Formelsammlung

zur Klausur

Beschreibende Statistik

Statistische Daten

Qualitative Daten

Nominal skalierte Merkmalsausprägungen (Unterscheidungsmerkmale)

- können nicht durch Auszählen oder Messen ermittelt werden.
- haben keine natürliche Reihenfolge.
- liefern keine Abstände oder Verhältnisse.

Ordinal skalierte Merkmalsausprägungen (Rangmerkmale)

- können in eine natürliche Reihenfolge (auf- oder absteigende Ordnung) gebracht werden.
- sind keine absoluten, sondern nur relative Werte.
- liefern keine Abstände oder Verhältnisse.

Quantitative Daten

Metrisch skalierte Merkmalsausprägungen (Abstandsmerkmale)

- sind messbar oder abzählbar (reelle Zahlen).
- sind absolute Werte.
- liefern Abstände und Verhältnisse.
- sind **diskret**, wenn es nur endlich viele Ausprägungen geben kann (**zählen**).
- sind **stetig**, wenn sie jeden beliebigen reellen Wert zumindest in einem bestimmten Intervall annehmen können (**messen**).

Häufigkeitsverteilung

Absolute Häufigkeit

Die Anzahl h_i (i = 1, 2, ..., k) der statistischen Einheiten mit der Merkmalsausprägung x_i bezeichnet man als **absolute Häufigkeit**. Es gilt:

$$h_1 + h_2 + h_3 + ... + h_k = \sum_{i=1}^{k} h_i = n$$

Relative Häufigkeit

Dividiert man die absoluten Häufigkeiten h_i durch die Anzahl n der statistischen Einheiten, so erhält man die **relativen Häufigkeiten f_i**.

$$f_i = \frac{h_i}{n}$$

Die Größen 100 * f_i % heißen **prozentuale Häufigkeiten**.

Es gilt: $\sum_{i=1}^{k} f_i = 1 = 100$

Regel für Klassenbildung

- Zu viele Klassen machen das Bild unübersichtlich
- Zu wenige Klassen lassen Informationen verloren gehen
- In der Regel 5 – 20 Klassen, jedoch weniger als n (n ist Größe der Stichprobe)
- Es soll eine obere und untere Klassengrenze fest gelegt werden
- In der Regel gleich breite Klassen verwenden
- Ungleiche Klassenbreiten nur, wenn viele Beobachtungen in kleinem Bereich und geringer Rest in weitem Bereich.

$$\text{Häufigkeitsdichte} = \frac{\text{Häufigkeit}}{\text{Klassenbreite}}$$

Häufigkeitssummenverteilung

Aufsummierte Häufigkeiten. Die Addition der Häufigkeiten erfolgt nach der natürlichen Reihenfolge der Ausprägungen von der kleinsten zur größten. Die Summenverteilung ist nur für Rang- und Abstandsmerkmale sinnvoll.

Man summiert die Häufigkeiten aller Ausprägungen bis zu einem bestimmten Wert. Die Häufigkeitssummenverteilung gibt also an, wie viele Einheiten einer Gesamtheit einen bestimmten Wert nicht überschreiten.

Die aufsummierten Häufigkeiten werden durch die Summenkurve grafisch veranschaulicht.

Mittelwerte oder Lageparameter

Das arithmetische Mittel

$$\textit{Arithmetisches Mittel} \quad \bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = \frac{\textit{Summe der Einzelwerte}}{\textit{Anzahl der Einheiten}}$$

Das gewogene arithmetische Mittel

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

Das arithmetische Mittel klassierter Daten

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} (x_i^* \cdot h_i) \qquad \textit{mit} \quad x_i^* \ \textit{Klassenmitte der Klasse} \ i$$

Modalwert oder häufigster Wert

Modalwert = diejenige Merkmalsausprägung die am häufigsten vorkommt.

Zentralwert oder Median

Der Zentralwert ist diejenige Merkmalsausprägung, die in der Mitte der in eine Rangfolge gebrachten Einzelausprägungen steht.

Anzahl der Elemente ungerade: Median an der Stelle $\frac{n+1}{2}$

Anzahl der Elemente gerade: Median arithmetisches Mittel der Elemente an den Stellen $\frac{n}{2}$ und $\frac{n}{2}+1$

Quartile

Quartile geben zusammen mit dem Median Hinweise auf die Verteilung der Daten: Links des unteren Quartils ($x_{0,25}$) liegen etwa 25% der Daten und rechts des oberen Quartils ($x_{0,75}$) ebenfalls etwa 25% der Daten. Im mittleren Bereich liegen die restlichen 50%.

