

Formelsammlung Statistik

Inhalt

Beschreibende Statistik	2
Relative Häufigkeiten	2
Bedingte Häufigkeiten	2
Arithmetisches Mittel	2
Geometrisches Mittel	2
Median	2
Quartile	3
Modus	3
Varianz	3
Standardabweichung	3
Variationskoeffizient	3
Gini-Koeffizient	3
Chiquadrat	4
Cramers V	4
Kovarianz	4
Korrelation	4
Spearman'sche Rangkorrelation	4
Regressionsgerade	4
Bestimmtheitsmaß	5
Wahrscheinlichkeitstheorie	5
Rechnen mit Wahrscheinlichkeiten	5
Rechnen mit bedingten Wahrscheinlichkeiten	5
Wahrscheinlichkeiten von unabhängigen Ereignissen	5
Binomialkoeffizient	5
Fakultät	5
Urnenmodell – Ziehen ohne Zurücklegen – Hypergeometrische Verteilung	5
Urnenmodell – Ziehen mit Zurücklegen - Binomialverteilung	6
Normalverteilung	6
Standard-Normalverteilung	6
Rechnen mit Standard-Normalverteilung	7
Schließende Statistik	7

Schätzen und Testen einer relativen Häufigkeit	7
Schätzen und Testen eines Mittelwerts	9
p-Wert	10
Testen von Hypothesen über Regressionskoeffizienten	10
Erforderlicher Stichprobenumfang in großen Grundgesamtheiten	11
Standardnormalverteilung	12
Binomialkoeffizient.....	13
Fakultäten.....	14

Beschreibende Statistik

Relative Häufigkeiten

$$p_{ij} = \frac{h_{ij}}{N}$$

Mit i=Zeilen und j=Spalten

Bedingte Häufigkeiten

$$p_{j|i=k} = \frac{h_{kj}}{N_{k.}} \quad \text{oder} \quad p_{i|j=l} = \frac{h_{il}}{N_{.l}}$$

Mit k als einer bestimmten Zeile bzw. l als einer bestimmten Spalte

Arithmetisches Mittel

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

$$\bar{x} = \frac{\sum_{i=1}^k x_i \cdot h_i}{N} = \sum_{i=1}^k x_i \cdot \frac{h_i}{N} = \sum_{i=1}^k x_i \cdot p_i$$

Mit k= Anzahl der verschiedenen Merkmalsausprägungen, h_i = absolute Häufigkeit der Merkmalsausprägungen und p_i = relative Häufigkeit der Merkmalsausprägungen

Geometrisches Mittel

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

Median

$$\tilde{x} = \begin{cases} x_{\frac{n+1}{2}} & \text{für n ungerade} \\ \frac{1}{2} \cdot \left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1} \right) & \text{für n gerade} \end{cases}$$

Quartile

$$Q_p = \begin{cases} x_{[n \cdot p]} & \text{für } n \cdot p \text{ nicht ganzzahlig} \\ \frac{1}{2} \cdot (x_{n \cdot p} + x_{n \cdot p + 1}) & \text{für } n \cdot p \text{ ganzzahlig} \end{cases}$$

Mit $[n \cdot p]$ die kleinste ganze Zahl, die größer oder gleich $n \cdot p$ ist (einfach $n \cdot p$ aufrunden bis zur nächsten vollen Zahl). Mit p als p -tes Quartil (z.B. $Q_{0,25}$, $Q_{0,75}$). Achtung. $Q_{0,50}$ =Median.

Modus

x_{mod} = Merkmalsausprägung mit der größten relativen Häufigkeit

Varianz

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

aus (absoluten) Häufigkeiten $s^2 = \frac{\sum_{i=1}^k (x_i - \bar{x})^2 \cdot h_i}{N}$

aus relativen Häufigkeiten $s^2 = \sum_{i=1}^k (x_i - \bar{x})^2 \cdot p_i$

Standardabweichung

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

Variationskoeffizient

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

Mit s = Standardabweichung und \bar{x} als Mittelwert des Merkmals

Gini-Koeffizient

Normierter Gini-Koeffizient = (Fläche unter der Diagonale – Fläche unter der Lorenzkurve) / Fläche Maximalkonzentration

Wobei die Fläche der Maximalkonzentration $= \frac{1}{2} \cdot \left(1 - \frac{1}{N}\right)$ mit N =Anzahl der Erhebungseinheiten

Chiquadrat

$$\chi^2 = N \cdot \sum \frac{(p_{ij}^b - p_{ij}^e)^2}{p_{ij}^e}$$

Mit b=beobachtet und e=erwartet.

