STAT 6200 — Design of Experiments for Research Workers
Lab 6 – Due Thursday, December 2

Example — Coffee and Occurrence of Myocardial Infarction

- In a random sample of 200 60–64 year old males from the general population, the following data were obtained:

Coffee Consumption		MI in past 5 years? Yes	No	
	< 2 cups/day	10	100	110
	≥ 2 cups/day	20	70	90
		30	170	200

1. Using significance level $\alpha = .01$, test for an association between coffee consumption and occurrence of MI in the past 5 years.

Answer: Provided that the expected cell counts are sufficiently large, the chi-square test is appropriate for this situation.

The test statistic can be computed via the formula

$$X^2 = \frac{n(ad - bc)^2}{(a+c)(b+d)(a+b)(c+d)} = \frac{200[(10)(70) - (20)(100)]^2}{(30)(170)(110)(90)} = 6.694$$

The .01 level critical value for this test may be obtained from Table A.8 of our book, or from Minitab. Select

Calc $\rightarrow$ Probability Distributions $\rightarrow$ Chi-Square...

Then select "Inverse cumulative probability", set "Noncentrality parameter:" to 0, "Degrees of freedom:" to 1, select "Input constant:" and set that constant to $1 - .01 = .99$. Then hit OK. The answer is found to be

$$\chi^2_{1-.01}(1) = 6.635$$

Since $X^2 = 6.694$ exceeds 6.635, we reject the hypothesis of independence and conclude that there is an association between coffee consumption and occurrence of MI.

- The p-value for this test is given by

$$P(\chi^2(1) > 6.694) = 1 - P(\chi^2(1) < 6.694) = 1 - .9903 = .0097$$

where $P(\chi^2(1) < 6.694)$ can be obtained by selecting

Calc → Probability Distributions → Chi-Square...

Then select "Cumulative Probability", set "Degrees of freedom:" to 1, and set "Input constant:" to 6.694. Hitting OK gives the answer, .9903.

The entire chi-square test can be done in Minitab via the following steps:

a. First enter the data. The easiest way to do that is just to enter values for three variables: COFFEE (defines the rows of the table), MI (defines the columns), FREQ (gives the cell counts).

   That is, the data should look like this:

COFFEE	MI	FREQ
1	1	10
1	2	100
2	1	20
2	2	70

   An example worksheet is provided in the file coffee.mtw on the course web site. If you can't figure out how to enter the data yourself, open that worksheet in minitab before proceeding.

b. Now select

   Stat → Tables → Cross Tabulation and Chi-Square...

   Then under "For rows:" select variable COFFEE, under "For columns:" select variable MI, and under "Frequencies are in:" select variable FREQ.

c. (This step is not necessary to perform the basic test, but it does give some useful additional information.) Now click on "Chi-Square...". In the new dialog box that comes up, place checks next to "Chi-Square analysis", "Expected cell counts", and "Each cell's contribution to the Chi-square statistic". Then hit OK twice to give the results of the computations.

The results of these steps should look something like this:

```
Results for: coffee.MTW
Tabulated statistics: COFFEE, MI
Using frequencies in FREQ
Rows: COFFEE   Columns: MI
             1       2     All
1           10     100     110
         16.50   93.50  110.00
        2.5606  0.4519       *
2           20      70      90
         13.50   76.50   90.00
        3.1296  0.5523       *
All         30     170     200
         30.00  170.00  200.00
             *       *       *
Cell Contents:       Count
                     Expected count
                     Contribution to Chi-square
Pearson Chi-Square = 6.694, DF = 1, P-Value = 0.010
Likelihood Ratio Chi-Square = 6.717, DF = 1, P-Value = 0.010
```

- Note that the contingency table is reproduced. In addition, for each cell, the expected cell count is given. For example, the expected cell count for the first cell is

$$E_{11} = \frac{(a+b)(a+c)}{n} = \frac{110(30)}{200} = 16.5$$

- In addition, the third value in each cell gives the contribution of that cell to the X^2 statistic. That is, the test statistic can be computed by summing $(O-E)^2/E$ for each cell. In the $(1,1)$th cell, for example,

$$\frac{(O_{11} - E_{11})^2}{E_{11}} = \frac{(10 - 16.50)^2}{16.50} = 2.5606$$

- The entire test statistic is the sum of these contributions. That is,

$$X^2 = 2.5606 + .4519 + 3.1296 + .5523 = 6.694$$

2. Compute Fisher's Exact Test of independence between coffee consumption and occurrence of MI during the last 5 years. Again, use significance level $\alpha = .01$.

