

I. Deskriptive Statistik

1. Grundbegriffe der Datenerhebung

Die wichtigsten Grundbegriffe der beschreibenden Statistik sollen anhand der folgenden Beispiele erläutert werden:

Begriff	Beispiele						
Grundgesamtheit	Bevölkerung von D			Ratten		Motoren	
Merkmalsträger	Ein Einwohner			Eine Ratte		Ein einzelner Motor	
Merkmal	Augenfarbe	Größe	Name	Gewicht	Geschlecht	Verbrauch	Lebensdauer
Merkmalsausprägung	blau	1,7 m	Max	1,2 kg	weiblich	6,5l	120.000 km

Statistische Mess-Skalen:

Nominalskala:

Die Merkmalsausprägungen sind Namen oder Bezeichnungen, die nur nach dem Kriterium „gleich oder verschieden“ geordnet werden können. Man spricht hier auch von qualitativen Merkmalen. (Beispiel: Augenfarbe, Geschlecht).

Ordinalskala oder Rangskala:

Die Merkmalsausprägungen bringen zusätzlich eine Rangfolge zum Ausdruck. (Beispiel Platzierung im Sport, Hotel- Güteklassen)

Kardinalskala oder metrische Skala:

Zusätzlich zu den Eigenschaften der vorherigen Skalen ist es sinnvoll, Differenzen und Verhältnisse der Merkmalsausprägungen zu berechnen. Man spricht hier von quantitativen Merkmalen. (Beispiel: Körpergröße, Gewicht)

Quantitative Merkmale nennt man **diskret**, wenn die Ausprägungen nur isolierte Zahlenwerte annehmen können (Beispiel: Kinderzahl), und **stetig**, wenn die Ausprägungen alle Zahlen auf einem bestimmten Intervall annehmen können (Beispiel: Körpergröße).

Verschiedene Typen statistischer Erhebungen:

- a) **Vollerhebung** oder Totalerhebung: z.B. Volkszählung
Teilerhebung: z.B. Wahlumfragen
- b) **Primärerhebung**: Es wird Datenmaterial eigens für die geplante Untersuchung erhoben.
Sekundärerhebung: Es wird auf bereits vorhandenes (möglicherweise für andere Zwecke gesammeltes) Datenmaterial zurückgegriffen (z.B. Lohnsteuerkarten für die Untersuchung von Einkommensverteilung)
- c) **Befragung**: z.B. durch Fragebögen
- d) **Beobachtung**: z.B. Verkehrszählung
- e) **Experiment**: z.B. Betrieb von Motoren, Untersuchung von neuen Medikamenten.

2. Darstellung von eindimensionalem Datenmaterial

Definition:

Ein Merkmal X werde an n Merkmalsträgern einer Grundgesamtheit beobachtet

- (1) Das n -Tupel (x_1, \dots, x_n) der beobachteten Merkmalsausprägungen nennt man eine **Urliste** oder **Erhebungsliste** oder **Realisation** einer Stichprobe vom Umfang n .
- (2) Für jede mögliche Merkmalsausprägung a bezeichnet $h(a)$ die Anzahl der Merkmalswerte x_k der Urliste, die gleich a sind. Man nennt $h(a)$ die **(absolute) Häufigkeit** von a in der Stichprobe (x_1, \dots, x_n) .
- (3) Der Anteilswert $f(a) = \frac{1}{n}h(a)$ heißt **relative Häufigkeit** von a in der Stichprobe (x_1, \dots, x_n) .

Folgerung:

Sind a_1, \dots, a_r die verschiedenen in der Stichprobe (x_1, \dots, x_n) vorkommenden Merkmalswerte, so gelten die Beziehungen

$$\sum_{k=1}^r h(a_k) = n \quad \text{und} \quad \sum_{k=1}^r f(a_k) = 1$$

Beispiel:

Für die Augenfarben-Stichprobe (blau, braun, braun, grau, grau, braun) gilt:

$$h(\text{blau}) = 1, \quad h(\text{braun}) = 3, \quad h(\text{grau}) = 2, \quad h(\text{grün}) = 0$$

$$f(\text{blau}) = 1/6, \quad h(\text{braun}) = 3/6, \quad h(\text{grau}) = 2/6, \quad h(\text{grün}) = 0$$

Eine Darstellung der erhobenen Daten erfolgt üblicherweise durch:

- Eine **Häufigkeitstabelle**
- Ein **Kreisdiagramm**
- Ein **Stabdiagramm**
- Ein **Histogramm**

Definition:

Gegeben sei eine Stichprobe (x_1, \dots, x_n) eines quantitativen Merkmals X . Die Zahlen a_1, \dots, a_r seien die dabei auftretenden Merkmalswerte. Dann bezeichnet man als

- (4) **absolute kumulierte Häufigkeitsverteilung** der Stichprobe die Funktion

$$H(x) = \sum_{a_k \leq x} h(a_k)$$

- (5) **relative kumulierte Häufigkeitsverteilung oder empirische Verteilungsfunktion** der Stichprobe die Funktion

$$F(x) = \sum_{a_k \leq x} f(a_k)$$

Folgerung:

Die Funktionen $H(x)$ und $F(x)$ sind für jede Zahl x definiert, und es gilt stets die Beziehung:

$$F(x) = \frac{1}{n}H(x)$$

3. Lageparameter

Definition:

Gegeben sei eine Stichprobe (x_1, \dots, x_n) eines Merkmals X . Die verschiedenen Merkmalswerte der Stichprobe seien mit (a_1, \dots, a_r) bezeichnet.

