Stata Handout #2

The purpose of this lab is twofold. First, I want to give you some basic familiarity with independent samples hypothesis testing commands in Stata. Second, I want to give you the opportunity to generate and interpret some independent samples hypothesis tests using a real social science dataset. I should note that I am writing these descriptions as a Mac user with a Mac version of Stata. Some things may be slightly different on a Windows machine, but we will cross that bridge when we get to it.

Section 1a: Setting Up Some Data
Open Stata 14 (not the 64-bit version). From the command line, type the following:

```
cd N:/soc_30903/
```

If you haven't done so already, download the 2014 GSS dataset located in the "Handout Data" sub-folder under the "Stata Handouts" tab in Sakai. Load it up with the following command:

```
use data/GSS2014.dta, clear
```

You should see the GSS variables populate the "Variables" pane in the main Stata window. Once you have done that, start up your do-file for this session. Recall that you can either use the "Do-File" icon on the top ribbon, or you can simply type `-doedit-` from the command line (without the dashes).

You will turn in this do-file at the end of the lab, so go ahead and put your name on it. You should use this do-file to write and execute the remainder of your commands for the day.

Section Ib: Two-Sample *t*-tests and difference-of-proportions tests
You already have some familiarity with the *t*-test in Stata from your third homework. The syntax for the command—like all of the commands we will cover today—are very simple. The basic structure for the two-sample *t*-test is as follows:

```
ttest DV, by(IV)
```

where `-DV-` is the interval-ratio dependent variable and `-IV-` is the categorical independent variable that consists of two categories.

Let's try a simple example. Let's say we want to know if, in the adult U.S. population, height (in inches) varies between the sexes.

> **Thought experiment #1a:** What are the dependent and independent variables? Have we made the necessary assumptions and met the proper requirements? What is the null hypothesis? What is the alternative hypothesis? What α-level do we want to use?

Now let's run the test:

```
ttest height, by(gender)
```

> **Thought experiment #1b:** What is our t-statistic, and what is its associated p-value? What do these numbers mean? What is the conclusion of our test?

We can also use multinomial variables as long as we select only two categories from the variable. For example, I want to know if Protestant adults are likely to have more children than Catholic adults in the U.S. population.

> **Thought experiment #2a:** What are the dependent and independent variables? Have we made the necessary assumptions and met the proper requirements? What is the null hypothesis? What is the alternative hypothesis? What α-level do we want to use?

Now we select the appropriate categories using the following syntax:

```
ttest childs if relig==1 | relig==2, by(relig)
```

The `-if-` portion of the command is used to specify that we want to run the test *if* the following conditions (up to the comma) are met—in this case, that the respondent is either Protestant (coded as 1) or Catholic (coded as 2). The "pipe" bar (|) means "or," as we want to run the test *if* the respondent is Protestant *or* Catholic. This effectively eliminates all non-Protestants and non-Catholics from the analysis.

Go ahead and run the test.

> **Thought experiment #2b:** What is our t-statistic, and what is its associated p-value? What do these numbers mean? What is the conclusion of our test?

The independent samples difference-of-proportions z-test functions the same way, but now with the `-prtest-` instead of the `-ttest-` command. I want to know if people who are 50 or older are more likely to report that they have diabetes than those who are 49 or younger.

> **Thought experiment #3a:** What are the dependent and independent variables? Have we made the necessary assumptions and met the proper requirements? What is the null hypothesis? What is the alternative hypothesis? What α-level do we want to use?

Now run the test:

```
prtest diabetes2, by(age2)
```

> **Thought experiment #3b:** What is our z-statistic, and what is its associated p-value? What do these numbers mean? What is the conclusion of our test?

You can "condition on" the `-prtest-` just as you did with `-ttest-`.

