PES University, Bangalore

(Established under Karnataka Act No. 16 of 2013)

UE19CS203 – STATISTICS FOR DATA SCIENCE

Unit-4 - Hypothesis and Inference

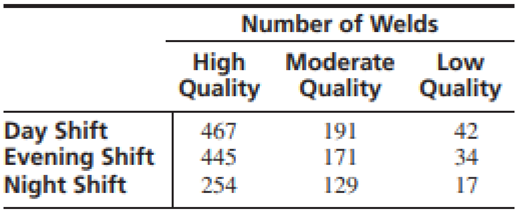
QUESTION BANK

Chi-square Test

Exercises for section 6.10: [Text Book Exercise 6.10– Pg. No. [466 – 468]]

1. Fasteners are manufactured for an application involving aircraft. Each fastener is categorized either as conforming (suitable for its intended use), downgraded (unsuitable for its intended use but usable for another purpose), and scrap (not usable). It is thought that 85% of the fasteners are conforming, while 10% are downgraded and 5% are scrap. In a sample of 500 fasteners, 405 were conforming, 55 were downgraded, and 40 were scrap. Can you conclude that the true percentages differ from $85\%, 10\%$, and 5%?
   a. State the appropriate null hypothesis.
   b. Compute the expected values under the null hypothesis.
   c. Compute the value of the chi-square statistic.
   d. Find the $P - value$. What do you conclude?

2. At an assembly plant for light trucks, routine monitoring of the quality of welds yields the following data:

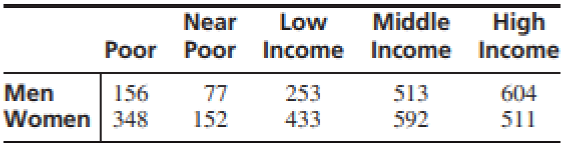


	Number of Welds		
	High Quality	Moderate Quality	Low Quality
Day Shift	467	191	42
Evening Shift	445	171	34
Night Shift	254	129	17

a. State the appropriate null hypothesis.

b. Compute the expected values under the null hypothesis.
c. Compute the value of the chi-square statistic.
d. Find the P-value. What do you conclude?

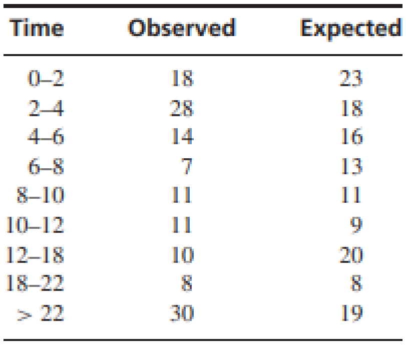
3. The article “Inconsistent Health Perceptions for US Women and Men with Diabetes” (M. McCollum, L. Hansen, et al., Journal of Women’s Health, 2007:1421–1428) presents results of a survey of adults with diabetes. Each respondent was categorized by gender and income level. The numbers in each category (calculated from percentages given in the article) are presented in the following table.



	Poor	Near Poor	Low Income	Middle Income	High Income
Men	156	77	253	513	604
Women	348	152	433	592	511

Can you conclude that the proportions in the various income categories differ between men and women?

4. The article “Analysis of Time Headways on Urban Roads: Case Study from Riyadh” (A. Al-Ghamdi, Journal of Transportation Engineering, 2001: 289–294) presents a model for the time elapsed between the arrival of consecutive vehicles on urban roads. Following are 137 arrival times (in seconds) along with the values expected from a theoretical model.

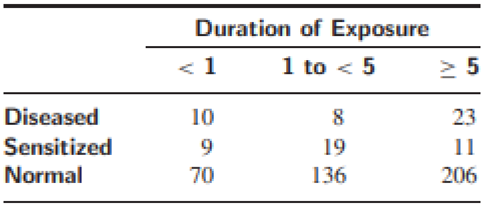


Time	Observed	Expected
0–2	18	23
2–4	28	18
4–6	14	16
6–8	7	13
8–10	11	11
10–12	11	9
12–18	10	20
18–22	8	8
> 22	30	19

Can you conclude that the theoretical model does not explain the observed values well?