- (1) Diejenigen a_k , welche größte Häufigkeit aufweisen, werden als Modalwerte der Stichprobe bezeichnet. Gibt es für die Stichprobe nur einen **Modalwert**, dann wird er mit x_{Mod} bezeichnet und auch **häufigster Wert** oder **Modus** genannt.

Ist X ein quantitatives Merkmal, dann definiert man

- (2) den Median oder Zentralwert der Stichprobe folgendermaßen: Man ordne die Stichprobe, so dass

$$x_1 \leq x_2 \leq \dots \leq x_n$$

wird. Ist n eine ungerade Zahl, dann setzt man

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

Ist n eine gerade Zahl, dann setzt man

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

also die Mitte der beiden Beobachtungspunkte

- (3) das **arithmetische Mittel** oder den **Durchschnittswert** oder **Mittelwert** der Stichprobe als die Zahl

$$\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k$$

- (4) für den Fall, dass kein x_k negativ ist, das geometrische Mittel der Stichprobe als die Zahl

$$x_{Geom} = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}$$

Eigenschaften:

1. Natürlich gilt auch

$$\bar{x} = \sum_{k=1}^r a_k f(a_k) = \frac{1}{n} \sum_{k=1}^r a_k h(a_k)$$

und mit dieser Formel lässt sich der Mittelwert meistens schneller berechnen als mit der Definition

2. Transformiert man die Beobachtungsdaten x_i gemäß

$$y_i = a + b x_i$$

linear, so transformieren sich die jeweiligen arithmetischen Mittel wie folgt

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

4. Streuungsparameter

Definition:

Gegeben sei eine Stichprobe (x_1, \dots, x_n) eines quantitativen Merkmals X .

Merkmalwerte der Stichprobe seien mit (a_1, \dots, a_r) bezeichnet.

- (1) Die **Spannweite** der Stichprobe

$$SP = \max_{k=1, \dots, n} x_k - \min_{k=1, \dots, n} x_k$$

ist die Differenz zwischen dem größten und dem kleinsten Beobachtungswert

- (2) Die **durchschnittliche Abweichung vom Median** ist die Zahl

$$\bar{s} = \frac{1}{n} \sum_{k=1}^n |x_k - x_{\text{Med}}|$$

also das arithmetische Mittel der Abstände aller Beobachtungswerte vom Median

- (3) Die **mittlere quadratische Abweichung oder Varianz** der Stichprobe ist die Zahl

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

also das arithmetische Mittel der quadrierten Abstände aller Beobachtungswerte vom Mittelwert

- (4) Die **Standardabweichung**

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

ist die nicht-negative Wurzel aus der mittleren quadratischen Abweichung.

- (5) Für positives \bar{x} heißt der Quotient aus Standardabweichung und arithmetischem Mittel

$$V = \frac{s}{\bar{x}}$$

der **Variationskoeffizient** der Stichprobe.

Eigenschaften:

1. Natürlich gilt wieder

$$\bar{s} = \frac{1}{n} \sum_{k=1}^r |a_k - x_{\text{Med}}| h(a_k) = \sum_{k=1}^r |a_k - x_{\text{Med}}| f(a_k)$$

und

$$s^2 = \frac{1}{n} \sum_{k=1}^r (a_k - \bar{x})^2 h(a_k) = \sum_{k=1}^r (a_k - \bar{x})^2 f(a_k)$$

2. Transformiert man die Beobachtungsdaten x_k gemäß

$$y_k = a + b x_k$$

linear, so gilt für die zugehörige mittlere quadratische Abweichung s_x^2, s_y^2 :

$$s_y^2 = b^2 s_x^2 \text{ und auch } s_y = |b| s_x$$

3. Es gilt das „Verschiebungsgesetz“

$$s^2 = \frac{1}{n} \sum_{k=1}^n x_k^2 - \bar{x}^2 = \frac{1}{n} \sum_{k=1}^r a_k^2 h(a_k) - \bar{x}^2$$

4. Die drei Streuungsparameter SP , \bar{s} und s besitzen dieselbe Dimension (die der Beobachtungswerte) und es gilt stets
- $$\bar{s} \leq s \leq SP$$
5. In der Praxis stellt man bei vielen Stichproben quantitative Merkmale fest:
- Im Intervall $[\bar{x} - s; \bar{x} + s]$ liegen etwa 68% aller Beobachtungswerte x_k
- Im Intervall $[\bar{x} - 2s; \bar{x} + 2s]$ liegen etwa 95% aller Beobachtungswerte x_k
- Im Intervall $[\bar{x} - 3s; \bar{x} + 3s]$ liegen praktisch alle Beobachtungswerte x_k

5. Mehrdimensionale Stichprobe

Definition:

Gegeben sei eine Stichprobe um Umfang n

des Merkmals X : x_1, \dots, x_n

Und eine Stichprobe vom gleichen Umfang

des Merkmals Y : y_1, \dots, y_n

- (1) Man nennt die Liste von Merkmalswertpaaren $(x_1, y_1), \dots, (x_n, y_n)$ eine **zweidimensionale Stichprobe**.

