1 Einführung und erste Grundbegriffe

1.1 Vorbemerkungen zur Organisation, Bedeutung und Struktur der Veranstaltung

1.2 Was soll Statistik (nicht)?

1.3 Literatur

1.4 Grundbegriffe

Notation Merkmale werden typischerweise mit Großbuchstaben bezeichnet (X, Y, Z, etc.), Ausprägungen mit dem zugehörigen Kleinbuchstaben (x, y, z). Der Wertebereich wird mit W_x, W_y, W_z bzw. W bezeichnet.

Formal ist jedes Merkmal eine Funktion.

$$\begin{array}{rccc} X: & \Omega & \to & W \\ & \omega & \mapsto & X(\omega) \end{array}$$

Merkmalstypen

- Stetige, quasi-stetige und diskrete Merkmale
- Skalenniveaus
- Qualitative und quantitative Merkmale

2 Häufigkeitsverteilungen

Ausgangssituation An n Einheiten $\omega_1, \ldots, \omega_n$ sei das Merkmal X beobachtet worden. Die *verschiedenen* potentiell möglichen Merkmalsausprägungen werden mit $a_1, \ldots, a_k$ bezeichnet.

2.1 Häufigkeiten

Absolute Häufigkeiten der Merkmalsausprägungen Für jedes a_j, $j = 1, \ldots, k$, bezeichnen h_j und $h(a_j)$ die *absolute Häufigkeit* der Ausprägung a_j, d.h. die Anzahl der x_i aus $x_1, \ldots, x_n$ mit $x_i = a_j$.
Formal:

$$h_j := h(a_j) := |\{\omega \in \Omega \mid X(\omega) = a_j\}|.$$

Es gilt:

$$\sum_{j=1}^{k} h_j = n.$$

Relative Häufigkeiten der Merkmalsausprägungen Für jedes a_j, $j = 1, \ldots, k$, bezeichnen f_j und $f(a_j)$ die *relative Häufigkeit* der Ausprägung a_j, also

$$f_j := f(a_j) := \frac{h_j}{n}.$$

$f_1, f_2, \ldots, f_k$ nennt man die *relative Häufigkeitsverteilung.*
Es gilt:

$$\sum_{j=1}^{k} f_j = 1.$$

Häufigkeitstabelle

j	a_j	h_j	f_j
1	a_1	h_1	f_1
2	a_2	h_2	f_2
3	a_3	h_3	f_3
$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	a_k	h_k	f_k
$\sum$		n	1

2.2 Grafische Darstellung

2.3 Histogramm

2.4 Kumulierte Häufigkeiten und empirische Verteilungsfunktion

Definition Gegeben sei die Urliste $x_1, \ldots, x_n$ eines (mindestens) ordinalskalierten Merkmals mit der Häufigkeitsverteilung $h_1, \ldots, h_k$ bzw. $f_1, \ldots, f_k$.
Dann heißt

$$H(x) = \sum_{j:a_j \leq x} h(a_j) = \sum_{j:a_j \leq x} h_j$$

absolute kumulierte Häufigkeitsverteilung und

$$F(x) = \sum_{j:a_j \leq x} f(a_j) = \frac{1}{n} \sum_{j:a_j \leq x} h(a_j) = \frac{H(x)}{n}$$

relative kumulierte Häufigkeitsverteilung bzw. *empirische Verteilungsfunktion.*

Gruppierte Daten

- k Klassen $[c_0, c_1), \ldots, [c_{j-1}, c_j), \ldots, [c_{k-1}, c_k]$, h_j Häufigkeit in j-ter Klasse, $j = 1, \ldots, k$
- Verwende bei einem x aus der Klasse $[c_{j-1}, c_j)$ als Approximation für $H(x)$ folgenden, aus der linearen Interpolation gewonnenen, Punkt:

$$H(x) \approx H(c_{j-1}) + \frac{h_j}{(c_j - c_{j-1})} \cdot (x - c_{j-1})$$

3 Lage- und Streuungsmaße

3.1 Arithmetisches Mittel und Varianz

Definition (Arithmetisches Mittel) Sei $x_1, \ldots, x_n$ die Urliste eines (mindestens) intervallskalierten Merkmals X. Dann heißt

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

das *arithmetische Mittel* der Beobachtungen $x_1, \ldots, x_n$.

