Deskriptive Statistik

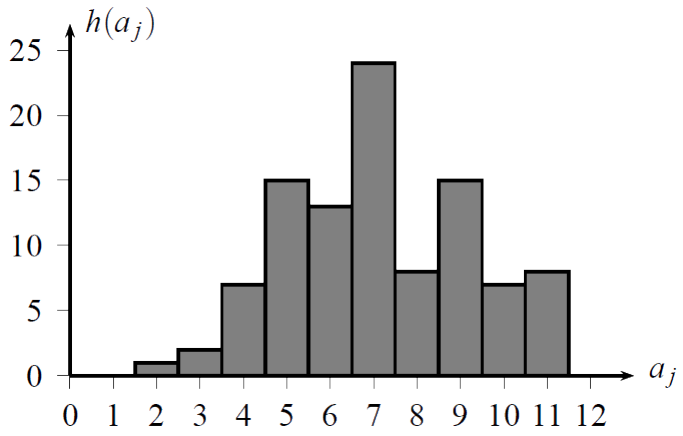
Contents

1 Graphische Darstellung

1.1 Stamm-Blatt-Diagramm

4	|	3,4,6,1,2,3,8,7,8,8		77 ⇒	0|77	⟶0|8
5	|	4,7,6		104 ⇒	1|04	⟶1|0
6	|	1,2,5,4,3,3	oder mit runden	132 ⇒	1|32	⟶1|3
7	|	9,0,0,2,1,8,4		227⇒	2|27	⟶2|3
8	|	1,2				

1.2 Histogramme



1.2.1 Häufigkeiten

Umfang der Stichprobe: n

absolute Häufigkeit: $h(a_j) =$ Anzahl der Ausprägungen in der Beobachtungsmenge

relative Häufigkeit: $f(a_j) = \frac{Anzahl\ der\ Ausprägungen\ in\ der\ Beobachtungsmenge}{Grösse\ der\ Beobachtungsmenge\ (Umfang\ der\ Stichprobe)} = \frac{h(a_j)}{n}$

absolute Summenhäufigkeit: $G(x) = \sum_i^n h(x)$

Verteilungsfunktion: $H(x) = \sum f(x) = \frac{1}{n} \sum h(x) = \frac{G(x)}{n}$

Beispiel

Zwei Würfel werden 1000 mal geworfen:

Ausprägung x	Häufigkeit $h(x)$	relative Häufigkeit $f(x)$	absolute Summenhäufigkeit $G(x)$	Verteilungsfunktion $H(x)$
2	12	0.012	12	0.012
3	46	0.046	58	0.058
4	83	0.083	141	0.141
5	103	0.103	244	0.244
6	160	0.160	404	0.404
7	180	0.180	584	0.584
8	159	0.159	743	0.743
9	125	0.125	868	0.868
10	77	0.077	945	0.945
11	43	0.043	988	0.988
12	12	0.012	1000	1.000

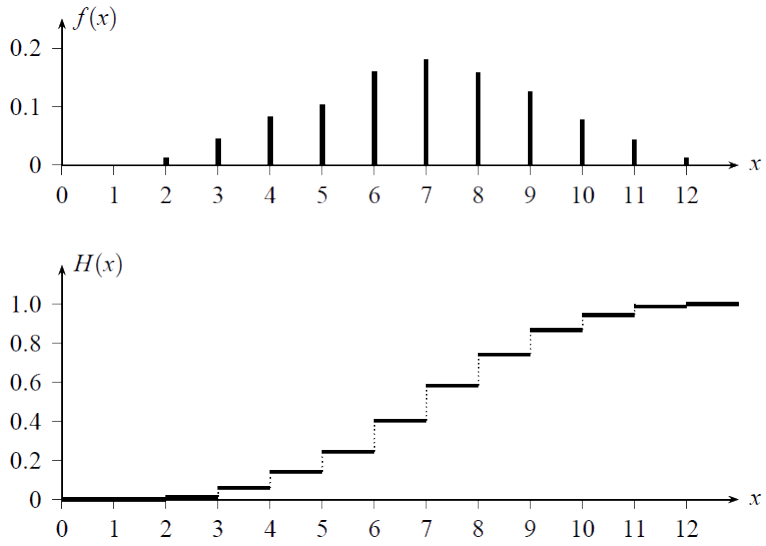
1.2.2 Klassen, Klassenbreite

Umfang der Stichprobe: n

Anzahl Klassen: $\sqrt{n}$

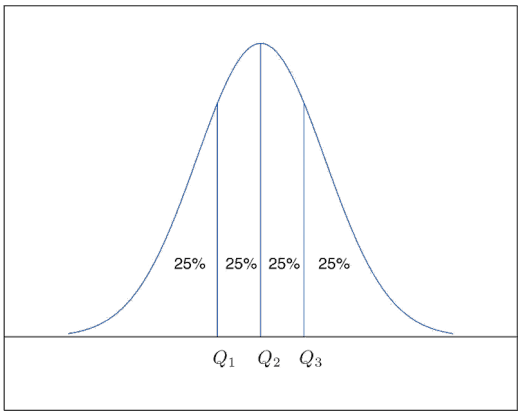
Klassenbreite: $\sqrt{n}$ oder $10log_{10}(n)$

1.3 Stabdiagramm



2 Arbeiten mit Quartilen

Quartile teilen die Grundgesamtheit in 4 gleich grosse Teile.