Es sei a_1, \dots, a_r eine Liste der verschiedenen unter x_i auftretenden Merkmalswerten und b_1, \dots, b_s eine Liste der verschiedenen unter den y_i auftretenden Merkmalswerten.

- (2) Für jedes Paar von möglichen Merkmalsausprägungen (a_i, b_j) bezeichne $h(a_i, b_j)$ die Anzahl der Paare (x, y) mit $x = a_i$ und $y = b_j$. Man nennt $h(a_i, b_j)$ die **(absolute) Häufigkeit** von (a_i, b_j) in der zweidimensionalen Stichprobe.

- (3) Der Anteilswert

$$f(a_i, b_j) = \frac{1}{n} h(a_i, b_j)$$

heißt die relative Häufigkeit von (a_i, b_j) in der zweidimensionalen Stichprobe.

- (4) Die Werte

$$h_X(a_i) = \sum_{j=1}^s h(a_i, b_j) \text{ bzw. } h_Y(b_j) = \sum_{i=1}^r h(a_i, b_j)$$

nennt man **Randhäufigkeiten**.

- (5) Die Werte

$$f_X(a_i | b_j) = \frac{h(a_i, b_j)}{h_Y(b_j)} \text{ bzw. } f_Y(b_j | a_i) = \frac{h(a_i, b_j)}{h_X(a_i)}$$

heißen **bedingte relative Häufigkeiten**.

- (6) Die bedingten relativen Häufigkeiten

$$f_X(a_1|b_j), f_X(a_2|b_j), \dots, f_X(a_r|b_j)$$

definieren die bedingte Verteilung des Merkmals X unter der gegebenen Ausprägung b_j des Merkmals Y . Entsprechend definieren die bedingten Häufigkeiten

$$f_Y(b_1|a_i), f_Y(b_2|a_i), \dots, f_Y(b_s|a_i)$$

die bedingte Verteilung des Merkmals Y unter der gegebenen Ausprägung a_i des Merkmals X .

6. Kovarianz, Korrelation, Regression

Definition:

Gegeben sei eine zweidimensionale Stichprobe $(x_1, y_1), \dots, (x_n, y_n)$ zweier quantitativer Merkmale X und Y mit den arithmetischen Mitteln \bar{x} bzw. \bar{y} und den Standardabweichungen s_X bzw. s_Y

- (1) Die Zahl

$$\text{Cov}(X, Y) = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y}) = \frac{1}{n} \sum_{k=1}^n (x_k y_k) - \bar{x} \bar{y} .$$

heißt **Kovarianz** der Stichprobe.

- (2) Man nennt die Zahl

$$r = \frac{\text{Cov}(X, Y)}{s_X s_Y} .$$

- falls definiert - den **Korrelationskoeffizienten** der Stichprobe.

- (3) Sind

$$a = \frac{\text{Cov}(X, Y)}{s_X^2} \text{ und } b = \bar{y} - a \bar{x} .$$

definiert, so heißt die Gerade mit der Gleichung $y = a x + b$ die **Regressions- oder Ausgleichsgerade** der Stichprobe.

II. Induktive Statistik - Wahrscheinlichkeitsrechnung

1. Kombinatorische Probleme

Viele Glücksspiele bestehen darin, unter N möglichen „gleichwahrscheinlichen“ Spielausgängen den Richtigen zu erraten. Beispielsweise:

- (i) Würfeln ($N=6$)
- (ii) Münzwurf ($N=2$)

Die Gewinnchance hängt also von der Gesamtzahl N aller Möglichkeiten ab. Zur Berechnung von N kann in etwas komplizierteren Fällen das folgende Modell angewandt werden:

Das Urnenmodell

Aus einer Urne werden n Kugeln, welche von 1 bis n durchnummeriert sind, wird genau k mal eine Kugel gezogen. Nach jedem Zug wird die Nummer der entnommenen Kugel notiert.

Frage: Wie viele Möglichkeiten dies zu tun gibt es?

Fallunterscheidung:

1. mit Zurücklegen: Jede gezogene Kugel wird nach Notieren ihrer Nummer in die Urne zurückgelegt, kann also mehrfach gezogen werden.

2. ohne Zurücklegen: Jede gezogene Kugel wird nach Notieren ihrer Nummer nicht mehr in die Urne zurückgelegt, kann also nur einmal gezogen werden.

A. mit Anordnung: Die Kugeln werden entsprechend der zeitlichen Reihenfolge ihrer Ziehung angeordnet

A. ohne Anordnung: Die zeitlichen Reihenfolge der Ziehung der Kugeln spielt keine Rolle.

Die sich ergebende Anzahl entnimmt man der folgenden Tabelle

Anzahl der Möglichkeiten beim Urnenmodell; Ziehung	1. mit Zurücklegen	2. ohne Zurücklegen
A) mit Anordnung	n^k	$\frac{n!}{(n-k)!}$
B) ohne Anordnung	$\binom{n+k-1}{k}$	$\binom{n}{k}$

Beispiel: Anzahl der Möglichkeiten beim Lotto (6 aus 49): $N = \binom{49}{6} = \frac{49!}{43!6!} = 13.983.816$

2. Wahrscheinlichkeitsräume

Im folgenden bezeichne Ω stets eine nicht-leere Menge, etwa die „Ergebnismenge“ eines Experiments mit zufälligem Ausgang. Z.B.

