Ausarbeitung Prüfung Statistik und Wahrscheinlichkeitstheorie (Universität Wien)

Prüfung 02.06.2003

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

Beispiel 1: Theorie

Welche grafischen Darstellungsformen für Häufigkeiten kennen Sie?
Was sollte man bei der Darstellung diskreter bzw. stetiger Merkmale beachten?

Mögliche eindimensionale Darstellungsformen sind:
- Balkendiagramm (Stabdiagramm): Balken berühren sich nicht;
  für absolute/relative Häufigkeiten gleich
- Histogramm: Balken berühren einander;
  bei ungleich breiten Klassen, ist die Fläche(NICHT die Höhe) Maß für die Häufigkeit;
  Berechnung der Höhe: Häufigkeit durch Klassenbreite
- (Empirische) Verteilungsfunktion

Mögliche mehrdimensionale Darstellungsformen sind:
- Mosaic-Plots:

Mosaic-Plots: Für die Darstellung mehrdimensionaler statistischer Werte
Relative Häufigkeiten werden durch proportionale Flächen dargestellt

Kerndichteschätzer: Verallgemeinerung des gleitenden Histogramms mit normaler stetiger Kernfunktion
Kernfunktion: K(x) >= 0 K(0) = max(K(x))
$K_d(x) = |1-x|$ für $|x| <= 1$ 0 sonst

Beispiel 2: Münzwerfen

Wie hoch ist die Wahrscheinlichkeit, dass beim 10maligen Werfen einer fairen Münze höchstens zweimal Kopf oder höchstens zweimal Adler kommt?

Das bedeutet, dass Kopf 0, 1 oder 2mal ODER Adler 0,1 oder 2mal kommt (die beiden Ereignisse können gleichzeitig sowieso nicht auftreten, deshalb kann man sie ohne weiteres addieren)

P(höchstens zweimal Adler)

$$= \binom{10}{0} * \left(\frac{1}{2}\right)^0 * \left(\frac{1}{2}\right)^{10} + \binom{10}{1} * \left(\frac{1}{2}\right)^1 * \left(\frac{1}{2}\right)^9 + \binom{10}{2} * \left(\frac{1}{2}\right)^2 * \left(\frac{1}{2}\right)^8$$

$$= \left[\binom{10}{0} + \binom{10}{1} + \binom{10}{2}\right] * \left(\frac{1}{2}\right)^{10} = 0{,}0545$$

Dasselbe gilt für P(höchstens zweimal Kopf). Aus diesem Grund gilt:

$$P = \left[\binom{10}{0} + \binom{10}{1} + \binom{10}{2}\right] * \left(\frac{1}{2}\right)^{10} * 2 = 0{,}109$$

Beispiel 3: T-Test

Welche Annahmen wurden gemacht? Worauf muss geachtet werden?
Als Generalvoraussetzung muss angenommen werden, dass es sich um eine Zufallsstichprobe handelt. Es muss auf Beobachtungsgleichheit und Strukturgleichheit geachtet werden. Dies ergibt sich daraus, dass es sehr wichtig ist, dass die Zufallsvariablen als normalverteilt angenommen werden.

Strukturgleichheit:
Die Gruppen müssen bezüglich aller wesentlichen Merkmale (mit Ausnahme des zu untersuchenden Einflussfaktors) identisch sein.
Dies kann am ehesten erreicht werden indem die Stichprobe aufgrund eines Zufallsverfahrens ausgewählt wird.
Da die Randomisierung auch nicht Strukturgleichheit garantiert, ist es gut zusätzlich vor der Randomisierung Schichten (oder Strata) zu bilden, welche aus Beobachtungseinheiten bestehen, die sich bezüglicher wichtiger Merkmale gleichen oder zumindest ähneln.(z.B. nach Altersgruppe und Geschlecht der Hühner)

Beobachtungsgleichheit:
Die Gruppen müssen in derselben Weise untersucht bzw. beobachtet werden, d.h. die Beobachtungseinheiten in beiden Gruppen müssen von denselben Personen, ungefähr im selben Zeitraum und mit denselben Methoden beobachtet werden.
Dadurch soll verhindert werden, dass der Beobachter oder das Huhn (bewusst / unbewusst) die Therapieformen unterschiedlich beurteilen. Um dies zu optimieren, kann Blindung eingesetzt werden. In einer doppelblinden Studie sind weder die Beobachter noch die Hühner über die Studie informiert. Bei einer einfachblinden Studie weiß nur der Beobachter Bescheid. Wissen alle über die Therapieform Bescheid, so nennt man dies eine offene Studie.