2.1 1. Quartil

$$Q_1 = \frac{1}{2} \cdot (1 + \frac{n+1}{2}) = \frac{n+3}{4}$$

Sollte das Quartil zwischen zwei Indizes liegen ($n_1 \leq Q_1 \leq n_2$) so gilt:

$$Q_1 = (x_{n2} - x_{n1}) \cdot \frac{n+3}{4} + x_{n1} \cdot n_2 - x_{n2} \cdot n_1$$

2.2 2. Quartil (Median)

n gerade:

$$Q_2 = \frac{1}{2} \cdot (x_{\frac{n}{2}} + x_{\frac{n}{2}+1})$$

n ungerade:

$$Q_2 = x_{\frac{n+1}{2}}$$

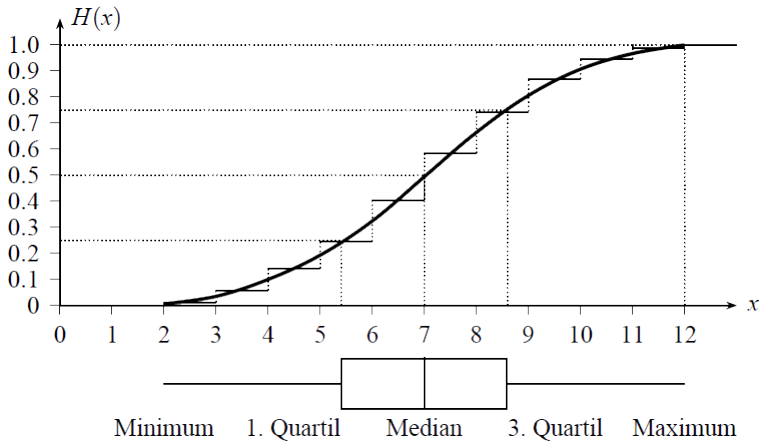
2.3 3. Quartil

$$Q_2 = \frac{1}{2} \cdot (\frac{n+1}{2} + n) = \frac{3n+1}{4}$$

Sollte das Quartil zwischen zwei Indizes liegen ($n_1 \leq Q_3 \leq n_2$) so gilt:

$$Q_1 = (x_{n2} - x_{n1}) \cdot \frac{3n+1}{4} + x_{n1} \cdot n_2 - x_{n2} \cdot n_1$$

2.4 Boxplot



3 Statistische Masszahlen

3.1 Arithmetisches Mittel

$$\bar{x} = \frac{1}{n}\sum x_i = \frac{1}{n}(n_1 a_1 + ... + n_k a_k) = \mu$$

Für Häufigkeitsdaten mit Ausprägungen $a_1, ..., a_k$ und relativen Häufigkeiten $f(a_1), ..., f(a_k)$ gilt:

$$\bar{x} = \sum_{i=1}^{k} a_i \cdot f(a_i) = a_1 \cdot f(a_1) + ... + a_k \cdot f(a_k) = \mu$$

Beispiel mit den Würfeln:

$$\bar{x} = \sum_{x=2}^{12} x f(x) = 7.013$$

3.2 Modalwert (Modus)

Der Modalwert gibt an, welche Ausprägung am häufigsten vorkommt.

- x_{mod} bei diskreten Merkmalen: Ausprägung mit grösster Häufigkeit
- x_{mod} bei stetigen Merkmalen: Maximum der Parabel, resp. die Mitte der stärkst besetzten Klasse

4 Streungsmasse

4.1 Empirische Varianz

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

4.2 Empirische Standardabweichung

$$s = \sqrt{s^2} = \sigma$$

Dies hat folgende bedeutung:

- etwa 68.3% aller Messwerte liegen im Intervall $[\bar{x} - s, \bar{x} + s]$
- etwa 95.5% aller Messwerte liegen im Intervall $[\bar{x} - 2s, \bar{x} + 2s]$
- etwa 99.7% aller Messwerte liegen im Intervall $[\bar{x} - 3s, \bar{x} + 3s]$

5 Dichtekurven und Normalverteilung

5.1 Normalverteilung

$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\frac{(x-\mu)^2}{2\sigma^2}}$
$\sigma = s$, $\mu = \bar{x}$

Je kleiner σ ist, desto steiler ist der Abfall und umso enger ist die Kurve um das Mittel μ konzentriert.
Je grösser σ ist, desto mehr Fläche liegt weiter links oder rechts von μ und umso grösser ist die Streeung der x-Werte.

