Ausarbeitung Prüfung Statistik und Wahrscheinlichkeitstheorie (Universität Wien)

Prüfung 16.12.2003

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

Beispiel 1: Würfelwerfen

Beschreiben Sie die Wahrscheinlichkeit bei 10maligem Wurf mit einem fairen Würfel mindestens zweimal einen Sechser zu bekommen!

Mindestens 2mal eine 6 bekommen bedeutet NICHT kein- oder einmal eine 6 bekommen.

$$P(0 \text{ mal } 6) = \left(\frac{1}{6}\right)^0 * \left(\frac{5}{6}\right)^{10} = 0{,}16$$

$$P(1 \text{ mal } 6) = \binom{10}{1} * \left(\frac{1}{6}\right)^1 * \left(\frac{5}{6}\right)^9 = 0{,}32$$

Daher gilt:

$$P(\text{min 2mal } 6) = 1 - \left[\left(\frac{1}{6}\right)^0 * \left(\frac{5}{6}\right)^{10} + \binom{10}{1} * \left(\frac{1}{6}\right)^1 * \left(\frac{5}{6}\right)^9\right] = 1 - 0{,}48 = 0{,}52$$

Beispiel 2: Theorie

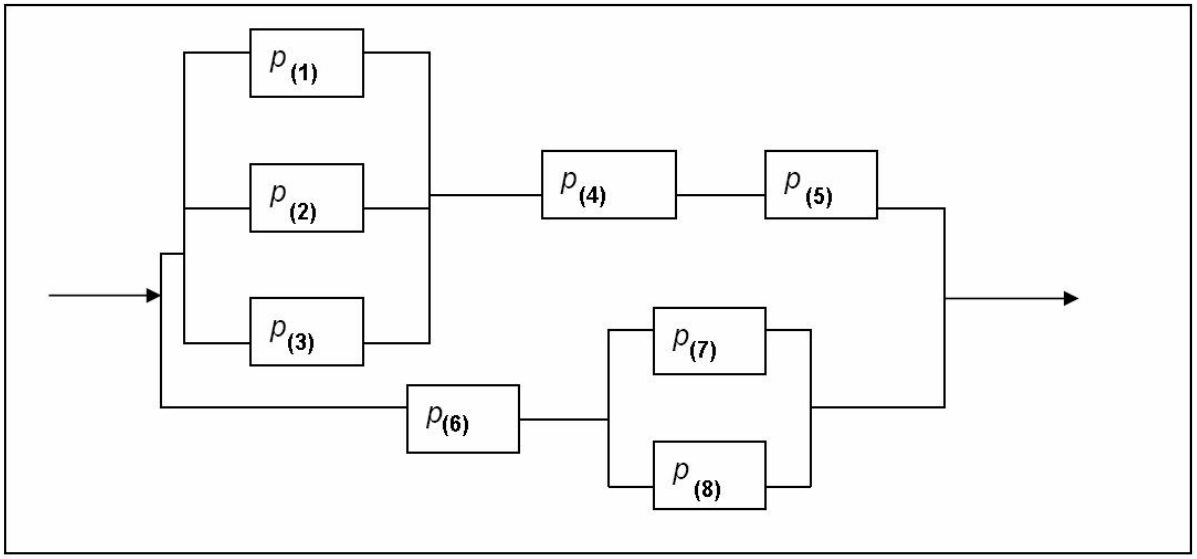
Erklären Sie den Unterschied zwischen Lagemaßen, Mittelwerten und Streuungsmaßen. Welche Invarianzeigenschaften haben diese Maßzahlen?
Welches mit Invarianz verwandte Konzept wird noch zur Charakterisierung von statistischen Maßzahlen verwendet?

Lagemaße:	liegen zwischen Minimum und Maximum der Daten Wenn alle Daten der selben linearen Transformation unterworfen werden, dann macht auch das Lagemaß diese Transformation mit Harmonisches und Geometrisches Mittel sind keine Lagemaße im strengen Sinn
Mittelwerte:	Beziehen sich direkt auf die stochastischen Werte und deren Durchschnitt/Mittel.
Streuung:	Beantworten die Frage nach der „Variabilität der Daten“. Wie weit liegen die Daten auseinander? Die Grundidee dahinter: Man berechnet die mittlere Abweichung von der Mitte.s

(HIER FEHLEN NOCH ANTWORTEN!)

