Mathematical Sciences 405/605
Review Exercises in Categorical Data

I. Suppose that a response can fall in one of k = 5 categories with probabilities π_1, π_1, ..., π_5, respectively, and that n= 300 responses produced the following category counts: Conduct a test to

Category	1	2	3	4	5
Observed count	47	63	74	51	65

determine if there is a difference the proportion of counts that fall in each of the categories. Use both an hypothesis testing approach (with α = 0.01) and a significance testing approach ie, determine a *p*-value to make your decision.

II. Gregor Mendel was the first to describe a theory of genetics used is determining genotypes of offspring. The Mendelian theory states that the number of peas of a certain type falling into the classifications i) round and yellow, ii) wrinkled and yellow, iii) round and green, and iv) wrinkled and green should be in the ratio 9:3:3:1. Suppose that 100 such peas revealed 56, 19, 17, and 8 in the respective classes. Do these data disagree with the Mendelian theory ? Use both an hypothesis testing approach (with α = 0.05) and a significance testing approach ie, determine a *p*-value to make your decision.

III. Medical statistics show that deaths due to four major diseases - call them disease A, disease B, disease C, and disease D, account for 15, 21, 18, and 14 percent, respectively, of all non-accidental deaths. A study of the cases of 308 non-accidental deaths at a hospital gave the following counts of patients dying of disease A, disease B, disease C, and disease D:

Disease	**Number of Deaths**
A	43
B	76
C	85
D	21
Others	83

Do these data provide sufficient evidence to indicate that the proportion of people dying of diseases A, B, C, and D at this hospital differ from the proportions accumulated for the population at large ? Use both an hypothesis testing approach (with α = 0.025) and a significance testing approach ie, determine a *p*-value to make your decision.

IV. Computer systems crash for a number of different reasons, among them are software failures, hardware failures, operator errors, and system overloads. It is believed that 10% of all crashes are due to software failure, 5% to hardware failure, 25% to operator error, and 40% to system overloading. Over an extended period of time 150 computer crashes were monitored with the following results: 13 crashes due to software failures, 10 to hardware failures, 42 to operator errors, 65 to system overloading, and the rest to other causes. Do these data lead us to suspect the accuracy of the stated percentages ? Use both an hypothesis testing approach (with α = 0.05) and a significance testing approach ie, determine a *p*-value to make your decision.

V. Although white has long been the most popular car color, recent trends in fashion and home design have signaled the emergence of green as the new color of the 1990s. The growth in the popularity of green hues stems partially from an increased interest in the environment and increased feelings of uncertainty. According to an article in the Press-Enterprise ("White Cars Still Favored," 1993),"green symbolizes harmony and counteracts emotional stress." The article cites the top five colors and the percentage of the market share for four different classes of cars. These data are given below for the truck-van category:

		Medium/Dark			
Color	**White**	**Red**	**Green**	**Red**	**Black**
Percentage	29.72	11.00	9.24	9.08	9.01

In an attempt to verify the accuracy of these figures, we take a random sample of 250 trucks and vans and record their color. Suppose that the number of vehicles falling in each of the five categories above were 82, 22, 27, 21, and 20, respectively.

- Is there any category that is missing in the above classification? How many cars and trucks fell in that category?
- Is there sufficient evidence to indicate that the percentages of trucks and vans differ from those given above? Find the approximate *p*-value for the test.

VI. Research has suggested a link between the prevalence of schizophrenia and birth during certain months of the year in which viral infections are prevalent. Suppose you are working on a similar problem and you suspect a linkage between a disease observed later in life and month of birth. You have records of 400 cases of the disease and you classify them according to month of birth. The data appears in the table below. Do the data present sufficient evidence to indicate that the proportion of cases of the disease per month varies from month to month ? Use both an hypothesis testing approach (with α = 0.10) and a significance testing approach ie, determine a *p*-value to make your decision.

Month	Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec
Number of births	38	31	42	46	28	31	24	29	33	36	27	35

VII. A commercial nuclear power plant contains one or more nuclear power units, the term *nuclear power plant* usually refers to a single nuclear unit. The Oconee nuclear cluster would have three units and hence be considered three plants. In a study of the amount of failures in plants similar to the Oconee plant over the failure history of the plant (*time since first failure*), the number of failures reported for 7 plants were considered with the following results:

Plant	**Failure History Period**	**Number of Failures in Failure History Period**
A	12/82-12/88	35
B	1/78-2/86	16
C	5/76-7/86	18
D	2/83-1/87	9
E	8/83-10/86	13
G	11/78-6/84	8
H	4/84-2/91	11

Ignoring the failure history period, does there appear to be sufficient evidence that the number of failures is different across all plants ? Use both an hypothesis testing approach (with α = 0.10) and a significance testing approach ie, determine a *p*-value to make your decision. Conduct the test again, after removing plant A.

