz vor.

Der Median

- Im Gegensatz zum arithmetischen Mittel reagiert der **Median** nicht sensibel auf Ausreißer in Datensätzen.
- Der Median teilt einen Datensatz derart auf, dass gleich viele Merkmalsausprägungen kleiner bzw. größer als er selbst sind.

Für n gegebene Merkmalsausprägungen x_i $(i = 1, \ldots, n)$ ist der Median:

$$\bar{x}^{\text{Med}} := \begin{cases} x_{\left[\frac{n+1}{2}\right]}\,, & \text{für } n \text{ ungerade}\,, \\ \frac{1}{2}\left(x_{\left[\frac{n}{2}\right]} + x_{\left[\frac{n}{2}+1\right]}\right)\,, & \text{für } n \text{ gerade}\,. \end{cases}$$

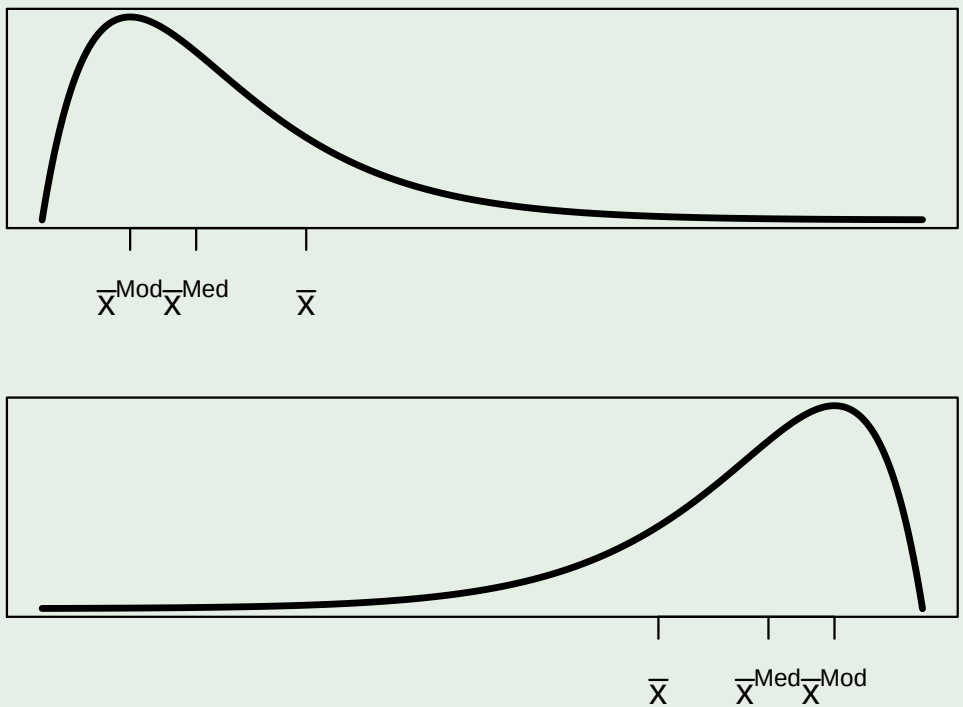
Vergleich der Lagemaße (1)

Für metrisch skalierte Merkmale kann sowohl das arithmetische Mittel, der Modus als auch der Median bestimmt werden, um den gegebenen Datensatz auf **Symmetrie** bzw. **Schiefe** zu untersuchen.

- Symmetrische Verteilung: $\bar{x} \approx \bar{x}^{\text{Med}} \approx \bar{x}^{\text{Mod}}$
- Linkssteile (rechtsschiefe) Verteilung: $\bar{x} > \bar{x}^{\text{Med}} > \bar{x}^{\text{Mod}}$
- Rechtssteile (linksschiefe) Verteilung: $\bar{x} < \bar{x}^{\text{Med}} < \bar{x}^{\text{Mod}}$

Vergleich der Lagemaße (2)

In der folgenden Abbildung ist eine linkssteile (rechtsschiefe) und eine rechtssteile (linksschiefe) Verteilung mit entsprechender Kennzeichung des arithmetischen Mittels, des Modus und des Median gegeben.



Quantile (1)

- Neben den vorgestellten Lage- und Streuemaßen bilden **Quantile** ein vergleichbar mächtiges Instrument, um Datensätze in einigen wenigen Kennzahlen/Maßzahlen zu beschreiben.
- Der Wert des ***p*-Quantils** gibt dabei an, dass mindestens $p \cdot 100\%$ der Merkmalsausprägungen kleiner oder gleich dem p-Quantil sind.