5. The article “Chronic Beryllium Disease and Sensitization at a Beryllium Processing Facility” (K. Rosenman, V. Hertzberg, et al., Environmental

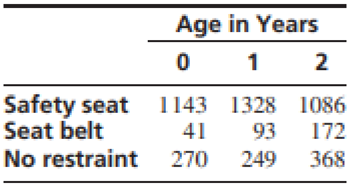
Health Perspectives, 2005:1366–1372) discusses the effects of exposure to beryllium in a cohort of workers. Workers were categorized by their duration of exposure (in years) and by their disease status (chronic beryllium disease, sensitization to beryllium, or no disease). The results were as follows:



	Duration of Exposure		
	< 1	1 to < 5	≥ 5
Diseased	10	8	23
Sensitized	9	19	11
Normal	70	136	206

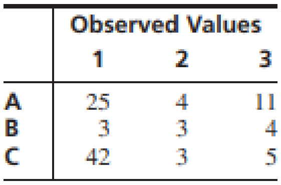
Can you conclude that the proportions of workers in the various disease categories differ among exposure levels?

6. The article "The Effectiveness of Child Restraint Systems for Children Aged 3 Years or Younger During Motor Vehicle Collisions: 1996 to 2005" (T. Rice and C. Anderson, American Journal of Public Health, 2009:252–257) studied a large number of automobile accidents involving small children. Following are the numbers of infants who used various types of restraints, categorized by age.



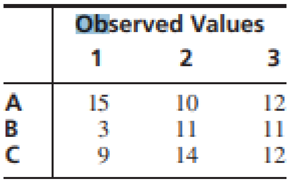
	Age in Years		
	0	1	2
Safety seat	1143	1328	1086
Seat belt	41	93	172
No restraint	270	249	368

7. For the given table of observed values,
   a. Construct the corresponding table of expected values.
   b. If appropriate, perform the chi-square test for the null hypothesis that the row and column outcomes are independent. If not appropriate, explain why.



	Observed Values		
	1	2	3
A	25	4	11
B	3	3	4
C	42	3	5

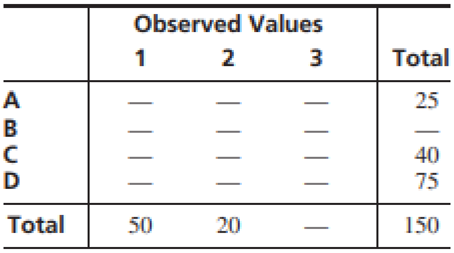
8. For the given table of observed values,
    a. Construct the corresponding table of expected values.
    b. If appropriate, perform the chi-square test for the null hypothesis that the row and column outcomes are independent. If not appropriate, explain why.



	Observed Values		
	1	2	3
A	15	10	12
B	3	11	11
C	9	14	12

9. Fill in the blank: For observed and expected values,
    i. The row totals in the observed table must be the same as the row totals in the expected table, but the column totals need not be the same.
    j. The column totals in the observed table must be the same as the column totals in the expected table, but the row totals need not be the same.
    k. Both the row and the column totals in the observed table must be the same as the row and the column totals, respectively, in the expected table.
    l. Neither the row nor the column totals in the observed table need be the same as the row or the column totals in the expected table.

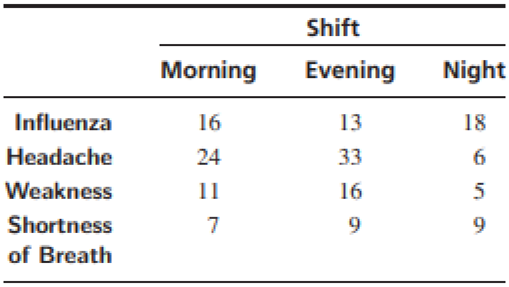
10. Because of printer failure, none of the observed values in the following table were printed, but some of the marginal totals were. Is it possible to construct the corresponding table of expected values from the information given? If so, construct it. If not, describe the additional information you would need.



	Observed Values			
	1	2	3	Total
A	—	—	—	25
B	—	—	—	—
C	—	—	—	40
D	—	—	—	75
Total	50	20	—	150

11. Plates are evaluated according to their surface finish, and placed into four categories: premium, conforming, downgraded, and unacceptable. A quality engineer claims that the proportions of plates in the four categories are $10\%, 70\%, 15\%$, and 5%, respectively. In a sample of 200 plates, 19 were classified as premium, 133 were classified as conforming, 35 were classified as downgraded, and 13 were classified as unacceptable. Can you conclude that the engineer's claim is incorrect?