Beispiel 3: Netzwerk

Gegeben ist unten skizziertes Netzwerk (Anm.: p ist überall gleich, Zahlen dienen dem besseren Verständnis, an welchen Schaltungen wir gerade rechnen) [Die Grafik in der Angabe ist fehlerhaft, deshalb wurde die Grafik aus der Prüfung vom 20.1.2004 wiederverwendet, es wird angenommen, dass es sich hierbei um genau dasselbe Beispiel handelt]



a) Berechnung der Zuverlässigkeit, zunächst allgemein mit Ausfallwahrscheinlichkeit p, dann mit p = 0,2

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 $z = 1-p$

Daraus ergibt sich:

z1 o z2: $1- (1 - z1) * (1 - z2) = 1-(1-(1-p))(1-(1-p)) = 1-(p*p) = 1 – p^2$

[z1 o z2] o z3: $1- (1 - [z1 o z2]) * (1 – z3) = 1- (1 – (1-p^2)) * (1 – (1-p)) = 1-(p^2*p) = 1 – p^3$

z4 o z5: $z4 * z5 = (1-p)*(1-p) = (1-p)^2 = 1 – 2p + p^2$

[z1 o z2 o z3] o [z4 o z5]: [z1 o z2 o z3] * [z4 o z5]
$= (1 – p^3) * (1 – 2p + p^2) = 1 – 2p +p^2 - p^3 + 2p^4 – p^5$

z7 o z8: $1- (1 – z7) * (1 – z8) = 1- (1 – (1-p)) * (1 - (1-p)) = 1- (p * p) = 1 – p^2$

z6 o [z7 o z8]: $z6 * [z7 o z8] = (1-p) * (1-p^2) = 1 - p - p^2 + p^3$

[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 – (1 – 2p +p² - p³ + 2p^4 – p^5)) * (1 – (1 - p - p² + p³))
= 1- (2p -p² + p³ - 2p^4 + p^5) * (p + p² - p³)
= 1 - 2p² +p³ - p^4 + 2p^5 - p^6 - 2p³ +p^4 - p^5 + 2p^6 – p^7 + 2p^4 – p^5 + p^6 - 2p^7 + p^8
= 1 - 2p² -p³ +2p^4 + 2p^6 - 3p^7 + p^8

Setzt man p=0,2 ein, ergibt dies
z(gesamt) = 1 - 2p² -p³ +2p^4 + 2p^6 - 3p^7 + p^8 = 0,915

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 Stichproben randomisiert werden.

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.

Wie lauten sinnvolle Nullhypothesen bzw. Alternativen, wenn Sie die unten angeführte statistische Auswertung verwenden wollen?
Als sinnvolle Nullhypothese wäre anzunehmen, dass der Dünger keinen Unterschied macht. Eine sinnvolle Alternativhypothese wäre, dass der Dünger die Erträge steigert.

H_0: $\mu_D \leq 0$
H_A: $\mu_D > 0$

Was ergibt sich als Aussage dieser statistischen Auswertung? Warum?
Laut Angabe gilt $T = -1{,}02 < 1{,}833$
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: ANOVA

a) Vervollständige die Tabelle und bestimme den Wert der F-Statistik zum Test der Nullhypothese, dass die Mittelwerte der Gruppen gleich sind.

SST = Sum of Squares Total
SSM = Sum of Squares Model
SSE = Sum of Squares Error
MSM=Mean of Squares Model

SST = SSM+SSE => SSE = SST-SSM = 241-190 = 51
n-1=44
n-r=41
MSM = SSM/(r-1) = 190/3 = 63,4
MSE = SSE/(n-r) = 51/41 = 1,24
F-Wert = MSM/MSE = 51,07

	Quadratsummen	Freiheitsgrade	Mittlere Quadratsummen	F-Wert	p-Wert
zw. d. Gruppen	190	3	63,4	51,07	7,7*10^-11
innerhalb d. Gr.	51	41	1,24		
Total	241	44			

b)Liegt auf Grund der Daten genügend Evidenz vor, dass zwischen den Gruppen ein Unterschied besteht (Signifikanzniveau α = 0,01)?
F(0,99;3;41) = 4,3126
99 % liegen im Bereich kleiner gleich 4,31; ab einem Wert von 4,31 wird daher die Nullhypothese verworfen.
F = 51,07 legt genügend Evidenz vor, sodass ein Unterschied zwischen den Gruppen besteht.

