

1 Beschreibende/Deskriptive Statistik

1.1 Begriffe

]

1.1.1 Beschreibende/Deskriptive Statistik

Beobachtete Daten werden durch geeignete statistische Kennzahlen charakterisiert und durch geeignete Grafiken anschaulich gemacht.

1.1.2 Schließende/Induktive Statistik

Aus beobachtete Daten werden Schlüsse gezogen und diese im Rahmen vorgegebener Modelle der Wahrscheinlichkeitstheorie bewertet.

1.1.3 Grundgesamtheit

Ω : Grundgesamtheit ω : Element oder Objekt der Grundgesamtheit diskret (< 30 Ausprägungen), stetig (≥ 30 Ausprägungen), univariat ($p=1$), multivariat ($p>1$)

1.2 Lagemaße

1.2.1 Modalwerte x_{mod}

Am häufigsten auftretende Ausprägungen (insbesondere bei qualitativen Merkmalen)

1.2.2 Mittelwert

R: $mean(x)$
Schwerpunkt der Daten.
Empfindlich gegenüber Ausreißern.

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

1.3 Median

R: $median(x)$
Liegt in der Mitte der sortierten Daten x_i .
Unempfindlich gegenüber Ausreißern.

$$x_{0.5} = \begin{cases} x_{\frac{n+1}{2}}, & \text{falls } n \text{ ungerade} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}), & \text{falls } n \text{ gerade} \end{cases} \quad (1)$$

1.4 Streuungsmaße

1.4.1 Spannweite

$$\max x_i - \min x_i$$

1.4.2 Stichprobenvarianz s^2

R: $var(x)$

Verschiebungssatz:

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

$n\bar{x}^2$) Gemittelte Summe der quadratischen Abweichung vom Mittelwert

1.4.3 Stichprobenstandardabweichung

R: $sd(x)$

$s = \sqrt{s^2}$ Streuungsmaß mit gleicher Einheit wie beobachteten Daten x_i . \bar{x} minimiert die "quadratische Verlustfunktion" oder die Varianz gibt das Minimum der Fehlerquadrate an.

1.5 p-Quantile

R: $quantile(x, p)$. Teilt die sortierten Daten x_i ca. im Verhältnis p : $(1-p)$ d.h. $\hat{F}(x_p) \approx p$; 1. Quartil = 0.25-Quantil; Median = 0.5-Quantil; 3. Quartil = 0.75-Quantil;

1.6 Interquartilsabstand I

$I = x_{0.75} - x_{0.25}$. Ist ein weiterer Streuungsparameter.

1.7 Chebyshev

$\frac{N(S_k)}{n} > 1 - \frac{1}{k^2}$, für alle $k \geq 1$ \bar{x} der Durchschnitt, $s > 0$ die Stichproben-Standardabweichung von Beobachtungswerten x_1, \dots, x_n . Sei $S_k = \{i, 1 \leq i \leq n : |x_i - \bar{x}| < k \cdot s\}$; Für eine beliebige Zahl $k \geq 1$ liegen mehr als $100 \cdot (1 - \frac{1}{k^2})$ Prozent der Daten im Intervall von $\bar{x} - ks$ bis $\bar{x} + ks$. **Speziell:** Für $k=2$ liegen mehr als 75% der Daten im 2s-Bereich um \bar{x} . Für $k=3$ liegen mehr als 89% der Daten im 3s-Bereich um \bar{x} . **Komplement Formulierung:** $\bar{S}_k = \{i || x_i - \bar{x}| \geq k \cdot s\}$; $\frac{N(\bar{S}_k)}{n} \leq \frac{1}{k^2}$;

Die Ungleichheit liefert nur eine sehr grobe Abschätzung, ist aber unabhängig von der Verteilung der Daten. **Empirische Regeln** 68% der Daten im Bereich um $\bar{x} \pm s$. 95% um $\bar{x} \pm 2s$. 99.7% um $\bar{x} \pm 3s$.

1.8 Korrelation

Grafische Zusammenhang zwischen multivariaten Daten x und y durch ein Streudiagramm. Kennzahlen zur Untersuchung des Zusammenhangs:

1.8.1 Empirische Kovarians

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

1.8.2 Empirische Korrelationskoeffizient r

R: $cor(x, y)$; $r = \frac{s_{xy}}{s_x s_y}$; Näherungsweise lin. Zusammenhang zw. x und y , falls $|r| \approx 1$.

