Angewandte Stochastik

Prof. Dr. Evgeny Spodarev | Vorlesungskurs |

2. Thema

Heutiges Thema

- Beschreibende Statistik

Verteilungen und ihre Darstellungen

- Sei unser statistisches Merkmal (ZV) $X \sim F$. Auf den folgenden Folien werden wir Methoden zur statistischen Beschreibung und grafischen Darstellung der (unbekannten) Verteilung F betrachten.
- Sei X diskret verteilt mit Zähldichte p, d.h. $P(X \in \{a_1, \ldots, a_k\}) = 1$ und damit $p_j = P\left(X = a_j\right)$.
- Alternativ sei X absolut stetig verteilt mit Dichte f, d.h. $F(x) = \int_{-\infty}^{x} f(y)dy, \ x \in \mathbb{R}$.
- Sei dann $(x_1, \ldots, x_n)$ eine konkrete Stichprobe mit n unabhängigen Realisierungen von X.

Diagramme und Histogramme

- Falls das quantitative Merkmal X eine endliche Anzahl von Ausprägungen $\{a_1, \ldots, a_k\}$, $a_1 < a_2 < \ldots < a_k$, besitzt, also
$$P\big(X \in \{a_1, \ldots, a_k\}\big) = 1 \,,$$
dann kann eine Schätzung der Zähldichte $p_i = P(X = a_i)$ von X aus den Daten $(x_1, \ldots, x_n)$ grafisch dargestellt werden.
- Ähnliche Darstellungen sind für die Dichte $f(x)$ von absolut stetigen Merkmalen X möglich, wobei ihr Wertebereich C sich in k Klassen aufteilen lässt: $(c_{i-1}, c_i] \,, \quad i = 1, \ldots, k$, wobei $c_0 = -\infty$, $c_1 < \ldots < c_{k-1}$, $c_k = \infty$ ist.
- Dann kann die Zähldichte $p_i = P\big(X \in (c_{i-1}, c_i]\big)$ gegeben durch
$$p_i = \int_{c_{i-1}}^{c_i} f(x)\, dx \,, \qquad i = 0, \ldots, k$$
betrachtet werden.

Diagramme und Histogramme

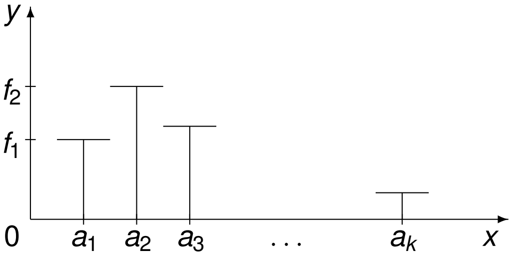
- Man unterscheidet bei der Betrachtung der Häufigkeit einer Merkmalsausprägung im Allgemeinen zwei Fälle:

1. Die *absolute Häufigkeit* von Merkmalsausprägung a_i bzw. die Klasse $(c_{i-1}, c_i]$, $i = 1, \ldots, k$ ist
$n_i = \#\{x_j, j = 1, \ldots, n : x_j = a_i\}$ bzw.
$n_i = \#\{x_j, j = 1, \ldots, n : \ x_j \in (c_{i-1}, c_i]\}$.
2. Die *relative Häufigkeit* von Merkmalsausprägung a_i bzw. Klasse $(c_{i-1}, c_i]$ ist $f_i = \frac{n_i}{n}$, $i = 1, \ldots, k$.

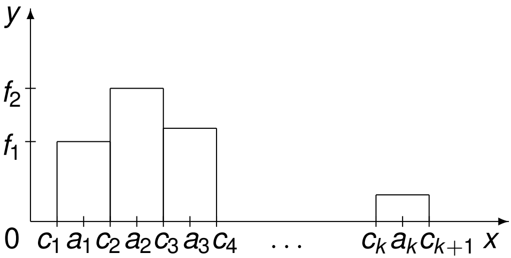
Diagramme und Histogramme

- Es gilt offensichtlich
  $n = \sum_{i=1}^{k} n_i\,, \quad 0 \le f_i \le 1\,, \quad \sum_{i=1}^{k} f_i = 1.$
- Die absoluten und relativen Häufigkeiten werden oft in Häufigkeitstabellen zusammengefasst.
- Zu ihrer Visualisierung dienen so genannte *Diagramme*.
- Eine wichtige Klasse von Diagrammen stellen *Histogramme* dar.
- Diese werden gebildet, indem man die Paare (a_i, f_i) (bzw. $({}^1\!/\!{}_2(c_1 + x_{(1)}), f_1)$, $\big({}^1\!/\!{}_2(c_{i-1} + c_i), f_i\big)$, $i = 2, \dots, k-1$, $({}^1\!/\!{}_2(c_{k-1} + x_{(n)}), f_k)$ im absolut stetigen Fall, wobei hier die Bezeichnung $a_i = {}^1\!/\!{}_2(c_{i-1} + c_i)$ verwendet wird und $x_{(1)} < c_1, \quad x_{(n)} > c_{k-1}$ angenommen wird.) auf der Koordinatenebene (x, y) folgendermaßen aufträgt:

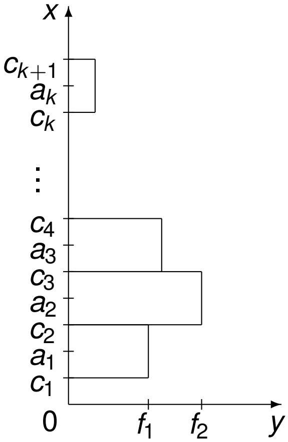
- *Stabdiagramm:* f_i wird als Höhe des senkrechten Strichs über a_i dargestellt:



- *Säulendiagramm:* genauso wie ein Stabdiagramm, nur werden Striche durch Säulen der Form $(c_{i-1}, c_i] \times f_i$ ersetzt, wobei im diskreten Fall die Aufteilung der reellen Achse $-\infty = c_0 < c_1 < c_2 < \ldots < c_{k-1} < c_k = \infty$ in Intervalle beliebig vorgenommen werden kann.



- *Balkendiagramm:* genauso wie Säulendiagramm, nur mit vertikaler statt horizontaler x-Achse.



Empirische Verteilungsfunktionen

- Es sei eine konkrete Stichprobe $(x_1, \ldots, x_n)$ gegeben, die eine Realisierung des statistischen Modells $(X_1, \ldots, X_n)$ ist, wobei $X_1, \ldots, X_n$ unabhängige identisch verteilte Zufallsvariablen mit Verteilungsfunktion F_X : $X_i \stackrel{d}{=} X \sim F_X$ sind.
- Wie kann die unbekannte Verteilungsfunktion F_X aus den Daten $(x_1, \ldots, x_n)$ rekonstruiert (die Statistiker sagen "geschätzt") werden?
- Dies ist mit Hilfe der sogenannten empirischen Verteilungsfunktion möglich:

Definition 6.3.1.

1. Die Funktion
   $\hat{F}_n(x) = \#\{x_i : x_i \leq x , i = 1, \ldots , n\}/n , \quad \forall x \in \mathbb{R}$ heißt
   empirische Verteilungsfunktion der konkreten Stichprobe $(x_1, \ldots , x_n)$.

   Dabei gilt $\hat{F}_n : \mathbb{R}^{n+1} \to [0, 1]$, weil $\hat{F}_n(x) = \varphi(x_1, \ldots , x_n, x)$.
2. Die mit $x \in \mathbb{R}$ indizierte Zufallsvariable $\hat{F}_n : \Omega \times \mathbb{R} \to [0, 1]$

   heißt
   empirische Verteilungsfunktion der Zufallsstichprobe $(X_1, \ldots , X_n)$, wenn

   $$\hat{F}_n(x, \omega) = \hat{F}_n(x) = \frac{1}{n}\#\{X_i, i = 1, \ldots , n : X_i(\omega) \leq x\} , \quad x \in \mathbb{R} .$$

- Äquivalent zur obigen Definition kann man

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \le x)\,, \quad x \in \mathbb{R}$$

schreiben, wobei

$$I(x \in A) = \begin{cases} 1, & x \in A \\ 0, & \text{sonst.} \end{cases}$$

- Es gilt

$$\hat{F}_n(x) = \begin{cases} 1, & x \ge x_{(n)}\,, \\ \frac{i}{n}, & x_{(i)} \le x < x_{(i+1)}\,, \quad i = 1, \ldots, n-1\,, \\ 0, & x < x_{(1)}\,. \end{cases}$$

für $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$.

- Dabei ist die Höhe des Sprungs an Stelle $x_{(i)}$ gleich der relativen Häufigkeit f_i des Wertes $x_{(i)}$.
- Falls $x_{(i)} = x_{(i+1)}$ für ein $i \in \{1, \ldots, n\}$, so tritt der Wert $\frac{i}{n}$ nicht auf.
- In der folgenden Abbildung sieht man, dass $\hat{F}_n(x)$ eine rechtsstetige monoton nichtfallende Treppenfunktion ist, für die $\hat{F}_n(x) \underset{x \to -\infty}{\longrightarrow} 0$, $\hat{F}_n(x) \underset{x \to \infty}{\longrightarrow} 1$ gilt.

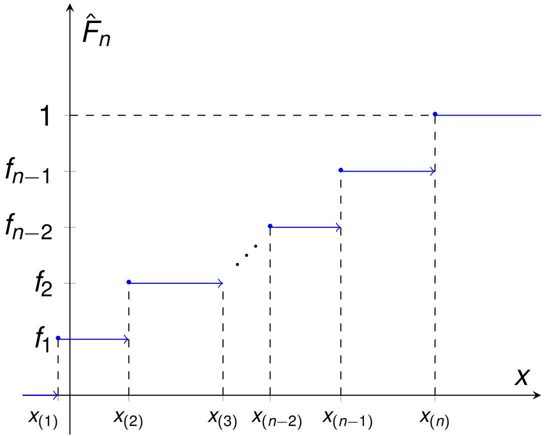


Figure: Eine typische empirische Verteilungsfunktion

Übungsaufgabe 6.3.2.

Zeigen Sie, dass $\hat{F}_n(x)$ eine Verteilungsfunktion ist.