Grundlagen Statistik

April 13, 2016

1 Grundlagen der Wahrscheinlichkeitsrechnung und Statistik

Dieser Abschnitt befasst sich mit den Grundlagen der Wahrscheinlichkeitstheorie in Hinblick auf die Zeitreihenanalyse. Zufallsvariablen sind darin Platzhalter für Messwerte eines Experiments. Wahrscheinlichkeitsdichteverteilungen beschreibt die theoretische Verteilungsfunktion. Realisierungen eines Experimentes werden als Stichproben bezeichnet. Aus einer Anzahl von Realisierungen ergibt sich die empirische Häufigkeitsverteilung.

1.1 Zufallsvariable

Sei $x(k)$ eine Menge von Zufallsvariablen. Die Zählvariable k bezeichnet ein bestimmtes Ereignis. Die Zufallsvariable $x(k)$ beschreibt eine Messung mit einem zufälligen Ergebnis. Der Ausgang eines N-mal wiederholten Experiments ist eine Reihe von Punkten bzw. Messwerten. Die Messwerte werden Stichproben (Samples) bzw. Realisierungen genannt. Die Anzahl der Stichproben wird hier mit N bezeichnet.

1.2 Wahrscheinlichkeitsverteilungs- und dichtefunktion

Die Wahrscheinlichkeitsverteilungsfunktion $P(x)$ wird definiert als die Wahrscheinlichkeit dafür, dass ein Ereignis $x(k)$ den Wert $x(k) \leq x$ annimmt.

$$P(x) = \text{Prob}(x(k) \leq x)$$

Es gilt

$$P(a) \leq P(b)$$

wenn $a \leq b$.
Der Wertebereich von P ist $[0, 1]$ weil gilt

$$P(-\infty) = 0, P(+\infty) = 1$$

Wenn der Wertebereich der Zufallsvariable $x(k)$ kontinuierlich ist, dann wird die Wahrscheinlichkeitsdichtefunktion $p(x)$ wie folgt definiert

$$p(x) = \lim_{\Delta x \to 0} \left(\frac{\text{Prob}(x < x(k) \leq x + \Delta x)}{\Delta x} \right)$$

Es gilt

$$p(x) \geq 0$$

$$\int_{-\infty}^{\infty} p(x)dx = 1$$

$$P(x) = \int_{-\infty}^{x} p(\zeta)d\zeta$$

$$\frac{dP(x)}{dx} = p(x)$$

Die Wahrscheinlichkeitsdichtefunktionen diskreter Zufallsvariablen, wie z.B. der Würfelfunktion, müssen mit Hilfe von Delta-Funktionen $\delta(x)$ oder als Summe beschrieben werden.

$$\text{Prob}(x < x(k) \leq x + \Delta x) = \sum_{x}^{x+\Delta x} p_k(x)$$

mit $p_k = \text{Prob}(x(k) = x_k)$.

Die Wahrscheinlichkeitsverteilungsfunktion $P(x)$ wird auch kumulative Verteilungsfunktion genannt. Sie ist definiert als das Integral über die Wahrscheinlichkeitsdichtefunktionen $p(x)$. Oft genutzte Abkürzungen sind PDF und CDF für Probability Density Function $p(x)$ und Cumulative (Probability) Density Function $P(x)$.

1.3 Erwartungswert

Der Erwartungswert einer Zufallsvariablen entspricht dem Mittelwert μ_x bei unendlicher Wiederholung eines Experiments. Der Erwartungswert errechnet sich aus der Zufallsvariablen mittels Gewichtung mit der Wahrscheinlichkeit

$$\text{E}(x(k)) = \int_{-\infty}^{\infty} xp(x)dx = \mu_x$$

Für diskrete Zufallsvariablen ist das Integral durch eine Summe zu ersetzen.

$$\text{E}(x) = \sum_{k} xp_k(x)$$

Die Varianz ist definiert als

$$\text{E}(x(k) - \text{E}(x(k))^2) = \sigma_x^2$$

1.3.1 Regeln

Der Erwartungswert ist ein linearer Operator

$$\text{E}(aX_1 + bX_2) = a\,\text{E}(X_1) + b\,\text{E}(X_2)$$

1.3.2 Beispiel Zufallsvariable Würfel

Sei die Zufallsvariable $x(k)$ der Ausgang eines Würfelexperimentes.

Beim idealen Würfel sind alle sechs Seiten gleichwahrscheinlich.

$$P(x = 1) = P(x = 2) = P(x = 3) = P(x = 4) = P(x = 5) = P(x = 6) = \frac{1}{6}$$

Damit ergibt sich für den Erwartungswert

$$\text{E}(x) = \sum_{k=1}^{6} x_k p_k(x_k)$$

mit $p_k(x) = \frac{1}{6}$ für alle $x = 1, 2, .., 6$

$$\text{E}(x) = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = 3.5$$

1.3.3 Übung

Berechnen Sie Mittelwert und Varianz der Zufallsvariable $x(k)$ "Würfel" theoretisch und experimentell mit einem Zufallsgenerator. Skizzieren Sie die kumulative Wahrscheinlichkeitsverteilungsfunktion und $P(x_k)$ Wahrscheinlichkeitsdichtefunktionen $p(x_k)$.