- (i) Würfeln: $\Omega = \{1,2,3,4,5,6\}$
- (ii) Münzwurf: $\Omega = \{K, W\}$ – K Kopf, W Wappen

Es werden folgende Bezeichnungen für Teilmengen A, B, \dots von Ω verwendet.

- (1) $A \cap B = \{x \mid x \in A \text{ und } x \in B\}$
- (2) $A \cup B = \{x \mid x \in A \text{ oder } x \in B\}$
- (3) $A \subseteq B$ bedeutet, dass jedes $x \in A$ auch in B enthalten ist.
- (4) Ist $A \subseteq B$, so definiert man $B \setminus A = \{x \mid x \in B \text{ und } x \notin A\}$
- (5) \emptyset und $\{\}$ bezeichnen die leere Menge.
- (6) $A^c = \Omega \setminus A$ nennt man das Komplement oder Gegenteil von A .
- (7) Man nennt Teilmengen A_1, A_2, A_3, \dots von Ω **paarweise disjunkt**, wenn für $i \neq k$ stets $A_i \cap A_k = \emptyset$ ist.
- (8) $|A|$ bezeichne die Anzahl der Elemente von A .

Definition:

Gegeben sei eine nicht-leere Menge Ω und gewisse Teilmengen von Ω , die man Ergebnisse nennt. Eine Vorschrift P , welche einem Ereignis A eine Zahl $P(A)$ zuordnet, nennt man **Wahrscheinlichkeitsmaß**, wenn folgende Bedingungen erfüllt sind.

- (1) $P(\Omega) = 1$ und für jedes Ereignis A gilt: $0 \leq P(A) \leq 1$
- (2) Sind A_1, A_2, A_3, \dots endlich abzählbar viele paarweise disjunkte Ereignisse, so gilt:
 $P(A_1 \cup A_2 \cup A_3 \cup \dots) = P(A_1) + P(A_2) + P(A_3) + \dots$
 Man nennt das Paar (Ω, P) einen **Wahrscheinlichkeitsraum**.

Definition:

In einem Wahrscheinlichkeitsraum (Ω, P) nennt man

- (i) \emptyset das unmögliche Ereignis.
- (ii) Ω das **sichere Ereignis** oder die **Ereignismenge** oder den **Stichprobenraum**.
- (iii) Jedes Ereignis der Gestalt $\{\omega\}$ ($\omega \in \Omega$) ein **Elementarereignis**.

Definition:

Ein Wahrscheinlichkeitsraum (Ω, P) heißt **diskret**, wenn endlich oder abzählbar ist und jede einelementige Teilmenge $\{\omega\}$ von Ω eine Ereignis ist.

Satz :

In jedem diskreten Wahrscheinlichkeitsraum (Ω, P) gilt für jedes Ereignis:

$$P(A) = \sum_{\omega \in A} P(\{\omega\})$$

Satz :

Besitzt der Wahrscheinlichkeitsraum (Ω, P) N gleichwahrscheinliche Elementarereignisse, so gilt:

- (i) $P(\{\omega\}) = \frac{1}{N}$ für jedes $\omega \in \Omega$
- (ii) Für jedes Ereignis A ist:
- $$P(A) = \frac{|A|}{|\Omega|}$$

3. Regeln für Wahrscheinlichkeiten

In jedem Wahrscheinlichkeitsraum (Ω, P) mit Ereignissen $A, B, A_1, A_2, A_3, \dots$ gelten folgende Regeln:

- (1) $P(\emptyset) = 0$ und $P(\Omega) = 1$
- (2) Aus $A \subseteq B$
 $P(A) \leq P(B)$ und $P(B \setminus A) = P(B) - P(A)$
- (3) $P(A^C) = 1 - P(A)$
- (4) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- (5) $P(A_1 \cup A_2 \cup A_3 \cup \dots) = P(A_1) + P(A_1^C \cap A_2) + P(A_1^C \cap A_2^C \cap A_3) + \dots$

4. Bedingte Wahrscheinlichkeiten**Definition:**

Sind A, H Ereignisse eines Wahrscheinlichkeitsraum (Ω, P) mit $P(H) > 0$, so nennt man die Zahl

$$P(A|H) = \frac{P(A \cap H)}{P(H)}$$

die **bedingte Wahrscheinlichkeit** von A unter der **Bedingung** oder **Hypothese** H .

