Deskriptive Statistik

Merkmalsträger

- Qualitativ/Kategoriell
  - Nominal: Parteien
  - Ordinal: Prüfungresultat schlecht, mittel, gut
- Quantitativ/Metrisch
  - Diskret: Würfel 1-6, Anzahl (es gibt Lücken)
  - Stetig: Geschwindigkeit

Verteilungsfunktionen

- **PMF** distrekte Merkmale
- **PDF** stetige Merkmale
- **CDF** kumulative Verteilungsfunktion

Klasse	[100,200[	[200,400[
hi	5	15
fi	5/20	15/20
PDF	5/20/100	15/20/200
CDF	5/20	20/20

Durchschnitt

= $\bar{x}$, arithmetisches Mittel, Mittelwert, Erwartungswert = fi * klassenmitte_i

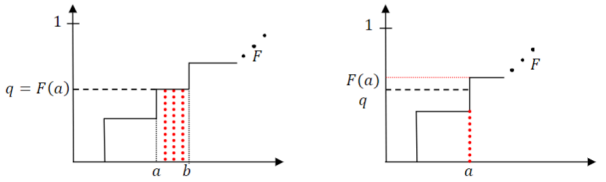
q-Quantil mit n Stichproben

Ist $n \cdot q$ ganze Zahl dann $R_q = \frac{1}{2}(x_{n \cdot q} + x_{n \cdot q+1})$

Sonst $R_q = x_{\lceil n \cdot q \rceil}$

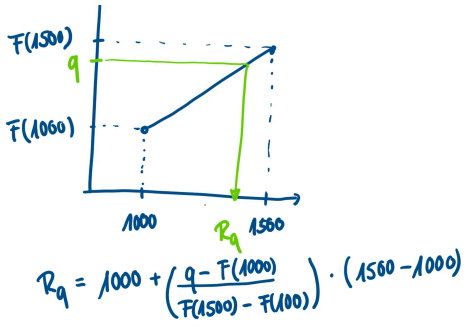
0.5-Quantil = 2. Quantil, Medianwert, Zentralwert, x_{med}

Quantile aus CDF: $\frac{a+b}{2}$



$$R_q = \frac{b-a}{F(a) - F(b)} \cdot (q - F(a-1)) + a$$

$$q = F(a-1) + \frac{R_q \cdot (F(a) - F(a-1))}{b-a}$$



Boxplot

Box: 1., 2. und 3. Quantil

Whisker: Maximal 1.5 x Interquartilsabstand entwernt von Q1/Q3

Ausreisser: alle Ausserhalb Whisker

Modalwert / Modus / x_{mod}

= Der Wert der am häufigsten vorkommt

Varianz

x=stichproben, a=merkmalwerte

$$\tilde{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{m} h_i \cdot (a_i - \bar{x})^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \bar{x}^2 = \left(\frac{1}{n}\sum_{i=1}^{m} a_i^2 \cdot h_i\right) - \bar{x}^2 = \left(\sum_{i=1}^{m} a_i^2 \cdot f_i\right) - \bar{x}^2 = \bar{x^2} - \bar{x}^2$$

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

Standardabweichung

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

Form der Verteilung

- Rechtsschief: $x_{\text{mod}} < x_{\text{med}} < \tilde{x}$ = Maximum auf linker Seite
- Linksschief: $x_{\text{mod}} > x_{\text{med}} > \tilde{x}$
- Symmetrisch: $x_{\text{mod}} = x_{\text{med}} = \tilde{x}$
- unimodal = 1 Maximum, bimodal = 2 Maxima etc.