Alternative Berechnung basierend auf Häufigkeiten Hat das Merkmal X die Ausprägungen $a_1, \ldots, a_k$ und die (relative) Häufigkeitsverteilung $h_1, \ldots, h_k$ bzw. $f_1, \ldots, f_k$, so gilt:

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{k} a_j h_j = \sum_{j=1}^{k} a_j f_j.$$

Definition (Varianz) Sei $x_1, \ldots, x_n$ die Urliste eines intervallskalierten Merkmals X. Dann heißen

$$\tilde{s}_X^2 := \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

die *(empirische) Varianz* oder *Stichprobenvarianz* und

$$\tilde{s}_X := \sqrt{\tilde{s}_X^2}$$

die *empirische Streuung, Stichprobenstreuung* oder *Standardabweichung von* X.

Alternative Berechnung basierend auf Häufigkeiten Sind die Ausprägungen $a_1, \ldots, a_k$ mit (relativer) Häufigkeitsverteilung $h_1, \ldots, h_k$ bzw. $f_1, \ldots, f_k$ gegeben, so gilt

$$\tilde{s}_X^2 = \frac{1}{n} \sum_{j=1}^{k} h_j (a_j - \bar{x})^2 = \sum_{j=1}^{k} f_j (a_j - \bar{x})^2.$$

Verschiebungssatz Es gilt

$$\begin{aligned}
\tilde{s}_X^2 &= \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right)^2 = \overline{x^2} - (\bar{x})^2, \\
&= \left(\frac{1}{n} \sum_{j=1}^{k} (a_j^2) \cdot h_j \right) - \left(\frac{1}{n} \sum_{j=1}^{k} a_j \cdot h_j \right)^2 \\
&= \sum_{j=1}^{k} (a_j^2) \cdot f_j - \left(\sum_{j=1}^{k} a_j \cdot f_j \right)^2
\end{aligned}$$

Korrigierte empirische Varianz Sei $x_1, \ldots, x_n$ die Urliste eines intervallskalierten Merkmals X. Dann heißt

$$s_X^2 := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

die *korrigierte empirische Varianz* oder *korrigierte Stichprobenvarianz von* X.

Satz (Arithmetisches Mittel und lineare Transformationen) Gegeben sei die Urliste $x_1, \ldots, x_n$ eines (mindestens) intervallskalierten Merkmals X. Betrachtet wird das (linear transformierte) Merkmal $Y = a \cdot X + b$ und die zugehörigen Ausprägungen $y_1, \ldots, y_n$.
Dann gilt für das arithmetische Mittel $\bar{y}$ von Y:

$$\bar{y} = a \cdot \bar{x} + b.$$

Satz (Varianz und lineare Transformationen) Sei $x_1, \ldots, x_n$ die Urliste eines mindestens intervallskalierten Merkmals X mit $\tilde{s}_X > 0$ und $y_1, \ldots, y_n$ die zugehörige Urliste des Merkmals $Y = a \cdot X + b$. Dann gilt

$$\tilde{s}_Y^2 = a^2 \cdot \tilde{s}_X^2$$

und

$$\tilde{s}_Y = |a| \cdot \tilde{s}_X.$$

Definition (Arithmetisches Mittel bei gruppierten Daten) Sei X ein intervallskaliertes Merkmal, das in gruppierter Form mit k Klassen $[c_0, c_1), [c_1, c_2), \ldots, [c_{k-1}, c_k]$ erhoben wurde. Mit h'_ℓ, $\ell = 1, \ldots k$, als absoluter Häufigkeit der $\ell-$ten Klasse, f'_l als zugehöriger relativer Häufigkeit und $m_\ell := \frac{c_\ell + c_{\ell-1}}{2}$ als der jeweiligen Klassenmitte definiert man als *arithmetisches Mittel für gruppierte Daten*

$$\bar{x}_{\text{grupp}} := \frac{1}{n} \sum_{\ell=1}^{k} h'_\ell m_\ell = \sum_{\ell=1}^{k} f'_\ell m_\ell.$$