1.8.3 Regressionsgerade y

$$y = mx + t \text{ mit } m = r \cdot \frac{s_y}{s_x} \text{ und } t = \bar{y} - m \cdot \bar{x}$$

2 Wahrscheinlichkeitsrechnung

2.1 Begriffe

Ergebnisraum Ω : Menge aller möglichen Ergebnisse eines Experiments

Elementarereignis $\omega \in \Omega$: einzelnes Element von Ω

Ereignis $E \subseteq \Omega$: beliebige Teilmenge des Ergebnisraums Ω heißt sicheres Ereignis, \emptyset heißt unmögliches Ereignis

Vereinigung $E \cup F$: Ereignis E oder Ereignis F treten ein. $\bigcup_{i=1}^n E_i$: mindestens ein Ereignis E_i tritt ein.

Schnitt $E \cap F$: Ereignis E und Ereignis F treten ein.

$\bigcap_{i=1}^n E_i$ alle Ereignisse E_i treten ein. **Generereignis** $\bar{E} = \Omega / E$: Ereignis E tritt nicht ein (Komplement von E)

Disjunkte Ereignisse E und F: $E \cap F = \emptyset$

2.2 De Morgan'schen Regeln

$$\overline{E_1 \cup E_2} = \bar{E}_1 \cap \bar{E}_2$$

$$\overline{E_1 \cap E_2} = \bar{E}_1 \cup \bar{E}_2$$

2.3 Wahrscheinlichkeit

$$0 \leq P(E) \leq 1; P(\Omega) = 1;$$

$$P(\bigcup_{i=1}^n E_i) = \sum_{i=1}^n P(E_i), \text{ falls } E_i \cap E_j = \emptyset \text{ für } i \neq j$$

2.3.1 Satz 2.1

$$P(\bar{E}) = 1 - P(E)$$

$$P(E \cup F) = P(E) + P(F) - P(E \cap F)$$

(Übungsaufgabe!!! Ergänzen)

2.4 Laplace-Experiment

Zufallsexperimente mit n gleich wahrscheinlichen Elementarereignissen. Dann berechnet sich die Wahrscheinlichkeit $P(E)$ für $E \subseteq \Omega$ aus:

$$P(E) = \frac{\text{Anzahl der für } E \text{ günstigen Ereignisse}}{\text{Anzahl der möglichen Ereignisse}} = \frac{\text{Mächtigkeit von } E}{\text{Mächtigkeit von } \Omega} = \frac{|E|}{|\Omega|} \text{ text}$$

2.5 Kombinatorik

2.5.1 Allgemeines Zählprinzip

Anzahl der Möglichkeiten für ein k -stufes Zufallsexperiment mit n_i Varianten im i -ten Schritt: $n_1 \cdot n_2 \cdot \dots \cdot n_k$

2.5.2 Permutationen

Anzahl einer n -elementigen Menge n -maliges Ziehen ohne Zurücklegen mit Beachtung der Reihenfolge: **n unterscheidbare Elemente:** $n! = n \cdot (n-1) \cdot \dots \cdot 2 \cdot 1$
k Klassen mit je n_i nicht unterscheidbaren Elementen $n = \sum_{i=1}^k n_i$; $\frac{n!}{n_1! \cdot n_2! \cdot \dots \cdot n_k!}$

2.5.3 Abzähl k-elementigen Teilmengen einer n-elementigen Menge k-maliges Ziehen aus einer n-elementigen Menge

ohne Zurücklegen = $k \leq n$.

mit Zurücklegen = $k > n$ möglich.

