3.2 Median & Quantile

3.2.1 Median

- Wie lässt sich ein Mittelwert bei ordinalskalierten Merkmalen definieren?
- Das arithmetische Mittel besitzt die Schwerpunkteigenschaft

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

- Eine andere mögliche Schwerpunkteigenschaft: Rechts und links des „mittleren Wertes“ $x_{0.5}$ liegen jeweils mit dem Wert selbst (mindestens) 50% der Daten. Dies ergibt den $Median$.

Definition 3.7.

Gegeben sei die Urliste $x_1, \ldots, x_n$ eines (mindestens) ordinalskalierten Merkmals X. Jede Zahl x_{med} mit

$$\frac{|\{i|x_i \leq x_{med}\}|}{n} \geq 0.5 \quad \text{und} \quad \frac{|\{i|x_i \geq x_{med}\}|}{n} \geq 0.5$$

heißt Median.

Beispiel: Klausurnoten

$\underbrace{1,1,1,\ldots,1}$	$\underbrace{2,2,2,\ldots,2}$	$\underbrace{3,3,3,\ldots,3}$	$\underbrace{4,4,4,\ldots,4}$	$\underbrace{5,5,5,\ldots,5}$
65 mal	96 mal	91 mal	78 mal	53 mal
17%	25,1%	23,8%	20,4%	13,8%

3.2.2 Quantile

Definition 3.8. *Gegeben sei die Urliste* $x_1, \ldots, x_n$ *eines (mindestens) ordinalskalierten Merkmals* X *und eine Zahl* $0 < \alpha < 1$. *Jede Zahl* x_α *mit*

$$\frac{|\{i|x_i \leq x_\alpha\}|}{n} \geq \alpha \quad \textit{und} \quad \frac{|\{i|x_i \geq x_\alpha\}|}{n} \geq 1 - \alpha$$

heißt $\alpha \cdot 100\%$*-Quantil.*

Spezielle Quantile:

- Median: $x_{0.5} = x_{med}$.
- Quartile: $x_{0.25}$, $x_{0.75}$.
- Dezile: $x_{0.1}$, $x_{0.2}$, . . . , $x_{0.8}$, $x_{0.9}$.

Beispiel Klausurnoten:

$$x_{0.25} = \qquad x_{0.1} =$$

Bemerkungen:

- Alternative Definition des Medians über die $geordnete$ Urliste (z.B. Fahrmeir et al., 2010)
  $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$:

$$x_{med} := \begin{cases} \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) & \text{für } n \text{ gerade} \\ x_{\left(\frac{n+1}{2}\right)} & \text{für } n \text{ ungerade} \end{cases}$$

  Ähnlich für andere Quantile möglich.

- Diese Definition ist insofern inkonsequent, als sie auf die bei ordinalen Daten streng genommen nicht zulässige Addiditionen rekurriert. Bei intervallskalierten Daten hingegen spricht vieles für diese Definition.

- Andererseits können in gewissen Grenzfällen Quantile im Sinne der ursprünglichen Definition **??** nicht eindeutig sein:

- Beide Definitionen sind letztlich in vielen praktisch relevanten Fällen miteinander verträglich. Für n ungerade fallen sie stets zusammen, für n gerade stimmen sie überein, falls $x_{\left(\frac{n}{2}\right)} = x_{\left(\frac{n}{2}+1\right)}$
- Man kann Quantile einfach an der empirischen Verteilungsfunktion ablesen:

- Bei linearer Interpolation für gruppierte intervallskalierte Merkmalen definiert man die Quartile analog über den Schnittpunkt mit der Verteilungsfunktion:
  Man beachte aber, dass es sich bei der Interpolation um eine Approximation handelt und damit auch der so ermittelte Median nur eine Approximation darstellt.

3.2.3 Verhalten unter Transformationen:

Wie ändert sich der Median bei Transformation der Daten?

Satz 3.9.

Sei $x_1, x_2, \ldots, x_n$ die Urliste eines (mindestens) ordinalskalierten Merkmals X mit Median x_{med}, und g eine streng monotone Funktion. Mit $y_1 = g(x_1), \ldots, y_n = g(x_n)$ als Urliste des Merkmals $Y = g(X)$ gilt für den Median y_{med} von Y:

$$y_{med} = g(x_{med}).$$

Fordert man zusätzlich, dass $g(\cdot)$ streng monoton steigend ist, so gilt die entsprechende Aussage für beliebige Quartile.

Beispiel: Drei quadratische Zimmer

Bem. 3.10.