Wie lauten sinnvolle Nullhypothesen bzw. Alternativen, wenn Sie die unten angeführte statistische Auswertung verwenden wollen?
Als sinnvolle Nullhypothese wäre anzunehmen, dass das Mastmittel keinen Unterschied macht.
Eine sinnvolle Alternativhypothese wäre, dass die Hühner durch das neue Mastmittel an mehr Gewicht zunehmen als durch herkömmliche Nahrung.
H_0: $\mu_D \leq 0$
H_A: $\mu_D > 0$

Was ergibt sich als Aussage dieser statistischen Auswertung? Warum?
Laut Angabe gilt $T = 0{,}6741 < 2{,}01$ (Achtung: Angabefehler, T ist beim Nachrechnen positiv!)
Da der T-Wert kleiner als der kritische t-Wert beim einseitigen t-Test ist, wird die Nullhypothese beibehalten. Das Mastmittel steigert also nicht nachweislich das Gewicht.

Beispiel 4: Hazardwerttabelle

Vervollständigen Sie die Tabelle.
Wir rechnen mit den folgenden Formeln:

Hazard-Werte:

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

Aktueller Bestand:

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

Monat	P(Verkauf)	Hazard	Bestand
0	0,4	0.40	450
1	0,2	0.33	300
2	0,1	0.25	225
3	0,1	0.33	150
4	0,1	0.50	75
5	0,1	1.00	0

a) Wie viele Harddisks kaufte der Händler ein?
Aus denselben Formeln ergibt sich: Zu Beginn des Monats 0 hat der Händler 750 Stück.

b) Wann kann er erwarten, dass eine beliebige Harddisk verkauft wird?

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

Wir rechnen unter Annahme, dass eine Harddisk genau in der Mitte des Monats verkauft wird, da wir es nicht genauer wissen, mit den Monatswerten + 0,5
E(X) = 0.4*0.5 + 0.2*1.5 + 0.1*2.5 + 0.1*3.5 + 0.1*4.5 + 0.1*5.5 = 2.1
Somit kann er im Monat 2 (genauer am 3. Tag des Monats 2, 0,1*30) einen Verkauf erwarten.

c) Angenommen eine Harddisk ist zu Anfang des Monats 4 noch vorhanden.
Wann kann erwartet werden, dass diese verkauft wird?
p(4) = p(5) = 0.1
p(4)+p(5) = 0.2
P(X=4|X ≥ 4) = 0.1/0.2 = 0.5
P(X=5|X ≥ 4) = 0.1/0.2 = 0.5

E(X|X ≥ 4) = 4.5*0.5 + 5.5*0.5 = 5
Am Anfang des Monats 5 kann der Verkauf erwartet werden.

d) Der Händler bestellt Harddisks nach, sobald sein Bestand auf die Hälfte gesunken ist.
In welchem Monat muss er nachbestellen?
Anfang Monat 1: 40 % verkauft
Ende Monat 1: 60 % verkauft -> Mitte des Monats 50 %
Er muss also mitten im ersten Monat (genauer am 15. Tag des Monats 1) nachbestellen.

e) Der Händler rechnet für die Nachbestellung eine Lieferzeit von 14 Tagen ein.
In welchem Monat muss er dann nachbestellen, wenn ein Monat 30 Tage hat?
Monat 1: 15.Tag minus 14 Tage Lieferzeit -> Monat 1: erster Tag

Beispiel 5: Chi² Test (Odds)

In einer Umfrage unter 300 Personen wird die Zustimmung der Bevölkerung zu einer geplanten Maßnahme erhoben. Es ergaben sich die folgenden Ergebnisse:

Zustimmung	Keine Zustimmung	Gesamt
Anhänger der Regierungspartei	57	78	135
Anhänger der Opposition	34	131	165
Gesamt	91	209	300

a) Welche grafischen Darstellung der Daten würden Sie empfehlen (Absolutwerte, verschiedene Prozentwerte, ...)?
Optimal zur Darstellung wären Säulendiagramme mit prozentueller Angabe bezüglich der Spalten oder Mosaic-Plots.

b) Liegt auf Grund der Daten genügend Evidenz vor, dass die Zustimmung bei den Anhängern der Regierungspartei höher ist? (Signifikanzniveau α = 0,05; kritischer Wert = 3,84)

$$T^2 = X^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = n_{..}\frac{(n_{11}n_{22}-n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}} = 300\cdot\frac{(57*131-34*78)^2}{91*209*135*165} = 16{,}4176$$

Den kritischen Wert ermittelt man durch Ablesen des Wertes in der Chi²-Tabelle beim Zeilenwert 1- α = 0,95
Bei einem Freiheitsgrad (Zustimmung oder nicht = 2-1) 1 => 3,8415

|Chi²| = 16,41 > 3,84 Die Zustimmung ist also abhängig von der Partei. Da die Odds-ratio > 1 (s. c) (wenn Odds-Ratio >1 dann sind die Fakten, die auf der Hauptdiagonale liegen, wahrscheinlicher) ist die Zustimmung bei Anhängern der Regierungspartei höher.

c) Berechnen Sie die Odds-Ratio und interpretieren Sie diese. Welchen Vorteil hat die Odds-Ratio gegenüber der Differenz der Anteile?