Bivariate Daten

- Zwei Kategorien: Mosaikplot
- 1 Kategorie, 1 Metrisch: Boxplot
- 2 Metrisch: Scatterplot

Pearson-Korrelationskoeffizient

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

$$r_{xy} = \frac{\tilde{s}_{xy}}{\tilde{s}_x \cdot \tilde{s}_y} = \frac{\bar{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\bar{x^2} - \bar{x}^2} \cdot \sqrt{\bar{y^2} - \bar{y}^2}}$$

Nahe bei 1 = hohe Korrelation

x	1	2	1.5
y	4	-1	1.5

x	1	2	1.5
x2	1	4	2.25
y2	16	1	2.25
xy	4	-2	2.25

Spearman-Rangkorrelationskoeffizient

x	5	2	
y	9	11	
rgx	2	1	1.5
rgy	1	2	1.5
rgx - avg rgx	0.5	-0.5	
rgy - avg rgy	-0.5	0.5	

$$R_{\mathrm{Sp}} = \frac{\sum_{i=1}^{n} \left(rg\left(x_i\right) - \overline{rg(x)}\right)\left(rg\left(y_i\right) - \overline{rg(y)}\right)}{\sqrt{\sum_{i=1}^{n} \left(rg\left(x_i\right) - \overline{rg(x)}\right)^2} \cdot \sqrt{\sum_{i=1}^{n} \left(rg\left(y_i\right) - \overline{rg(y)}\right)^2}}$$

Kombinatorik

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

$$\binom{n}{0} = 1$$

Variation (mit Reihenfolge)		Kombination (ohne Reihenfolge)	
Mit Wiederholung	Ohne Wiederholung	Mit Wiederholung	Ohne Wiederholung
n^k	$\frac{n!}{(n-k)!}$	$\binom{n+k-1}{k}$	$\binom{n}{k}$

1. Nummerncode
2. Platzierung Wettkampf
3. x objekte aus y schalen
4. x zahlen im lotto ziehen

k	Teilmengen mit k Elementen	Anzahl
0	{} «leere Menge»	$\binom{4}{0} = 1$
1	{1},{2},{3},{4}	$\binom{4}{1} = 4$
2	{1,2},{1,3}, {1,4}{2,3},{2,4},{3,4}	$\binom{4}{2} = 6$
3	{1,2,3},{1,2,4},{1,3,4},{2,3,4}	$\binom{4}{3} = 4$
4	{1,2,3,4}	$\binom{4}{4} = 1$

Elementare Wahrscheinlichkeit

- E(X) = μ (Lagemass)
- V(X) = σ^2 (Streumass)
- S(X) = σ

Bedingte Wahrscheinlichkeit

	X	Y	Summe
A	2	4	6
B	3	1	4
Summe	5	5	10

Ereignisbaum

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

2 Ereignisse sind stochastisch unabhängig wenn gilt:

$$P(X = x \wedge Y = y) = P(X = x) \cdot P(Y = y)$$

Bei 3 Ereignissen müssen alle Teilmengen unabhängig sein. Für stochastisch unabhängige Ereignisse gilt:

$$E(X \cdot Y) = E(X) \cdot E(Y)$$

und

$$V(X + Y) = V(X) + V(Y)$$

Z.b. P(X&A) = 2/10. P(X) = 1/2. P(A) = 6/10. nein.

Verteilungen

	diskrete Zufallsvariablen	**stetige Zufallsvariablen**
Dichtefunktion / PMF bzw. PDF	$f(x) = P(X = x)$	$f(x) = F'(x) \neq P(X = x)$!!!
Kumulative Verteilungsfunktion/ CDF	$F(x) = P(X \leq x) = \sum_{x \leq X} f(x)$	$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du$
Wahrscheinlichkeiten	$P(a \leq X \leq b) = \sum_{a \leq x \leq b} f(x)$	$P(a \leq X \leq b) = \int_{a}^{b} f(x)dx$
Graphische Darstellung von f	Stabdiagramm	Graph
Erwartungswert	$E(X) = \sum_{x \in \mathbb{R}} f(x) \cdot x$	$E(X) = \int_{-\infty}^{\infty} f(x) \cdot x \, dx$
Varianz	$V(X) = \sum_{x \in \mathbb{R}} f(x) \cdot (x - E(X))^2$	$V(X) = \int_{-\infty}^{\infty} f(x) \cdot (x - E(X))^2 \, dx$

Für diskrete und stetige Zufallsvariablen X und Y gelten die folgenden Regeln:

(1) **Linearität des Erwartungswertes**:
$E(X + Y) = E(X) + E(Y)$ und $E(\alpha X) = \alpha E(X)$ mit $\alpha \in \mathbb{R}$.

