Statistik (Business Statistics) / Studium der Verteilung

Deskriptive Statistik: Darstellung + Zusammenfassung beobachteter Daten

Induktive Statistik: (Inferenz) Analyse mittels math. Modellen

Wahrscheinlichkeitstheorie: mathematisches Fundament

Variable: spezielle Eigenschaft eines Individuums

⇒ Variablenwerte: • Kategorisch / qualitativ

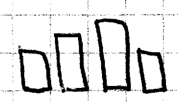
⇒ vorgegebene Liste von Werte

• numerisch / quantitativ

⇒ eine Zahl mit der man umgeht

- Name
- Werte
- Masseinheit

Diagramme

Blockdiagramm 

Kuchendiagramm 

Histogramm ⇒ • symmetrisch / asymmetrisch

- Unimodal (nur 1 MaxWert)
- Bimodal (zwei MaxWerte)
- Rechtsschief (links > rechts)
- Linksschief (rechts > links)
- Uniform / Gleichverteilt (alles gleich hoch)

Zeit-Diagramm (x: zeit, Y: Wert der stichprobe)

Kennzahlen

- Lage ("Mittelpunkt" der Verteilung)
- Streuung (Abweichung vom Mittelpunkt)

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

Median: Wert in der Mitte der sortierten Daten

$\bar{x} < m$: Verteilung linksschief

$\bar{x} > m$: Verteilung rechtsschief

$\bar{x} \approx m$: Verteilung $\approx$ symmetrisch

Perzentile: Verallgemenerung des Medians

$\Rightarrow Q_1$, 25%, erstes Quartil

Q_2, 50%, Median (zweites Quartil)

Q_3, 75%, drittes Quartil

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

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

Interquartilabstand: $IQA = Q_3 - Q_1$

Aussreisser: Wert ausserhalb von

$[Q_1 - 1{,}5 \times IQA,\ Q_3 + 1{,}5 \times IQA]$

Fünf-Kennzahlen-Zusammenfassung: $FKZ = (Min, Q_1, m, Q_3, Max)$

$\Rightarrow$ Boxplot: Visualisierung FKZ 

Sammeln von Daten

Stichprobendesign:

- Bevölkerung: gesamtheit der Indinduen, über die eine statistische Aussage gemacht werden soll
- Stichprobe: Teil der Bevölkerung, über den Informationen gesammelt wurden

$\Rightarrow$ Stichprobe muss repräsentativ sein!

• verzerrt (biased): bevorzugt/benachteiligt Segmente der
VS. Bevölkerung systematisch
• unverzerrt (unbiased): Keine Segmente bevorzugt/benachteiligt

Schlechte Stichproben:

• Freiwillige-Antwort: nur interessierte Teilnehmer mit Zugang zur Stichprobe antworten, Negatives fällt oft aus (da keine Lust zu beantworten usw.)
• Ich-mach's-mir-leicht: nicht zufällig, sondern Situationsabhäng ausgewählt, um die Arbeit zu erleichtern (zb. vor Ort, spezielles Medium, ...)

Datenherkunft: • Beobachtungsdaten (durch Befragung gesammelt ohne die Individuen vorher zu manipulieren)
• Experimentelle Daten (durch Erzeugung spezieller Situation gewonnen)

Parameter (Kennzahl der gesamten Bevölkerung)
Statistik (Kennzahl der einen Stichprobe)

Bevölkerung $\xrightarrow[\text{Auswahl}]{}$ Stichprobe $\xrightarrow[\text{Auswertung}]{}$ Schätzer θ

Stichprobenverteilung ($\Rightarrow$ Aussagekraft der Statistik)

• Verzerrung des Schätzers: Abstand des Schätzermittels $E(\theta)$ zum tatsächlichen θ

$\Rightarrow$ unverzerrt: $E(\theta) = \theta$, sonst verzerrt

• Variabilität des Schätzers: Abweichung vom Parameter
$\uparrow$ Stichproben $= \downarrow$ Variabilität

Wahrscheinlichkeitstheorie

- Zufallsexperiment: Vorgang der nach einer bestimmten Vorschrift ausgeführt wird, beliebig oft wiederholbar ist und zufallsabhängig ein Ergebnis liefert
- Elementarereignis / Ergebnis: möglicher Ausgang des Experiments
- Ergebnisraum: Menge aller möglichen Ergebnisse des Experiments
- Ereignis: Teilmenge des Ergebnisraums