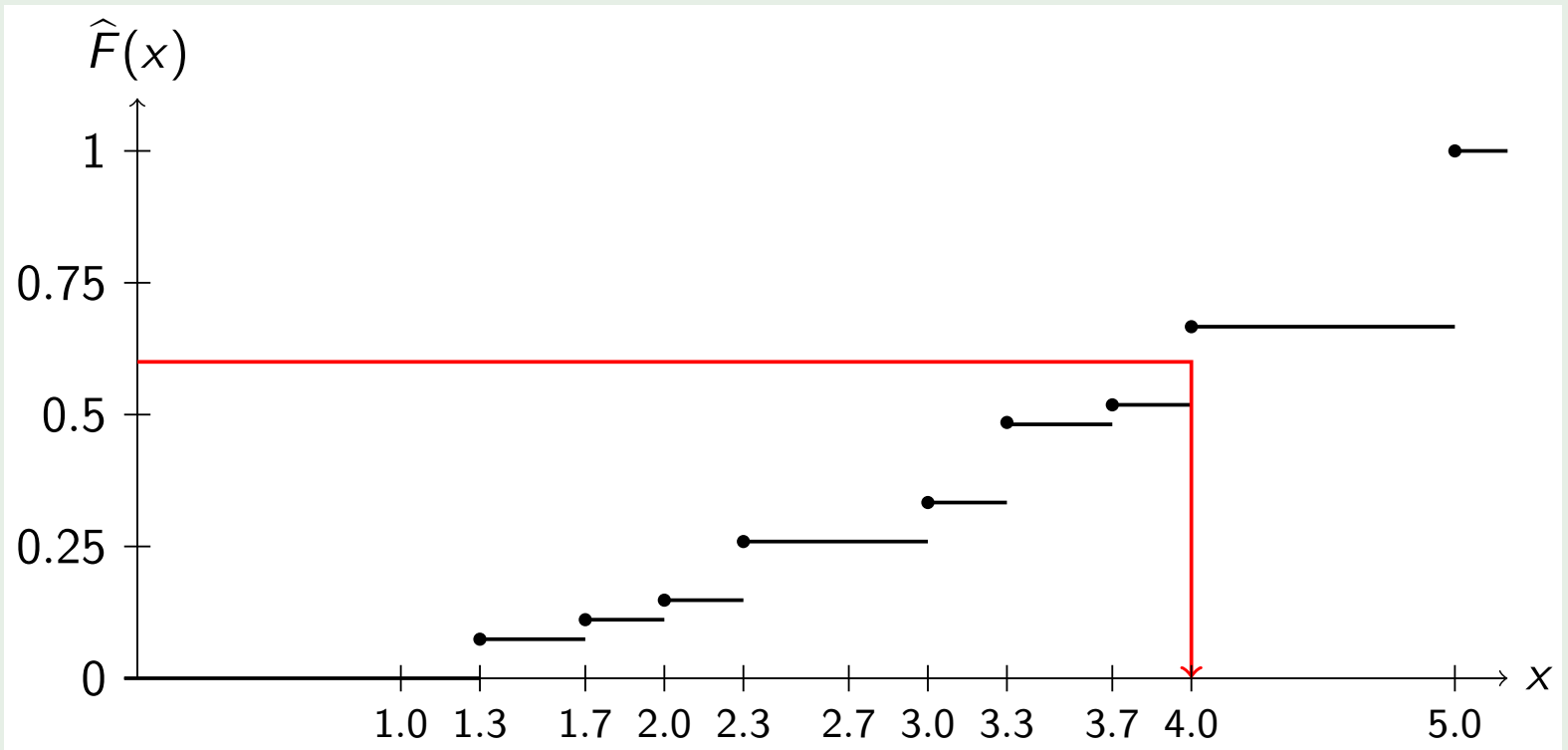
Quantile (2)

- Für $n = 27$ Klausurteilnehmer einer Statistikklausur ergaben sich folgende empirische Verteilungsfunktionswerte:

x_k	1.3	1.7	2	2.3	3	3.3	3.7	4	5
$\widehat{F}(x_k)$	0.07	0.11	0.15	0.26	0.33	0.48	0.52	0.67	1.00

- Man erhält somit bspw. für das 50%-Quantil: $z_{0.5} = x_7 = 3.7$. Das 25%-Quantil ist: $z_{0.25} = x_4 = 2.3$.
- Dies lässt sich grafisch, am Beispiel des 60%-Quantils ($z_{0.6} = 4$), darstellen:

Quantile (3)



Der Interquartilsabstand

Ein Maß für die Streuung, basierend auf Quantilen, ist der sogenannte **Interquartilsabstand**. Dieser misst den Abstand zwischen dem 75%-Quantil und dem 25%-Quantil.

