Ausarbeitung Prüfung Statistik und Wahrscheinlichkeitstheorie (Universität Wien)

Prüfung 28.10.2003

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

Beispiel 1: Theorie

Welche grafischen Darstellungsformen für Häufigkeiten kennen Sie?
Erklären Sie insbesondere auch, wofür man Kerndichteschätzer und Mosaic-plots verwenden kann.

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: Würfelwerfen

Wie hoch ist die Wahrscheinlichkeit, dass wenn man 20mal hintereinander würfelt, die Augenzahl höchstens 22 oder mindestens 118 beträgt?

Höchstens 22 kann folgendes bedeuten:

20mal die 1 (P = $\binom{20}{0} * \left(\frac{1}{6}\right)^{20} = 2{,}7 * 10^{-16}$) ODER

19mal die 1 und einmal die 2 (P = $\binom{20}{1} * \left(\frac{1}{6}\right)^{20} = 5{,}4 * 10^{-15}$)ODER

19mal die 1 und einmal die 3 (P = $\binom{20}{1} * \left(\frac{1}{6}\right)^{20} = 5{,}4 * 10^{-15}$) ODER

18mal die 1 und zweimal die 2 (P = $\binom{20}{2} * \left(\frac{1}{6}\right)^{20} = 5{,}2 * 10^{-14}$)

Das ergibt als Gesamtwahrscheinlichkeit:

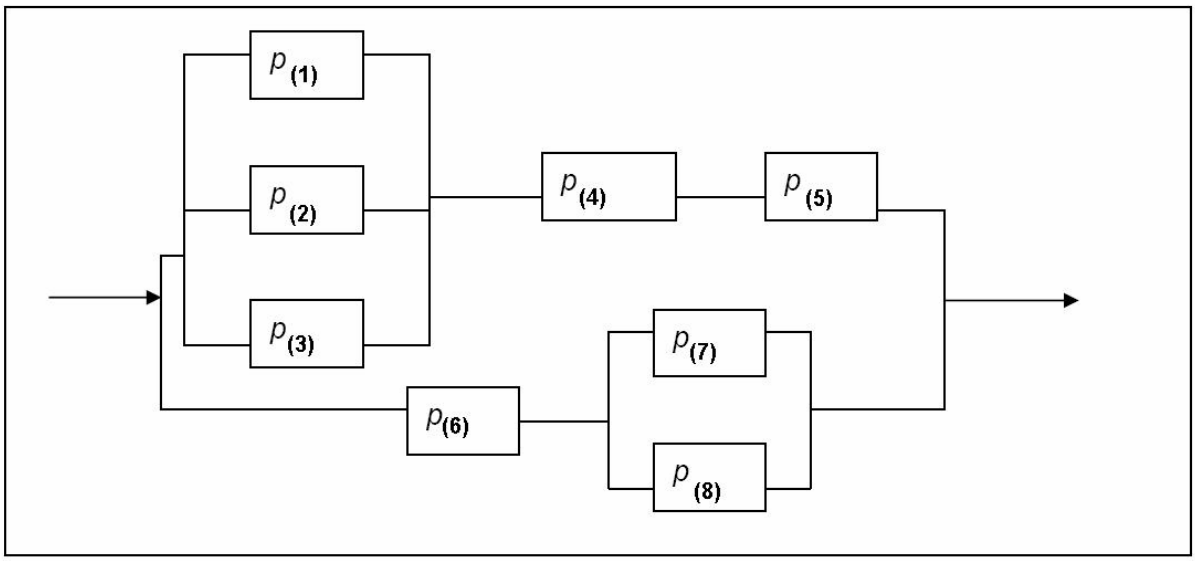
P(höchstens 22) = $\left[\binom{20}{0}+2*\binom{20}{1}+\binom{20}{2}\right]*\left(\frac{1}{6}\right)^{20} = 6{,}3*10^{-14}$

Für P(mindestens 118) ergibt sich aus denselben Überlegungen (mit anderen Würfelzahlen, aber gleichen Wahrscheinlichkeiten) dieselbe Wahrscheinlichkeit. Addiert ergibt das daher:

P = $2*\left[\binom{20}{0}+2*\binom{20}{1}+\binom{20}{2}\right]*\left(\frac{1}{6}\right)^{20} = 1{,}3*10^{-13}$