wobei $p_{ij}^e = p_j^b \cdot p_i^b$ (Produkt der beobachteten Randhäufigkeiten)

Cramers V

$$V = \sqrt{\frac{\chi^2}{N \cdot (\min(k, l) - 1)}}$$

Mit k und l als Anzahl der Merkmalsausprägungen der beiden Merkmale

Kovarianz

$$s_{xy} = \frac{\sum_{i=1}^N (x_i - \bar{x}) \cdot (y_i - \bar{y})}{N}$$

Korrelation

$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

Mit s_x und s_y als Standardabweichung der Merkmale x und y.

Spearmanische Rangkorrelation

$$r = \frac{s_{uv}}{s_u \cdot s_v}$$

Mit s_u und s_v als Standardabweichung der Ränge u und v zweier Merkmale. Und s_{uv} als Kovarianz der Ränge u und v zweier Merkmale

Regressionsgerade

$$y = b_1 \cdot x + b_2$$

Mit dem Regressionskoeffizienten (Steigung): $b_1 = \frac{s_{xy}}{s_x^2}$

Und der Konstante (Achsenabschnitt): $b_2 = \bar{y} - b_1 \cdot \bar{x}$

Y wird häufig als Regressand oder abhängige Variable bezeichnet, x als Regressor oder unabhängige Variable, b_1 als Regressionskoeffizient und b_2 als Konstante.

Bestimmtheitsmaß

$$B = r^2$$

Mit r als Korrelationskoeffizient

Wahrscheinlichkeitstheorie

Rechnen mit Wahrscheinlichkeiten

$$\Pr(x = a) = \frac{\text{Anzahl der günstigen Ereignisse}}{\text{Anzahl möglicher Ereignisse}}$$

Rechnen mit bedingten Wahrscheinlichkeiten

$$\Pr(x = a \mid x \neq b) = \frac{\text{Anzahl der günstigen Ereignisse}}{\text{Anzahl möglicher Ereignisse} - \text{Anzahl der unmöglichen Ereignisse (b)}}$$

Mit a = günstiges Ereignis und b=unmögliche Ereignis

Wahrscheinlichkeiten von unabhängigen Ereignissen

$$\Pr(x = a \text{ \& } y = c) = \Pr(x = a) \cdot \Pr(y = c)$$

Binomialkoeffizient

$$\binom{a}{b} = \frac{a!}{b! \cdot (a-b)!}$$

Fakultät

$$a! = \prod_{k=1}^a k = 1 \cdot 2 \cdot \dots \cdot a$$

Urnenmodell – Ziehen ohne Zurücklegen – Hypergeometrische Verteilung

$$\Pr(x = a) = \frac{\binom{A}{a} \cdot \binom{N-A}{n-a}}{\binom{N}{n}}$$

N Kugeln, A weiße, N-A schwarze, n werden zufällig gezogen, x = a ... Anzahl der gezogenen weißen Kugeln, n-a Anzahl der gezogenen schwarzen Kugeln

$$\text{Theoretischer Mittelwert: } \mu = n \cdot \frac{A}{N}$$

$$\text{Theoretische Varianz: } \sigma^2 = n \cdot \frac{A}{N} \cdot \left(1 - \frac{A}{N}\right) \cdot \frac{N-n}{N-1}$$

Achtung: Bei großen Stichprobenumfängen ($n > 100$) nähert sich die hypergeometrische Verteilung der Normalverteilung an → zentraler Grenzwertsatz der Statistik → Erwartungswert und theoretische Varianz dieser Normalverteilung entsprechen jenen der hypergeometrischen Verteilung

Urnenmodell – Ziehen mit Zurücklegen - Binomialverteilung

$$Pr(x = a) = \binom{n}{a} \cdot \pi^a \cdot (1 - \pi)^{n-a}$$

Mit $\pi = A/N$ und $1 - \pi = 1 - A/N$

N Kugeln, A weiße, N-A schwarze, n werden zufällig gezogen, x = a ... Anzahl der gezogenen weißen Kugeln, n-a Anzahl der gezogenen schwarzen Kugeln

Theoretischer Mittelwert: $\mu = n \cdot \pi$

Theoretische Varianz: $\sigma^2 = n \cdot \pi \cdot (1 - \pi)$

Achtung: Binomialverteilung als Näherungslösung für Hypergeometrische Verteilung bei großen Grundgesamtheiten und kleinen Stichproben