- unmögliches Ereignis $A = \emptyset$
- sicheres Ereignis: $A = S$
- Komplementäre Ereignis: $A^c = S \setminus A$
- Durchschnitt: $A \cap B$
- Vereinigung: $A \cup B$
- Disjunkte Ereignisse: A, B, $A \cap B = \emptyset$
- Ausschöpfende Ereignisse: A, B, $A \cup B = S$
- Partition: des Ergebnisraums, eine Folge $A_1, \dots, A_n$ von ausschöpfen und paarweise disjunkten Ereignissen

Wahrscheinlichkeit $P(A)$, P: Wahrscheinlichkeitsfunktion, A: Ereignis

- LAPLACE: alle Ergebnisse gleich Wahrscheinlich $\Rightarrow$ Zählregel

$$P(A) = \frac{\text{Anzahl der für A günstigen Fälle}}{\text{Anzahl aller möglichen Fälle}} = \frac{\#A}{\#S}$$

- STATISTISCH: Grenzwert der Häufigkeiten wenn die Anzahl Versuche gross wird

$$P(A) = \lim_{\text{Versuche} \to \infty} \frac{\text{Anzahl der für A günstigen Versuche}}{\text{Anzahl Versuche}}$$

- AXIOMATISCH / KOLMOGOROFF: abstrakte Funktion mit:
  1) $P(A) \in \mathbb{R}$, $P(A) \geq 0$
  2) $P(S) = 1$
  3) disjunkt A, B gilt: $P(A \cup B) = P(A) + P(B)$

- Normierung: $0 \leq P(A) \leq 1$
- unmögliches Ereignis $P=0$, sicheres Ereignis $P=1$
- $P(A^c) = 1 - P(A)$
- Monotonie: $A \subseteq B \Rightarrow P(A) \leq P(B)$
- Zählregel: $P(A) = \sum_{s \in A} P(\{s\})$
- Zerlegung: $P(A) = P(A \cap B) + P(A \cap B^c)$
- Additionssatz: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Bedingte Wahrscheinlichkeit: $P(B|A) = \dfrac{P(A \cap B)}{P(A)}$

Wahrscheinlichkeitsbaum

$\Rightarrow$ Multiplikationssatz: $P(A \cap B) = P(A) \cdot P(B|A)$

Bayes'sche Formel: $P(A|B) = \dfrac{P(A) \cdot P(B|A)}{P(B)}$

Unabhängige Ereignisse: A, B unabhängig falls:

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

Zufallsvariable: Funktion $X: S \to \mathbb{R}$, die jedem Elementarereignis eine Zahl zuordnet

Realisationen: Werte einer Zufallsvariable

$\Rightarrow$ Wahrscheinlichkeit $P(X=x)$, dass die ZV X den Wert x ann

- diskrete ZV: endlich oder abzählbar viele Werte
- stetige ZV: alle Werte in einem Intervall

Wahrscheinlichkeitsfunktion $P(X=x_i)$ von ZV X

$0 \leq P(X=x_i) \leq 1$ zwischen 0 und 1

$\sum_i P(X=x_i) = 1$ Summe aller immer 1

Verteilungsfunktion: $F(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i)$

Mittelwert und Varianz charakterisieren eine ZV.

$\Rightarrow$ genau wie bei statistischen Variablen, aber mit der Wahrscheinlichkeit dass ZV X den Wert x_i annimmt

Erwartungswert: $\mu_X = E(X) = \sum_i x_i \cdot P(X = x_i)$

Varianz: $\sigma_X^2 = Var(X) = E((X-\mu)^2) = \sum_i (x_i - \mu_X)^2 \cdot P(X = x$

Standardabweichung: $\sigma_X = \sqrt{\sigma_X^2}$

$\hookrightarrow \sigma^2 = E(X^2) - (E(x))^2$

Gesetz der grossen Zahlen: (stat. Var. durch ZV ersetzen)

geht der Stichprobenumfang n gegen ∞,
so wird $\bar{x} \approx \mu_X$ und $s \approx \sigma_X$

- **Bernoulliverteilung:** Münzwurf, abhängig von p (W. Erfolg)

