

3.3. Independence and Categorical Variables

Purpose: Learn what it means for two categorical variables to be independent. Also, explore what can and cannot be concluded about the independence of two categorical variables for population data based only on sample data. This background will be useful when tests of independence are considered in STT 265.

Reading Assignment: Read through Section 3.7.

Step 0: Move two files on your Novell account. For this lab, the populations you'll sample from are "generated" by Minitab "programs" (collections of Minitab commands) which are stored in the files named "pop1.MTB" and "pop2.MTB" in the folder "STT 264 Data". To use the files, you must first move them out of the folder. On your Novell account window, open the folder "STT 264 Data" to a window, then drag these two files from the folder into the window of your Novell account.

Step 1: Execute the commands stored in "pop1.MTB"
Enter the following command, then see what happens.

```
MTB > exec 'pop1'
```

Columns c1 and c2 now contain 1000 rows of data that constitute a population of size 1000. Each person in the population was asked to indicate a drink preference and a cookie preference for an afternoon snack. The three drink choices were 1=coffee, 2=tea, or 3=milk. The three cookie choices were 1=sugar cookie, 2=chocolate chip cookie, or 3=oatmeal cookie.

The Table command provides the number (count) and percentage of people in each of the nine cells of the two-way cross-classification table for this population, as well as the row, column, and overall total counts and percentages.

No doubt you are wondering if there is any sort of relationship between these two categorical variables! First, let's see if the events "prefers coffee" and "prefers sugar cookies" are independent. They are if, when selecting one person at random, the probability of the intersection equals the product of the probabilities. Since 4% of the people in the population prefer coffee and sugar cookies, the probability that a randomly selected person is in the intersection is 0.04. Also, 20% prefer coffee, and 20% prefer sugar cookies, so the product of these probabilities is $(20\%) \times (20\%) = (0.20) \times (0.20) = 0.04 = 4\%$, exactly equal to the probability of the intersection. Hence, the events "prefers coffee" and "prefers sugar cookies" are independent.

So, one of the drink preferences (i.e. coffee) is independent of one of the cookie preferences (i.e. sugar cookie). If each of the drink preferences is independent of each of the cookie preferences, then the qualitative variables "drink preference" and "cookie preference" are said to be **independent**. However, if any drink preference and any cookie preference are not independent, then they are dependent and, as a consequence, the two variables drink preference and cookie preference are said to be **dependent**.

STOP AND THINK: Are the variables "drink preference" and "cookie preference" independent for population 1? (For each of the 9 cells of the table, check whether the percentage in that cell is the same as the product of the corresponding row and column percentages. You will be asked to provide these and similar subsequent details in your lab report.)

STOP AND THINK: A common statistical problem is to use a random sample from a population to test if two categorical variables are independent. The categorical variables "drink preference" and "cookie preference" for population 1 are independent, but suppose all we had were sample data. Suppose we take a random sample of size 500 from our population. For the sample data, do you think that "coffee" and "sugar cookie" will be independent? Do you think that "drink preference" and "cookie preference" will be independent?

Step 2: Let's do it! Take a random sample of size 500 from our population, sampling without replacement, and determine if the two categorical variables are independent **for the sample data**. To randomly sample 500 rows of c1 and c2, put the sample into columns c3 and c4, and tabulate the results, give the following Minitab commands.

```
MTB > sample 500 c1-c2, put into c3-c4.  
MTB > table c3-c4;  
SUBC> counts;  
SUBC> totpercents.
```

STOP AND THINK: Are the variables "drink preference" and "cookie preference" independent for this sample data from population 1? (For each of the 9 cells of the table, check whether the percentage in that cell is exactly the same as the product of the corresponding row and column percentages.) Are the variables nearly independent? (For each of the 9 cells of the table, is the percentage in that cell nearly equal to the product of the corresponding row and column percentages?)

Step 3: Repeat step 2, taking another random sample of size 500.

STOP AND THINK: Are the variables "drink preference" and "cookie preference" independent for this second sample from population 1? Are they nearly independent?

STOP AND THINK: If two categorical variables are independent with respect to population data, should we expect them to still to be independent for the data of a random sample? Should they be nearly independent?

Let's look at another population.

Step 4: Execute the commands stored in "pop2"
To do so, enter the following command.

```
MTB > exec 'pop2'
```

Now columns c1 and c2 contain 1000 rows of data that constitute a second population of size 1000, call it population 2. As for population 1, each person in population 2 was asked to indicate a drink preference and a cookie preference for an afternoon snack. The three drink choices were again

1=coffee, 2=tea, or 3=milk, and the three cookie choices were again 1=sugar cookie, 2=chocolate chip cookie, or 3=oatmeal cookie.

STOP AND THINK: Are the variables "drink preference" and "cookie preference" independent for population 2? Are they nearly independent? Which drink preference and cookie preference are furthest from being independent? (In other words, for which cell is the "cell-%" most different from the corresponding product "(row-%)x(column-%)"?)

Step 5: Next we will take a random sample of size 500 from population 2, sampling without replacement, and determine if the two categorical variables are independent **for the sample data**. Give the following Minitab commands to take a random sample of size 500 and tabulate the results.

```
MTB > sample 500 c1-c2, put into c3-c4.  
MTB > table c3-c4;  
SUBC> counts;  
SUBC> totpercents.
```

STOP AND THINK: Are the variables "drink preference" and "cookie preference" independent for this sample data from population 2? (For each of the 9 cells of the table, check whether the percentage in that cell is the same as the product of the corresponding row and column percentages.) Are the variables nearly independent? (For each of the 9 cells of the table, is the percentage in that cell nearly equal to the product of the corresponding row and column percentages.)

Step 6: Repeat step 5, taking another random sample of size 500.

STOP AND THINK: Are the variables "drink preference" and "cookie preference" independent for this second sample from population 2? Are they nearly independent?

STOP AND THINK: Suppose there is a third population, population 3 say, and suppose we do not know if the variables "drink preference" and "cookie preference" are independent for population 3. Furthermore, suppose you are given the tabulated data for a random sample of size 500 from population 3. Using what you have observed for your samples from populations 1 and 2, what can you hope to learn from the sample data about population 3? Specifically, can the sample data possibly convince you that the two variables are independent for the population data? Why or why not? Can the sample data possibly convince you that the two variables are dependent for the population data? Explain!

LAB REPORT: Your lab report should include the contents of your Minitab session window. Annotate the output to show the products "(row-%)x(column-%)" for each cell of each table, for comparison with the corresponding cell percentage. Answer all STOP AND THINK questions. (As always, annotate and append your Minitab output to your report, cross-referencing the output in your report as appropriate.)