Normalverteilung

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

Mittelwert: μ

Varianz: σ^2

$$Pr(x \leq x_0) = \int_{-\infty}^{x_0} f(x) dx = \int_{-\infty}^{x_0} \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot e^{-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}} dx$$

Standard-Normalverteilung

Dichte:
$$f(u) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2} u^2}$$

Mittelwert: $\mu = 0$

Varianz: $\sigma^2 = 1$

$$Pr(u \leq u_0) = \int_{-\infty}^{u_0} f(u) du = \int_{-\infty}^{u_0} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2} u^2} du$$

Standardisierung der Normalverteilung durch $u_0 = \frac{x_0 - \mu}{\sigma}$

Mit μ = theoretischer Mittelwert und σ = Standardabweichung der Zufallsvariable x

Rechnen mit Standard-Normalverteilung

Gegenwahrscheinlichkeit: $Pr(u > u_0) = 1 - Pr(u \leq u_0)$

Negative Werte: $Pr(u \leq -u_0) = Pr(u > u_0) = 1 - Pr(u \leq u_0)$

Intervalle: $Pr(u_1 \leq u \leq u_2) = Pr(u \leq u_2) - Pr(u \leq u_1)$

Schließende Statistik

Punktschätzer (aus der Stichprobe)	Parameter (in der Grundgesamtheit)
Relative Häufigkeit p	Relative Häufigkeit π
Mittelwert \bar{x}	Mittelwert μ
Stichprobenvarianz s^2	Varianz σ^2
Differenz zweier relativer Häufigkeiten (oder Mittelwerte) d	Differenz zweier relativer Häufigkeiten (oder Mittelwerte) δ
Chiquadrat χ_{err}^2	Chiquadrat χ^2
Korrelationskoeffizient r	Korrelationskoeffizient ρ
Regressionskoeffizient b_1, b_2	Regressionskoeffizient β_1, β_2

Schätzen und Testen einer relativen Häufigkeit

Hypergeometrische Verteilung

Theoretischer Erwartungswert: $\mu = n \cdot \pi$

Theoretische Varianz: $\sigma^2 = n \cdot \pi \cdot (1 - \pi) \cdot \frac{N-n}{N-1}$

Theoretische Varianz bei großen Grundgesamtheiten: $\sigma^2 = n \cdot \pi \cdot (1 - \pi)$

Mit $\pi = A/N$ als Häufigkeit in der Grundgesamtheit und N=Grundgesamtheit

Konfidenzintervall für Punktschätzer π zur Sicherheit $1-\alpha$ in großen Grundgesamtheiten und großen Stichproben

obere Intervallgrenze :
$$\pi_o = p + u_{1-\alpha/2} \cdot \sqrt{\frac{p \cdot (1-p)}{n}}$$

untere Intervallgrenze:
$$\pi_u = p - u_{1-\alpha/2} \cdot \sqrt{\frac{p \cdot (1-p)}{n}}$$

Mit $p = a/n$ als relativer Häufigkeit in der Stichprobe und n =Stichprobenumfang

Zweiseitiges Testen von Hypothesen zum Signifikanzniveau α über eine relative Häufigkeit

$H_0: \pi = \pi_0$ und $H_1: \pi \neq \pi_0$

Zweiseitiger Test: obere Schranke :
$$p_o = \pi + u_{1-\alpha/2} \cdot \sqrt{\frac{\pi \cdot (1-\pi)}{n}}$$

untere Schranke:
$$p_u = \pi - u_{1-\alpha/2} \cdot \sqrt{\frac{\pi \cdot (1-\pi)}{n}}$$

Beibehaltung von H_0 wenn gilt $p \in [p_u; p_o]$, Akzeptanz von H_1 wenn $p \notin [p_u; p_o]$

Einseitiges Testen von Hypothesen zum Signifikanzniveau α über eine relative Häufigkeit

$H_0: \pi \leq \pi_0$ und $H_1: \pi > \pi_0$

Einseitiger Test: obere Schranke :
$$p_o = \pi + u_{1-\alpha} \cdot \sqrt{\frac{\pi \cdot (1-\pi)}{n}}$$

Beibehaltung von H_0 wenn gilt $p \leq p_o$, Akzeptanz von H_1 wenn $p > p_o$

Einseitiges Testen von Hypothesen zum Signifikanzniveau α über eine relative Häufigkeit