Beispiel 3: Netzwerk

Gegeben ist unten skizziertes Netzwerk (Anm.: Die verschiedenen ps sind nicht in der Angabe, sie dienen dem besseren Verständnis, an welchen Schaltungen wir gerade rechnen)



a) Berechnung der Zuverlässigkeit mit den angegebenen Werten

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)

Wir wissen, dass z = 1 - p

Daher gilt:

z1 = 1- p1 = 1- 0,05 = 0,95

z2 = 1- p2 = 1- 0,04 = 0,96

z3 = 1- p3 = 1- 0,01 = 0,99

z4 = 1- p4 = 1- 0,05 = 0,95

z5 = 1- p5 = 1- 0,01 = 0,99

z6 = 1- p6 = 1- 0,1 = 0,9

z7 = 1- p7 = 1- 0,2 = 0,8

z8 = 1- p8 = 1- 0,1 = 0,9

Daraus ergibt sich:

z1 o z2: 1- (1 - z1) * (1 - z2) = 1- (1 – 0,95) * (1 – 0,96) = 0,998

[z1 o z2] o z3: 1- (1 - [z1 o z2]) * (1 – z3) = 1- (1 - 0,998) * (1 – 0,99) = 0,99998

z4 o z5: z4 * z5 = 0,95 * 0,99 = 0,9405

[z1 o z2 o z3] o [z4 o z5]: [z1 o z2 o z3] * [z4 o z5]
= 0,99998 * 0,94 = 0,940481

z7 o z8: 1- (1 – z7) * (1 – z8) = 1- (1 – 0,8) * (1 – 0,9) = 0,98

z6 o [z7 o z8]: z6 * [z7 o z8] = 0,9 * 0,98 = 0,882

[z1 o z2 o z3 o z4 o z5] o [z6 o z7 o z8]:
1- (1 - [z1 o z2 o z3 o z4 o z5]) * (1 - [z6 o z7 o z8])
= 1- (1 - 0,940481) * (1 - 0,882) = 0,992977

b) Berechnung der Gesamtlebensdauer, wenn jede Komponente eine Lebensdauer von G(x) besitzt

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)

Daraus ergibt sich nach denselben Überlegungen wie a):

G1 o G2: G(X)^2

[G1 o G2] o G3: G(X)^3

G4 o G5: 2G(X) – G(X)^2

[G1 o G2 o G3] o [G4 o G5]: 2G(X) – G(X)^2 + G(X)^3 – 2G(X)^4 + G(X)^5

G7 o G8: G(X)^2

G6 o [G7 o G8]: G(X) + G(X)^2 - G(X)^3

[G1 o G2 o G3 o G4 o G5] o [G6 o G7 o G8]:
2G(X)^2 + G(X)^3 – 2G(X)^4 – 2G(X)^6 + 3G(X)^7 – G(X)^8

Beispiel 4: 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 Teilnehmer einer Studie auf Grund eines Zufallsverfahrens (Randomisierung) auf die Gruppen verteilt werden.
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)

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 Arzt oder die Patienten (bewusst / unbewusst) die Therapieformen unterschiedlich beurteilen. Um dies zu optimieren, kann Blindung eingesetzt werden. In einer doppelblinden Studie sind weder die Ärzte noch die Patienten über die Studie informiert. Bei einer einfachblinden Studie weiß nur der Arzt 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 neue Medikament keine blutdrucksenkende Wirkung hat.
Eine sinnvolle Alternativhypothese wäre, dass das Medikament den Blutdruck senkt.
H_0: $\mu_D \leq 0$
H_A: $\mu_D > 0$
Was ergibt sich als Aussage dieser statistischen Auswertung? Warum?
Laut Angabe gilt T = 1,7657 < 1,8595
Errechnung T-Wert (optional):

$$s = \sqrt{\frac{1}{n-1} * \sum_{i=1}^{n} (dif_i - \overline{dif})^2} = \sqrt{\frac{1}{8} * (34^2 + 14^2 + 13^2 + 29^2 + 16^2 + 26^2 + 19^2 + 31^2 + 36^2)} = \sqrt{739}$$

$$T = \sqrt{n}\,\frac{\overline{X} - \mu}{s} = \sqrt{n}\,\frac{\overline{X} - \overline{Y}}{s} = \sqrt{9}\,\frac{16}{\sqrt{739}} = 1{,}7657 \quad < 1{,}8595$$

Da der T-Wert kleiner als der kritische t-Wert beim einseitigen t-Test ist, wird die Nullhypothese beibehalten. Das Medikament senkt also nicht nachweislich den Blutdruck.
Ist dieser Test sinnvoll?
Nein, der Test ergibt eigentlich wenig Sinn, da ein T-test grundsätzlich nur dann durchgeführt werden sollte, wenn die Varianz unbekannt ist. Man verschenkt in diesem Fall Information.

