04b Sample Examination Problems Chapter 14 SOLUTIONS

1. (a) Why are the degrees of freedom for a test of independence of row and column classifications in an $r \times c$ contingency table equal to $(r-1)(c-1)$?

DF : (r -1)(c -1) ?
In a contingency table , we have r rows and c columns , i.e. rc cells .
We loose 1 DF because of the constraint that : $\sum O_{ij} = \sum E_{ij}$
Moreover, the probabilities π_i **(row) and** π_j **(columns) are not specified, we only need to estimate (r-1) + (c -1) distinct probability parameters as** $\sum \pi_i = \sum \pi_j$ **= 1 and hece the last one in any row or column can be fixed.**

DF = rc – 1 – (r-1) – (c-1) = rc – r – c + 1 = (r -1)(c -1)

(b) The table below shows the number of units sold by three sales operatives for three different products.

	Product		
Sales Operative	A	B	C
Alpha	14	12	4
Beta	21	16	8
Gamma	15	5	10

i. Is there any difference in the patterns of sales for different Sales Operatives?

ii. Display the information in the table in column profile form, and comment on any association displayed.

i. H_0 **: No association between product and sales operatives (Independent)**
H_1 **: there is association (Dependent)**

	A		B		C		Total
α	14	14.2857	12	9.4286	4	6.2857	30
β	21	21.4286	16	14.1429	8	9.4286	45
γ	15	14.2857	5	9.4286	10	6.2857	30
Total	50		33		22		105

$$E_{ij} = \frac{RowToatal \times ColumnTotal}{OverallTotal}$$

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim X^2_{(r-1)(c-1)}$$

= 0.005714 + 0.701299 + 0.831169
+ 0.008571 + 0.243867 + 0.21645
+ 0.035714 + 0.080087 + 0.194805 = 6.32

Let the significance level $\alpha = 0.05$ **, one tailed test(upper)**
DF = (r-1)(c-1) = (3-1)(3-1) = 4
$X^2_{0.05,4} = 9.488$ **, TS Value = 6.32 < 9.488 , does not fall within the rejection region and Therefore we do not reject** H_0**.That is , there is no evidence at The 5% level of any difference in sales pattern for different operatives, i.e. no evidence of association between the variables.**

ii. Column profile form : percentage of A , B, C sold due to α, β, γ **for e.g. 50 is the total sales of A of which 14 due to** α **:**
Percentage of A sales due to α **= (14/50)(100) = 28%**
The complete table :

	A	**B**	**C**
α	**28%**	**36.4%**	**18.2%**
β	**42%**	**48.5%**	**36.4%**
γ	**32%**	**15.2%**	**45.5%**
	100%	**100%**	**100%**

Any association between the variables would be reflected in differences in the percentages of the rows :
If you look at the γ **row : 32% , 15.2% , 45.5% , there is considerable variations so there is a difference between the operatives.**

2. (a) Why would it usually be unwise to carry out both a chi-squared test for independence of the row and column classifications of a table and a two-way analysis of variance for the same table.

(a) **The** X^2 **test for independence is designed to test for association between two factors.**
In two – way ANOVA : test for independence by seeing If there is an interaction between the variables in an explanatory way rather than in statistical sense.
Therefore , it is not wise to perform both tests to the same table.

(b) The table below shows the numbers of piston ring failures in each of three legs of four compressors.

Compressor	Compressor legs North	Centre	South
1	17	17	12
2	11	9	13
3	11	8	19
4	14	7	28

i. Is there any difference in the pattern of failures over different legs for different compressors?

ii. By looking at the contributions to χ^2, or profiles, give a qualitative description of any difference that you find.

i. H_0 : No association between compressor and legs (Independent)

H_1 : there is association (Dependent)

	North		Center		South		Total
1	17	14.68675	17	11.36145	12	19.95181	46
2	11	10.53614	9	8.15060	13	14.31325	33
3	14	12.13253	8	9.38554	19	16.48193	38
4	14	15.64458	7	12.10241	28	21.25301	49
Total	53		41		72		166

$$E_{ij} = \frac{RowToatal \times ColumnTotal}{OverallTotal} \quad \textbf{and} \quad X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim X^2_{(r-1)(c-1)}$$

$$\begin{aligned} = \ & 0.3644 + 2.7983 + 3.1692 \\ & + 0.0204 + 0.0885 + 0.1205 \\ & + 0.1057 + 0.2045 + 0.3847 \\ & + 0.1729 + 2.1512 + 2.1414 = 11.7223 \end{aligned}$$

Let the significance level $\alpha = 0.05$, one tailed test(upper)
DF = (r-1)(c-1) = (4-1)(3-1) = 6

$X^2_{0.05,6} = 12.59$ **, TS Value =** 11.7223**< 12.59 , does not fall within the rejection region and Therefore we do not reject H_0.That is , there is no evidence at The 5% level of any difference in compressors for different legs. i.e. no evidence of association between the variables.**

ii. Column profile form : See 1(b)ii.