Bestimmung der Quartile Q_1, Q_2 und Q_3:

Q_2 entspricht dem Median. Bestimmung siehe oben.

Zur Bestimmung von Q_1 und Q_3 wird die untere bzw. die obere Hälfte der Daten nach der gleichen Methode wie bei der Bestimmung des Medians nochmals unterteilt.

Geometrisches Mittel

Immer dann, wenn es um die Ermittlung durchschnittlicher Wachstumsraten geht, versagt das arithmetische Mittel. An seiner Stelle wird das geometrische Mittel verwendet.

$$GM = \sqrt[n]{x_1 \cdot x_2 \cdot ... \cdot x_n} \quad \text{für } x_i > 0$$

Streuungsmaße

Spannweite oder Variationsbreite

Spannweite = Differenz zwischen dem grössten (x_{max}) und dem kleinsten (x_{min}) Wert.

Quartilsabstand und Boxplot

Der Quartilsabstand ist die Differenz zwischen dem ersten und dritten Quartil. Er umfasst den Bereich der mittleren 50% der Werte.

Das Box- and Wiskersdiagramm stellt eine Häufigkeitsverteilung dar: Zwischen dem 1. und 3. Quartil wird ein Kasten aufgebaut. In diesem Bereich liegen 50% der Beobachtungen.

Die seitlich angesetzten Schnurrhaare vermitteln einen Eindruck, wie weit die restlichen 50% der Werte streuen. Wie weit die Schnurrhaare ausgezogen werden ist unterschiedlich. Die gebräuchlichsten Verfahren gehen bis zu den Extremwerten bzw. bis zum 10. und 90. Perzentil.

Mittlere lineare Abweichung

Mittlere lineare Abweichung bezeichnet das arithmetische Mittel der absoluten Abweichungen der Merkmalswerte von einem Mittelwert (arithmetisches Mittel oder Median).

$$d = \frac{1}{n}\sum_{i=1}^{n}|x_i - M| \qquad \text{M = Mittelwert}$$

Varianz

$$s^2 = \frac{\textit{Summe aller Abweichungsquadrate}}{\textit{Zahl der Meßwerte}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - AM)^2$$

(AM = arithmetisches Mittel)

Standardabweichung

Die Standardabweichung s ist die Wurzel aus der Varianz.

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - AM)^2}$$

Variationskoeffizient

Variationskoeffizient = Quotient aus Standardabweichung und arithmetischem Mittel.

$$v = \frac{s}{AM}$$

Der Variationskoeffizient v ist eine dimensionslose Zahl. Er gibt an, wie viel Prozent vom arithmetischen Mittelwert die Standardabweichung beträgt.

Der Variationskoeffizienten ist ein relatives oder größenunabhäniges Streuungsmaß. Er ist daher geeignet, die Streuung mehrerer Verteilungen mit unterschiedlichen Mittelwerten zu vergleichen.

Regression und Korrelation

Lineare Regression

Bestimmung der Geradengleichung y = ax + b.

Berechnungsverfahren 1:

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$b = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Arbeitstabelle:

x_i	y_i	$x_i \cdot y_i$	x_i^2
$\sum x_i$	$\sum y_i$	$\sum x_i y_i$	$\sum x_i^2$

Berechnungsverfahren 2:

$$a = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{Kovarianz}{Varianz\,der\,x-Werte}$$

$$b = \bar{y} - a\,\bar{x}$$

Arbeitstabelle:

x_i	y_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(x_i-\bar{x})(y_i-\bar{y})$
$\sum x_i$	$\sum y_i$	$\sum x_i-\bar{x}$	$\sum (x_i-\bar{x})^2$	$\sum y_i-\bar{y}$	$\sum (x_i-\bar{x})(y_i-\bar{y})$

Korrelationskoeffizient von Pearson

Der Korrelationskoeffizient von Pearson liefert ein Maß für die Abhängigkeit der beiden Merkmale x und y. Er kann die Werte zwischen -1 und +1 annehmen.

r = 1: Alle Beobachtungswerte liegen auf einer steigenden Geraden.

r = –1: Alle Beobachtungswerte liegen auf einer fallenden Geraden.

r > 0: Merkmale positiv korreliert, d.h. die Regressionsgerade ist steigend.

r < 0: Merkmale negativ korreliert, d.h. die Regressionsgerade ist fallend.

r = 0: Die Merkmale sind unkorreliert, d.h. es besteht kein linearer Zusammenhang.