Man beachte, dass in obiger Situation geschichteter Gesamtheiten (vgl. Satz **??**) eine korrekte Bestimmung des Gesamtmedians aus den Medianen in den einzelnen Schichten nicht möglich ist.

Welche dieser Transformationen sind linear, welche streng monoton? Was lässt sich über die Beziehungen zwischen den arithmetischen Mitteln der Merkmale sagen, was über ihre Mediane und Quantile?

3.3 Modus

- Gesucht: geeignetes Lagemaß bei auf Nominalskala gemessenen Daten?
- Der exakte Wert der als Merkmalsausprägungen vergebenen Zahlen ist inhaltlich völlig bedeutungslos, d.h, etwas formaler: beliebige eineindeutige Transformationen verändern die inhaltliche Aussage nicht (z.B. Parteienpräferenz: ob man die Partei alphabetisch durchnummeriert oder anhand ihrer Stimmenanteile bei der letzten Wahl ändert nichts).
- Als Lagemaß dient der $häufigste\ Wert$: genauer jede Ausprägung a_j mit der größten Häufigkeit h_j.

Definition 3.11.

Sei $x_1, \ldots, x_n$ die Urliste eines nominalskalierten Merkmals mit den Ausprägungen $a_1, \ldots, a_k$ und der Häufigkeitsverteilung $h_1, \ldots, h_k$, so heißt a_{j^*} $Modus$ x_{mod} genau dann, wenn $h_{j^*} \geq h_j$, für alle $j = 1, \ldots, k$.

Bemerkungen:

- Der Modus wird auch als Modalwert bezeichnet.
- Existieren mehrere Ausprägungen mit der gleichen größten Häufigkeit, so ist der Modus nicht eindeutig.
- Der Modus bleibt unter beliebigen eineindeutigen Transformationen erhalten: Betrachtet man das Merkmal X, eine eineindeutige Transformation g und das Merkmal $Y = g(X)$, so gilt

$$y_{mod} = g(x_{mod}).$$

3.4 Ein kurzer Vergleich der Lagemaße und einige Bemerkungen

- Bei intervallskalierten Daten darf man auch den Modus oder den Median anwenden, man verschenkt (bei alleiniger Verwendung) aber meist viel Information.
- Der Median geht nur auf die Ordnung der Beobachtungen und nicht auf die Abstände ein, der Modus gibt nur die am stärksten vertretende Ausprägung an.
- Anschaulich gesprochen ist der Median der mittlere Wert, und wird deshalb oft umgangssprachlich auch als Mittelwert bezeichnet. Vorsicht bei nicht statistischen Veröffentlichungen! (Etwa Nachrichtenmeldungen im Rundfunk zum Armutsbericht.)
- Im Gegesatz zum arithmetischen Mittel sind Median und Modus unempfindlich gegenüber Ausreißern. Wird die größte Beobachtung verhundertfacht, so ändern sich Median und Modus nicht, das arithmetische Mittel reagiert dagegen stark.

Beispiel: Einkommensverteilung

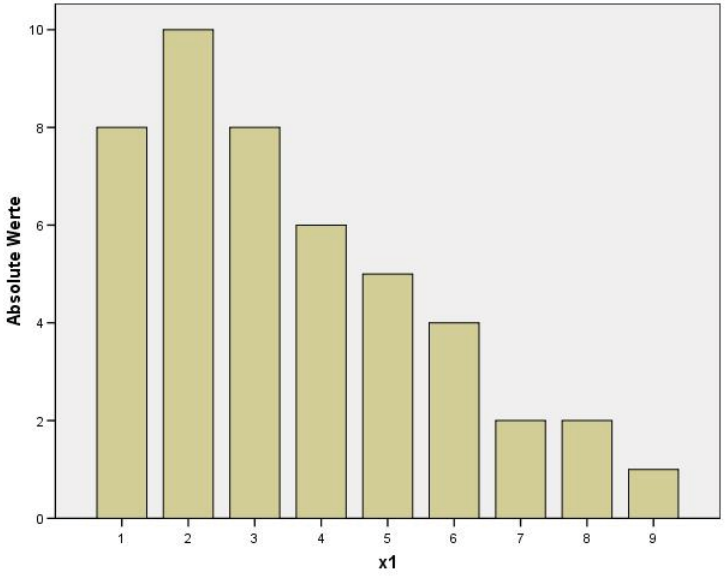
Generell ist bei der Betrachtung von Einkommen das arithmetische Mittel meist deutlich größer als der Median.