(2) **Verschiebungssatz für die Varianz**: $V(X) = E(X^2) - (E(X))^2$
mit $E(X^2) = \int_{-\infty}^{\infty} f(x) \cdot x^2 \, dx$

(3) $V(\alpha X + \beta) = \alpha^2 \cdot V(X)$ mit $\alpha, \beta \in \mathbb{R}$.

(4) Sind X und Y unkorreliert ($COV(X, Y) = 0$), so gilt: $V(X + Y) = V(X) + V(Y)$.

E(Z) = 2E(X) + E(Y), z = 2x+y V(Z) = 4V(X) + V(Y)

Hypergeometrische Verteilung

Es gibt M Merkmalsträger in N, x Merkmalsträger in n Stichproben ohne zurücklegen

$$P(X = x) = \frac{\binom{M}{x} \cdot \binom{N - M}{n - x}}{\binom{N}{n}}$$

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

$$VAR(X) = n \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right) \cdot \frac{N - n}{N - 1}$$

Bernoulliverteilung

- P(X = 1) = p, P(X = 0) = q
- E(X) = p
- E(X^2) = p *1* + *q* 0 = p
- V(X) = pq

Binomialverteilung

P(X = 1) tritt x-Mal ein bei n Wiederholungen mit zurücklegen. zb 3x Kopf

$$\binom{n}{x} \cdot p^x \cdot q^{n-x}$$

Poissonverteilung

Wahrscheinlichkeit dass Ereignis in einem Intervall i x-Mal vorkommt. Einheit von λ ist in Ereignisse/i

$$P(X = x) = \frac{\lambda^x}{x!} \cdot e^{-\lambda}, \lambda > 0$$

Annäherung der Binominalverteilung: $\lambda = n \cdot p$

E(X) = VAR(X) = lambda

Gaussverteilung

$$\varphi(\mu, \sigma) = \frac{1}{2\pi \cdot \sigma} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$E(X) = \mu$$

$$V(X) = \sigma^2$$

Normalverteilung

= $\varphi(0, 1)$

Bei der Normalverteilung liegen:

- ca. 68% zwischen $\mu - \sigma$ und $\mu + \sigma$
- ca. 95% zwischen $\mu - 2\sigma$ und $\mu + 2\sigma$
- ca. 99.7% zwischen $\mu - 3\sigma$ und $\mu + 3\sigma$

$$U = \frac{X - \mu}{\sigma}$$

$$P(X \leqslant x) = P(U \leqslant u)$$

$$P(\mu - e \leqslant X \leqslant \mu + e) = 2 * \phi\left(\frac{e}{\sigma}\right) - 1$$

Zentraler Grenzwertsatz

Eine Grösse ist näherungsweise normalverteilt, wenn sie von einer Überlagerung von vielen unabhängigen zufälligen Einflüssen abhängt.

Annäherung der Binominalverteilung: $\mu = np$ und $\sigma^2 = npq$ wenn $npq > 9$

Stetigkeitskorrektur: P(5 < X <= 10) = P(5.5 <= X <= 10.5)

Annäherung der Poissonverteilung: $\mu = \lambda$ und $\sigma^2 = \lambda$

Methode der kleinsten Quadrate

Zusammenfassung Regressionsgerade:
Die *Regressionsgerade* $g(x) = mx + d$ mit den Parametern m und d ist die Gerade, für die die *Residualvarianz* $\tilde{s}_\epsilon{}^2$ minimal ist.

Die Regressionsgerade hat die Steigung

$$m = \frac{\tilde{s}_{xy}}{\tilde{s}_x{}^2}$$

und den y-Achsenabschnitt

$$d = \bar{y} - m\bar{x}$$

Für die zugehörige (minimale) Residualvarianz gilt:

$$\tilde{s}_\epsilon{}^2 = \tilde{s}_y{}^2 - \frac{\tilde{s}_{xy}{}^2}{\tilde{s}_x{}^2}$$

mit:
Varianz der x_i-Werte

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

Varianz. der y_i-Werte

$$\tilde{s}_y{}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \left(\frac{1}{n}\sum_{i=1}^{n} y_i{}^2\right) - \bar{y}^2$$