Für n gegebene Merkmalsausprägungen x_i $(i = 1, \ldots, n)$ ist

$$d_Q = z_{0.75} - z_{0.25}$$

der Interquartilsabstand.

- Das 75%-Quantil und das 25%-Quantil werden auch **oberes** und **unteres Quartil** genannt.
- $z_{0.1}$, $z_{0.2}$,..., $z_{0.9}$ werden **Dezile** genannt.

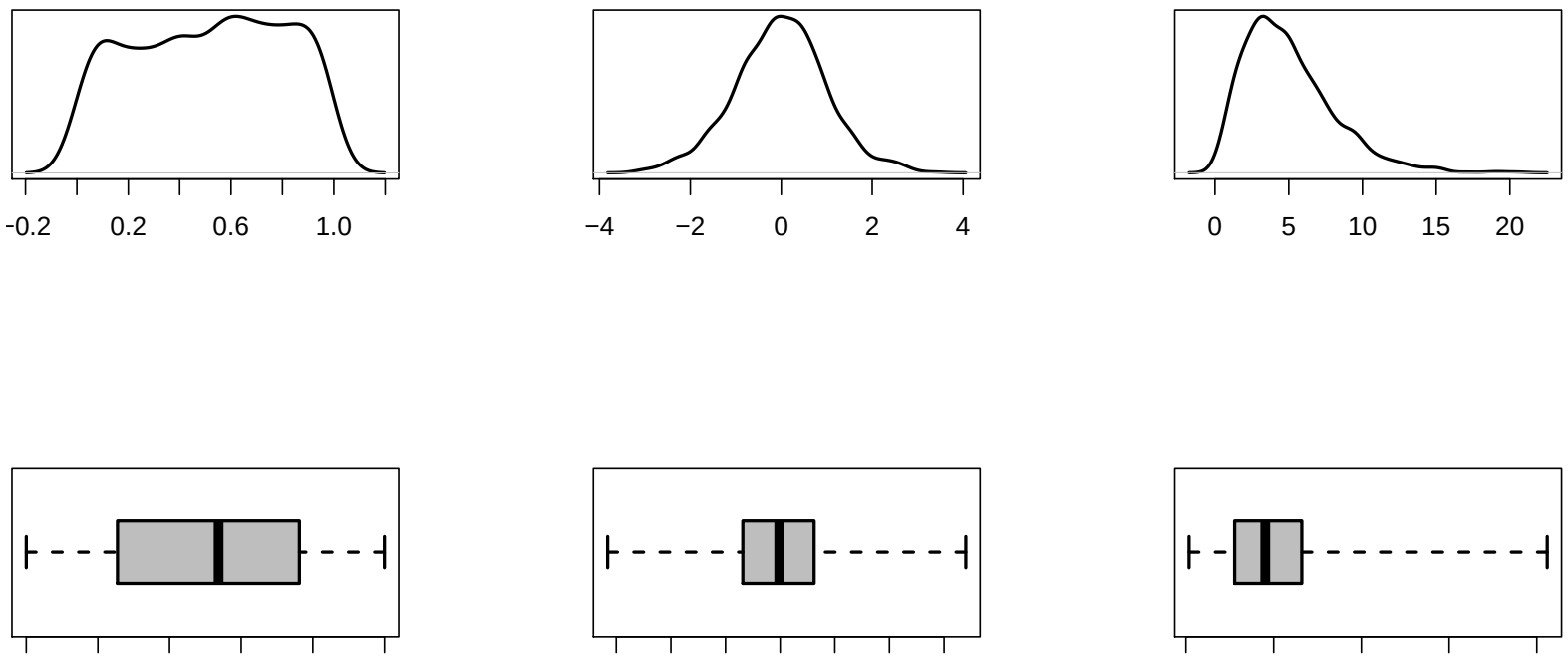
Box-Plots (1)

- Eine Möglichkeit, einen gegebenen Datensatz grafisch mit Hilfe von Quantilen darzustellen, ist der sogenannte **Box-Plot**.
- Hierbei wird das 25%- und das 75%-Quantil, der Median sowie der kleinste und der größte Wert in einer Grafik dargestellt.