$X \sim Bern(p)$

$P(X=1) = p, \; P(X=0) = 1-p$

$\mu_X = p$

$\sigma_X^2 = p(1-p)$

$\sigma_X^2 = \sqrt{p(1-p)}$

- **Binomialverteilung:** Bernoulli, repetiert n Mal

$X \sim Bin(n,p)$

$P(X=k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$ $\left(\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!} \right.$

$\mu_X = np$

$\sigma_X^2 = np(1-p)$

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

• **Poissonverteilung:** Erfolg pro Einzelversuch gering, aber Versuche unendlich oft wiederholt werden

$X \sim Po(\mu)$

$P(X=x) = \frac{\mu^x \cdot e^{-\mu}}{x!}$ $\left(\mu = \frac{n}{p} \right.$ um Binomial zu ersetzen $\left. \right)$

$\mu_X = \mu$

$\sigma_X^2 = \mu$

$\sigma_X = \sqrt{\mu}$

Stetige Wahrscheinlichkeitsverteilung (Intervall)

stetige Verteilung $\Rightarrow$ Dichtefunktion

$$f(x) \geq 0 \text{ und } \int_{-\infty}^{\infty} f(x)\,dx = 1$$

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx = F(b) - F(a)$$

Verteilungsfunktion: $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt$

$\Rightarrow f(x) = F'(x) \Leftrightarrow$ dichte $=$ d Verteilung

Erwartungswert: $\mu_X = E(x) = \int_{-\infty}^{\infty} x \cdot f(x)\,dx$

Varianz: $\sigma_X^2 = Var(x) = E((X-\mu)^2) = \int_{-\infty}^{\infty} (x-\mu_X)^2 \cdot f(x)\,dx$

Standardabweichung: $\sigma_X = \sqrt{\sigma_X^2}$

Perzentil: $p\% = \int_{-\infty}^{x_{p\%}} f(x)\,dx$

$$P(X \leq x_{p\%}) = F(X_{p\%}) = p\%$$

Median: 50% Perzentil mit $F(m) = \frac{1}{2}$

~~Gleichverteilung:~~

$$\cancel{X \sim U[a,b]}$$

$$\cancel{\mu_X = \frac{1}{2}(b+a)}$$

$$\cancel{\sigma_X^2 = \frac{1}{12}(b-a)^2}$$

$$\cancel{\sigma_X = \sqrt{\frac{1}{12}}\,(b-a)}$$

Rechenregeln: X, Y sind ZV

$E(X+Y) = E(X) + E(Y)$ addition Erwartungswerte

$E(a + b \cdot X) = a + b \cdot E(X)$ skalierung Erwartungswerte

$Var(X+Y) = Var(X) + Var(Y)$ addition Varianzen (unabhäng

$Var(a + b \cdot X) = b^2 \cdot Var(X)$ skalierung Varianzen (quadra

$$\sigma_{X+Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}\ , \quad \sigma(a + bX) = |b| \cdot \sigma(X)$$

Standardisierung: $Z = \dfrac{X - \mu_X}{\sigma_X}$, $E(Z) = 0$, $\sigma_Z = 1$

- **Gleichverteilung:** $X \sim U[a,b]$

$$f(x) = \frac{1}{b-a} \quad \text{im Intervall } [a,b]$$

$$\mu_X = \frac{1}{2}(b+a)$$

$$\sigma_X^2 = \frac{1}{12}(b-a)^2$$

$$\sigma_X = \sqrt{\frac{1}{12}} \cdot (b-a)$$

- **Normalverteilung:** $X \sim N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \Rightarrow \text{Gauss'sche Glockenkurve}$$

$$\mu_x = \mu$$
$$\sigma_x^2 = \sigma^2$$
$$\sigma_x = \sigma = \sqrt{\sigma^2}$$

$$\Phi(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt$$

$\Rightarrow$ Tabellen benutzen, zuerst aber standardisieren!

$$P(X \leq a) = P\left(Z \leq \frac{a-\mu}{\sigma}\right) = \Phi\left(\frac{a-\mu}{\sigma}\right)$$
$$P(X \geq a) = 1 - P(X \leq a) = 1 - \Phi\left(\frac{a-\mu}{\sigma}\right)$$
$$P(b \leq X \leq a) = \Phi\left(\frac{a-\mu}{\sigma}\right) - \Phi\left(\frac{b-\mu}{\sigma}\right)$$