Beispiel: Statistikbücher. Häufigkeitsverteilung und zur graphischen Veranschaulichung ein maßstabtreues „Pseudostabdiagramm“:

	Häufigkeiten
$a_1 = 0$	$h_1 = 2$
$a_2 = 1$	$h_2 = 2$
$a_3 = 2$	$h_3 = 4$
$a_4 = 3$	$h_4 = 1$
$a_5 = 12$	$h_5 = 1$

Allgemeiner gilt: Die relative Lage von $\bar{x}, x_{med}, x_{mod}$ zueinander kann zur Charakterisierung von Verteilungen herangezogen werden. Unter Regularitätsbedingungen gilt:

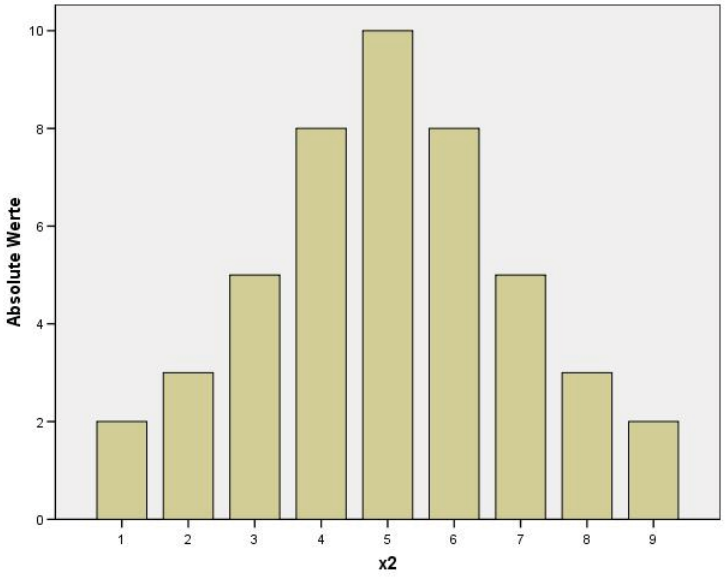
symmetrisch:	$\bar{x}$	$\approx$	x_{med}	$\approx$	x_{mod}
linkssteil:	$\bar{x}$	$>$	x_{med}	$>$	x_{mod}
rechtssteil:	$\bar{x}$	$<$	x_{med}	$<$	x_{mod}



$\bar{x} = 3.57$

$x_{med} = 3$

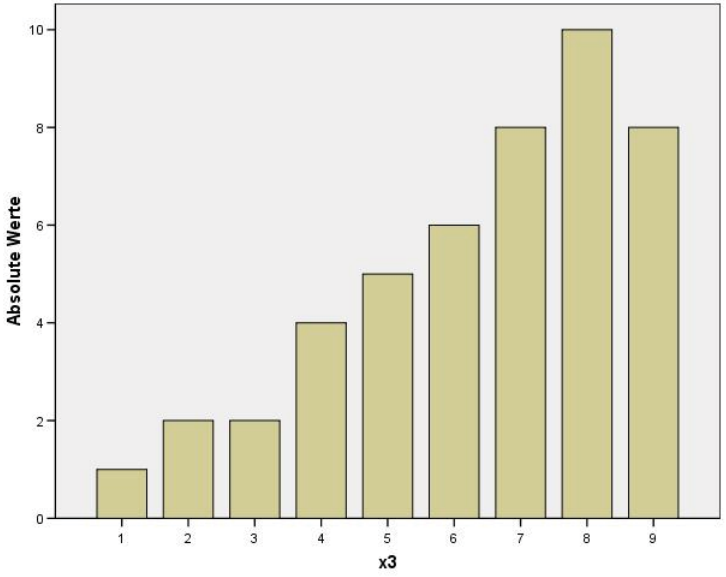
$x_{mod} = 2$



$$\bar{x} = 5$$

$$x_{med} = 5$$

$$x_{mod} = 5$$



$$\bar{x} = 6.43$$

$$x_{med} = 7$$

$$x_{mod} = 8$$

Exkurs: Lagemaße als Lösung eines Optimierungsproblems Alternative Möglichkeit, Lagemaße zu begründen, die später in der Regressionsanalyse verallgemeinert wird. (Typische statistische Sicht: Verfahren als in einem gewissen Sinn optimale Datenbeschreibung.)

