Statistik und Wahrscheinlichkeitstheorie (Universität Wien)

Kochrezepte

Ausgearbeitet von Murmel (Murmel.vienna@gmx.at)

Kombinatorik

Binomialverteilung B(n,p)
n: Gesamtanzahl Versuche
p: Wahrscheinlichkeit, dass das gewünschte Ereignis eintritt
q: Gegenwahrscheinlichkeit (1-p)
k: Anzahl der Versuche, in denen man das gewünschte Ereignis will

$$B(n,p) = \binom{n}{k} \cdot p^k \cdot q^{n-k}$$

Mindestens k-mal ein Ereignis kann man auf NICHT höchstens k-mal ein Ereignis umformen und umgekehrt.

Anzahl Möglichkeiten:
Ohne Zurücklegen:

$$\binom{n}{k}$$

Mit Zurücklegen:

$$\binom{n+k-1}{k}$$

Erwartungswert:

$E(X) = \sum_{i=1}^{n} p_i \cdot x_i$ mit n Anzahl verschiedener Möglichkeiten

$E(a*X) = a*E(X)$

Hazardwerte

Hazard-Wert:

$$h_i = \frac{p_1}{1 - \sum_{k=1}^{i-1} p_k}$$

Monat	P(Verkauf)	Hazard	Bestand
0	p0	h0	b0
1	p1	h1	b1
2	p2	h2	b2

Aktueller Bestand:

$$b_i = b_{i-1} - b_{i-1} h_i$$

Erwartungswert $E(X) = \sum_{i=1}^{N} p_i \cdot Monat$

Netzwerk

Gesamtzuverlässigkeit:
Es geht hier darum, mittels geeigneter Formeln jeweils so lange jeweils zwei hintereinander/parallel geschaltete Komponenten zusammenzufassen, bis nur noch eine einzige übrig bleibt.
Wir verwenden dazu die folgenden Formeln (mit z = neue Gesamtzuverlässigkeit, z1 = Zuverlässigkeit der ersten Komponente, z2 = Zuverlässigkeit der zweiten Komponente):
Für eine Serienschaltung: z = z1 * z2
Für eine Parallelschaltung: z = 1- (1 - z1) * (1 - z2)

Bezeichnet p die Ausfallwahrscheinlichkeit, so wissen wir, dass z = 1 – p

Gesamtlebensdauer:
Dieser Teil löst sich analog zur Zuverlässigkeit, jedoch mit anderen Formeln.
Es geht hier darum, mittels geeigneter Formeln jeweils so lange jeweils zwei hintereinander/parallel geschaltete Komponenten zusammenzufassen, bis nur noch eine einzige übrig bleibt.
Wir verwenden dazu die folgenden Formeln (mit G(X) = neue Gesamtlebensdauer, G1(X) = Zuverlässigkeit der ersten Komponente, G2(X) = Zuverlässigkeit der zweiten Komponente):
Für eine Serienschaltung: G(X) = 1- (1 - G1(X)) * (1 - G2(X))
Für eine Parallelschaltung: G(X) = G1(X) * G2(X)

Berechnung der erwarteten Lebensdauer des Gesamtsystems:
Da X nur positive Werte annehmen kann, lässt sich das ganze über den Erwartungswert der Verteilungsfunktion V(X) des Systems errechnen. Hierbei gilt:

$$E(X) = \int_0^\infty (1 - V(X))dx$$

Markovketten

Die Zeilensummen innerhalb von Markovketten und ihren Anfangsverteilungen sind immer gleich 1.

Absorbierende Zustände sind solche, welche einen Wert auf der Hauptdiagonale beinhalten, welcher gleich 1 ist (man kommt von dem Zustand nicht mehr weg).

Markovketten sind irreduzibel, wenn man von jedem Knoten der Kette zu jedem anderen Knoten der Kette in einer endlichen Anzahl Schritte gelangen kann.

Ein Zustand ist transient, wenn man, wenn man ihn verlässt, ihn nie wieder erreicht.

Ein Zustand ist rekurrent, wenn man, wenn man ihn verlässt, ihn wieder erreichen kann (aber nicht muss!).

Nach der Definition einer stabilen Anfangsverteilung π von P muss für diese nach Punkt (iii) des Existenzsatzes gelten: $\pi * P = \pi$
Mittels Matrizenmultiplikation lässt sich so ein lösbares Gleichungssystem aufstellen.
Hinzuzufügen ist auch, dass die Zeilensumme von π gleich 1 sein muss.