$H_0: \pi \geq \pi_0$ und $H_1: \pi < \pi_0$

Einseitiger Test: untere Schranke :
$$p_u = \pi - u_{1-\alpha} \cdot \sqrt{\frac{\pi \cdot (1-\pi)}{n}}$$

Beibehaltung von H_0 wenn gilt $p \geq p_u$, Akzeptanz von H_1 wenn $p < p_u$

Schätzen und Testen eines Mittelwerts

Bei großen Stichproben annähernd normalverteilt

Theoretischer Erwartungswert: $\mu = \bar{x}$

Theoretische Varianz: $\sigma_x^2 = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$

Theoretische Varianz bei großen Grundgesamtheiten: $\sigma_x^2 = \frac{\sigma^2}{n}$

Mit N = Anzahl der Beobachtungen in der Grundgesamtheit und n=Stichprobenumfang

Konfidenzintervall für Punktschätzer μ zur Sicherheit $1-\alpha$ in großen Grundgesamtheiten und großen Stichproben

obere Intervallgrenze : $\mu_o = \bar{x} + u_{1-\alpha/2} \cdot \sqrt{\frac{s^2}{n}}$

untere Intervallgrenze: $\mu_u = \bar{x} - u_{1-\alpha/2} \cdot \sqrt{\frac{s^2}{n}}$

Mit \bar{x} = Mittelwert in der Stichprobe, s^2 als Stichprobenvarianz und n=Stichprobenumfang

Zweiseitiges Testen von Hypothesen zum Signifikanzniveau α über einen Mittelwert

$H_0: \mu = \mu_0$ und $H_1: \mu \neq \mu_0$

Zweiseitiger Test: obere Schranke : $\bar{x}_o = \mu + u_{1-\alpha/2} \cdot \sqrt{\frac{\sigma^2}{n}}$

untere Schranke: $\bar{x}_u = \mu - u_{1-\alpha/2} \cdot \sqrt{\frac{\sigma^2}{n}}$

Ersetzen mit
Stichproben-
varianz s^2 wenn σ^2
unbekannt ist

Beibehaltung von H_0 wenn gilt $\bar{x} \in [\bar{x}_u; \bar{x}_o]$, Akzeptanz von H_1 wenn $\bar{x} \notin [\bar{x}_u; \bar{x}_o]$

Einseitiges Testen von Hypothesen zum Signifikanzniveau α über einen Mittelwert

$H_0: \mu \leq \mu_0$ und $H_1: \mu > \mu_0$

Einseitiger Test: obere Schranke : $\bar{x}_o = \mu + u_{1-\alpha} \cdot \sqrt{\frac{\sigma^2}{n}}$

Beibehaltung von H_0 wenn gilt $\bar{x} \leq \bar{x}_o$, Akzeptanz von H_1 wenn $\bar{x} > \bar{x}_o$

Ersetzen mit
Stichproben-
varianz s^2 wenn σ^2
unbekannt ist

Einseitiges Testen von Hypothesen zum Signifikanzniveau α über einen Mittelwert

$H_0: \mu \geq \mu_0$ und $H_1: \mu < \mu_0$

Einseitiger Test: untere Schranke : $\bar{x}_u = \mu - u_{1-\alpha} \cdot \sqrt{\frac{\sigma^2}{n}}$

Beibehaltung von H_0 wenn gilt $\bar{x} \geq \bar{x}_u$, Akzeptanz von H_1 wenn $\bar{x} < \bar{x}_u$

p-Wert

p-Wert ist die Irrtumswahrscheinlichkeit, die gemacht wird wenn man aufgrund der Daten aus der Stichprobe die H_1 akzeptiert obwohl die H_0 zutrifft.

p-Wert bei zweiseitigen Fragestellungen: α_2

p-Wert bei einseitigen Fragestellungen: α_1

Beziehung zwischen beiden p-Werten: $\alpha_1 = \alpha_2 / 2$

Testen von Hypothesen über Regressionskoeffizienten

Regressionsgerade in der Grundgesamtheit $y = \beta_1 \cdot x + \beta_2$

Mit y =abhängige Variable, x =unabhängige Variable, β_1 =Steigung, β_2 =Achsenabschnitt (Konstante)

Zweiseitiges Testen von Hypothesen zum Signifikanzniveau α über die Steigung β_1

$H_0: \beta_1 = 0$ und $H_1: \beta_1 \neq 0$

Obere und untere Schranken $b_{1o} = u_{1-\alpha/2} \cdot s_{b1}$ und $b_{1u} = -u_{1-\alpha/2} \cdot s_{b1}$ sind t-verteilt
→ in großen Stichproben annähernd normalverteilt

zweiseitiger Test: Obere Schranke: $b_{1o} = +u_{1-\alpha/2} \cdot \sqrt{\frac{1-r^2}{n-2} \cdot \frac{s_y^2}{s_x^2}}$