Das 25%-Quantil ($z_{0.25}$) und das 75%-Quantil ($z_{0.75}$), der Median ($\bar{x}^{\text{Med}}$) sowie der kleinste ($x_{[1]}$) und der größte ($x_{[n]}$) Wert eines gegebenen Datensatzes x_i $(i = 1, \ldots, n)$ werden derart dargestellt, dass zwischen $z_{0.25}$ und $z_{0.75}$ eine „Box“ gezeichnet wird.

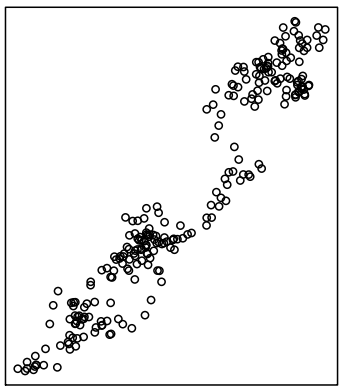
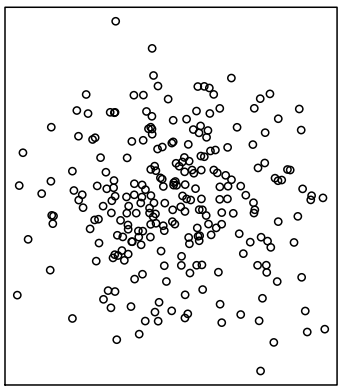
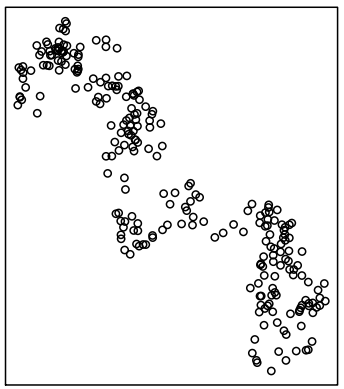
Box-Plots (2)

Für verschiedene empirische Datensätze sind in der folgenden Abbildung die Verteilungen und die zugehörigen Box-Plots dargestellt.



Kapitel 7.4
Beschreibung bivariater Datensätze

Streudiagramme

- Das linke Schaubild zeigt einen relativ hohen **positiven** Zusammenhang: Steigende Werte in x-Richtung gehen tendenziell mit steigenden Werten in y-Richtung einher.
- Im mittleren Schaubild ist augenscheinlich **kein** Zusammenhang erkennbar.
- Im rechten Schaubild erkennt man einen **negativen** Zusammenhang: Steigende Werte in x-Richtung gehen mit fallenden Werten in y-Richtung einher.

Die Kovarianz

Die **Kovarianz** misst für zwei verhältnisskalierte Merkmale die durchschnittliche Summe der Abweichungsprodukte.

Für zwei verhältnisskalierte Merkmale X und Y mit den zugehörigen Merkmalsausprägungen x_i und y_i $(i = 1, \ldots, n)$ ist

$$s_{x,y} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \left(\frac{1}{n} \sum_{i=1}^{n} x_i y_i \right) - \bar{x}\bar{y}$$

die Kovarianz der Merkmale X und Y.

Der Bravais-Pearson-Korrelationskoeffizient

Für zwei Merkmale X und Y mit zugehörigen Merkmalsausprägungen x_i bzw. y_i $(i = 1, \ldots, n)$ und Varianzen ungleich Null ($s_x^2 \neq 0$, $s_y^2 \neq 0$) ist

$$r_{x,y} := \frac{s_{x,y}}{s_x \cdot s_y}$$

der Bravais-Pearson-Korrelationskoeffizient.

Der Korrelationskoeffizient ist im Gegensatz zur Kovarianz (oder auch zur Varianz) dimensionslos, misst jedoch ebenfalls lediglich lineare Abhängigkeiten.

Kapitel 7.5
Das lineare Regressionsmodell

Das lineare Regressionsmodell

- Für zwei Merkmale X und Y mit jeweils n Merkmalsausprägungen x_i bzw. y_i $(i = 1, \ldots, n)$ ist

$$y_i = a + b \cdot x_i + \varepsilon_i \,,$$

mit Achsenabschnitt $a \in \mathbb{R}$ und Steigungskoeffizient $b \neq 0$ das empirische **lineare Modell**.