Erwartungswert

$E(X) = \mu = \bar{x}$

Varianz

$VAR(X) = \sigma^2$

5.2 Dichtekurve

Eine Funktion $f(x)$ ist eine Dichtekurve, wenn $f(x) \geq 0$ ist und die von $f(x)$ überdeckte Gesamtfläche gleich 1 ist, also gilt.

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

Erwartungswert

$$E(X) = \int_{-\infty}^{\infty} t \cdot f(t)dt$$

Varianz

$Var(X) = E(X^2) - [E(X)]^2$

Beispiel

$$f(x) = \left\{ \begin{array}{c|c} \frac{11}{54}x & 0 \leq x \leq 2 \\ -\frac{8}{27}x + 1 & 2 \leq x \leq 3 \\ -\frac{2}{27}x + \frac{1}{2} & 3 \leq x \leq 4.5 \\ 0 & sonst \end{array} \right\}$$

Die Funktion $f(x)$ ist eine Dichte, denn $f(x) \geq 0$ für $0 \leq x \leq 4.5$ und

$$\int_{-\infty}^{\infty} f(x)dx = \int_{0}^{2} \frac{11}{54}xdx + \int_{2}^{3} (-\frac{8}{27}x + 1)dx + \int_{3}^{4.5} (-\frac{2}{27}x + \frac{1}{2})dx = 1$$

6 Mehrdimensionale Verteilungen

6.1 Lineare Regression

Finde eine gerade $y = mx + c$ die möglichst eng von der Punktewolke umschlossen wird.

$$m = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \text{ und } c = \bar{y} - m\bar{y}$$

6.2 Empirische Kovarianz

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

dies entspricht auch:

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

6.3 Empirischer Korrelationskoeffizient

$$r(x,y) = \frac{1}{n}\sum_{i=1}^{n} \frac{(x_i-\bar{x})}{s(x)} \cdot \frac{(y_i-\bar{y})}{s(y)}$$

wobei $s(x)$ und $s(y)$ die numerische Standartabweichug ist.

7 Binomialverteilung

Wird gebraucht wenn nur 2 Ereignisse eintreten können.

$$q(x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x}$$

n = Anzahl wiederholungen

p = Wahrscheinlichkeit des Ereignisses

Erwartungswert

$$E(X) = \sum_{x \in \Omega'} q(x) \cdot x = n \cdot p$$

Streuung

$$\sigma = \sqrt{np(1-p)}$$

8 Geometrische Verteilung

Wird verwendet für Lebensdauer oder Wartezeiten

$$q(x) = (1-p)^{x-1} p$$

Erwartungswert

$$E(X) = \frac{1}{p}$$

Streuung

$$\sigma = \sqrt{\frac{1-p}{p^2}}$$

9 Hypergeometrische Verteilung

Verwendung:
Eine Menge bestehe aus N Teilen unter denen sich M Teile mit einem Merkmal A befinden.
Man entnimmt n Teile ohne Zurücklegen. x zählt die Anzahl der Teile in n mit Merkmal A.

$$q(x) = \frac{\left(\begin{array}{c} M \\ x \end{array} \right) \left(\begin{array}{c} N-M \\ n-x \end{array} \right)}{\left(\begin{array}{c} N \\ n \end{array} \right)}$$

Erwartungswert

$$E(X) = n\frac{M}{N}$$

Streuung

$$\sigma = \sqrt{n\frac{M}{N}(1-\frac{M}{N})(1-\frac{n-1}{N-1})}$$

10 Poisson Verteilung

Verwendung:

- Anzahl Unfälle in einer bestimmten Region und pro Zeiteinheit
- Anzahl Anrufe pro Zeiteinheit
- Irgendetwas pro Zeiteinheit

$$q(x) = \frac{\mu^x}{x!} e^{-\mu}$$

Erwartungswert

$$E(X) = \mu$$

Streuung

$$\sigma^2 = \mu$$

11 Gleichverteilung

Unter der Gleichverteilung (Rechteckverteilung/uniforme Verteilung) über dem Intervall (a,b) versteht man die Verteilung einer Variablen X, deren Dichte gegeben ist durch.

$$f(x) = \left\{ \begin{array}{cc} \frac{1}{b-a} & für\ a < x < b \\ 0 & sonst \end{array} \right\}$$

Erwartungswert

$$E(X) = \frac{a+b}{2}$$

Varianz

$$Var(x) = \frac{(b-a)^2}{12}$$

12 Punktschätzer

Unbekannte Parameter einer Zufallsvariable

- Erwartungswert
- Varianz

12.1 Erwartungswert schätzen

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

12.2 Varianz schätzen

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