Satz :

Gegeben seien endlich viele Ereignisse H_1, \dots, H_n eines Wahrscheinlichkeitsraum (Ω, P) . Die Ereignisse H_1, \dots, H_{n-1} mögen positive Wahrscheinlichkeiten besitzen. Dann gilt:

$$P(H_1 \cap \dots \cap H_n) = P(H_1) \cdot P(H_2 | H_1) \cdot P(H_3 | H_1 \cap H_2) \cdot \dots \cdot P(H_n | H_1 \cap \dots \cap H_{n-1})$$

Satz :

Gegeben seien endlich (oder abzählbar) viele Ereignisse A, H_1, H_2, \dots eines Wahrscheinlichkeitsraum (Ω, P) mit $P(H_k) > 0$. Die Ereignisse H_k seien paarweise disjunkt. Ihre Vereinigung sei gleich Ω . Dann gilt:

- (1) (**Satz für die totale Wahrscheinlichkeit**)
- $$P(A) = \sum_{k \geq 1} P(H_k) \cdot P(A | H_k)$$

(2) **(Bayes'sche Formel)**

$$P(H_n|A) = \frac{P(H_n) \cdot P(A|H_n)}{\sum_{k \geq 1} P(H_k) \cdot P(A|H_k)} = \frac{P(H_n) \cdot P(A|H_n)}{P(A)}$$

für jedes H_n , falls $P(A) > 0$

Definition:

Zwei Ereignisse A, B eines Wahrscheinlichkeitsraum (Ω, P) heißen **unabhängig**, wenn gilt:

$$P(A \cap B) = P(A) \cdot P(B)$$

Definition:

Man nennt eine Menge von Ereignissen unabhängig, wenn für jeweils endlich viele dieser Ereignisse etwa A_1, \dots, A_r , stets gilt: $P(A_1 \cap \dots \cap A_r) = P(A_1) \cdot \dots \cdot P(A_r)$

5. Zufallsvariable, Verteilungsfunktion**Definition:**

Es sei (Ω, P) ein Wahrscheinlichkeitsraum

- (1) Eine Vorschrift X , welche jedem $\omega \in \Omega$ eine reelle Zahl $X(\omega)$ zuordnet (die Realisation von X an der Stelle ω), nennt man eine **Zufallsvariable**, wenn für jedes $x \in \mathbb{R}$ die Menge $\{\omega \in \Omega \mid X(\omega) \leq x\}$ ein Ereignis ist.

- (2) Ist X eine Zufallsvariable, so nennt man die durch

$$F_X(x) = P(\{\omega \in \Omega \mid X(\omega) \leq x\}) \quad \text{kurz: } F_X(x) = P(X \leq x)$$

für alle reellen Zahlen x definierte Funktion $F_X(x)$ die **Verteilungsfunktion** von X .

Eigenschaften von Verteilungsfunktionen:

- (1) $F_X(x) \leq F_X(y)$ für $x \leq y$
 (2) $F_X(-\infty) = 0$; d.h. $\lim_{x \rightarrow -\infty} F_X(x) = 0$
 $F_X(+\infty) = 1$; d.h. $\lim_{x \rightarrow +\infty} F_X(x) = 1$
 (3) $F_X(x)$ ist in jedem Punkt rechtsseitig stetig.

6. Diskrete Zufallsvariable**Definition:**

Eine Zufallsvariable X heißt **diskret**, wenn sie nur endlich oder abzählbar viele Werte annimmt. Ist X eine diskrete Zufallsvariable, so nennt man die durch

$$f_X(x) = P(\{\omega \in \Omega \mid X(\omega) = x\}) \quad \text{kurz: } f_X(x) = P(X = x)$$

Definierte Funktion die **Wahrscheinlichkeitsfunktion** von X .

Folgerung:

Ist $\{x_0, x_1, x_2, \dots\}$ der Wertebereich von X , so ist für jede Reelle Zahl x

$$F_X(x) = \sum_{x_k \leq x} f_X(x_k)$$

7. Stetige Zufallsvariable

Definition:

Eine Zufallsvariable X heißt **stetig**, wenn es eine Funktion $f : \mathbb{R} \rightarrow [0, \infty)$ gibt, so dass die Verteilungsfunktion $F_X(x)$ für jedes $x \in \mathbb{R}$ die Darstellung

$$F_X(x) = \int_{-\infty}^x f_X(t) dt$$

besitzt. Die Funktion $f_X(x)$ nennt man **Dichte** von X bzw. von $F_X(x)$

Die Wahrscheinlichkeit, dass eine stetige Zufallsvariable X mit Dichte $f_X(x)$ Werte zwischen zwei Zahlen a und b mit $-\infty \leq a \leq b \leq +\infty$ annimmt, ist stets gleich dem Integral $\int_a^b f_X(x) dx$.

Dabei ist es unerheblich, ob die Grenzen a, b mitberücksichtigt werden oder nicht, d.h. es gilt:

$$P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b) = \int_a^b f_X(x) dx$$

Die Wahrscheinlichkeit wird durch die Fläche repräsentiert, die oberhalb des Intervalls $[a; b]$ zwischen x -Achse und Dichtefunktion liegt.

8. Parameter von Verteilung

In Analogie zu den Lage- und Streuungsparametern in der deskriptiven Statistik gilt die folgende

Definition:

Ist X eine diskrete Zufallsvariable mit dem Wertebereich $\{x_1, x_2, x_3, \dots\}$ bzw. ist X eine stetig Zufallsvariable mit der Dichte $f_X(x)$, dann heißt

(1) Die Zahl

$$E(X) = \mu_X = \begin{cases} \sum_{k \geq 1} x_k f_X(x_k) \\ \text{bzw.} \\ \int_{-\infty}^{+\infty} x f_X(x) dx \end{cases}$$

der Erwartungswert von X bzw. $F_X(x)$

(2) Die Zahl

$$\text{Var}(X) = \begin{cases} \sum_{k \geq 1} (x_k - \mu_X)^2 f_X(x_k) \\ \text{bzw.} \\ \int_{-\infty}^{+\infty} (x - \mu_X)^2 f_X(x) dx \end{cases}$$

die **Varianz** von X bzw. $F_X(x)$

(3) Die Zahl

$$\sigma_X = \sqrt{\text{Var}(X)}$$

die **Standardabweichung** von X bzw. $F_X(x)$

Bemerkungen:

(i) Es gilt das Verschiebegesetz

$$\text{Var}(X) = \sum_{k \geq 1} x_k^2 f_X(x_k) - \mu_X^2$$

bzw.

$$\text{Var}(X) = \int_{-\infty}^{+\infty} x^2 f_X(x) dx - \mu_X^2$$

(ii) Ist der zugrunde liegende Wahrscheinlichkeitsraum (Ω, P) endlich oder abzählbar, dann kann man $E(X)$ oder $\text{Var}(X)$ auch folgendermaßen berechnen:

a. $E(X) = \sum_{\omega \in \Omega} X(\omega) \cdot P(\{\omega\})$

b. $\text{Var}(X) = \sum_{\omega \in \Omega} (X(\omega) - E(X))^2 \cdot P(\{\omega\}) = \sum_{\omega \in \Omega} X(\omega)^2 \cdot P(\{\omega\}) - E(X)^2$

c. Wenn der Wertebereich der Zufallsvariablen X unendlich ist, so sind $E(X)$ und $\text{Var}(X)$ laut Definition der Wert einer unendlichen Reihe bzw. eines uneigentlichen Integrals. Eine solche Reihe bzw. ein solches Integral kann auch divergent sein, d.h. $E(X)$ bzw. $\text{Var}(X)$ können undefiniert sein.

9. Der zentrale Grenzwertsatz

Für kleine n und bestimmte Werte von p ist die Binomialverteilung tabelliert. Sind $p, q > 0$ und es gilt $p + q = 1$ sowie x reell, dann gilt für große n der Satz.

Satz (Zentraler Grenzwertsatz)

$$B_{n,p}(x) \approx N_{0,1} \left(\frac{x - np}{\sqrt{npq}} \right)$$

wobei $N_{0,1}$ die tabellierte **Standardnormalverteilung** darstellt.

III. Schließende Statistik

Die zentrale Frage der schließenden Statistik besteht darin, Aussagen aufgrund einer Stichprobe für die Grundgesamtheit zu treffen.

1. Schätzung von Parametern

Das Problem, einen unbekannt Parameter einer Stichprobe zu schätzen wird folgendermaßen definiert:

Definition:

Eine Stichprobenfunktion, deren Realisierung (Schätzer) $\hat{\gamma}$ als Näherung eines Parameters γ einer Stichprobe angesehen werden kann, heißt **Punktschätzung** von γ .

Definition:

Eine Schätzung heißt **erwartungstreu**, wenn ihr Erwartungswert gleich dem zu schätzenden Parameter ist. Es gilt $E(\hat{\gamma}) = \gamma$

Definition:

Eine Schätzung $\hat{\gamma}_1$ heißt **effizient** (wirksam), wenn für zwei erwartungstreue Schätzer $\hat{\gamma}_1$ und $\hat{\gamma}_2$ für γ gilt $\text{Var}(\hat{\gamma}_1) < \text{Var}(\hat{\gamma}_2)$.

Definition:

Ein Schätzer heißt **konsistent**, wenn er für sehr große Stichproben gegen den wahren Wert der Grundgesamtheit konvergiert.

2. Methoden zur Gewinnung von Schätzungen

Definition:

Für Parameter, die sich aus den Momenten zusammensetzen, gewinnt man Schätzungen, indem man die Momente durch die empirischen Momente ersetzt. Diese Methode heißt **Momentenmethode**. Als empirisches k-tes Moment bezeichnet man die Stichprobenfunktion

$\frac{1}{n} \sum_{i=1}^n X_i^k$. Als empirisches zentrales Moment der Ordnung k bezeichnet man die

Stichprobenfunktion $\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^k$

Eine Punktschätzung liefert aus der vorgelegten Stichprobe einen Schätzwert des betreffenden Parameters. Zur Genauigkeit und Sicherheit der Schätzung liefern Konfidenzschätzungen Ergebnisse.

Definition:

Sei eine mathematische Stichprobe (X_1, \dots, X_n) aus einer Grundgesamtheit gegeben, wobei der Parameter γ der Stichprobe geschätzt werden soll. Ferner seien Schätzer $\hat{\gamma}_1$ und $\hat{\gamma}_2$ zwei Schätzer derart, dass bei beliebigem γ gilt $P(\hat{\gamma}_1 < \gamma < \hat{\gamma}_2) = 1 - \alpha$.

Dann heißt das Intervall $[\hat{\gamma}_1; \hat{\gamma}_2]$ eine **Konfidenzschätzung** oder **Konfidenzintervall** von γ zum Konfidenzniveau $1 - \alpha$.

Sei eine normalverteilte Grundgesamtheit gegeben mit bekannter Varianz σ^2

Das Intervall

$$\left[\bar{x} - \frac{\hat{\sigma}}{\sqrt{n}} z_{1-\alpha/2}; \bar{x} + \frac{\hat{\sigma}}{\sqrt{n}} z_{1-\alpha/2} \right]$$

ist dann das symmetrisches Konfidenzintervall für Erwartungswert μ zum Konfidenzniveau $1-\alpha$.