Kovarianz

$$\tilde{s}_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y}) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \bar{x}\cdot\bar{y}$$

arithmetische Mittelwerte

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

(in y richtung)

Nichtlineares Verhalten

Statt mit den originalen Werten kann z.B. mit dem Logarithmus der Werte gerechnet werden:

Ausgangsfunktion	Transformation
$y = q \cdot x^m$	$\log(y) = \log(q) + m \cdot \log(x)$
$y = q \cdot m^x$	$\log(y) = \log(q) + \log(m) \cdot x$
$y = q \cdot e^{m \cdot x}$	$\ln(y) = \ln(q) + m \cdot x$
$y = \frac{1}{q + m \cdot x}$	$V = q + m \cdot x; \quad V = \frac{1}{y}$
$y = q + m \cdot \ln(x)$	$y = q + m \cdot U; \quad U = \ln(x)$
$y = \frac{1}{q \cdot m^x}$	$\log\left(\frac{1}{y}\right) = \log(q) + \log(m) \cdot x$

Residuenplot

$\epsilon_i = y_i - \hat{y}_i$

$\hat{y}_i = g(x_i) = mx_i + d$

Bestimmtheitsmass

Zusammenfassung Bestimmtheitsmass:
Die Totale Varianz setzt sich zusammen aus der Residualvarianz und der Varianz der prognostizierten Werte:

$$\tilde{s}_y{}^2 = \tilde{s}_\epsilon{}^2 + \tilde{s}_y{}^2 \text{ bzw. } s_y^2 = s_\epsilon^2 + s_{\hat{y}}^2$$

Das *Bestimmtheitsmass* R^2 beurteilt die globale Anpassungsgüte einer Regression über den Anteil der prognostizierten (erklärten) Varianz $s_{\hat{y}}^2$ an der totalen Varianz s_y^2:

$$R^2 = \frac{\tilde{s}_{\hat{y}}{}^2}{\tilde{s}_y{}^2} \text{ bzw. } R^2 = \frac{s_{\hat{y}}^2}{s_y^2}$$

Das *Bestimmtheitsmass* R^2 stimmt überein mit dem Quadrat *des Korrelations-koeffizienten (nach Bravais-Pearson)*

$$R^2 = \frac{\tilde{s}_{xy}{}^2}{\tilde{s}_x{}^2\tilde{s}_y{}^2} = r_{xy}{}^2 \text{ bzw. } R^2 = \frac{s_{xy}{}^2}{s_x{}^2 s_y{}^2} = r_{xy}{}^2$$

R^2 % der Gesamtvarianz in den y-Daten kann durch die Regressionsgerade erklärt werden

Mehrere Variablen

Gegeben sind:

Messwerte der Kategorie C: $y = \begin{pmatrix} y_1 \\ \dots \\ y_n \end{pmatrix}$

Andere Kategorien: $X = \begin{pmatrix} A_1 & B_1 & 1 \\ \dots & \dots & 1 \\ A_n & B_n & 1 \end{pmatrix}$

Lösbar mit $p = \left(X^T X\right)^{-1} X^T y$ für $y = Xp + \epsilon$

$\hat{y} = b_1 x_1 + b_2 x_2 + a$

$R^2 = \frac{\Sigma\left(\hat{y}_i - \bar{y}\right)^2}{\Sigma\left(y_i - \bar{y}\right)^2}$

Parameter- und Intervallschätzung

Schätzungen für $\hat{\theta}$

$\hat{\mu} = \bar{x}$ oder x_{med}

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

$\hat{p} = fi = \frac{M}{N}$ bei M Merkmalsträgern in M

Maximum-Likelihood-Schätzung

Für Normalverteilung: $\hat{\mu} = \bar{x}$ und $\hat{\sigma}^2 = \tilde{s}^2$

Für Poissonverteilung: $\lambda = \frac{1}{\bar{x}}$

Vertrauensintervalle

Für Vertrauensintervall $P\left(\Theta_u \leqslant \theta \leqslant \Theta_o\right) = y$

