# STA 312f10 Assignment 2

Do this review assignment in preparation for the quiz on Friday, Sept. 24th. The non-computer parts are practice for the quiz, and are not to be handed in. But please bring your $R$ printout from Question 9a, the beer study. It may be handed in (or maybe not).

1. Ten friends have a party right after graduating from university. At the time, none of them has ever been married. The party includes a visit by a fortune teller, who says "Five years from now, 3 of you will still be unmarried, 3 of you will be married for the first time, 2 will be divorced, one will be married for the second time, and one will be widowed."

   How many ways are there for this to happen? The answer is a number. Show your work.

2. Students entering U of T have to choose a division: Humanities, Social Sciences, or Sciences.

   (a) Of the 25 students from a particular high school, how many ways are there for 8 to choose the Humanities, 14 to choose the Social Sciences and 3 to choose the Sciences? The answer is a number. Show your work.

   (b) Of the 3 students from another high school, how many ways are there for 1 to choose the Humanities, 1 to choose the Social Sciences and 1 to choose the Sciences? The answer is a number. Show your work.

3. Customers arrive at a Tim Hortons according to a Poisson process with rate $\lambda = 30$ per hour. What is the probability that exactly 8 customers arrive during a 10-minute period?

4. A fair die is tossed 8 times. What is the probability of observing the numbers 3 and 4 twice each, and the others once each? The answer is a number.

5. A box contains 5 red, 3 white and two blue marbles. A sample of six marbles is drawn with replacement. Find the probability that

   (a) 3 are red, 2 are white and one is blue

   (b) 2 are red, 3 are white and 1 is blue

   (c) 2 of each colour appears.

   All the answers are numbers.

6. Let $X_1, \ldots, X_k$ be independent Poisson random variables with respective parameters $\lambda_1, \ldots, \lambda_k$. Given that $Y = \sum_{i=1}^{k} X_i = n$, what is the joint probability distribution of $X_1, \ldots, X_k$? Show your work. Start by stating the range of possible values for $X_1, \ldots, X_k$ given that $Y = n$. This is the support of the conditional distribution for which you are being asked. Hint: See Problem 12 of Assignment 1.

   The lesson of this problem is the following. For many data sets, the number of observations (cases) is in fact a random quantity, not determined in advance. In many such cases, it makes sense to model the events of different types as occurring according to independent Poisson processes. Now, if you are willing to conduct the analysis *conditionally* upon the number of data cases, the distribution is multinomial, and you can use familiar tools without worry.

7. Let $\mathbf{X} = (X_1, X_2, \ldots, X_k) \sim M(1, (p_1, p_2, \ldots, p_k))$. Given that either outcome 1 or outcome 2 has occurred, what is the probability that it was outcome 1? The answer is a symbolic expression. Show your work.

8. Let $\mathbf{X} = (X_1, X_2, \ldots, X_k) \sim M(n, (p_1, p_2, \ldots, p_k))$, with $n > 20$. What is the probability distribution of $X_1$ given that $X_1 + X_2 = 20$? The answer is a symbolic expression. Show your work. Start by stating the range of possible values for $X_1$ given that $X_1 + X_2 = 20$. Hint: Remember how nicely you can combine multinomial categories – see lecture notes.

   The lesson of this problem is that conditional multinomials are still multinomial, so it is okay to restrict an analysis to cases where a certain subset of outcomes have occurred — like only survey respondents who live in urban areas. Methods based on the multinomial distribution still apply, and all your results automatically refer to conditional probabilities. Like Question 6, this is a formal justification of common-sense data analysis practice.

9. Under carefully controlled conditions, 120 beer drinkers each tasted 6 beers and indicated which one they liked best. Here are the numbers preferring each beer.

| | Preferred Beer | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| Frequency | 30 | 24 | 22 | 28 | 9 | 7 |

   (a) The first question is whether preference for the 6 beers is different in the population from which this sample was taken.

      i. State the null hypothesis in symbols. It is a statement about the $p_j$s. Please be specific. The research question allows you to give a specific numerical value for each $p_j$ under $H_0$.

      ii. What are the degrees of freedom of the test? The answer is a number.

      iii. What are the expected frequencies under $H_0$? Your answer is a set of 6 numbers. Are these estimated expected values, or exact expected values?

      iv. Using the "Observed" and "Expected" formula from lecture, calculate the likelihood ratio test statistic $G^2$. Show your work. Your answer is a number. This is something you should be able to do with a calculator if necessary on the quiz.

      v. Now calculate $G^2$ again using $R$.

      vi. Obtain the critical value at $\alpha = 0.05$? with $R$. Check your answer in Appendix III of the text. We will never use the text's approximation formula.

      vii. Calculate the $p$-value using $R$. Print out all the $R$ output and bring it to the quiz.

      viii. Do you reject the null hypothesis at $\alpha = 0.05$? Answer Yes or No.

      ix. It is tempting to ask you to state your conclusion in words. But all you can conclude without further testing is that preference for all the beers is not equal. It *looks* like preference for beers 1 through 4 is greater than preference for 5 and 6, and this is what you would tell your management or client in a job situation.

      x. Calculate the Pearson chi-square statistic $X^2$ for these data. Your answer is a number. This is something you should be able to do with a calculator if necessary on the quiz.

      xi. Do you reject the null hypothesis at $\alpha = 0.05$? Answer Yes or No.