Anova

Quadratsummen zwischen den Gruppen:

$$\bar{y}_{..} = \frac{1}{r} \cdot \sum_{i=1}^{r} y_i$$

$$\text{SSM} = \sum_{i=1}^{r} n_i \cdot (\bar{y}_{i.} - \bar{y}_{..})^2$$

Quadratsummen innerhalb der Gruppen (=Fehler):

$$\text{SSE} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (n_i - 1) \cdot (y_{ij} - y_{i.})^2$$

Summe:
SST = SSM + SSE

Freiheitsgrad:
f(mod) = Anzahl Gruppen – 1
f(feh) = Summe einzelner Gruppenanzahlen (jeweils -1)

Mittlere Quadratsumme:
m(mod) = SSM / f(mod)
m(feh) = SSE / f(feh)

F-Statistik:
F = m(mod) / m (feh)
(Die F-Statistik ist eher klein, wenn die Nullhypothese zutrifft)

Tabelle:

	Quadratsummen	Freiheitsgrade	Mittlere Quadratsummen	F	Signifikanz
Model	SSM	f(mod)	m(mod)	F(mod)	α
Fehler	SSE	f(feh)	m(feh)		
Total	SST	f(fmod)+f(feh)			

(hierbei kann es auch mehrere Modellzeilen geben, die Formeln bleiben aber dieselben)

F-werte eines Tests liest man aus der F-Tabelle über $F_{f;\ SSE;\ 1-\alpha}$ wobei f = f(mod) (Freiheitsgrad), SSE = Fehler und 1-α = Signifikanzniveau.
Hierbei sieht man in der 1-α F-Tabelle, Zeile SSE, Spalte f nach.
Ist F > F(mod), so wird die Nullhypothese beibehalten.
Ist F < F(mod), so wird die Alternativhypothese angenommen.

T-Test

Hier muss immer angenommen werden, dass die Zufallsvariablen normalverteilt sind. Der Test macht außerdem erst Sinn, wenn die Varianz unbekannt ist.

Einseitiger Test Hypothesen:
H_0: $\mu_D \leq 0$
H_A: $\mu_D > 0$
(bzw andersrum, je nach Interesse)

Zweiseitiger Test Hypothesen:
H_0: $\mu_D = 0$
H_A: $\mu_D \neq 0$

Die Werte vom einseitigen Test werden verwendet, wenn einen eine Zu- oder Abnahme interessiert, die vom zweiseitigen Test, wenn man generell an einem Unterschied interessiert ist. Beim zweiseitigen Test muss immer mit Signifikanz α/2 gerechnet werden.

T kann man berechnen:

$$s = \sqrt{\frac{1}{n-1} * \sum_{i=1}^{n} (dif_i - \overline{dif})^2}$$

$$T = \sqrt{n}\,\frac{\overline{X} - \mu}{s} = \sqrt{n}\,\frac{\overline{X} - \overline{Y}}{s}$$ (für eine Differenz)

Ist |T | < T(kritisch, je nach Test) dann wird die Nullhypothese beibehalten.
Ist |T | > T(kritisch, je nach Test) dann wird die Alternativhypothese angenommen.

P kann man nur sehr schwer berechnen, heutzutage verwendet man Statistikprogramme, früher Tabellen.
Ist P(je nach Test) > α wird die Nullhypothese beibehalten.
Ist P(je nach Test) < α dann wird die Alternativhypothese angenommen.

Je nach angenommenen Alphawert, gilt für die Signifikanz des P-Wertes dann:

Alpha	Signifikanz
0,100	schwachsignifikant
0,050	signifikant
0,010	hochsignifikant
0,001	höchstsignifikant

Chi² Test

Der Chi² Test basiert auf dem Chi² Vergleichsprinzip, d.h. man berechnet die Verteilung der Summe der Quadrate von k unabhängigen standardisierten normal verteilten zufälligen Größen.

Nullhypothese: Klassen gleich (etwas macht keinen Unterschied)
Alternativhypothese: Klassen nicht gleich (etwas macht einen Unterschied)

Bei Nicht-4feldertafeln:
H_0: Alle Methoden sind gleichwertig.
$p_{ij} = p_{1.} * p_{.j}$ für alle Paare i,j

H_1: Mindestens eine Methode liefert ein besseres Ergebnis.
$p_{ij} \neq p_{1.} * p_{.j}$ für mindestens ein Paar i,j

Vierfeldertafel allgemein:

n11	n12	n1.
n21	n22	n2.
n.1	n.2	n..