$\phi\left(c\right) = \frac{1+y}{2}$

$e = c \cdot \frac{\sigma}{\sqrt{n}}$

$\Theta_u = \bar{X} - e$ und $\Theta_o = \bar{X} + e$

P($\bar{x}$ max um a abweicht)

$$X \sim N\left(\mu, \sigma\right) \rightarrow \bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) P\left(-a \leqslant \bar{x} - \mu \leqslant +a\right) = P\left(-\frac{a}{\frac{\sigma}{\sqrt{n}}} \leqslant \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \leqslant \frac{a}{\frac{\sigma}{\sqrt{n}}}\right) = \phi\left(?\right) - \phi\left(-?\right) = ?? - (1 - ??)$$

Weitere Verteilungen

t-verteilung: bei unbekannter varianz → s^2 der stichprobe

- E(T) = 0
- Var(t) = n / (n - 2)
- t(n, a) = -t(n, 1-a)

χ2-Verteilung: σ schätzen

- chi(n) n = anzahl frei veränderbarer parameter
- n zufallsvars → summe der quadrierten zsv

Die χ2-Verteilung findet Anwendung, wenn man die *empirische Varianz bestimmt* hat und die Schätzung des Vertrauensintervalls ermitteln möchte, das den (unbekannten) Wert der Varianz der Grundgesamtheit mit einer gewissen Wahrscheinlichkeit einschliesst.

Übersicht über verschiedene Vertrauensintervalle zum Niveau γ

	(1) Verteilung der Grundgesamtheit	(2) Param.	(3) Schätzfunktionen	(4) zugehörige standardisierte Zufallsvariable	(5) Verteilung und benötigte Quantile	(6) Zufallsvariablen für Intervallgrenzen
1	Normalverteilung (Varianz σ^2 bekannt)	μ	$\bar{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$	$U = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$	**Standardnormalverteilung** (Tabelle 2) $c = u_p$ mit $p = \frac{1+\gamma}{2}$	$\Theta_u = \bar{X} - c \cdot \frac{\sigma}{\sqrt{n}}$ $\Theta_o = \bar{X} + c \cdot \frac{\sigma}{\sqrt{n}}$
2	Normalverteilung (Varianz σ^2 unbekannt und $n \leq 30$; sonst Fall 1 mit s als Schätzwert für σ)	μ	$\bar{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$ $S^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (X_i - \bar{X})^2$	$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$	***t*-Verteilung** (Tabelle 4) mit $f = n - 1$ $c = t_{(p;f)}$ mit $p = \frac{1+\gamma}{2}$	$\Theta_u = \bar{X} - c \cdot \frac{S}{\sqrt{n}}$ $\Theta_o = \bar{X} + c \cdot \frac{S}{\sqrt{n}}$
3	Normalverteilung	σ^2	$\bar{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$ $S^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (X_i - \bar{X})^2$	$Z = (n-1)\frac{S^2}{\sigma^2}$	**Chi-Quadrat-Verteilung** (Tabelle 3) mit $f = n - 1$ $c_1 = z_{(p_1;f)}$ mit $p_1 = \frac{1-\gamma}{2}$ $c_2 = z_{(p_2;f)}$ mit $p_2 = \frac{1+\gamma}{2}$	$\Theta_u = \frac{(n-1) \cdot S^2}{c_2}$ $\Theta_o = \frac{(n-1) \cdot S^2}{c_1}$
4	Bernoulli-Verteilung Anteilsschätzung (mit $n\hat{p}(1-\hat{p}) > 9$)	p	$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ X_i 0/1-wertig mit $P(X_i = 1) = p$	$U = \frac{\bar{X} - p}{\sqrt{p(1-p)/n}}$	**Standardnormalverteilung** (näherungsweise), Tabelle 2 $c = u_q$ mit $q = \frac{1+\gamma}{2}$	$\Theta_u = \bar{X} - c \cdot \sqrt{\frac{\bar{X} \cdot (1-\bar{X})}{n}}$ $\Theta_o = \bar{X} + c \cdot \sqrt{\frac{\bar{X} \cdot (1-\bar{X})}{n}}$
5	beliebig mit $n > 30$	μ, σ^2	wie im Fall 1 (gegebenenfalls mit s als Schätzwert für σ) bzw. im Fall 3			