mit Beachtung der Reihenfolge, ohne

Zurücklegen: $\frac{n!}{(n-k)!}$

ohne Beachtung der Reihenfolge, ohne

Zurücklegen: $\binom{n}{k} = \frac{n!}{(n-k)! \cdot k!}$

mit Beachtung der Reihenfolge, mit Zurücklegen: n^k

ohne Beachtung der Reihenfolge, mit Zurücklegen: $\binom{n+k-1}{k}$

2.6 Bedingte Wahrscheinlichkeit

$$P(E|F) = P_F(E) = \frac{|E \cap F|}{|F|} = \frac{P(E \cap F)}{P(F)}$$

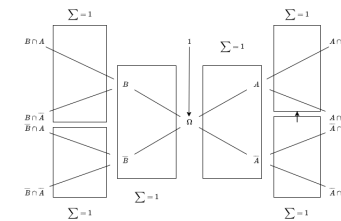
$$P(E \cap F) = P(F) \cdot P(E|F)$$

$$P(E \cap F) = P(F|E) \cdot P(E)$$

2.6.1 Satz 2.2

$$P(E \cap F) = P(E|F) \cdot P(F)$$

$$P(E \cap F) = P(F|E) \cdot P(E)$$



2.6.2 Satz der totalen Wahrscheinlichkeit

Sei $\Omega = \bigcup_{i=1}^n E_i$ mit $E_i \cap E_j = \emptyset$ für $i \neq j$ d.h. die Ereignisse bilde eine disjunkte Zerlegung bzw. eine Partition von Ω . Somit gilt:

$$P(F) = \sum_{i=1}^n P(F \cap E_i) = \sum_{i=1}^n P(F|E_i) \cdot P(E_i)$$

Summe der Äste des Wahrscheinlichkeitsbaums zu allen Schnitten $F \cap E_i$



2.6.3 Vierfeldertafel

$$P(F) = P(F \cap E) + P(F \cap \bar{E})$$

$$P(\bar{F}) = P(\bar{F} \cap E) + P(\bar{F} \cap \bar{E})$$

	E	\bar{E}	
F	$P(F \cap E)$	$P(F \cap \bar{E})$	$P(F)$
\bar{F}	$P(\bar{F} \cap E)$	$P(\bar{F} \cap \bar{E})$	$P(\bar{F})$
	$P(E)$	$P(\bar{E})$	1

Satz 2.2 $P(E \cap F) = P(E) \cdot P(F|E)$
 $P(F|E) = P(F) \cdot P(E|F)$ **Tafel** $P(F) = P(F \cap E) + P(F \cap \bar{E})$
 $P(\bar{F}) = P(\bar{F} \cap E) + P(\bar{F} \cap \bar{E})$; $P(\bar{F}|E) = 1 - P(F|E)$

2.6.4 Formel von Bayes

Hilfreich, wenn man man $P(F|E_i)$ kennt, aber nicht $P(E_k|F)$ **Satz 2.4** $P(E_k|F) =$

$$\frac{P(F|E_k) \cdot P(E_k)}{\sum_{i=1}^n P(F|E_i) \cdot P(E_i)}$$

Nur Nenner! $P(F)$ aus dem Satz der totalen Wahrscheinlichkeit.

2.6.5 Stochastische Unabhängigkeit

Uebung Die Ereignisse E und F heißen (stochastisch) unabhängig, wenn die Information über das Eintreten des einen Ereignisses die Wahrscheinlichkeit für das Eintreten des anderen Ereignisses nicht ändert, d.h. falls

$$P(E|F) = P(E) \text{ oder } P(E \cap F) = P(E) \cdot P(F)$$

$$= \frac{P(E \cap F)}{P(F)}$$

Es gilt Falls die Ereignisse E, F unabhängig sind, dann sind auch:

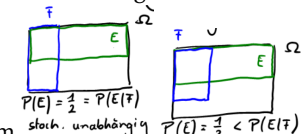
$$E, \bar{F}$$

$$\bar{E}, F$$

$$\bar{E}, \bar{F}$$

unabhängig **Bemerkung**

- Stochastische Unabhängigkeit bedeutet nicht notwendigerweise eine kausale Abhängigkeit
- Veranschaulichung mit Venn Diagramm



gramm stoch. unabhängig $P(E) = \frac{1}{2} < P(E|F)$

- $A, B \neq \emptyset$ und $A \cap B = \emptyset$
 $P(A \cap B) \stackrel{?}{=} P(A) \cdot P(B)$
 $\emptyset \neq P(A) \cdot P(B)$ da $P(A) > 0$ und $P(B) > 0$
 $\Rightarrow A, B$ stochastisch abhängig

3 Zufallsvariable

Abbildung des abstrakten Ergebnisraums Ω auf \mathbb{R} . Eine Abbildung $X: \Omega \rightarrow \mathbb{R}$, $\omega \mapsto X(\omega)$ heißt Zufallsvariable (ZV). $x \in \mathbb{R}$ heißt Realisation der ZV X .