Perzentile: α-te Perzentile, z_α

$P(X \leq x) = 10\% = 0{,}1$

X standardisieren $\Rightarrow Z$ (zb. $N(3,4)$)

$z_\alpha = z_{0,1} = -z_{0,9} = -1{,}28$

$0{,}1 = 10\% = P(Z \leq z_{0,1}) = P(Z \leq -1{,}28)$

standardisierung rückgängig

$$P\left(\frac{X-3}{2} \leq -1{,}28\right) = P(X \leq 0{,}44)$$

Approximation der Binomialverteilung:

gute Approximation bei $n \cdot p \geq 10$, $np(1-p) \geq 10$

Bin(n,p) mit Stetigkeitskorrektur:

$$X \sim Bin(n,p) \Rightarrow Y \sim N(np, np(1-p))$$

$$\Rightarrow P(X = x) \approx P(x - 0{,}5 \leq Y \leq x + 0{,}5) = P\left(\frac{x - 0{,}5 - \mu}{\sigma} \leq Z \leq \frac{x + 0{,}5 - \mu}{\sigma}\right)$$

Stichprobenverteilung:

Stichprobenerhebung $\Rightarrow$ Erkenntnisse über die Verteilung
$\Rightarrow$ schätzen von μ, σ

Verteilung des Mittelwerts:

Zentraler Grenzwertsatz: $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ strebt mit grossem gegen eine Normalverteilung mit:

$$\mu_{\overline{x}} = \mu$$

$$\sigma_{\overline{x}}^2 = \frac{\sigma^2}{n}$$

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

Induktive Statistik (Stichprobe $\Rightarrow$ Bevölkerung)

Parameter θ, Schätzer $\hat{\theta}$

$\Rightarrow$
- Verzerrung: $Bias = E(\hat{\theta}) - \theta$
  unverzerrt bei $E(\hat{\theta}) = \theta$
- Variabilität: $Var(\hat{\theta})$
- Erwartete quadratische Abweichung:
  $(\hat{\theta} - \theta)^2 = Bias^2 + Var(\hat{\theta})$

Verteilungsmittelwert $\mu \Rightarrow$ Stichprobenmittel $\overline{X} \Rightarrow Var(\overline{X}) = \frac{\sigma^2}{n}$

Verteilungsvarianz $\sigma^2 \Rightarrow$ Stichprobenvarianz $\bar{\sigma}^2 \Rightarrow Var(\bar{\sigma}^2) = \frac{\sigma^2}{n}$

Eigenschafts-Anteil $p \Rightarrow$ Anteil in der Stichprobe $\hat{p} \Rightarrow Var(\hat{p}) = \frac{p(1-p)}{n}$

$$\Rightarrow Var(\hat{p}) = \frac{p(1-p)}{n}\left(\cdot\frac{1}{b^2}\right), SA(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}\left(\cdot\frac{1}{b}\right)$$

wenn man SA um b verkleinern will, muss n um b^2 multipliz werden!

Konfidenzintervalle

$\Rightarrow$ Intervall um Schätzer $\hat{\theta}$, indem Parameter θ mit einer gewissen Konfidenz enthalten ist.

Konfidenzniveau $1-\alpha$

Konfidenzgleichung:

$\Rightarrow$ Konfidenzintervall $=$ Schätzer $\pm$ FM $=$ Schätzer $\pm z_{1-\frac{\alpha}{2}} \times SF$

FM $=$ Fehlermarge $= z_{1-\frac{\alpha}{2}} \times SF$

SF $=$ Standardfehler $= \frac{s}{\sqrt{n}}$

$s =$ Standardabweichung

$n =$ Stichprobenzahl

$Z =$ Perzentilfunktion der Standardnormalverteilung

$z = z_{1-\frac{\alpha}{2}}$, so dass $P(-z \leq Z \leq z) = 1-\alpha$

$\alpha =$ Signifikanz

Schätzung des Mittelwerts:

Gegeben $1-\alpha, n, s, \overline{x}$

$SF = \frac{s}{\sqrt{n}}$

$FM = z_{1-\frac{\alpha}{2}} \times SF$, Perzentil z der Standardnormalverteilung

$$[a,b] = \overline{x} \pm z_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

Schätzung eines Anteils einer Eigenschaft:

Gegeben $1-\alpha$, n, $\hat{p}$

$$SF = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$[a,b] = \hat{p} \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Schätzung einer Differenz von Mittelwerten:

$$\Delta = \mu_1 - \mu_2$$

$$SF = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$[a,b] = \bar{x}_1 - \bar{x}_2 \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Schätzung einer Differenz von Anteilen:

$$\Delta = p_1 - p_2$$

$$SF = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$[a,b] = \hat{p}_1 - \hat{p}_2 \pm z_{1-\frac{\alpha}{2}} \cdot SF$$

Frage nach der Stichprobenzahl:

Gegeben FM, s, $1-\alpha$

$$n \geq \left(\frac{z_{1-\frac{\alpha}{2}} \cdot s}{FM}\right)^2$$

Hypothesentests $\Rightarrow$ Aussage über Wert des Parameters bestätigen

Hypothese: H_0, Nullhypothese
H_A, Alternativhypothese

	H_0 korrekt	H_0 falsch
H_0 verworfen	Fehler Typ 1	OK
H_0 beibehalten	OK	Fehler Typ 2

Signifikanz α kontrolliert nur Fehler Typ 1.

Güte eines Tests: $1 - P(\text{Fehler Typ 2})$

- grosses $\alpha \Rightarrow$ grosse Güte
- Je weiter μ von μ_0 entfernt, grössere Güte
- Je kleiner Standardabweichung s, grössere Güte
- Je grösser Probenanzahl n, grössere Güte
- einseitige Tests grössere Güte als zweiseitige

Hypothesentests für Mittelwert μ $\quad H_0: \mu = \mu_0$

Testvariable: $Z = \dfrac{\text{Schätzer-Vermutung}}{\text{Standardfehler}} = \dfrac{\bar{X} - \mu_0}{s/\sqrt{n}}$

Teststatistik: $z = \dfrac{\text{Konkretes Ergebnis-Vermutung}}{\text{Standardfehler}} = \dfrac{\bar{z} - \mu_0}{s/\sqrt{n}}$

p-Wert:
$$PW = \begin{cases} P(Z \geq z) & \text{falls } H_A: \mu > \mu_0 \\ P(Z \leq z) & \text{falls } H_A: \mu < \mu_0 \\ 2P(Z \geq |z|) & \text{falls } H_A: \mu \neq \mu_0 \end{cases}$$

Testentscheidung: $PW \leq \alpha \Rightarrow H_0$ verwerfen zugunsten H_A
$PW > \alpha \Rightarrow H_0$ behalten

Hypothesentests für Anteil p $\quad H_0: p = p_0$

Testvariable: $Z = \dfrac{\bar{P} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$

Teststatistik: $z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$

p-Wert: analog zu "Hypothesentests für Mittelwert μ"

Testentscheidung: analog zu "Hypothesentests für Mittelwert μ"

Hypothesentests für Differenzen von Mittelwerten $\quad H_0: \Delta = \Delta_0$

Testvariable: $Z = \dfrac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$

Teststatistik: $z = \dfrac{\bar{x}_1 - \bar{x}_2 - \Delta_0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$

p-Wert: analog zu "Hypothesentests für Mittelwert μ"

Testentscheidung: analog zu "Hypothesentests für Mittelwert μ"

Hypothesentests für Differenz von Anteilen $H_0: \Delta = \Delta_0$

Testvariable: $Z = \dfrac{\bar{p}_1 - \bar{p}_2 - \Delta_0}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$

Teststatistik: $z = \dfrac{\hat{p}_1 - \hat{p}_2 - \Delta_0}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$

p-Wert: analog zu "Hypothesentests für Mittelwert μ"

Testentscheidung: analog zu "Hypothesentests für Mittelwert μ"

Anwendungen der X^2-Verteilung

H_0: X hat eine vorgegebene Verteilung

$\Rightarrow$ endlich viele Klassen (je mehr, je genauer, aber komplizierter)

h_j: absolute Häufigkeit der Stichprobenelemente in der Klasse j

e_j: erwartete Häufigkeit laut H_0

$$x^2 = \sum_{j=1}^{K} \frac{(h_j - e_j)^2}{e_j}$$

wenn x_2 zu gross, lehnt man H_0 ab!

$\Rightarrow$ Freiheitsgrade, Tabelle