Ausarbeitung Prüfung Statistik und Wahrscheinlichkeitstheorie (Universität Wien)

Prüfung 30.03.2004

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

Beispiel 1: Kombinatorik

Eine Gruppe aus 12 Personen wählt ein Komitee mit 5 Stimmen. Wieviele Möglichkeiten gibt es – mit mehreren Stimmen pro Person? –mit einer?

Es gibt eine Wiederholung (jedes Mitglied kann auch mehrere Stimmen haben) daher gilt für die Anzahl der Möglichkeiten A:

$$A = \binom{n+k-1}{k} = \binom{16}{5} = 4368$$

Gibt es keine Wiederholung, so gilt ganz einfach:

$$A = \binom{n}{k} = \binom{12}{5} = 792$$

Beispiel 2: Würfel

Man wirft 10mal einen unfairen (p(6)=0,125) Würfel
a) Wahrscheinlichkeit min eine 6 zu erhalten? Wie wäre sie bei einem fairen Würfel?

Mindestens eine 6 bedeutet: NICHT keine 6. Dafür gilt:
$P(\text{min eine 6}) = 1 - (0{,}875)^{10} = 0{,}74$

Für einen fairen Würfel würde gelten:
$$P(\text{min eine 6}) = 1 - \left(\frac{5}{6}\right)^{10} = 0{,}84$$

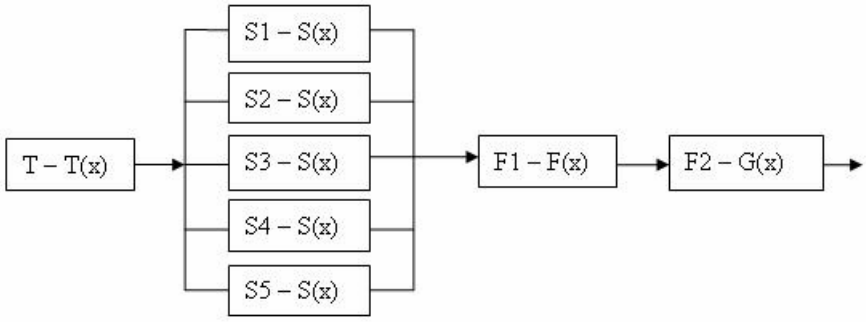
b) Durchschnittsaugensumme bei 3mal Würfeln?

Der Erwartungswert bei einmaligem Würfeln ist:
$$E(X) = 6 * 0{,}125 + \left(\frac{1+2+3+4+5}{5} * 0{,}875\right) = 3{,}375$$

Es gilt: E(3X) = 3E(X) = 3 * 3,375 = 10,125
Man könnte also erwarten, eine 10 zu würfeln.

Beispiel 3: Netzwerk

a) Zeichnen Sie den Schaltplan des beschriebenen Netzes
(eine mögliche Lösung, wie Punkt c zeigen wird ist es egal, in welcher Reihenfolge man die Komponenten anreiht)



b) Berechnen Sie die Lebensdauerverteilung obigen Netzes
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:

S1 o S2 o S3 o S4 o S5: S(X)^5

F o G: 1- (1 - F(X)) * (1 - G(X))

T o [S1 o S2 o S3 o S4 o S5]: 1- (1 - T(X)) * (1 - S(X)^5)

[T o [S1 o S2 o S3 o S4 o S5]] o [F o G] :
1- (1 – [1- (1 - T(X)) * (1 - S(X)^5)]) * (1 – [1- (1 - F(X)) * (1 - G(X))])
=1- (1 - T(X)) * (1 - S(X)^5) * (1 - F(X)) * (1 - G(X))

c) Ist es für die Lebensdauerverteilung wichtig, in welcher Reihenfolge die Firewalls geschaltet sind?
Nein denn die Gesamtlebensdauer wäre dieselbe:
1- (1 - F(X)) * (1 - G (X)) = 1- (1 - G(X)) * (1 - F(X))

d) Wie berechnet man die erwartete Lebensdauer des Gesamtsystems?
(ACHTUNG: Lösung unsicher!)
Da X nur positive Werte annehmen kann, lässt sich das ganze über den Erwartungswert der Verteilungsfunktion V(X) des Systems errechnen. Hierbei gilt:

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

Beispiel 4: Wahlanalyse

Dieses Beispiel ist laut Angabe von Herrn Prof. Grossmann fehlerhaft, es fehlen die nötigen Angaben, um das Beispiel lösen zu können. Dies war bei dieser Prüfung auch die einzig richtige Antwort (so etwas sollte aber nicht mehr vorkommen).

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 deskriptiven Statistiken und Grafiken sollte man jedenfalls bei der Varianzanalyse betrachten?
Man sollte die Mittelwerte und Standardabweichungen der einzelnen Gruppen betrachten. Optimal zu deren Darstellung wären Whisker-Box-Plots oder Strip-charts.

e) Welche Gruppen von Hypothesen gibt es bei der zweifachen Varianzanalyse?
Folgende Fragestellungen sind bei der zweifachen Varianzanalyse von Interesse:

- Unterschiede bezüglich Faktor A;
- Unterschiede bezüglich Faktor B;
- Wechselwirkungen zwischen den Faktoren A und B.

Beispiel 6: Chi² Test (Odds)

In einer Umfrage unter 300 Personen wird die Zustimmung der Bevölkerung zum Bau einer Tiefgarage erhoben. Die Ergebnisse der Befragung, gegliedert nach Geschlecht der Befragten, sind in folgender Tabelle zu finden

	Zustimmung	Keine Zustimmung	Gesamt
Frauen	34	131	165
Männer	61	74	135
Gesamt	95	205	300

f) Welche grafische Darstellung der Daten würden Sie in Hinblick auf die Fragestellung empfehlen (Absolutwerte, welche Prozentwerte)?
Optimal zur Darstellung wären Säulendiagramme mit prozentueller Angabe bezüglich der Spalten oder Mosaic-Plots.

g) Man erkläre den Unterschied zwischen Homogenitätshypothese und Unabhängigkeitshypothese.
Siehe Kapitel „Analyse von Häufigkeitsdaten“ - Seiten 10, 14, 15

Homogenitätshypothese:
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.

Unabhängigkeitshypothese:
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.

In unserem Fall:
Die Homogenitätshypothese bezieht sich auf den Anteil der Geschlechter in der jeweiligen Zustimmungsgruppe (spaltenweise) und die Unabhängigkeitshypothese auf den Anteil der Zustimmung beim jeweiligen Geschlecht (zeilenweise).

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

$$T^2 = X^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = n_{..}\frac{(n_{11}n_{22}-n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}} = 300\cdot\frac{(34*74-131*61)^2}{165*135*95*205} = 20{,}72976232$$

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

|Chi²| = 20,73 > 3,84 Die Zustimmung ist daher abhängig vom Geschlecht. Da die Odds-ratio <1 (s. i) (wenn Odds-Ratio <1 dann sind die Fakten, die nicht auf der Hauptdiagonale liegen, wahrscheinlicher) ist die Zustimmung bei Männern höher.

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

$\Psi = \frac{n_{11}n_{22}}{n_{12}n_{21}}$ =34*74 / 131*61 = 0,315≠ 1 Die Alternativhypothese wird also angenommen.
Dementsprechend wird die Nullhypothese H_0, dass das Geschlecht 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.