Section Ic: ANOVA

The ANOVA (ANanalysis Of VAriance) test is also very easy to do in Stata. Unlike the two-sample tests above, the ANOVA test does not require us to specify only two groups on the independent variable. This means that we do not need the -by- portion of the code, nor do we need the conditional "if" statements. In general, you just specify that you want to use the -oneway- command (for "one-way ANOVA," which is the specific type of ANOVA test with which we are concerned), and then list your dependent variable followed by your independent variable. The syntax looks like this:

```
oneway DV IV
```

If you want the group means, standard deviations, and frequency distributions across the groups of your categorical independent variable, you can also follow up the command with a comma and then the -tab- option:

```
oneway DV IV, tab
```

Let's say we want to know if the number of hours worked in the average week is somehow dependent on the method by which a person is paid. The categorical variable, "waypaid2," consists of the following factors: salaried, paid by the hour, and other.

> **Thought experiment #4a:** What are the dependent and independent variables? Have we made the necessary assumptions and met the proper requirements? What is the null hypothesis? What is the alternative hypothesis? What α-level do we want to use?

Let's run the test:

```
oneway hrs2 waypaid2, tab
```

> **Thought experiment #4b:** What is our F-statistic, and what is its associated p-value? What do these numbers mean? What is the conclusion of our test?

Notice we are informed at the bottom of the test output that one of our categories—"other"—only consists of one observation and therefore exhibits no within-group variation. We want to take care of that by removing the category with an "if" statement:

```
oneway hrs1 waypaid2 if waypaid2!=3, tab
```

In the situation where the ANOVA test only has two factors, the F-statistic reduces to the squared t-statistic for that sample mean contrast. In this case: $2.261^2 = 5.11$, where 5.11 is the F-statistic associated with the model where we purposefully omit the "other" category.

Section Ic: Chi-Square Test

We also have the chi-square (χ^2) test at our disposal. Recall that the chi-square test is nonparametric, meaning that we do not have to make any assumptions about the mean or variance (i.e., "parameters") of the sampling distribution from which our test statistic was drawn. As such, this test is ideal for making inferences about variable relationships when the variables in question are categorical—i.e., either ordinal or nominal—especially when they consist of more than two categories.

To run a chi-square test, you first want to construct a bivariate table. This can be done in Stata using the `-tab-` (which stands for "tabulate") command. For instance, I want to make a bivariate table where the row variable is race-ethnicity and the column variable is political party identification:

```
tab race partyid
```

Now we can add some percentages to the cells to make the table a little more interpretable. Let's say we want to see the percentage distribution of the different racial-ethnic groups across the party categories:

```
tab race partyid, row
```

Or the percentage distribution of party categories across the racial-ethnic groups:

```
tab race partyid, col
```

Once we have our bivariate table, we can run the chi-square test by adding the `-chi2-` option after comma:

```
tab race partyid, chi2
```

Our χ^2-statistic and its associated p-value are located just below the table. We see here that our test statistic is statistically significant, meaning that we reject the null hypothesis that these two variables are not related in the U.S. adult population. We instead find support for the alternative hypothesis that there *is* an association in the population: it seems that knowing one's race-ethnicity helps us predict their political party and how strongly they identify with them.

Finally, we can use a combination of the row/column percentages and the "chi-square contributions" to figure out which cells contribute the most to our large test statistic. The chi-square contributions—which will add up to the overall test statistic—can be found with this:

```
tab race partyid, chi2 cchi2
```

The biggest contribution to the test statistic comes from the black-by-strong Democrat cell. That cell alone accounts for 45.33% of the test statistic ([170.4/375.9475]*100). It seems that black U.S. adults identify as strong Democrats at a significantly higher rate than we would anticipate under the assumption that this relationship is random.

Section Ic: Go Forth and Explore

Now try running some hypothesis tests on your own. Use the `-codebook-`, `-tab-`, and `-summarize-` commands to find some variables that interest you. Go through the Thought Experiment exercises again with these new variables (and their interrelationships) in mind. Type your codes into a do-file rather than from the command line in the main window.

Save your dataset and do-file when you are finished (you have to save them separately). You can do this directly from your do-file:

```
save "data/GSS2014.dta", replace
save "do_files/stata_handout_2.do", replace
```

Once you are done, execute the following to close you Stata session and exit the program.

```
clear
exit
```