- Das Merkmal Y wird als „abhängige Variable“ und das Merkmal X als „unabhängige Variable“ bezeichnet. Die einzelnen ε_i's bezeichnet man als Residuen.

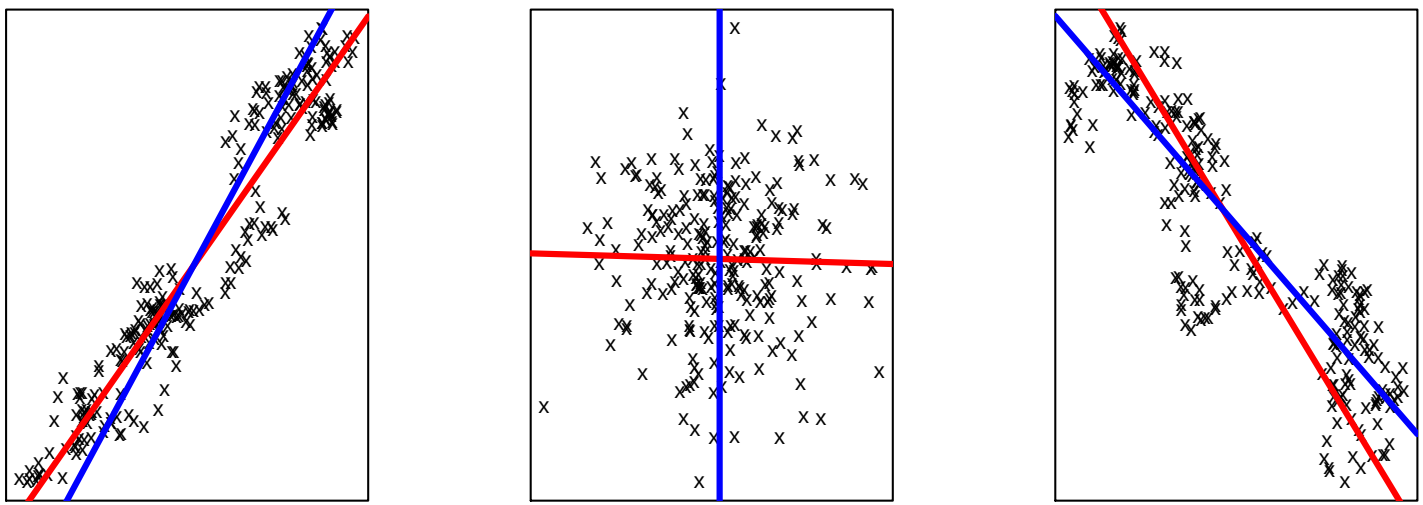
- Die Gerade

$$\hat{y} = a + b \cdot x$$

wird als **Ausgleichsgerade** oder **Regressionsgerade** bezeichnet.

Schätzung der Regressionskoeffizienten

- Ziel ist es nun, die beiden Parameter a und b so zu bestimmen / zu schätzen, dass die Gerade die vorliegenden Daten möglichst **gut** beschreibt.
- Die folgende Abbildung gibt verschiedene Beispiele für bivariate Datensätze und Vorschläge für Regressionsgeraden.



Die Kleinste-Quadrate-Methode

Mit dem Ansatz der **Kleinste-Quadrate-Methode** (**KQ-Methode**) erhält man für die Parameter a und b der Ausgleichsgeraden die nachfolgenden Schätzungen $\hat{a}$ und $\hat{b}$.

Für zwei Merkmale X und Y mit jeweils n Merkmalsausprägungen x_i bzw. y_i $(i = 1, \ldots, n)$, für die ein lineares Modell mit Y als abhängigem Merkmal und X als unabhängigem Merkmal unterstellt wird, gilt für die Parameter der Ausgleichsgeraden $\hat{y} = a + bx$ nach der KQ-Methode:

$$\hat{a} = \bar{y} - \hat{b}\bar{x} \qquad \hat{b} = \frac{s_{x,y}}{s_x^2}$$

Streuungszerlegung im Regressionsmodell

Im linearen Regressionsmodell mit beobachteten Werten y_i $(i = 1, \ldots, n)$ und zugehörigen theoretischen Werten $\hat{y}_i$ (ermittelt mittels KQ-Methode) gilt:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \,.$$