	North	Center	South
1	32.1%	41.5%	16.7%
2	20.8%	22%	18.1%
3	20.8%	19.5%	26.4%
4	26.4%	17.1%	38.9%
	100%	100%	100%

Any association (variations) between the variables would be reflected in differences in the percentages of the rows

If you look at the 1 and 4 rows :we can see for compressors 1 and 4 , there appears to be some association (41.5% , 16.7% variation in 1) and (17.1% , 38.9% in 4) between the Center and the South legs.

3. (a) Explain why the fitted values for a χ^2 test of association in a two-way table take the form that they do.

Like any hypothesis test , we assume the null hypothesis is true and we construct our test statistic under this assumption.

The expected values E_{ij} , are calculated under the null

Independent (no association) :

Independence under H_0 implies : $\pi_{ij} = \pi_i \times \pi_j$

i.e. the joint probability is equal to the product of the marginal probabilities, therefore the expected(fitted) values:

$$E_{ij} = \frac{RowToatal \times ColumnTotal}{OverallTotal}$$

(b) The table below shows the number of employees of a manufacturer of animal feeds and soap classified by gender, year of entrance, and length of service in months before resignation. Only those employees with lengths of service of less than 15 months are included.

	1950 Entrants		1951 Entrants	
Length of service	Male	Female	Male	Female
< 3	182	25	147	38
$> 3, < 6$	103	26	54	29
$> 6, < 9$	60	22	47	15
$> 9, < 12$	29	13	21	9
$> 12, < 15$	31	15	12	5

i. Is there any difference in the patterns of length of service over the different columns of the table?

ii. Would there be more or less association in the table if the results for female employees were excluded completely? Explain your answer.

i. H_0 : No association between length of service and years/gender (Independent)

H_1 : there is association (Dependent)

	Male 1950		Female1950		Male1951		Female1951		Total
<3	182	179.80	25	44.84	147	124.75	38	42.62	392
>3,<6	103	97.24	26	24.25	54	67.47	29	23.05	212
>6,<9	60	66.05	22	16.47	47	45.83	15	15.66	144
>9,<12	29	33.02	13	8.24	21	22.91	9	7.83	72
>12,<15	31	28.90	15	7.21	12	20.05	5	6.85	63
Total	405		101		281		96		883

$E_{ij} = \frac{RowToatal \times ColumnTotal}{OverallTotal}$ **and** $X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ **= 36.33**

Let the significance level $\alpha = 0.05$, one tailed test(upper)
DF = (r-1)(c-1) = (5-1)(4-1) = 12

$X^2_{0.05,12} = 21.03$ **, TS Value = 36.33 > 21.03 , falls within the rejection region and Therefore we reject H_0.That is , there is an evidence at the 5% level of association between the variables , the length od service and Years/gender.**

ii. Column profile form : See 1(b)ii.

	Male 1950	Female1950	Male1951	Female1951
<3	44.9 %	24.8 %	52.3 %	39.6 %
>3,<6	25.4 %	25.7 %	19.2 %	30.2 %
>6,<9	14.8 %	21.8 %	16.7 %	15.6 %
>9,<12	7.2 %	12.9 %	7.5 %	9.4 %
>12,<15	7.7 %	14.9 %	4.3 %	5.2 %
	100%	100%	100%	100%

Any association(Variation) between the variables would be reflected in differences in the percentages of the rows

If you look at the 1st and 5th rows :
there appears to be some association (24.8% variation in 1) and (14.9% in 5) between the Female 1950 and others.
Both are in the Female 1950 , hence excluding all females will reduce associations.

Equivalently , you can see this by conducting the test again without The Females :

	Male 1950		Male1951		Total
<3	182	194.23	147	134.77	329
>3,<6	103	92.64	54	64.31	157
>6,<9	60	63.17	47	43.83	107
>9,<12	29	29.52	21	20.48	50
>12,<15	31	25.39	12	17.61	43
Total	405		281		686

$E_{ij} = \frac{RowToatal \times ColumnTotal}{OverallTotal}$ **and** $X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim X^2_{(r-1)(c-1)}$

= 0.771 + 0.111
+ 1.147 + 1.653
+ 0.159 + 0.229
+ 0.009 + 0,013
+ 1.241 + 1.784 = 8.123

Let the significance level $\alpha = 0.05$ **, one tailed test(upper)**
DF = (r-1)(c-1) = (5-1)(2-1) = 4

$X^2_{0.05,4} = 9.488$ **, TS Value = 8.123 < 9.488 , does not fall within the rejection region and Therefore we do not reject** H_0**.That is , there is no evidence at The 5% level of any difference in length of service for different years/gender. i.e. no evidence of association between the variables.**