c)Wie groß ist die Anzahl der Beobachtungen in jeder Gruppe unter der Annahme, dass die Anzahl der Beobachtungen in allen Gruppen gleich ist?
Anzahl der Elemente / Anzahl der Gruppen = n/r = (44+1)/(3+1) = 11,25

d)Welche Möglichkeiten zur Darstellung der Daten würden Sie bei der Varianzanalyse empfehlen?
Optimal wären Whisker-Box-Plots oder Strip-charts.

e)Welche deskriptiven Statistiken sollte man jedenfalls bei der Varianzanalyse betrachten?
Man sollte die Mittelwerte und Standardabweichungen der einzelnen Gruppen betrachten.

Beispiel 6: Chi² Test

Zwei Gruppen, die aus jeweils 150 Personen bestehen leiden an einer Krankheit. Die Gruppe A wird mit einem neuen Mittel behandelt, die Gruppe B mit dem herkömmlichen Mittel. In Gruppe A werden dabei 105 Personen gesund, in Gruppe B werden 45 Personen gesund.

a) Man teste mittels des χ^2-Test die Hypothese, dass zwischen der Wirkung der Medikamente ein Unterschied besteht. (Vergleichswert $\chi^2_{1;\,0.95} = 3{,}84$)

allemeines Schema:

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

Konkrete Vierfeldertafel:

	geheilt	!geheilt	Total
Gruppe A	105	45	150
Gruppe B	45	105	150
Total	150	150	300

$$T^2 = X^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = n_{..}\frac{(n_{11}n_{22}-n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}} = 300*(105^2\text{-}45^2)^2/150\text{^}4 = 48$$

$|T^2| = 48 > 3{,}84$

Es wird daher H0 verworfen und H1 angenommen, das Medikament macht also einen Unterschied.

b) Bestimme die Odds-Ratio für die Wirkung der beiden Medikamente.

$$\Psi = \frac{n_{11}n_{22}}{n_{12}n_{21}} = (105/45)^2 = 5{,}44$$

c) Bestimme ein 95% Konfidenzintervall für die Differenz der Anteile der durch die zwei Behandlungen genesenen Patienten. ($z_{0{,}975} = 1{,}96$)

Siehe Kapitel „Analyse von Häufigkeitsdaten“ - Seite 8

$$p_1 = n_{11}/n_{1.} = 105/150 = 0{,}7$$
$$p_2 = n_{21}/n_{2.} = 45/150 = 0{,}3$$
$$\Delta = p_1 - p_2 = 0{,}7 - 0{,}3 = 0{,}4$$
$$z_{1-\alpha/2} = 1{,}96$$

Konfidenzintervall I

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

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

$$s_\Delta = \sqrt{\frac{0.7 \cdot 0{,}3}{150}} = \sqrt{0{,}014} = 0{,}0374$$

[0,7 - 1,96*0,0374;0,7 + 1,96*0,0374] = [0,63;0,77]

Konfidenzintervall II

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

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

$$s_\Delta = \sqrt{\frac{0.7 \cdot 0{,}3}{150}} = \sqrt{0{,}014} = 0{,}0374$$

[0,3 - 1,96*0,0374;0,3 + 1,96*0,0374] = [0,23;0,37]

Konfidenzintervall für die Differenz der Anteile:

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

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

$$s_\Delta = \sqrt{\frac{0{,}7 \cdot 0{,}3}{150} + \frac{0{,}7 \cdot 0{,}3}{150}} = \sqrt{2 \cdot \frac{0{,}7 \cdot 0{,}3}{150}} = \sqrt{0{,}0028} = 0{,}052$$

[0,4-1,96*0,052 ; 0,4+1,96*0,052] = [0,3 ; 0,5]

d) Welche Voraussetzungen müssen für die Berechnungen nach a) und c) gelten?
Die Ereignisse müssen voneinander unabhängig sein.

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

f) 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.

g) 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. h) (wenn Odds-Ratio >1 dann sind die Fakten, die auf der Hauptdiagonale liegen, wahrscheinlicher) besuchen Männer das Lokal eher Abends.

h) 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

e) 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			

f) 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.

g) 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.

h) 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.