VIII. A cancer researcher performs what is called a prospective be selecting a large group of individuals at random and following their progress for a long period of time. At the end of the study period each individual is classified according to whether or not lung cancer was present and according to whether the individual has been exposed to an identifiable source of airborne asbestos. The result of this classification yielded the following table:

		Exposure Status		**Totals**
		Exposed	**Unexposed**	
Cancer	**Yes**	10	40	50
	No	490	4460	4950
Totals		500	4500	5000

Do these data suggest an association of exposure to airborne asbestos and cancer development ? Use both an hypothesis testing approach (with α = 0.10) and a significance testing approach ie, determine a *p*-value to make your decision.

IX. A study is conducted to test for independence between air quality and air temperature. These data are obtained from records on 200 randomly selected days over the last few years. Do these data indicate an association between these variables ? Use both an hypothesis testing approach (with α = 0.10) and a significance testing approach ie, determine a *p*-value to make your decision.

	Air quality		
Temperature	**Poor**	**Fair**	**Good**
Below average	1	3	24
Average	12	28	76
Above Average	12	14	30

X. A new method for etching semiconductors is being studied. The quality of the etch is to be compared to that obtained using two older techniques. The results of the study are given in the table below. State the null hypothesis of homogeneity mathematically. Use both an hypothesis testing approach (with α = 0.10) and a significance testing approach ie, determine a *p*-value to make your decision.

	Quality				
Method	**Excellent**	**Good**	**Fair**	**Poor**	
High Pressure (old)	113	34	21	32	200
Reactive ion(old)	117	31	25	27	200
Magnetron(new)	130	40	20	10	200
					600

XI. Are baby-boomers more likely to increase their investing now that they are reaching middle age? A poll was conducted by Hal Riney & Partners (Los Angeles Times, June 11, 1990). in which 400 investors were classified according to their age group and their likely investment pattern over the next 5 years versus the last 5 years. The data are shown below. Notice that there were 200 investors included from each age group, ie., a fixed marginal. Do these data provide sufficient evidence to

Age Group	**More**	**Less**	**Same**
35-54	90	18	92
55+	40	60	100

conclude that the investing patterns of the baby-boomers age group differs from that of that of the older age group ? Use both an hypothesis testing approach (with α = 0.01) and a significance testing approach ie, determine a *p*-value to make your decision.

XII. A study of the purchase decisions for three stock portfolio managers A, B, and C was conducted to compare the rates of stock purchases that resulted in profits over a time period that was less than or equal to 1 year. One hundred randomly selected purchases obtained for each of the managers gave the following results:

	Manager		
	A	**B**	**C**
Purchases that resulted in a profit	63	71	55
Purchases that resulted in no profit	37	29	4
Total	100	100	100

Do the data provide evidence of differences among the rates of successful purchases for the three managers? Use both an hypothesis testing approach (with α = 0.05) and a significance testing approach ie, determine a *p*-value to make your decision.

XIII. A problem that sometimes occurs during surgical operations is the occurrence of infections during blood transfusions. An experiment was conducted to determine whether the injection of antibodies reduced the probability of infection. An examination of the records of 138 patients produced the data shown in the accompanying table. Do the data provide sufficient evidence to indicate that injections of antibodies affect the likelihood of transfusional infections? Use both an hypothesis testing approach (with α = 0.01) and a significance testing approach ie, determine a *p*-value to make your decision.

	Infection	**No Infection**
Antibody	4	78
No antibody	11	45

XIV. A recent study claims that an increasing proportion of engineering firms are purchasing liability insurance. This claim is based on a survey of 753 engineering firms. The status of each firm is recorded for the current and for the previous year. The data upon which the claim is based are shown in the table below. Do the data support the claim? Explain, based on the *p*-value of McNemar's test.

	This year		
Last year	**Insured**	**Uninsured**	
Insured	650	5	655
Uninsured	28	70	98
	678	75	753

XV. The following table shows the categorization of 204 men awaiting bypass heart surgery according to the relative degree of each man's coronary artery obstruction and according to his perceived level of discomfort due to angina pectoris (Jenkins et al., 1983). Do the data present sufficient evidence to indicate that the level of angina is dependent on the level of coroners artery obstruction? The authors report the *p*-value for a chi-square test to be $p = 0.01$.

- Compute the value of χ^2 for the data.
- Find the *p*-value for the test and compare with the authors' value $p = 0.01$.
- What conclusions would you reach based on your analysis ?

	Arteries Obstructed 75% or More			
Level of Angina	**0 or 1**	**2**	**3 to 6**	**Total**
None	3	21	20	44
Mild	2	12	9	23
Moderate	26	20	31	77
Moderate/severe	13	10	18	41
Severe	7	5	7	19