Berechnungsverfahren 1:

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{s_{xy}}{s_x \cdot s_y}$$

Arbeitstabelle:

x_i	y_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
$\sum x_i$	$\sum y_i$		$\sum (x_i-\bar{x})^2$		$\sum (y_i-\bar{y})^2$	$\sum (x_i-\bar{x})(y_i-\bar{y})$

Berechnungsverfahren 2:

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}$$

Arbeitstabelle:

x_i	y_i	$x_i \cdot y_i$	x_i^2	y_i^2
$\sum x_i$	$\sum y_i$	$\sum x_i y_i$	$\sum x_i^2$	$\sum y_i^2$

Rangkoeffizient von Spearman

Voraussetzung: es liegen zwei Merkmale vor, die mindestens eine Ordinalskala besitzen.

Die Merkmalswerte eines jeden Merkmals werden aufsteigend geordnet und es wird ihnen entsprechend ihrem Platz eine Rangzahl zugeordnet.

Für die weitere Berechnungen verwendet man nur noch die Rangzahlen, nicht mehr die tatsächlichen Merkmalswerte.

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

n = *Anzahl der statistischen Einheiten*

$d_i = Rang(x_i) - Rang(y_i)$

d_i^2 = *quadratische Differenz zwischen den Rängen der beiden Merkmale*

Mittlere quadratische Kontingenz

	y_1	...	y_j	...	y_r	
x_1	n_{11}	...	n_{1j}	...	n_{1r}	n_{1*}
...	...	...	...	...	...	
x_i	n_{i1}		n_{ij}		n_{ir}	n_{i*}
...	...	...	...	...	...	
x_m	n_{m1}	...	n_{mj}	...	n_{mr}	n_{m*}
	n_{*1}		n_{*j}		n_{*r}	n

Berechnungsverfahren 1:

$$C = \frac{1}{n}\left(\sum_{i=1}^{m}\sum_{j=1}^{r}\frac{\left(n_{ij} - \frac{n_{i*}\cdot n_{*j}}{n}\right)^2}{\frac{n_{i*}\cdot n_{*j}}{n}}\right)$$

Arbeitstabelle:

n_{ij}	$\frac{n_{i*}\cdot n_{*j}}{n}$	$n_{ij} - \frac{n_{i*}\cdot n_{*j}}{n}$	$\left(n_{ij} - \frac{n_{i*}\cdot n_{*j}}{n}\right)^2$	$\frac{\left(n_{ij} - \frac{n_{i*}\cdot n_{*j}}{n}\right)^2}{\frac{n_{i*}\cdot n_{*j}}{n}}$
				$\sum \frac{\left(n_{ij} - \frac{n_{i*}\cdot n_{*j}}{n}\right)^2}{\frac{n_{i*}\cdot n_{*j}}{n}}$

Berechnungsverfahren 2:

$$C = \left(\sum_{i=1}^{m} \sum_{j=1}^{r} \frac{n_{ij}^2}{n_{i*} \cdot n_{*j}} \right) - 1$$

Arbeitstabelle:

n_{ij}	n_{ij}^2	$n_{i*} \cdot n_{*j}$	$\frac{n_{ij}^2}{n_{i*} \cdot n_{*j}}$
			$\sum \frac{n_{ij}^2}{n_{i*} \cdot n_{*j}}$

Vierfelderkoeffizient

n_{11}	n_{12}	n_{1*}
n_{21}	n_{22}	n_{2*}
n_{*1}	n_{*2}	n

$$\varphi = \frac{n_{12} \cdot n_{21} - n_{11} \cdot n_{22}}{\sqrt{(n_{11}+n_{12})(n_{21}+n_{22})(n_{11}+n_{21})(n_{12}+n_{22})}}$$

$$\varphi_{korr} = \frac{n_{12} \cdot n_{21} - n_{11} \cdot n_{22}}{n \cdot min(n_{11,} n_{22}) + n_{12} n_{21} - n_{11} n_{22}} \quad wenn \quad n_{12} n_{21} \geqslant n_{11} n_{22}$$

$$\varphi_{korr} = \frac{n_{12} \cdot n_{21} - n_{11} \cdot n_{22}}{n \cdot min(n_{21,} n_{12}) - n_{12} n_{21} + n_{11} n_{22}} \quad wenn \quad n_{12} n_{21} < n_{11} n_{22}$$