Answer: For now, we're not going to say too much about exactly how Fisher's exact test is computed or why those computations make sense. At this point, all I want you to know is that Fisher's Exact Test is the exact version, of the two sample test for equal population proportions.

- That is, the chi-square and z tests are essentially equivalent ways to test for equal population proportions in a two sample problem. They are based upon a normal approximation to the binomial distribution.
- The exact version of these tests is Fisher's Exact Test.
- Fisher's Exact test is based upon the *hypergeometric distribution*, which is a probability distribution that gives the probability of observing a particular combination of cell frequencies in a 2×2 contingency table, if we assume that the row and column are fixed (i.e., conditional on the observed row and column margins).
- Fisher's exact test computes a p value by determining how extreme the observed cell frequencies are compared to what one would expect under the null hypothesis of no association between rows and columns.
- How "extreme" is quantified by summing up hypergeometric probabilities associated with all tables that have the same margins as the observed table and which are at least as unexpected as the observed table if the null hypothesis were true.
  - The details of this calculation are not important.
- It is only necessary to use Fisher's exact test when the chi-square test of independence is inappropriate because of small sample sizes. That is, the Fisher exact test should be used whenever any of the expected cell counts are < 5.

- Fisher's exact test can be computed in Minitab as follows: Repeat step (b) as given above (or simply hit ctrl-E to edit the last dialog box that you used in part 1). Then select "Other Stats..." and place a check next to "Fisher's exact test for 2×2 tables". Hitting OK twice should yield the following results:

```
Results for: coffee.MTW
Tabulated statistics: COFFEE, MI
Using frequencies in FREQ
Rows: COFFEE   Columns: MI
             1       2     All
1           10     100     110
         16.50   93.50  110.00
        2.5606  0.4519       *
2           20      70      90
         13.50   76.50   90.00
        3.1296  0.5523       *
All         30     170     200
         30.00  170.00  200.00
             *       *       *
Cell Contents:      Count
                    Expected count
                    Contribution to Chi-square
Pearson Chi-Square = 6.694, DF = 1, P-Value = 0.010
Likelihood Ratio Chi-Square = 6.717, DF = 1, P-Value = 0.010
Fisher's exact test: P-Value =  0.0158415
```

- The p-value for Fisher's exact test is $p = .01584$, which is somewhat higher than the approximate p-value provided by the chi-square test. Based upon a significance level of $\alpha = .01$, we would fail to reject H_0 according to Fisher's exact test in this example.

Exercise — Salt and Cardiovascular Disease

- Suppose that a retrospective study is done among men aged 50–54 in a specific county who died over a 1-month period. The investigators attempt to include approximately equal numbers of men who died from cardiovascular disease (CVD, the cases), and men who died from other causes (the controls).

- It is found that of 35 cases, 5 were on a high salt diet before they died, whereas of 25 controls, 2 were on such a diet. These data are summarized in the table below:

		Type of diet		
		High Salt	Low Salt	
Cause of death	Non-CVD	2	23	25
	CVD	5	30	35
		7	53	60

a. According to the hypothesis that there is no association between the level of salt in the diet and death from CVD, what is the minimum expected cell count in this table?

b. Based on your answer to part (a), which test should you use to test the hypothesis of no association between the level of salt in the diet and death from CVD? Fisher's exact test or the Pearson chi-square test?

c. Compute the p value for whichever test is more appropriate here. Based upon this p-value and a significance level of $\alpha = .05$ what do you conclude?