Beispiel 5: Chi² Test (Odds)

Ein Kellner arbeitet sowohl tagsüber als auch abends in einem Lokal und führt eine Statistik über das Geschlecht der Besucher. Im Laufe einer Woche ergeben sich die folgenden Daten:

	Tagsüber	Abends	Gesamt
Weiblich	35	52	87
Männlich	33	124	157
Gesamt	68	176	244

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 abends mehr Männer in das Lokal kommen? (Signifikanzniveau α = 0,05; kritischer Wert = 3,84)

$$T^2 = X^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = n_{..}\frac{(n_{11}n_{22}-n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}} = 244\cdot\frac{(35*124-52*33)^2}{68*176*87*157} = 10{,}2773$$

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

|Chi²| = 10,27 > 3,84 Die Besuchszeit ist also abhängig vom Geschlecht. Da die Odds-ratio >1 (s. c) (wenn Odds-Ratio >1 dann sind die Fakten, die auf der Hauptdiagonale liegen, wahrscheinlicher) besuchen Männer das Lokal eher Abends.

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

$\Psi = \frac{n_{11}n_{22}}{n_{12}n_{21}}$ =35*124 / 52*33 = 2,53 $\neq 1$ Daher wird die Alternativhypothese angenommen.
Dementsprechend wird die Nullhypothese H_0, dass das Geschlecht keinen Einfluss auf die Besuchszeit 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: ANOVA

a) Ergänze die Tabelle.

$x_1 = 944,76 / 3 = 314,92$
$x_2 = 206,42 / 1 = 206,42$
$x_3 = 458,98 / 77,59 = 5,924$

	Quadratsummen	Freiheitsgrade	Mittlere Qu.Summe	F	Signifikanz
Gesamtmodell	944,76	3	x1	4,06	0,008
Zeit	458,98	1	458,98	x3	0,016
Geschlecht	206,42	1	x2	2,66	0,104
Zeit*Geschl	7,30	1	7,3	0,09	0,759
Fehler	18622,33	240	77,59		
Gesamt	19567,09	243			

b) Kann man daraus schließen, dass der Rechnungsbetrag tagsüber und abends unterschiedlich ist? (Begründung)

$F_{1;240;0,984} = 6,8509$
6,8 > 5,92 => keine Relevanz der Tageszeit. Man kann daher nicht schließen, dass der Rechnungsbetrag je nach Tageszeit unterschiedlich ist.

Alternative: Signifikanz = p-Wert. Daraus ergibt sich bei Signifikanzniveau 0,05 (wie im vorigen Beispiel genannt) P-Wert(Zeit) = 0,0016 < 0,05 und daher IST der Rechnungsbetrag tatsächlich verschieden je nach Tageszeit.

c) Kann man daraus schließen, dass der Rechnungsbetrag bei Männern und Frauen unterschiedlich ist? (Begründung)?

$F_{1;240;0,896} = 2,7487$

2,75 > 2,66 => keine Relevanz des Geschlechts. Man kann daher nicht schließen, dass der Rechnungsbetrag je nach Geschlecht unterschiedlich ist.

Alternative: Signifikanz = p-Wert. Daraus ergibt sich bei Signifikanzniveau 0,05 (wie im vorigen Beispiel genannt) P-Wert(Geschlecht) = 0,105 > 0,05. Ergibt ebenfalls Beibehaltung der Nullhypothese.

d) Kann man daraus schließen, dass es beim Rechnungsbetrag einen Effekt gibt, der sich aus der Kombination von Tageszeit und Geschlecht zusammen setzt? (Begründung)

$F_{1;240;0,241} = \ldots?$

3 Gründe für die Absenz der Relevanz:
- Keine Tabelle für F(0,241) => Fangfrage ???
- Kein Unterschied zwischen Mann und Frau, Kein unterscheid zwischen Tag und Nacht => Kein Unterschied in der Kombination
- 0,09 als F-Wert ist sehr klein => kann leicht überschritten werden

Alternative: Signifikanz = p-Wert. Daraus ergibt sich bei Signifikanzniveau 0,05 (wie im vorigen Beispiel genannt) P-Wert(Zeit*Geschlecht) = 0,795 > 0,05. Ergibt ebenfalls Beibehaltung der Nullhypothese.