   Just so you can check your answer to this question, my $p$-value for $X^2$ is 0.0002479085.

(b) It seems that the first 4 beers are lagers and the last two are ales. No one would expect preference for lagers and ales to be the same.[1] So let's test whether preference for the 4 lagers is different, and at the same time, whether preference for the 2 ales is different.

i. State the null hypothesis in symbols. It is a statement about the $p_j$s.

ii. What are the degrees of freedom of the test? The answer is a number.

iii. Differentiating the log likelihood function, obtain the maximum likelihood estimator of the parameter $\mathbf{p}$, under the null hypothesis. Show all your work. The answer is a symbolic expression, a vector of length 6. If it is not absolutely the most natural thing you can imagine, then either it's wrong or you have not simplified enough.

iv. Give the maximum likelihood estimate of $\mathbf{p}$ under the null hypothesis for this particular data set. The answer is a set of 6 numbers. Note the difference between an estimator and an estimate.

v. What are the expected frequencies under $H_0$? Your answer is a set of 6 numbers. Are these estimated expected values, or exact expected values?

vi. Using the "Observed" and "Expected" formula from lecture, calculate the likelihood ratio test statistic $G^2$. Show your work. Your answer is a number. This is something you should be able to do with a calculator if necessary on the quiz. You would be silly not to check it with *R*.

vii. Calculate the Pearson $X^2$ statistic for these data. The answer is a number.

viii. What is the critical value for this test at $\alpha = 0.05$? The answer is a number.

ix. Do you reject the null hypothesis at $\alpha = 0.05$ based on the likelihood ratio test? Answer Yes or No.

x. Do you reject the null hypothesis at $\alpha = 0.05$ based on the Pearson Chisquare test? Answer Yes or No.

xi. Based on this analysis, are you able to conclude that there is any difference in preference among the 4 lagers or between the 2 ales? Answer Yes or No. (In applied situations, beware of concluding that there is no effect, or absolutely no difference at all in the population.)

Just so you can check your answer to this question, my $p$-value for the likelihood ratio test is 0.7735209.

[1] Actually, I am making all this up with only a vague idea of what these terms mean.

10. A sample of 150 students each try to solve two difficult logic problems. The problems are complicated, but it's multiple choice, so the answers are either right or wrong. For each student, the data file indicates whether he or she got Question 1 right, and whether he or she got Question 2 right. It is expected that students who get one question correct will also tend to get the other correct; this is not the issue. The issue is whether the one question is more difficult than the other.

    At first glance, this seems like just a problem of comparing two proportions, but there is a twist. Each student answers *both* questions. In a multinomial model, there are $N$ *independent* observations, so each case (person or whatever) must contribute only one frequency to the table.

    We will set up the problem as a multinomial with four categories, as follows.

| | Question 1 Correct | Question 1 Incorrect |
|---|---|---|
| Question 2 Correct | $p_1$ | $p_2$ |
| Question 2 Incorrect | $p_3$ | $p_4$ |

    The observed frequencies are:

| | Question 1 Correct | Question 1 Incorrect |
|---|---|---|
| Question 2 Correct | 66 | 41 |
| Question 2 Incorrect | 30 | 13 |

    (a) State the null hypothesis in symbols. It is a statement about the $p_j$s. Simplify.

    (b) What are the degrees of freedom for this test?

    (c) Differentiating the log likelihood function, obtain the maximum likelihood estimator of the parameter $\mathbf{p}$, under the null hypothesis. Show all your work. The answer is a symbolic expression, a vector of length 4.

    (d) Give the maximum likelihood estimate of $\mathbf{p}$ under the null hypothesis for this particular data set. The answer is a set of 4 numbers.

    (e) What are the expected frequencies under $H_0$? Your answer is a set of 4 numbers. Are these estimated expected values, or exact expected values?

    (f) Using the "Observed" and "Expected" formula from lecture, calculate the likelihood ratio test statistic $G^2$. Show your work. Your answer is a number. This is something you should be able to do with a calculator if necessary on the quiz.

    (g) Calculate the Pearson $X^2$ statistic for these data. The answer is a number. This is something you should be able to do with a calculator if necessary on the quiz.

    (h) What is the critical value for this test at $\alpha = 0.05$? The answer is a number.

    (i) Do you reject the null hypothesis at $\alpha = 0.05$ based on the likelihood ratio test? Answer Yes or No.

    (j) Do you reject the null hypothesis at $\alpha = 0.05$ based on the Pearson Chisquare test? Answer Yes or No.

    (k) What percent of students answered Question 1 correctly? The answer is a number.

    (l) What percent of students answered Question 2 correctly? The answer is a number.

    (m) Does this study provide solid evidence that the two questions differ in their difficulty? Answer Yes or No.