Gegeben sei die Urliste $x_1, \ldots, x_n$ eines intervallskalierten Merkmals X, die zu einer Zahl a^* zusammengefasst werden soll. Man könnte sagen, das beste a^* ist dasjenige, das so gewählt wird, dass der Gesamtabstand zwischen a^* und den Daten minimal wird. Misst man den Abstand

quadratisch	$(x - a^*)^2$	so ergibt sich für a^*	$\bar{x}$
linear durch den Absolutbetrag	$\lvert x - a^* \rvert$	so ergibt sich für a^*	x_{med}

Für alle anderen $a \in \bullet$ gilt:

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

$$\sum_{i=1}^{n} \mid x_i - x_{med} \mid \leq \sum_{i=1}^{n} |x_i - a|.$$

Den Modus erhält man durch eine Grenzwertbetrachtung über die sogenannte Toleranzverlustfunktion. Mit

$$\mathbb{I}_{\{x \neq a\}} := \begin{cases} 1 & x_i \neq a \\ 0 & \text{sonst,} \end{cases}$$

ist für alle a

$$\sum_{i=1}^{n} \mathbb{I}_{\{x_i \neq x_{mod}\}} \leq \sum_{i=1}^{n} \mathbb{I}_{\{x_i \neq a\}}.$$

3.5 Geometrisches und harmonisches Mittel

3.5.1 Das geometrische Mittel

Es gibt Fälle, bei denen das arithmetische Mittel bei verhältnisskalierten Merkmalen nicht angemessen ist, zum Beispiel für Wachstumsraten oder Geschwindigkeiten.

Definition 3.12.

Sei $\Omega = \{0, \ldots, n\}$ eine Menge von Zeitpunkten und $B(i) =: b_i$ ein zum Zeitpunkt i erhobenes Merkmal, z.B. das Bruttosozialprodukt.

Für $i = 1, \ldots, n$ heißt

$$x_i = \frac{b_i}{b_{i-1}}$$

der i-te *Wachstumsfaktor* und

$$r_i = \frac{b_i - b_{i-1}}{b_{i-1}} = x_i - 1$$

die i-te *Wachstumsrate*.
Dann bezeichnet man

$$\bar{x}_{geom} := \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}} = (x_1 \cdot x_2 \cdot \ldots \cdot x_n)^{\frac{1}{n}}$$

als das *geometrische Mittel der Wachstumsfaktoren* $x_1, \ldots, x_n$.

Beispiel: Wirtschaftwachstum gemessen zu drei Zeitpunkten.

Geometrisches Mittel der Wachstumsfaktoren:

$$\bar{x}_{geom} =$$

Bem. 3.13.

- Es gilt

$$b_n = b_0 \cdot (\bar{x}_{geom})^n$$

d.h. $\bar{x}_{geom}$ ist tatsächlich ein durchschnittlicher Wachstumsfaktor, also derjenige Wert, der sich aus b_n und b_0 ergäbe, wenn zu allen Zeitpunkten konstantes Wachstum geherrscht hätte. Im Beispiel gilt in der Tat:

- Das geometrische Mittel kann auch zur Prognose (unter der Stabilitätsannahme, dass das durchschnittliches Wachstum gleich bleibt) verwendet werden:

$$b_{n+q} = b_n \cdot (\bar{x}_{geom})^q, \quad q \in \bullet \ .$$

- Logarithmieren liefert:

$$\ln \bar{x}_{geom} = \frac{1}{n} \sum_{i=1}^{n} \ln x_i.$$

  Das geometrische Mittel ist also ein arithmetisches Mittel auf der logarithmierten Skala.

- Man kann zeigen:

$$\bar{x}_{geom} \leq \bar{x}$$

  Da typischerweise $\bar{x}_{geom} \neq \bar{x}$,würde im Allgemeinen also die Angabe von $\bar{x}$ erhöhte Wachstumsraten vortäuschen.

3.5.2 Harmonisches Mittel

Beispiel: Die Entfernung von A nach B sei 99 km. Herr K. humpelt von A nach B mit konstant 1 km/h und fährt zurück mit konstant 99 km/h. Wie groß ist seine Durchschnittsgeschwindigkeit?

Naive Lösung: 50 km/h.

Definition 3.14.

Sei $x_1, \ldots, x_n$ mit $x_i \neq 0$ für alle i die Urliste eines verhältnisskalierten Merkmals X. Dann heißt

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

das *harmonische* Mittel der $x_1, \ldots, x_n$.

3.6 Weitere Streuungsmaße

3.6.1 Variationskoeffizient:

Definition 3.15.

Ist $\bar{x} > 0$, so heißt die Größe

$$v_X := \frac{\tilde{s}_X}{\bar{x}}$$

Variationskoeffizient des Merkmals X.

Bemerkungen:

- Gemessen wird hier die Streuung relativ zum Mittelwert. Insbesondere ist v_X dimensionslos.
- Der Variationskoeffizient erlaubt beispielsweise auch den Vergleich der Streuung von Preisen, die in verschiedenen Währungen gemessen wurden.