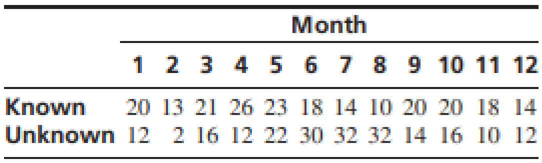
12. The article "Determination of Carboxyhemoglobin Levels and Health Effects on Officers Working at the Istanbul Bosphorus Bridge" (G. Kocasoy and H. Yalin, Journal of Environmental Science and Health, 2004:1129–1139) presents assessments of health outcomes of people working in an environment with high levels of carbon monoxide (CO). Following are the numbers of workers reporting various symptoms, categorized by work shift. The numbers were read from a graph.



	Shift		
	Morning	Evening	Night
Influenza	16	13	18
Headache	24	33	6
Weakness	11	16	5
Shortness of Breath	7	9	9

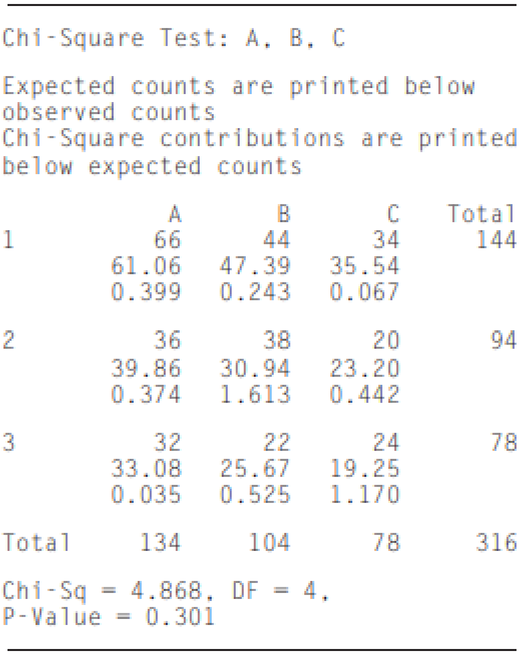
Can you conclude that the proportions of workers with the various symptoms differ among the shifts?

13. The article "Analysis of Unwanted Fire Alarm: Case Study" (W. Chow, N. Fong, and C. Ho, Journal of Architectural Engineering, 1999:62–65) presents a count of the number of false alarms at several sites. The numbers of false alarms each month, divided into those with known causes and those with unknown causes, are given in the following table. Can you conclude that the proportion of false alarms whose cause is known differs from month to month?



	Month											
	1	2	3	4	5	6	7	8	9	10	11	12
Known	20	13	21	26	23	18	14	10	20	20	18	14
Unknown	12	2	16	12	22	30	32	32	14	16	10	12

14. At a certain genetic locus on a chromosome, each individual has one of three different DNA sequences (alleles). The three alleles are denoted A, B, C. At another genetic locus on the same chromosome, each organism has one of three alleles, denoted 1, 2, 3. Each individual therefore has one of nine possible allele pairs: A1, A2, A3, B1, B2, B3, C1, C2, or C3. These allele pairs are called haplotypes. The loci are said to be in linkage equilibrium if the two alleles in an individual's haplotype are independent. Haplotypes were determined for 316 individuals. The following MINITAB output presents the results of a chi-square test for independence.



```
Chi-Square Test: A, B, C

Expected counts are printed below
observed counts
Chi-Square contributions are printed
below expected counts

             A        B        C    Total
1           66       44       34      144
         61.06    47.39    35.54
         0.399    0.243    0.067

2           36       38       20       94
         39.86    30.94    23.20
         0.374    1.613    0.442

3           32       22       24       78
         33.08    25.67    19.25
         0.035    0.525    1.170

Total      134      104       78      316

Chi-Sq = 4.868, DF = 4,
P-Value = 0.301
```

a. How many individuals were observed to have the haplotype B3?
b. What is the expected number of individuals with the haplotype A2?
c. Which of the nine haplotypes was least frequently observed?
d. Which of the nine haplotypes has the smallest expected count?
e. Can you conclude that the loci are not in linkage equilibrium (i.e., not independent)? Explain.

f. Can you conclude that the loci are in linkage equilibrium (i.e., independent)? Explain.