$\Psi = \frac{n_{11}n_{22}}{n_{12}n_{21}}$ =57*131 / 34*78 = 2,81 $\neq 1$ Daher wird die Alternativhypothese angenommen.
Dementsprechend wird die Nullhypothese H_0, dass die Parteizugehörigkeit keinen Einfluss auf die Zustimmung hat, verworfen.

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.

Beispiel 6: Markovkette

a) Untersuchen Sie ob die Markovkette absorbierende Zustände enthält.

Da $p_{3,3}=1$ und $p_{5,5}=1$ (beide in der Form $p_{j,j}=1$), sind die Zustände 3 und 5 absorbierende Zustände.

b) Ist die Markovkette irreduzibel? (Begründen Sie die Antwort)

Nein, die Kette ist reduzibel, da eine Markovkette mit absorbierenden Zuständen gar nicht irreduzibel sein kann: Von den Zuständen 3 und 5 kommt man in keinen anderen Zustand mehr, was der Definition einer irreduziblen Kette, in welcher von jedem Knoten jeder andere Knoten in einer endlichen Anzahl Schritte erreicht werden kann, widerspricht.

c) Zeigen Sie, dass die Anfangsverteilung π=(2/5;3/5;0;0;0) eine stabile Anfangsverteilung ist. Ist dies die einzige stabile Anfangsverteilung?

Nach der Definition einer stabilen Anfangsverteilung π von P muss für diese nach Punkt (iii) des Existenzsatzes gelten: π * P = π

Es handelt sich hierbei um die Multiplikation von Matrizen, wir multiplizieren also die Spaltenterme von π jeweils mit den Zeilentermen einer Spalte von P, addieren diese Terme und setzen sie mit dem jeweiligen Spaltenterm von π gleich. Schreiben wir π allgemein als π = (v,w,x,y,z) ergibt sich folgendes lösbares Gleichungssystem:

$$\begin{cases} \frac{1}{4}*v+\frac{1}{2}*w+0*x+0*y+0*z=v \\ \frac{3}{4}*v+\frac{1}{2}*w+0*x+0*y+0*z=w \\ 0*v+0*w+1*x+\frac{1}{3}*y+0*z=x \\ 0*v+0*w+0*x+\frac{2}{3}*y+0*z=y \\ 0*v+0*w+0*x+0*y+1*z=z \end{cases}$$

Ein weiteres wichtiges Detail ist, dass die Summe der Terme in π genau wie die Zeilensummen der Matrix selbstverständlich auch gleich 1 sein muss. Bezieht man dies ein und vereinfacht obige Gleichungen aus, erhält man:

$$\begin{cases} \frac{1}{4}*v+\frac{1}{2}*w=v \\ \frac{3}{4}*v+\frac{1}{2}*w=w \\ x+\frac{1}{3}*y=x \\ \frac{2}{3}*y=y \\ z=z \\ v+w+x+y+z=1 \end{cases}$$

Rechnet man dies aus, müssen für π = (v,w,x,y,z) also folgende Einschränkungen gelten, damit π eine stabile Anfangsverteilung von P ist:

1. 3v = 2w (folgt aus den ersten beiden Gleichungen)

2. y = 0 (folgt aus der dritten und vierten Gleichung)
3. Aus 3ter und 4ter Gleichung folgt: x und z beliebig, solange die letzte Gleichung erfüllt wird.

Damit hätten wir also eine allgemeine Formel. π = (2/5;3/5;0;0;0) erfüllt alle diese Kriterien (3* (2/5) = 2* (3/5); 0 = 0; 2/55 + 3/5 = 1) und ist damit eine stabile Anfangsverteilung. Wie man unschwer erkennen kann, sind diese Kriterien sehr allgemein gefasst, wir könnten jederzeit andere stabile Anfangsverteilungen generieren (wenn dies gefragt wäre). Daher ist π bestimmt nicht die einzige stabile Anfangsverteilung.

Für alle Skeptiker noch ein Beispiel: (2/9;1/3;1/3;0;1/9) wäre ebenfalls eine stabile Anfangsverteilung. Man kann so eine stabile Anfangsverteilung leicht finden, indem man die ersten 2 Bedingungen in die 3te einsetzt und dann für die übrig bleibenden Parameter passende Werte einsetzt, sodass kein negativer Wert herauskommt.