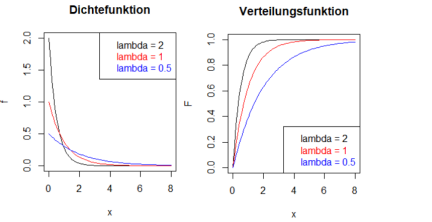
- Diskrete ZV: $X(\Omega) = x_1, \dots, x_2 (n \in \mathbb{N})$; z.B. X = "Augensumme beim Würfeln"
- Stetige ZV: $X(\Omega) \subseteq \mathbb{R}$; "z.B. Körpergröße eines Menschen"

3.1 Verteilungsfunktion-allg.

Die Wahrscheinlichkeit $P(B)$ für ein Ereignis B in \mathbb{R} wird zurückgeführt auf die Wahrscheinlichkeit der entsprechenden Ereignisse in Ω . Für jedes $X \in \mathbb{R}$ ist die

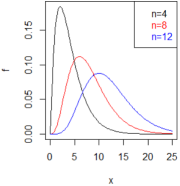
4.2.4 Exponentialverteilung

Modellierung von Lebensdauern, Wartezeiten Sei $Y_t \sim P_{\lambda t}$ im Intervall $[0, t]$ von t Zeiteinheiten, dann beschreibt die Exponentialverteilung die Wartezeit X bis zum Eintreten eines Ereignisses; **Dichte- und Verteilungsfunktion:** $f(x) = \lambda e^{-\lambda x} (x \geq 0)$ und $F(x) = 1 - e^{-\lambda x}$; **Verteilung:** $X \sim Exp_{\lambda}$; $E[X] = \frac{1}{\lambda} \Rightarrow$ Berechnung mit partieller Integration; $Var[X] = \frac{1}{\lambda^2}$; **R:** $dexp(x, \lambda) = f(x)$; $pexp(x, \lambda) = F(x)$; **Eigenschaft:** Eine exponentialverteilte ZV X ist gedächtnislos, d.h. $P(X > s + t) | X > t = P(X > s)$;



4.2.5 Chiquadrat-Verteilung

Z_1, \dots, Z_n seien unabhängige, standard-normalverteilte ZV $\Rightarrow X = Z_1^2 + \dots + Z_n^2$ hat Chiquadratverteilung mit n Freiheitsgraden; **Anwendungsmodell:** Summen unabhängiger, standardnormalverteilter ZV; **Verteilung:** $X \sim \chi_n^2$; $E[X] = n$; $Var[X] = 2n$; **R:** $dchisq(x, n) = f(x)$; $pchisq(x, n) = F(x)$; **Eigenschaft:** $X_1 \sim \chi_{n_1}^2$ und $X_2 \sim \chi_{n_2}^2 \Rightarrow X_1 + X_2 \sim \chi_{n_1+n_2}^2$



4.2.6 t-Verteilung

$Z \sim N_{0,1}$ und $X \sim \chi_n^2 \Rightarrow Y = \frac{Z}{\sqrt{X/n}}$ ist t-verteilt mit n Freiheitsgraden; **Anwendungsmodell:** Schätz- und Testverfahren bei unbekannter Varianz; **Verteilung:** $Y \sim t_n$; $E[Y] = 0$ für $n > 1$; $Var[Y] = \frac{n}{n-2}$ für $n > 2$; **R:** $dt(y, n) \hat{=} f(x)$; $pt(y, n) \hat{=} F(x)$; **Eigenschaften:** Für $n \rightarrow \infty: t_n \rightarrow N_{0,1}$; Achsensymmetrie der Dichtefunktion $\Rightarrow -y_p = x_{1-p}$

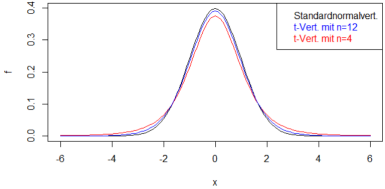


Abbildung Dichtefunktion

5 Zentraler Grenzwertsatz

$\mu \sigma^2$ bekannt aber nicht die Verteilung

5.1 ZGW

Seien $X_i (i = 1, \dots, n)$ unabhängige identische verteilte (**i.i.d**) ZV mit Erwartungswert μ und Varianz σ^2 . Dann gilt für hinreichend große n (> 30) und $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ näherungsweise:

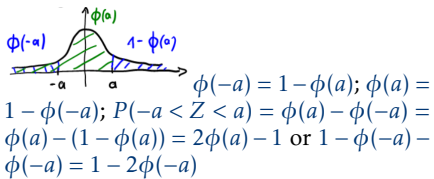
$$\sum_{i=1}^n X_i \sim N_{n\mu, n\sigma^2} \text{ \& }$$