Satz (Arithmetisches Mittel bei geschichteten Daten) Zerfällt die Grundgesamtheit in z Schichten, so kann $\bar{x}$ aus den Schichtmitteln $\bar{x}^{(\ell)}$, $\ell = 1, \ldots, z$, berechnet werden:

$$\bar{x} = \frac{1}{n} \sum_{\ell=1}^{z} n^{(\ell)} \bar{x}^{(\ell)}.$$

Dabei bezeichnet $n^{(\ell)}$ die Anzahl der Elemente in der ℓ-ten Schicht.

Satz (Varianz bei geschichteten Daten) – Varianzzerlegung / Streuungszerlegung

- Schicht $1, \ldots, \ell, \ldots, z$
- Besetzungszahlen $n^{(1)}, \ldots, n^{(\ell)}, \ldots, n^{(z)}$; $\quad \sum_{l=1}^{z} n^{(\ell)} = n$
- Mittelwerte $\bar{x}^{(1)}, \ldots, \bar{x}^{(\ell)}, \ldots, \bar{x}^{(z)}$
- Varianzen $\tilde{s}^{2^{(1)}}, \ldots, \tilde{s}^{2^{(\ell)}}, \ldots, \tilde{s}^{2^{(z)}}$

$$\text{Mit} \qquad \tilde{s}^2_{\text{innerhalb}} := \frac{1}{n} \sum_{\ell=1}^{z} n^{(\ell)} \tilde{s}^{2^{(\ell)}}$$

$$\text{sowie} \qquad \tilde{s}^2_{\text{zwischen}} := \frac{1}{n} \sum_{\ell=1}^{z} n^{(\ell)} (\bar{x}^{(\ell)} - \bar{x})^2$$

$$\text{gilt} \qquad \tilde{s}^2 = \tilde{s}^2_{\text{innerhalb}} + \tilde{s}^2_{\text{zwischen}}.$$

3.2 Median & Quantile

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

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

heißt *Median*.

Definition (Quantile) 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

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

heißt $\alpha \cdot 100\%$-*Quantil*.

Spezielle Quantile

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

Alternative Definition des Medians über die *geordnete* Urliste $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$:

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

Satz (Verhalten unter Transformation) Sei $x_1, x_2, \ldots, x_n$ die Urliste eines (mindestens) ordinalskalierten Merkmals X und g eine monotone Funktion.

i) Ist x_{med} ein Median von X, so gilt mit $y_1 = g(x_1), \ldots, y_n = g(x_n)$ als Urliste des Merkmals $Y = g(X)$:
$$y_{med} = g(x_{med})$$
ist ein Median von Y.

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

Bei gruppierten Daten gilt für alle $\alpha \in (0,1)$ und alle $\alpha-$Quantile x_α: Die Gruppe, in der x_α liegt, ist ein $\alpha-$Quantil für das gruppierte Merkmal X_{grupp}.

3.3 Modus

Definition 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 $j = 1, \ldots, k$.

3.4 Ein kurzer Vergleich der Lagemaße und einige Bemerkungen

3.5 Geometrisches und harmonisches Mittel

Definition (Geometrisches Mittel) Sei $\Omega = \{0, \ldots, n\}$ eine Menge von Zeitpunkten und $b_0, b_1, \ldots, b_n$ mit $b_i := B(i)$ die Urliste eines Merkmals B.
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
$$\bar{x}_{\text{geom}} := \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}} = (x_1 \cdot x_2 \cdot \ldots \cdot x_n)^{\frac{1}{n}}$$
als das *geometrische Mittel der Wachstumsfaktoren* $x_1, \ldots, x_n$.

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

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

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

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

3.6 Weitere Streuungsmaße

Variationskoeffizient Ist $\bar{x} > 0$, so heißt die Größe

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

Variationskoeffizient des Merkmals X.

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*.

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

Spannweite Die Größe

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

heißt *Spannweite* von X.

3.7 Boxplot