Voraussetzung für alle Berechnungen: Ereignisse unabhängig voneinander.
Optimal zur Darstellung einer solchen Vierfeldertafel: Säulendiagramme mit prozentueller Angabe bezüglich der Spalten oder Mosaic-Plots.

Anzahl Freiheitsgrade: f= (k-1)*(l-1) = (für 4feldertafel) (2-1)*(2-1) = 1

χ^2 Wert: $T^2 = X^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = n_{..} \frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}}$

Ist der Absolutbetrag dieses Wertes größer als der kritische Wert, so wird die Nullhypothese verworfen.
Den kritischen Wert $\chi^2_{f;\,1-\alpha}$ findet man, indem man in der χ^2 Tabelle in der Zeile f (Freiheitsgrade) und Spalte 1-α (Signifikanz) nachschaut.

Odds-Ratio: $\Psi = \frac{n_{11}n_{22}}{n_{12}n_{21}}$ Wenn $\Psi = 1$ kann die Nullhypothese angenommen werden, je weiter man von 1 abkommt, umso eher wird sie verworfen.
Ist $\Psi > 1$ so sind die Ereignisse in der Hauptdiagonale wahrscheinlicher.
Ist $\Psi < 1$ so sind die Ereignisse in der Hauptdiagonale unwahrscheinlicher.

Vorteil Odds-Ratio gegenüber Differenz der Anteile:
Die Odds-Ratio ist allgemein aussagekräftiger, ist sie doch quasi ein Faktor, wie stark das Verhältnis der einen Gruppe im Vergleich zur anderen Gruppe ist. Die Differenz der Anteile hingegen gibt nur eine Verbesserung an, die je nach Größe der Stichprobe viel oder wenig bedeuten kann. Beispiel: Eine Odds-Ratio von 5 bedeutet, dass durch die Veränderung die überprüfte Eigenschaft 5mal so stark ist, hat man gleichzeitig eine Differenz der Anteile von 0,4 sagt dies ohne weitere Informationen jedoch nichts aus.

Errechnung Konfidenzintervall Differenz der Anteile:

$p_1 = n_{11} / n_{1.}$

$p_2 = n_{21} / n_{2.}$

$\Delta = p_1 - p_2$

$z_{1-\alpha/2}$ wird durch Nachsehen in der t-Tabelle in der Zeile n (Gesamtsumme) und der Spalte 1-(α/2) (Signifikanz) herausgelesen.

Konfidenzintervall I:

$[p_1 - z_{1-\alpha/2} \cdot s_\Delta ; p_1 + z_{1-\alpha/2} \cdot s_\Delta]$

$s_\Delta = \sqrt{\frac{p_1 \cdot (1-p_1)}{n_{1.}}}$

Konfidenzintervall II:

$[p_2 - z_{1-\alpha/2} \cdot s_\Delta ; p_2 + z_{1-\alpha/2} \cdot s_\Delta]$

$s_\Delta = \sqrt{\frac{p_2 \cdot (1-p_2)}{n_{1.}}}$

Konfidenzintervall Differenz der Anteile:

$[\Delta - z_{1-\alpha/2} \cdot s_\Delta ; \Delta + z_{1-\alpha/2} \cdot s_\Delta]$

$s_\Delta = \sqrt{\frac{p_1 \cdot (1-p_1)}{n_{1.}} + \frac{p_2 \cdot (1-p_2)}{n_{2.}}}$

<u>Homogenitätshypothese:</u>

Wir sind an der Untersuchung einer Responsevariable in Abhängigkeit einer dichotomen erklärenden Variablen interessiert. Die Antwortvariable Y wird dabei als binäre Variable aufgefasst: Erfolg (Y=1) oder Misserfolg (Y=0). Die Werte der erklärenden Variablen X haben häufig die Interpretation Behandlung und Kontrolle.

<u>Unabhängigkeitshypothese:</u>

Gegeben sind zwei dichotome Variable und Y, die für insgesamt n Fälle beobachtet wurden. Zu untersuchen ist die Fragestellung, ob die beiden Merkmale unabhängig auftreten.

Die Homogenitätshypothese ist gleichbedeutend mit der Unabhängigkeitshypothese bedingt auf gegebene Zeilenwahrscheinlichkeiten.