Zweiseitiger Test: Untere Schranke: $b_{1u} = -u_{1-\alpha/2} \cdot \sqrt{\frac{1-r^2}{n-2} \cdot \frac{s_y^2}{s_x^2}}$

Beibehaltung von H_0 wenn $b_1 \in [b_{1u}; b_{1o}]$, Akzeptanz von H_1 wenn $b_1 \notin [b_{1u}; b_{1o}]$

Einseitiges Testen von Hypothesen zum Signifikanzniveau α über die Steigung β_1

$H_0: \beta_1 \leq 0$ und $H_1: \beta_1 > 0$

Einseitiger Test: Obere Schranke: $b_{1o} = +u_{1-\alpha} \cdot \sqrt{\frac{1-r^2}{n-2} \cdot \frac{s_y^2}{s_x^2}}$

Beibehaltung von H_0 wenn $b_1 \leq b_{1o}$, Akzeptanz von H_1 wenn $b_1 > b_{1o}$

Einseitiges Testen von Hypothesen zum Signifikanzniveau α über die Steigung β_1

$H_0: \beta_1 \geq 0$ und $H_1: \beta_1 < 0$

Einseitiger Test: Untere Schranke: $b_{1u} = -u_{1-\alpha} \cdot \sqrt{\frac{1-r^2}{n-2} \cdot \frac{s_y^2}{s_x^2}}$

Beibehaltung von H_0 wenn $b_1 \geq b_{1u}$, Akzeptanz von H_1 wenn $b_1 < b_{1u}$

Erforderlicher Stichprobenumfang in großen Grundgesamtheiten

ε : Tolerierte Schwankungsbreite

p : Parameter

n : Stichprobenumfang

$$n = \frac{u_{1-\alpha/2}^2}{\varepsilon^2} \cdot p \cdot (1 - p)$$

$$Pr(u \leq u_0) = \int_{-\infty}^{u_0} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}u_0^2} du$$
[illegible]

Tabelle: **Binomialkoeffizienten** $\binom{n}{k}$

$n \backslash k$	0	1	2	3	4	5	6	7	8	9	10
0	1										
1	1	1									
2	1	2	1								
3	1	3	3	1							
4	1	4	6	4	1						
5	1	5	10	10	5	1					
6	1	6	15	20	15	6	1				
7	1	7	21	35	35	21	7	1			
8	1	8	28	56	70	56	28	8	1		
9	1	9	36	84	126	126	84	36	9	1	
10	1	10	45	120	210	252	210	120	45	10	1
11	1	11	55	165	330	462	462	330	165	55	11
12	1	12	66	220	495	792	924	792	495	220	66
13	1	13	78	286	715	1287	1716	1716	1287	715	286
14	1	14	91	364	1001	2002	3003	3432	3003	2002	1001
15	1	15	105	455	1365	3003	5005	6435	6435	5005	3003
16	1	16	120	560	1820	4368	8008	11440	12870	11440	8008
17	1	17	136	680	2380	6188	12376	19448	24310	24310	19448
18	1	18	153	816	3060	8568	18564	31824	43758	48620	43758
19	1	19	171	969	3876	11628	27132	50388	75582	92378	92378
20	1	20	190	1140	4845	15504	38760	77520	125970	167960	184756

Beispiele: $\binom{8}{3} = 56$; $\binom{15}{12} = \binom{15}{15-12} = \binom{15}{3} = 455$

Fakultät

Jede natürliche Zahl n hat eine Fakultät. Sie ist das Produkt der natürlichen Zahlen, die kleiner oder gleich der Zahl n sind.

Man schreibt sie als $n! = 1 \cdot 2 \cdot 3 \cdot \dots \cdot (n-1) \cdot n$ und liest sie *n Fakultät*.

Es ist zweckmäßig, $1! = 1$ und auch $0! = 1$ zu definieren.

$0! = 1$	$5! = 120$	$10! = 3.628.800$
$1! = 1$	$6! = 720$	$11! = 39.916.800$
$2! = 2$	$7! = 5.040$	$12! = 479.001.600$
$3! = 6$	$8! = 40.320$	$13! = 6.227.020.800$
$4! = 24$	$9! = 362.880$	$14! = 8,717829120 \cdot 10^{10}$