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n} \cdot \sigma} \sim N_{0,1}$$

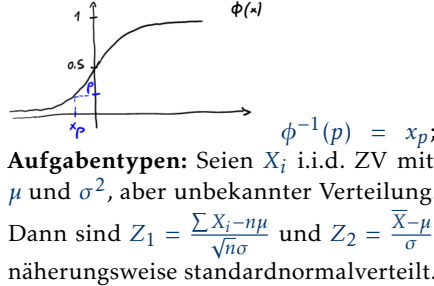
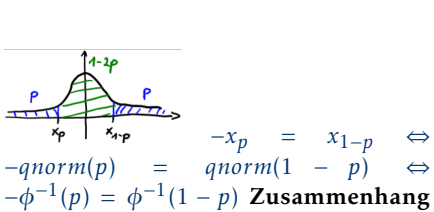
$\sum X_i$ bezieht sich auf Y ; $\sum X_i - n\mu$ bezieht sich auf X_i ; $\bar{X} \sim N_{\mu, \frac{\sigma^2}{n}}$ & $\frac{\bar{X} - \mu}{\sigma} \sim N_{0,1}$;

Der Satz gilt sogar allgemeiner, wenn die X_i abhängig und nicht identisch verteilt sind, vorausgesetzt kein X_i ist deutlich dominanter?! als die anderen. Für die Voraussetzung des ZGW ist, dass die X_i nicht normalverteilt sein müssen., damit $\sum_{i=1}^n X_i$ oder \bar{X} bei **hinreichend großem n** normalverteilt sind. Faustregel: **Je** schiefer die Verteilung der X_i , **desto** größer muss n sein: **$n > 30$:** falls die unbekannte Verteilung ohne markanten Ausreißer, aber schief ist (Exponentialverteilung); **$n > 15$:** falls die unbekannte Verteilung annähernd symmetrisch ist (Binomialverteilung); **$n \leq 15$:** falls die unbekannte Verteilung annähernd normalverteilt ist;

5.2 ϕ



5.3 $\phi(-1)$



Aufgabentypen: Seien X_i i.i.d. ZV mit μ und σ^2 , aber unbekannter Verteilung. Dann sind $Z_1 = \frac{\sum X_i - n\mu}{\sqrt{n}\sigma}$ und $Z_2 = \frac{\bar{X} - \mu}{\sigma}$ näherungsweise standardnormalverteilt.

- Es lassen sich Wahrscheinlichkeiten für $\sum X_i, \bar{X}, Z_1$ oder Z_2 berechnen.
- Es lässt sich **n** bestimmen, so dass, zu vorgegebener Schranke k und Wahrscheinlichkeit p gilt: $P(Z_i > k) \geq p$ or $P(-k \leq Z_i \leq k) \geq p$

5.4 Stichprobenverteilungen für normalverteilte Grundgesamtheiten

5.4.1 Stichprobenmittel

Die Stichprobenfunktion $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ ist eine erwartungstreu Schätzfunktion für Erwartungswert μ , d. h. $E[\bar{X}] = \mu$

5.4.2 Stichprobenvarianz

Die Stichprobenfunktion $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} (\sum_{i=1}^n X_i^2 - n\bar{X}^2)$ ist eine erwartungstreu Schätzfunktion für die Varianz σ^2 , d. h. $E[S^2] = \sigma^2$; $E[\bar{X}] = E[\frac{1}{n} \sum X_i] = \frac{1}{n} E[\sum X_i] = \frac{1}{n} \sum_{i=1}^n E[X_i] = \frac{1}{n} n\mu = \mu$; $Var[\bar{X}] = Var[\frac{1}{n} \sum X_i] = \frac{1}{n^2} Var[\sum X_i] = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$; Seien $X_i (i = 1, \dots, n)$ unabhängige normalverteilte ZV mit Erwartungswert μ und Varianz σ^2 . Dann gilt: **bei unbekannter Varianz:** $\frac{\bar{X} - \mu}{\sigma} \sqrt{n} \sim N_{0,1}$; $\frac{(n-1)S^2 = \sum (x - \bar{x})^2}{\sigma^2 \Rightarrow \text{Standardisierung}} \sim \chi_{n-1}^2$; **Bei unbekannter Varianz:** $\frac{\bar{X} - \mu}{S} \sqrt{n} \sim t_{n-1}$;