2 Kovarianz und Korrelation

Der Korrelationskoeffizient r beschreibt, wie eng zwei Zufallsvariablen in Raum oder Zeit zusammenhängen. Für zwei Zufallsvariablen $x = (x_1, x_2, ..., x_n)$ und $y = (y_1, y_2, ..., y_n)$ berechnet sich der Korrelationskoeffizient als

$$r = \frac{1}{N-1} \sum_{i=1}^{N} \frac{(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

oder ausgedrückt durch die Kovarianz C_{xy}

$$r = \frac{C_{xy}}{s_x s_y}$$

dabei gilt

$$C_{xy} = \mathrm{cov}(x, y) = \sigma^2_{x,y} \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$$

Der Korrelationskoeffizient errechnet sich aus der Kovarianz durch Normierung mit den Standardabweichungen s_x und s_y definiert durch

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

Äquivalent ist die Definition der Kovarianz durch den Erwartungswert

$$\mathrm{cov}(x, y) = \mathrm{E}\big[(x - \mathrm{E}(x)) \cdot (y - \mathrm{E}(y))\big]$$

Für statistisch unabhängige Zufallsvariablen x und y gilt $\mathrm{cov}(x, y) = 0$.

3 Normalverteilung (z-Verteilung)

Die Gaußverteilung ist gegeben durch

$$p(x) = \frac{1}{\sigma_x \sqrt{2\pi}} e^{-\frac{(x-\mu_x)^2}{2\sigma^2_x}}$$

Es gilt

$$E[x] = \mu_x$$

und

$$E[(x - \mu_x)^2] = \sigma^2_x$$

Substitution von $z = \frac{x-\mu_x}{\sigma_x}$ liefert die standardisierte Normalverteilung

$$p(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

für die gilt $E[x] = \mu_z = 0$ und $E[(x - \mu_z)^2] = \sigma^2_z = 1$

Die Wahrscheinlichkeit P berechnet sich aus dem Integral

$$P(z_\alpha) = \int_{-\infty}^{z_\alpha} p(z)dz = Prob[z < z_\alpha] = 1 - \alpha$$

bzw.

$$1 - P(z_\alpha) = \int_{z_\alpha}^{\infty} p(z)dz = Prob[z > z_\alpha] = \alpha$$

3.1 Gaußsche Fehlerfunktion

Als Fehlerfunktion oder gaußsche Fehlerfunktion bezeichnet man in der Theorie der Speziellen Funktionen das Integral

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-\tau^2} \, \mathrm{d}\tau$$

Aus der Fehlerfunktion errechnet sich, wieviele Werte bei einer gegebenen Normalverteilung innerhalb eines Wertebereichs, z.B. $\pm 1\sigma$ zu erwarten sind.

3.1.1 Beispiel erf

```
In [9]: erf(1/sqrt(2)) % Octave/Matlab

Out[9]: 0.68268949213708585

In [5]: from scipy.special import erf
        from numpy import sqrt
        erf(1/sqrt(2))

Out[5]: 0.68268949213708585

In [6]: erf(2/sqrt(2))

Out[6]: 0.95449973610364158

In [7]: erf(3/sqrt(2))

Out[7]: 0.99730020393673979
```

Im Intervall $\pm\sigma$ sind 68,27% aller Zufallswerte zu erwarten, in $\pm 2\sigma$ sind es 95,45%, in $\pm 3\sigma$ erwarten wir 99,73%.

3.2 Parameter-Schätzungen

Das Grundproblem der statistischen Analyse besteht darin, dass die Stichprobenanzahl begrenzt ist und die theoretischen Verteilungsfunktionen nicht bekannt sind. Stattdessen müssen statistische Parameter aus einer begrenzten Anzahl Stichproben möglichst genau geschätzt werden.

Gegeben sei eine Zufallsvariable x_i, z.B. eine stochastische Zeitserie. Im Folgenden wird der Stichproben-Index i weggelassen, um eine kompaktere Darstellung zu erhalten. Die beiden statistischen Parameter Mittelwert μ und Varianz σ^2

$$\mu_x = E[x] = \int_{-\infty}^{\infty} xp(x)dx$$

$$\sigma_x^2 = E[(x - \mu_x)^2] = \int_{-\infty}^{\infty} (x - \mu_x)^2 p(x)dx$$

werden mit Hilfe der Wahrscheinlichkeitsdichtefunktion $p(x)$ theoretisch berechnet. Der Operator $E[]$ bezeichnet den Erwartungswert. Im Allgemeinen ist die Anzahl der Stichproben begrenzt und die Wahrscheinlichkeitsdichtefunktion ist nicht bekannt.

Eine (von vielen) Möglichkeiten um Mittelwert und Varianz von x aus N unabhängigen Messungen zu schätzen, ist gegeben durch

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

$$s_b^2 = \hat{\sigma_x}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

Der Unterschied von $\hat{\mu_x}$ und μ_x besteht darin, dass $\hat{\mu_x}$ den geschätzen Wert und μ_x den wahren bzw. theoretisch richtigen Wert bezeichnet.