3.6.2 Inter-Quartils-Abstand:

Sind $x_{0.25}$ und $x_{0.75}$ das obere und das untere Quartil eines Merkmals, so heißt

$$d_{QX} := x_{0.75} - x_{0.25}$$

der *Interquartilsabstand*.

3.6.3 Median-Absolute-Deviation:

Der Median der Werte $|x_i - x_{med}|$, $i = 1, \ldots, n$ heißt *Median-Absolute-Deviation* von X (MAD_X).

3.6.4 Spannweite:

Die Größe

$$R_X := x_{(n)} - x_{(1)}$$

heißt *Spannweite* von X.

Bemerkungen

- Alle betrachteten Streuungsmaße sind nur für (mindestens) intervallskalierte Merkmalse sinnvoll definiert, da sie auf Abständen (typischerweise dem Abstand der Beobachtungen zu einem Lagemaß) beruhen.
- $\tilde{s}^2$, $\tilde{s}$, s^2, s sind die gebräuchlichsten Streuungsmaße.
- $\tilde{s}^2$, $\tilde{s}$, s^2, s sind sehr empfindlich gegenüber Ausreißern! Das Gleiche gilt für die Spannweite R. Die Kennzahlen MAD und d_Q hingegen entstammen der sogenannten robusten Statistik, die sich um ausreißerresistente Methoden bemüht.
- Gilt $x_1 = x_2 = \ldots = x_n$, so weisen alle Streungsmaße den Wert 0 auf. Mit Ausnahme von d_Q gilt auch die Umkehrung: Sind die Steuungsmaße (außer eben d_Q) $= 0$, so sind alle Werte der Urliste gleich.

- Nochmals der Hinweis: Eine häufige Ursache für Verwirrung und Missverständnisse liegt in der Tatsache, dass der Begriff „Streuung“ in der Statistik in einem doppelten Sinn gebraucht wird:

  – in einem allgemeinen Sinn: Streuung als Phänomen („Die Daten streuen stark“).

  – in einem speziellen Sinn: als $eine$ Maßzahl für dieses Phänomen.

Beispiel: Statistikbücher. Man berechne den Variationskoeffizienten, den Interquartilsabstand und die Spannweite.

Ausprägungen	h_j
0	2
1	2
2	4
3	1
12	1
$\sum$	10

3.7 Box-Plot

Ziele:

- einfache Darstellung von Verteilungen und ihrer Kennzahlen
- Identifikation von potentiellen Ausreißern
  $\Rightarrow$ nicht ausreißeranfällige Meßzahlen verwenden.

Idee:

i) markiere den Median

ii) symbolisiere Lage der „mittleren Werte“ durch eine Box

iii) wie weit reichen „weitere nicht atypische“ Werte?

iv) identifiziere potentielle Ausreißer: atypische (ungewöhnlich große, ungewöhnlich kleine) Werte, die genauerer Untersuchung bedürfen

zu ii) wähle die mittleren 50%: Die Box hat also Länge $dQ = x_{0.75} - x_{0.25}$

zu iii) als „nicht atypisch“ gelten alle Werte, die nicht weiter als $1.5dQ$ von der Box entfernt sind

Also bestimme:

- $x_{0.25}, \quad x_{0.50}, \quad x_{0.75}$.
- Interquartilsabstand: $d_{QX} = x_{0.75} - x_{0.25}$
- Zäune z_u, z_o, die am kleinsten bzw. größten Datenpunkt im Bereich $x_{0.25} - 1.5 \cdot d_{QX}$; $x_{0.75} + 1.5 \cdot d_{QX}$ liegen.
- Ausserhalb der Zäune werden $alle$ Punkte eingezeichnet; sie sind ausreißerverdächtig.

Vorsicht bei der Anwendung von Software! Vor allem außerhalb der Box sind auch andere Darstellungen üblich (z.B. Zäune immer bis $x_{(1)}$ und $x_{(n)}$).
Toutenburg (2002) beispielsweise unterscheidet zwischen Ausreißern ($1.5 \cdot d_{QX}$ bis $3 \cdot d_{QX}$ von Rändern der Box entfernt) und Extremwerten (mehr als $3 \cdot d_{QX}$ vom Rand entfernt).
Oft wird der Median durch einen dicken Punkt ausgedrückt. Der Box-Plot gibt einen kompakten Überblick über die Form der Verteilung (Zentrale Tendenz, Variabilität, Schiefe, extreme Werte).

Box-Plots können auch zum graphischen Vergleich von Verteilungen verwendet werden: