Notizen MANIT3 - Stochastik

Vorlesungsinhalt *14 Wochen*

- 25 Beschreibende Statistik *3.5 Wochen*
- 26 Elementare Wahrscheinlichkeitsrechnung *1 Wochen* + Kombinatorik *2 Wochen* **Teil 1 des Buches**
- 27 Zufallsvariablen *3 Wochen*
- 28 - 29 Verteilungen *1 Wochen* **ohne spezielle Verteilungen in Kapitel 28, ohne 29.3**

Kapitel 25 - Deskriptive Statistik

Klassifizierte Daten

- emprisiche Verteilungsfunktion bei klassierten daten: $F(x)$ = Fläche im Histogramm bis x
  - $0.5 = F(x_{0.5})$ wobei $x_{0.5}$ = Median
  - z.B. Daten gehen nur bis 2000: dann ist $F(5000) = 1$, da alle $x < 5000$
- Berechnung Median bei klassierten Daten: $y = ax + m$

Name	Notation	Berechnung	Berechnung bei klassierten Daten

- $a = \frac{\Delta y}{\Delta x} = \frac{\text{relativer Anteil der Klasse am Ganzen}}{\text{Klassenbreite}}$
- $m = y_1 - a * x_1$
- $0.5 \neq a \cdot x_{0.5} + m \Rightarrow x_{0.5} = (0.5 - m)/a$
- Boxplot: von x_{min} nach x_{max} mit Strichen; Box von $x_{0.25}$ mit Strich bei $x_{0.5}$ bis $x_{0.75}$

Lage-Kennwerte

Name	Notation	Berechnung	Berechnung bei klassierten Daten
Mittelwert	$\overline{x}$	$\frac{1}{n}\sum_{i=1}^{n} x_i$	$\frac{1}{n}\sum_{i=1}^{m} h_i \cdot x_i m < n$ $x_i : \text{Klassenmittel}, h_i : \text{Häufigkeit}, m : \text{Anzahl Klassen}, n : \text{Anza}$
Median	$x_{0.5}$	$x_{[np]}$ (sortieren nach Höhe; Messwert in der Mitte nehmen)	$0.5 = F(x_{0.5})$; nach $x_{0.5}$ auflösen (lineare Interpolation; $y = mx +$
Modus / Modalwert		zuerst klassieren, dann -->	wir suchen x_i mit $h_i = max(h)$; $x_i = argmax(h)$; nicht immer de

Streu-Kennwerte

Name	Notation	Berechnung	Berechnung bei klassierten Daten
Varianz, Stabw	s^2, s	$s^2 = \frac{1}{n-1} \cdot \sum (x_i - \bar{x})^2$	möglich, aber nicht so wichtig
IQR (inter-quartile range)	IQR	$x_{0.75} - x_{0.25}$	mit $F(x_{\dots})$

Lineare Interpolation in Matlab: `interp1(x, y, x_0)`.

Streudiagramm

- Zeigt ob und wie Daten zusammenhängen (e.g. linear curve fitting)
- verschiedene Ausprägungen: zusammenhängend, nicht zusammenhängend, gegensinning
- Lineare Korrelation $r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}$ wobei $s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_0 - \bar{x})(y_i - \bar{y})$ die Covarianz ist und $s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ ist die emprirische Standardabweichung; s_y ist analog
- Es gilt $-1 \leq r_{xy} \leq 1$, r_{xy} heisst linerarer Korrelationskoeffizient; Vorzeichen gibt Aufschluss über Trend (1. Ableitung), *siehe Satz 25.19*
- MATLAB: `corrcoef`

Lineare Regression (linear least squares)

- Modell: $y = a_1 x + a_0 + \varepsilon$; x ist genau bekannt, aber y ist mit einer unbekannten Abweichung ε gestreut

- Oft werden zur berechnung der Koeffizienten (Steigung a, Achsenabschnitt a_0) das Kriterium der least squares verwendet
- Die Summe der quadreirten Abweichungen in y-Richtung gilt es zu minimieren: $J(a_0, \dots, a_n) = \sum_{i=1}^{n} (y_i - (a_1 x_i + a_0))^2$
- Koeffizienten, die das Kriterium der least squares erfüllen werden mit $\widehat{a_0}, \widehat{a_1}, \dots$ bezeichnet: $\widehat{a_1} = r_{xy} = \frac{s_y}{s_x}, \widehat{a_0} = \overline{y} - \widehat{a_1} \cdot \overline{x}$
- $\hat{y}$ ist das y, das auf der Geraden liegt
- MATLAB: `p = polyfit(x, y, 1), polyval(p, x)`

Multivariate lineare Regression

- $y_i = a_1 x_i + a_0 + \varepsilon$ Regressionsgleichung für lineare Regression. Beobachtungen (x_i, y_i) ergeben n lineare Gleichungen -> $Y = Ak + \varepsilon$ mit $A = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} = (\vec{x} \quad \vec{1}), k = \begin{pmatrix} a_1 \\ a_0 \end{pmatrix}, Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \vec{y}$ und $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} = (\vec{\varepsilon})$
- k in MATLAB bestimmen:

```
x = [1 2 3]
y = x.^2
bar(x, y)
hold on

% calculating the linear regression automatically
p = polyfit(x, y, 1)
pp = polyval(p, x)
plot(x, pp)

% manual calculation
A = [x; ones(1, length(x))]'
k = A\y

% k and p' are identical
```

- **Aufgabe** Bestimme A, Y sd Regressionsgerade durch $(0, 0)$ geht: $y = a_i x + \varepsilon$; Achsenabschnitt ist $= 0$ (trivially), sprich $A = \vec{x}, Y = \vec{y}$
- **Aufgabe** Bestimme A, Y für ein quadratisches Modell $\vec{y} = a_0 + a_1 x + a_2 x^2 + \varepsilon$: $\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1^2 & x_1^1 & x_1^0 \\ x_2^2 & x_2^1 & x_2^0 \\ \vdots & \vdots & \vdots \\ x_n^2 & x_n^1 & x_n^0 \end{pmatrix} \cdot \begin{pmatrix} a_2 \\ a_1 \\ a_0 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$
- **Aufgabe** Exponentielles Modell für die Konzentration eines löslichen Stoffes in Abhängigkeit der Zeit ist $y(t) = ae^{-bt}$. Umformung in additive Gleichung (für Matrizen/MATLAB) mit Logarithmus (immer positiv, da Konzentration immer ≥ 0); Wertepaare (t_i, y_i);

$$\ln y = \ln(ae^{-bt}) = \ln a + \ln e^{-bt} = \ln a - bt \cdot \underbrace{\ln e}_{=1} = \underbrace{\ln a}_{=a_0} - \underbrace{b}_{=a_1} t = a_1 t + a_0 + \varepsilon$$

$$\Rightarrow \begin{pmatrix} \ln y_1 \\ \ln y_2 \\ \vdots \\ \ln y_n \end{pmatrix} = \begin{pmatrix} t_1 & 1 \\ t_2 & 1 \\ \vdots & \vdots \\ t_n & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_0 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Kapitel 26 - Wahrscheinlichkeitsrechnung

nicht mitgeschrieben bei:

- Zufallsmodelle
- Zufallsereignisse, Zufallsexperimente
- Ω
- Münzenwurf, Würfelwurf, $A = \{2, 4, 6\}, A \subseteq \Omega$

Axiomatische Wahrscheinlichkeit nach Kolomogorov

für stetige Räume

1. $\Omega, P(\Omega) = 1$
2. $A \subseteq \Omega : 0 \leq P(A) \leq 1$
3. Für $A \cap B = \{\} \Rightarrow P(A \cup B) = P(A) + P(B)$

Daraus folgt:

1. $P(\bar{A}) = 1 - P(A)$
2. $P(A_1 \cup A_2 \cup \dots A_n) = \sum_{k=1}^{n} P(A_k); A_k \text{ sind paarweise disjunkt}$
3. $A \subseteq B \Rightarrow P(A) \leq P(B)$

Schublade mit 6 roten, 8 blauen Socken. Zwei Socken werden gezogen. WS für...

1. zwei rote -> $\frac{\binom{6}{2}}{\binom{14}{2}} = 0.165$
2. zwei blaue -> $\frac{\binom{8}{2}}{\binom{14}{2}}$
3. zwei verschiedene -> $= 0.527$
4. zwei gleiche/passende -> $= 0.473$

Multiple-Choice, 4 Fragen, 3 Antworten, eine richtig. WS für

1. alle 4 Antworten richtig -> $(\frac{1}{3})^4$
2. genau eine Antwort richtig -> $4 \cdot \frac{1}{3} \cdot (\frac{2}{3})^3$

Sei $X \sim Exp(k)$

$P(X \leq x) = F(X) = \left\{ \begin{matrix} 1 - e^{-kx}, & x \geq 0 \\ 0, & x < 0 \end{matrix} \right\}$

$E(X) = k^{-1}, Var(X) = k^{-2}$

$P(|X - \frac{1}{k}| \geq c) = P(X \in]-\infty, \frac{1}{k} - c[\bigcup X \in]\frac{1}{k} + c, \infty[)$

$P(X \leq x) = F(x)$ wobei $\leq \Rightarrow X \in]-\infty, x[$

$= P(X \leq \frac{1}{k} - c) + P(X \geq \frac{1}{k} + c)$

$= F(\frac{1}{k} - c) + 1 - F(\frac{1}{k} + c)$ (Intervalle aufaddieren)

$\leq \frac{?k^2}{c^2} = \frac{1}{k^2 c^2}$

Exponentielle Verteilung

Für $t \geq 0$ **:** ist die Verteilungsfunktion $F(t) = 1 - e^{-c \cdot t}$ mit Parameter $c > 0$.

$1 - F(t)$ ist die Überlebenswahrscheinlichkeit des Zeitpunktes t

$F(t)$ ist die WSK, dass das System den Zeitpunkt t nicht erlebt.

$\mu = c^{-1}, \sigma^2 = c^{-2}$

1. Waschmaschine Lebenserwartung L 6 Jahre, $L \sim Exp(c)$. WSK länger betriebsfähig?
   $(L) = 6y = \mu_L \Rightarrow c = \frac{1}{6}; P(L \geq \mu_L) = 1 - F_L(6) = e^{-5/6}$.
2. Autotyp Lebensdauer $L, P(L \leq 8) = 1/2$. Erwartete Lebensdauer?
   $1/2 = 1 - e^{-c \cdot 8} \Leftrightarrow c = -\frac{\ln 1/2}{8} \Rightarrow E(L) = c^{-1}$

Gleichverteilung

1. Erwartete Wartedauer auf Bus ist 5 Minuten, Streuung 1 Minute.
   1. Intervall gleichverteilte Wartedauer?
      $a + b = 10, b - a = \sqrt{(12)} \Rightarrow a = \frac{10 - \sqrt{(12)}}{2} = 3.26, b = \frac{10 + \sqrt{(12)}}{2} = 6.73$
   2. Wie gross ist $P(4 \leq \text{Wartezeit} \leq 6)$? $\frac{6-4}{6.73-3.26}$
   3. $F(x) = \left\{ \begin{matrix} 1 & x > 6.73 \\ \frac{1}{12}, & x \in [3.26, 6.73] \\ 0 & x < 3.26 \end{matrix} \right\}$
2. Addieren zweier gleichverteilter Zufallsvariablen ergibt Dreiecksverteilung. Sei das Dreieck $x = 0..2$ somit ist $h = 1$ da die Fläche bei Gleichverteilung *immer* $= 1$. Sei $f(x)$ die Funktion, die das Dreieck beschreibt, $f(x) = \left\{ \begin{matrix} x & 0 \leq x < 1 \\ -x + 2, & 1 \leq x < 2 \end{matrix} \right\}$; $X \sim f(x)$. Dann $E(X) = 1$ (weil symmetrisch), $Var(X) = ?, \sigma_X = ?$ (Integration von $f(x)$ von 0 bis 1 und von 1 bis 2 gemäss Definition von $f(x)$).

> Allgemein: $Var(X) = \int_{-\infty}^{\infty} (x - \mu_X)^2 \cdot f(x) dx = E(X^2) - E(X)^2$

Portfolio

Titel	Anteil	Rendite / μ	Volatitlität / σ
A	0.3	6%	5%
B	0.2	2%	1%
C	0.5	4%	6%

Aktienkurse seien unabhängig

Bestimme Rendite und Volatilität des Portfolios $P = 0.3A + 0.2B + 0.5C$.

$E(P) = 4.2\%, \sqrt{Var(P)} = 3.36\%$

> Allgemein für X_k iid mit μ_X, σ_X mit $1 \leq k \leq n$:
>
> $\bar{X} = \frac{1}{n} \sum_{k=1}^{n} X_k$
>
> $E(\bar{X}) = \mu_X$
>
> $Var(\bar{X}) = Var(\sum_{i=1}^{n} \frac{1}{n} X_i) = \sum_{i=1}^{n} \frac{1}{n^2} Var(X_i) = \frac{\sigma_X}{n}$

Gesetz der grossen Zahlen

Wikipedia

Prüfung

Aufgabe 1

- Histogramm soll sich nicht verändern wenn die Klassenbreite verändert wird
- Schätzung Mittelwert = (sum (Häufigkeit * Klassenmitte))/count

Aufgabe 5

- Notation $P(X_i = j)$ für i Würfe Abbruch bei Wurf j
- z.B. für $n = 2$: $P(X_2 = 1) = 1/6, P(X_2 = 2) = 5/6$
- für $n = 3$: $P(X_3 = 1) = 1/6, P(X_3 = 2) = 5/6 \cdot 1/6, P(X_3 = 3) = (5/6)^2$
- allgemein: $P(X_i = 1 \ldots i-1) = (5/6)^{i-1} \cdot 1/6, P(X_i = i) = (5/6)^{i-1}$ -> Baum

Indikatorvariablen

- $1_A : X \to \{0, 1\}, 1_A(X) :== \begin{Bmatrix} 1 & x \in A \\ 0 & x \notin A \end{Bmatrix}$
- rechts-stetige Funktion
- *Erinnerung* Hauptsatz der Statistik: $\bar{F}(x) =$ Anzahl der Datenpunkte $x_k \leq x = \frac{1}{n} = \sum_{k=1}^{n} 1_{x_k \leq x}$
- Seien X_i iid ZV mit Verteilungsfunktion $F(x), x_i$ ist Zufallsstichprobe gem. der Verteilung von X_k mit empirischer $\bar{F}(x)$. Dann gilt für $\varepsilon > 0 : \lim_{n \to \infty} P(|\bar{F}(x) - F(x)| < \varepsilon) = 1$

Normalverteilung

$X \sim N(\mu, \sigma^2)$: $E(X) = \mu, Var(X) = \sigma^2$.

Standardnormalverteilung $\mu = 0, \sigma^2 = 1$ hat Dichte $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} = \phi(X)$, Verteilung $F(X) = \int_{-\infty}^{\infty} \phi(t)dt = \Phi(X)$

Standardisierung Sei $X \sim N(\mu, \sigma^2)$. Dann ist $Z = \frac{X-\mu}{\sigma}$ standardnormalverteile ZV, sprich $Z \sim N(0,1)$. Für die Verteilungsfunktion golt $F(x) = \Phi(\frac{x-\mu}{\sigma})$, für die p-Quantile: $Xp = \sigma Z_p + \mu$

Allgemein gilt für $X \sim N(\mu, \sigma^2)$, dass die Trafo $Y = aX + b \sim N(a\mu + b, a^2\sigma^2)$

Beispiel Gegeben Messwerte, normalverteilt, $\mu = 4, \sigma = 2$. WSK, dass Messwert:

1. ≤ 6 -> $P(X \leq 6) = F(6) = \Phi(\frac{6-4}{2}) = \Phi(1)$
2. ≥ 2 -> $1 - \Phi(2)$
3. $3.8 < x < 7$

Allgemein
$P(|X - \mu| \geq c) \leftrightarrow (-\infty, \mu - c] \cup [\mu + c, \infty) \Rightarrow 1 - (F(\mu + c) - F(\mu - c)) = 1 - (\Phi(\frac{c}{\sigma}) - \Phi(\frac{-c}{\sigma}))$ mit $F(x) = \Phi(\frac{x-\mu}{\sigma})$.

Berechnung von Quantilen Sei Abfüllgewicht von Paketen $\sim N(100g, (25g)^2)$. Berechne Bereich $\mu \pm c$ mit 90% der Pakete.

$Z = \frac{X-\mu}{\sigma} \sim N(0,1)$ mit $x_p = \sigma Z_p + \mu$. $\Phi(Z_p) = p$

$\Rightarrow 2\Phi(Z_p) = 0.9 \Rightarrow \Phi(Z_p) = \frac{1+0.9}{2} = 0.95$ Das gesuchte p-Quantil is also $z_{0.95}$ und e gilt $x_{0.95} = \sigma \cdot z_{0.95} + \mu$ ($z_{0.95}$ ist tabelliert, $= 1.644$). So folgt $5 \cdot 1.644 + 100 = 108.2$. Das Intervall ist $[100 - 8.2, 100 + 8.2]$.

**Allgemein* um Grenzen $[\mu + c, \mu - c]$ mit $P(|X - \mu| \leq c) = p, X \sim N(\mu, \sigma^2)$ zu finden, löse $2\Phi(\frac{c}{\sigma}) - 1 = p$ nach c $\Rightarrow c = \sigma Z_{\frac{1+p}{2}}$

Für $c = n \cdot \sigma, n \in \mathbb{N}$ sind die Werte tabelliert (68.3%, ...)!

Beispiel Preis ist $\mu = 850, \sigma = 150$

1. WSK Preis $\pm\sigma = 0.683$
2. WSK Preis $\pm 200 = F(\mu - 200) + 1 - F(\mu + 200) = \Phi(\frac{(\mu-200)-\mu}{\sigma}) + 1 - \Phi(\frac{(\mu+200)-\mu}{\sigma})$
3. Obere Schranke für die billigsten 10%:
   $\Phi(z_{0.1}) = 0.1, z_{0.1} = -1.28...; x_p = \sigma z_p + \mu \Rightarrow x_{0.1} = 150 \cdot z_{0.1} + 850 \approx 658$

```
zp = norminv(p, mu, sigma) % norminv(0.1,850,150)
```

Additionssatz

$X \sim N(\mu_x, \sigma_x), Y \sim N(\mu_y, \sigma_y) iid. \Rightarrow X + Y \sim N(\mu_x + \mu_y, \sigma_x + \sigma_y)$

Zentraler Grenzwertsatz

$X_1, \ldots, X_n$ iid mit μ_x, σ_x^2. $S_n = \sum X_i, \bar{X_n} = \frac{1}{n} S_n, Z_n = \frac{\bar{X_n} - \mu}{\sigma\sqrt{n}} = \frac{S_N - n\mu}{\sqrt{(n)}\sigma}$

$\Rightarrow \lim_{n\to\infty} P(Z_n \leq z) = \Phi(z)$

Übungen

Serie 1

- Häufigkeiten
- Histogramme
- Kennwerte
- Summenkurve
- Boxplot

Serie 2

- Varianz
- Kovarianz
- Standardisierung
- Standardabweichung

Serie 3

- Regressionsgerade
- erklärte Varianz

Serie 5

- Kombinatorik

Serie 6

- Kombinatorik
- Laplace WSK

Serie 7

- Klassische WSK
- Bedingte WSK

Serie 8

- Bedingte WSK
- Bayes

Serie 9

- ZV
- Verteilungsfunktion
- Exponentialverteilung

Serie 11

- Exponentialverteilung
- Erwartungswert/Varianz Linearität
- Kennwerte bestimmen