Wir wollen nun untersuchen, ob die Schätzung von Mittelwert und Varianz unsere Erwartungen erfüllt. Der Erwartungswert des Mittelwertes berechnet sich aus (Bendat und Piersol, 1986):

$$E[\bar{x}] = E\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = \frac{1}{N}E\left[\sum_{i=1}^{N} x_i\right] = \frac{1}{N}(N\mu_x) = \mu_x$$

und erfüllt unsere Erwartungen $\hat{\mu_x} = \bar{x}$ (erwartungstreu).

Der mittlere quadratische Fehler der Schätzung des Mittelwertes ergibt sich aus

$$E[(\bar{x} - \mu_x)^2] = E\left[\left(\frac{1}{N}\sum_{i=1}^{N} x_i - \mu_x\right)^2\right] = \frac{1}{N^2}E\left[\left(\sum_{i=1}^{N}(x_i - \mu_i)\right)^2\right]$$

Die Messungen x_i sind unabhängig, daher folgt

$$E[(\bar{x} - \mu_x)^2] = \frac{1}{N^2}E\left[\sum_{i=1}^{N}(x_i - \mu_i)^2\right] = \frac{1}{N^2}(N\sigma_x^2) = \frac{\sigma_x^2}{N}$$

Der Erwartungswert der Varianz s_b errechnet sich zu

$$E[s_b^2] = E[\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2] = \frac{1}{N}E[\sum_{i=1}^{N}(x_i - \bar{x})^2] =$$

mit $\sum_{i=1}^{N}(x_i - \bar{x})^2 = ... = \sum_{i=1}^{N}(x_i - \bar{x})^2 - N(x_i - \bar{x})^2$ und $E[(x_i - \mu_x)^2] = \sigma_x^2$ und $E[(\bar{x} - \mu_x)^2] = \frac{\sigma_x^2}{N}$ folgt (Herleitung siehe Bendat und Piersol, 1986)

$$E[s_b^2] = \frac{N-1}{N}\sigma_x^2$$

Die Varianz-Schätzung s_b bzw. $\hat{\sigma_x}$ ist offensichtlich nicht erwartungstreu, da $\hat{\sigma_x}^2 < \sigma_x^2$. Die Schätzung nennt sich darum verzerrt oder biased. Eine erwartungstreue/unverzerrte Schätzung der Varianz ist gegeben durch

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

der sogennanten empirischen Varianz oder Stichprobenvarianz. Der Unterschied macht sich insbesondere für eine kleine Anzahl von Stichproben bemerkbar und kann für eine große Anzahl meist vernachlässigt werden. Vorsicht ist jedoch geboten bei Hypothesentests, die auf einem Vergleich der Varianz basieren.

4 Hypothesenprüfungen

Bei einem statistischen Prüfverfahren wird einer sogenannten Nullhypothese eine (oder mehrere) Alternativhypothese gegenübergestellt. Im Sinne eines mathemathischen Widerspruchsbeweises wird die Nullhypothese statistisch widerlegt, um die Alternativhypothese zu beweisen.

Bei der Auswahl eines Prüfverfahrens muss die theoretische Wahrscheinlichkeitsverteilung und der Stichprobenumfang beachtet werden.

Das Ergebnis einer statistischen Prüfung ist immer im Zusammenhang mit dem sog. Signifikanzniveau zu nennen. Die zur Signifikanz komplementäre Größe ist die Irrtumswahrscheinlichkeit. Es gilt

$$Signifikanz = 1 - Irrtumswahrscheinlichkeit$$

oder

$$Si = 1 - \alpha$$

Wahrscheinlichkeiten werden entweder in Prozent oder als Zahl zwischen 0-1 angeben. Anstatt α wird üblicherweise auch die Bezeichnung p verwendet. Die Regel besagt, dass die Signifikanz groß ist, wenn p klein ist.

Aus der Angabe einer Wahrscheinlichkeit ergibt sich auch ein sog. Vertrauensbereich (auch Konfidenzintervall oder Mutungsbereich), der den Wertebereich für eine spezifische Wahrscheinlichkeit angibt.

4.1 Beschreibung von Wahrscheinlichkeitsbereichen (IPCC-Terminologie)

Bei einer Einschätzung der Unsicherheit bestimmter Ergebnisse mittels fachkundiger Beurteilung und statistischer Analyse eines Beweises (z.B. Beobachtungen oder Modellergebnisse) werden folgende Wahrscheinlichkeitsbereiche verwendet, um die geschätzte Eintrittswahrscheinlichkeit auszudrücken:

- praktisch sicher >99%
- höchst wahrscheinlich >95%
- sehr wahrscheinlich >90%
- wahrscheinlich >66%
- wahrscheinlicher als nicht >50%
- etwa so wahrscheinlich wie nicht 33% bis 66%
- unwahrscheinlich <33%
- sehr unwahrscheinlich <10%
- höchst unwahrscheinlich <5%
- außergewöhnlich unwahrscheinlich <1%

Übernommen aus IPCC AR4 (2007)