Sei eine normalverteilte Grundgesamtheit gegeben mit unbekannter Varianz

Das Intervall

$$\left[\bar{x} - \frac{s}{\sqrt{n}} t_{m;1-\alpha/2}; \bar{x} + \frac{s}{\sqrt{n}} t_{m;1-\alpha/2} \right]$$

ist dann das symmetrisches Konfidenzintervall für Erwartungswert μ zum Konfidenzniveau $1-\alpha$.

3. Prüfen von Hypothesen (Tests)

Mit Hilfe von Tests (Signifikanztests) werden Hypothesen über statistische Parameter überprüft.

Aufbau von statistischen Tests

- (1) Aufstellen der Nullhypothese H_0 und der Alternativhypothese H_1
- (2) Festlegung des Signifikanzniveaus und Festlegung des Nichtablehnungsbereichs
- (3) Berechnung der Teststatistik
- (4) Bestimmung des Ablehnungsbereichs
- (5) Testentscheidung

Signifikanztests für bestimmte Fragestellungen

Test für μ bei normalverteilter Grundgesamtheit und bekannter Varianz σ^2 (Gauß-Test)

Hypothesen	Teststatistik	Testentscheidung
$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$z^* = \frac{ \bar{x} - \mu }{\sigma} \sqrt{n}$	Ablehnung von H_0 für : $z^* > z_{1-\alpha/2}$

Test für μ bei normalverteilter Grundgesamtheit und unbekannter Varianz

Hypothesen	Teststatistik	Testentscheidung
$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$t^* = \frac{ \bar{x} - \mu }{s} \sqrt{n}$	Ablehnung von H_0 für : $t^* > t_{n-1, 1-\alpha/2}$

Test für σ^2 bei normalverteilter Grundgesamtheit

Hypothesen	Teststatistik	Testentscheidung
$H_0 : \sigma^2 = \sigma_0^2$ $H_1 : \sigma^2 \neq \sigma_0^2$	$\chi^{2*} = \frac{(n-1)s^2}{\sigma_0^2}$	Ablehnung von H_0 für : $\chi^{2*} > \chi^2_{n-1, 1-\alpha/2}$ oder $\chi^{2*} < \chi^2_{n-1, \alpha/2}$

Test für Vergleich zweier Varianzen bei unabhängigen Stichproben und normalverteilter Grundgesamtheit

Hypothesen	Teststatistik	Testentscheidung
$H_0 : \sigma_1^2 = \sigma_2^2$ $H_1 : \sigma_1^2 \neq \sigma_2^2$	$F^* = \frac{s_1^2}{s_2^2}$	Ablehnung von H_0 für : $F^* > F_{n_1-1, n_2-1, 1-\alpha/2}$

Tabelle der Standardnormalverteilung

Beispiel: $N_{0,1}(x) = N_{0,1}(2,36) = 0,99086$

	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,50000	0,50399	0,50798	0,51197	0,51595	0,51994	0,52392	0,52790	0,53188	0,53586
0,1	0,53983	0,54380	0,54776	0,55172	0,55567	0,55962	0,56356	0,56749	0,57142	0,57535
0,2	0,57926	0,58317	0,58706	0,59095	0,59483	0,59871	0,60257	0,60642	0,61026	0,61409
0,3	0,61791	0,62172	0,62552	0,62930	0,63307	0,63683	0,64058	0,64431	0,64803	0,65173
0,4	0,65542	0,65910	0,66276	0,66640	0,67003	0,67364	0,67724	0,68082	0,68439	0,68793
0,5	0,69146	0,69497	0,69847	0,70194	0,70540	0,70884	0,71226	0,71566	0,71904	0,72240
0,6	0,72575	0,72907	0,73237	0,73565	0,73891	0,74215	0,74537	0,74857	0,75175	0,75490
0,7	0,75804	0,76115	0,76424	0,76730	0,77035	0,77337	0,77637	0,77935	0,78230	0,78524
0,8	0,78814	0,79103	0,79389	0,79673	0,79955	0,80234	0,80511	0,80785	0,81057	0,81327
0,9	0,81594	0,81859	0,82121	0,82381	0,82639	0,82894	0,83147	0,83398	0,83646	0,83891
1,0	0,84134	0,84375	0,84614	0,84849	0,85083	0,85314	0,85543	0,85769	0,85993	0,86214
1,1	0,86433	0,86650	0,86864	0,87076	0,87286	0,87493	0,87698	0,87900	0,88100	0,88298
1,2	0,88493	0,88686	0,88877	0,89065	0,89251	0,89435	0,89617	0,89796	0,89973	0,90147
1,3	0,90320	0,90490	0,90658	0,90824	0,90988	0,91149	0,91309	0,91466	0,91621	0,91774
1,4	0,91924	0,92073	0,92220	0,92364	0,92507	0,92647	0,92785	0,92922	0,93056	0,93189
1,5	0,93319	0,93448	0,93574	0,93699	0,93822	0,93943	0,94062	0,94179	0,94295	0,94408
1,6	0,94520	0,94630	0,94738	0,94845	0,94950	0,95053	0,95154	0,95254	0,95352	0,95449
1,7	0,95543	0,95637	0,95728	0,95818	0,95907	0,95994	0,96080	0,96164	0,96246	0,96327
1,8	0,96407	0,96485	0,96562	0,96638	0,96712	0,96784	0,96856	0,96926	0,96995	0,97062
1,9	0,97128	0,97193	0,97257	0,97320	0,97381	0,97441	0,97500	0,97558	0,97615	0,97670
2,0	0,97725	0,97778	0,97831	0,97882	0,97932	0,97982	0,98030	0,98077	0,98124	0,98169
2,1	0,98214	0,98257	0,98300	0,98341	0,98382	0,98422	0,98461	0,98500	0,98537	0,98574
2,2	0,98610	0,98645	0,98679	0,98713	0,98745	0,98778	0,98809	0,98840	0,98870	0,98899
2,3	0,98928	0,98956	0,98983	0,99010	0,99036	0,99061	0,99086	0,99111	0,99134	0,99158
2,4	0,99180	0,99202	0,99224	0,99245	0,99266	0,99286	0,99305	0,99324	0,99343	0,99361
2,5	0,99379	0,99396	0,99413	0,99430	0,99446	0,99461	0,99477	0,99492	0,99506	0,99520
2,6	0,99534	0,99547	0,99560	0,99573	0,99585	0,99598	0,99609	0,99621	0,99632	0,99643
2,7	0,99653	0,99664	0,99674	0,99683	0,99693	0,99702	0,99711	0,99720	0,99728	0,99736
2,8	0,99744	0,99752	0,99760	0,99767	0,99774	0,99781	0,99788	0,99795	0,99801	0,99807
2,9	0,99813	0,99819	0,99825	0,99831	0,99836	0,99841	0,99846	0,99851	0,99856	0,99861
3,0	0,99865	0,99869	0,99874	0,99878	0,99882	0,99886	0,99889	0,99893	0,99896	0,99900
3,1	0,99903	0,99906	0,99910	0,99913	0,99916	0,99918	0,99921	0,99924	0,99926	0,99929
3,2	0,99931	0,99934	0,99936	0,99938	0,99940	0,99942	0,99944	0,99946	0,99948	0,99950
3,3	0,99952	0,99953	0,99955	0,99957	0,99958	0,99960	0,99961	0,99962	0,99964	0,99965
3,4	0,99966	0,99968	0,99969	0,99970	0,99971	0,99972	0,99973	0,99974	0,99975	0,99976
3,5	0,99977	0,99978	0,99978	0,99979	0,99980	0,99981	0,99981	0,99982	0,99983	0,99983
3,6	0,99984	0,99985	0,99985	0,99986	0,99986	0,99987	0,99987	0,99988	0,99988	0,99989
3,7	0,99989	0,99990	0,99990	0,99990	0,99991	0,99991	0,99992	0,99992	0,99992	0,99992
3,8	0,99993	0,99993	0,99993	0,99994	0,99994	0,99994	0,99994	0,99995	0,99995	0,99995
3,9	0,99995	0,99995	0,99996	0,99996	0,99996	0,99996	0,99996	0,99996	0,99997	0,99997

Tabelle der t-Verteilung (aus Wikibooks, der freien Wissensdatenbank)

Quantile der t-Verteilung nach ausgewählten Wahrscheinlichkeiten p und Freiheitsgraden					
	Wahrscheinlichkeit p				
Freiheitsgrade	0,900	0,950	0,975	0,990	0,995
1	3,078	6,314	12,706	31,821	63,656
2	1,886	2,920	4,303	6,965	9,925
3	1,638	2,353	3,182	4,541	5,841
4	1,533	2,132	2,776	3,747	4,604
5	1,476	2,015	2,571	3,365	4,032
6	1,440	1,943	2,447	3,143	3,707
7	1,415	1,895	2,365	2,998	3,499
8	1,397	1,860	2,306	2,896	3,355
9	1,383	1,833	2,262	2,821	3,250
10	1,372	1,812	2,228	2,764	3,169
p →	0,900	0,950	0,975	0,990	0,995
11	1,363	1,796	2,201	2,718	3,106
12	1,356	1,782	2,179	2,681	3,055
13	1,350	1,771	2,160	2,650	3,012
14	1,345	1,761	2,145	2,624	2,977
15	1,341	1,753	2,131	2,602	2,947
16	1,337	1,746	2,120	2,583	2,921
17	1,333	1,740	2,110	2,567	2,898
18	1,330	1,734	2,101	2,552	2,878
19	1,328	1,729	2,093	2,539	2,861
20	1,325	1,725	2,086	2,528	2,845
p →	0,900	0,950	0,975	0,990	0,995
21	1,323	1,721	2,080	2,518	2,831
22	1,321	1,717	2,074	2,508	2,819
23	1,319	1,714	2,069	2,500	2,807
24	1,318	1,711	2,064	2,492	2,797
25	1,316	1,708	2,060	2,485	2,787
26	1,315	1,706	2,056	2,479	2,779
27	1,314	1,703	2,052	2,473	2,771
28	1,313	1,701	2,048	2,467	2,763
29	1,311	1,699	2,045	2,462	2,756
30	1,310	1,697	2,042	2,457	2,750
1000	1,282	1,646	1,962	2,330	2,581