Alternativ schreibt man auch:

$$SST = SSE + SSR$$

- **SST**: „**S**um of **S**quares **T**otal“, als **Gesamtstreuung** (das n-fache der Varianz s_y^2)
- **SSE**: „**S**um of **S**quares **E**xplained“, durch das lineare Modell **erklärte Streuung**
- **SSR**: „**S**um of **S**quares **R**esiduals“, **Rest-** oder **Residualstreuung**, die Streuung, die das Modell **nicht** erklärt

Das Bestimmtheitsmaß

Die Überlegung ist nun, dass ein lineares Modell umso **„besser“** ist, je größer der Anteil der erklärten Streuung an der Gesamtstreuung ist. Ein lineares Modell sollte eine möglichst kleine Rest- bzw. Residualstreuung aufweisen.

Im linearen Regressionsmodell ist

$$R^2 := \frac{SSE}{SST} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

das sogenannte **Bestimmtheitsmaß**.

Es gilt: $0 \leq R^2 \leq 1$

Kapitel 7.6
Übungsaufgaben

43) Für die 18 Arbeitstage des Monats April wurden bei einem Vertreter hinsichtlich des Merkmals „Anzahl der Verkaufsabschlüsse pro Arbeitstag“ die folgenden Werte erhoben:

22 18 22 15 16 23 24 17 19 17 17 18 23 18 22 23 25 21

a) Stellen Sie für das betrachtete Merkmal eine Häufigkeitstabelle auf, die die absoluten, relativen und kumulierten relativen Häufigkeiten beinhaltet.
b) Zeichnen Sie ein Stabdiagramm und die empirische Verteilungsfunktion für das betrachtete Merkmal. Was können Sie aus dem Stabdiagramm über den Modus des betrachteten Merkmals folgern?
c) Berechnen Sie das arithmetische Mittel und die empirische Standardabweichung des betrachteten Merkmals.
d) Bestimmen Sie für das betrachtete Merkmal den Median, das 50%-Quantil und den Interquartilsabstand.
e) Treffen Sie auf Grundlage eines Vergleiches von arithmetischem Mittel, Median und Modus eine Aussage über die Schiefe der Verteilung des betrachteten Merkmals.

44) Ein Zeitschriftenhändler bestellt von einer wöchentlich erscheinenden Zeitschrift jeweils 20 Exemplare pro Ausgabe. Aus seinen Aufzeichnungen ermittelt er die folgende Häufigkeitsverteilung für den wöchentlichen Absatz:

Verkaufte Exemplare	9	10	11	12	13	14	15	16	17
Absolute Häufigkeit	1	3	6	8	10	10	7	4	1

a) Erstellen Sie eine vollständige Häufigkeitstabelle für die wöchentlich verkauften Exemplare.
b) Bestimmen Sie das arithmetische Mittel, den Median und den Modus der verkauften Exemplare pro Woche.
c) Bestimmen Sie das untere und das obere Quartil sowie den Interquartilsabstand der verkauften Exemplare pro Woche.
d) Stellen Sie den gegebenen Datensatz mit Hilfe eines Box-Plots grafisch dar.

45) Für $n = 40$ Besucher eines Erlebnisbades wurden die Merkmale „Alter in Jahren“ (X) und „Aufenthaltsdauer im Erlebnisbad in Stunden“ (Y) erhoben. Aus den Daten wurden folgende Summen berechnet:

$$\sum_{i=1}^{40} x_i^2 = 100840 \qquad \sum_{i=1}^{40} (x_i - \overline{x})^2 = 57840 \qquad \sum_{i=1}^{40} y_i^2 = 1998{,}4$$

$$\sum_{i=1}^{40} (y_i - \overline{y})^2 = 360 \qquad \sum_{i=1}^{40} (x_i - \overline{x})(y_i - \overline{y}) = 2592$$

a) Berechnen Sie für beide Merkmale die empirische Varianz und die empirische Standardabweichung.
b) Berechnen Sie die empirische Kovarianz und den empirischen Korrelationskoeffizienten nach Bravais-Pearson und interpretieren Sie Ihre Ergebnisse.
c) Stellen Sie eine sinnvolle Regressionsbeziehung zwischen den beiden Merkmalen auf, schätzen Sie den Steigungsparameter unter Verwendung der KQ-Methode und interpretieren Sie diesen.
d) Beurteilen Sie die Güte des von Ihnen aufgestellten linearen Regressionsmodells anhand eines geeigneten Kriteriums.