STA9750 – Lecture III

OUTLINE

1. Non-parametric one- and two-sample tests
    a. One-sample: Sign test
    b. Two-sample: Mann-Whitney-Wilcoxon
    c. Paired tests
2. Contingency tables: independence tests
3. One-way ANOVA and Kruskal-Wallis

1. NON-PARAMETRIC ONE- AND TWO-SAMPLE TESTS

One-sample tests

As you know, PROC UNIVARIATE gives a lot of details about the data. Among the stuff that it does by default, you'll be able to find the t-test, the sign test and the ranked signed test. The default test is 2-sided and it assumes the null hypothesis is that the population mean is 0. You can change the hypothesized value under the null as follows. Suppose that your data is called "onesamp" and you want to test whether the population mean is equal to 1. The following code will do that for you:

```
PROC UNIVARIATE data=onesamp mu0=1;
RUN;
```

Exercise: *A group of scientists recorded the following 10 measurements:*

-1 -2 2 0 -2 -4 2 0 2 -2

They want to know whether the population mean is -1 or if it isn't equal to -1.

Read in the data with SAS and

i) Summarize the data: create a plot and find the mean and standard deviation.

ii) Report the p-value of the appropriate t-test.

iii) Report the p-value of the analogous sign test.

iv) Is the t-test appropriate? Also, comment on the differences between the parametric and the nonparametric test.

Two-sample tests

You can do nonparametric 2-sample tests using PROC NPAR1WAY. If your outcome is called "outcome" and the variable that indicates which group each row comes from is called "group", the code below will do tests for you:

```
PROC NPAR1WAY data=twosamp;
      VAR outcome;
      CLASS group;
RUN;
```

Exercise: *Some scientists want to know whether people who drink coffee regularly are more depressed when they don't drink their morning coffee. They recorded the Hamilton depression score (HamD) of 5 participants who regularly drink coffee, but didn't that morning, and the score of 10 participants who regularly drink coffee and drank their morning coffee that day. The data can be found on the course website under the name "depression.csv".*

Read in the data with SAS and

i) Rename the columns of the dataset to something more meaningful.
ii) Summarize the data: create plots and find group means and standard deviations.
iii) Report the p-value of the appropriate t-test.
iv) Report the p-value of the analogous Mann-Whitney-Wilcoxon test.
v) Is the t-test appropriate? Comment on the differences between the parametric and the nonparametric test..

Paired tests

You can do paired tests by taking the difference between the paired variables and doing a one-sample test.

Exercise: *Download the dataset "zinc_conc.txt" from the course website.*

Read in the data with SAS and

i) Summarize the data: create a plot and find the mean and standard deviation.
ii) Report the p-value of a one-sided paired t-test where the alternative is that the difference "bottom-surface" is not equal 0.
iii) Report the p-value of the analogous sign test.

[Source: https://onlinecourses.science.psu.edu/stat500/node/51/]

Solution: If the name of the original dataset is "zinc", the following code will produce the output needed to answer the questions:

```
DATA zinc2;
      SET zinc;
      diff = bottom-surface;
RUN;

PROC UNIVARIATE data=zinc2;
      VAR diff;
RUN;
```

2. CONTINGENCY TABLES

If your data is called "tab" and the categorical variables that you're interested in are called "v1" and "v2", you can do a chi-square test as follows:

```
PROC FREQ data=tab;
TABLES v1*v2 / chisq;
RUN;
```

You can create stacked bar plots as follows:

```
PROC SGPLOT data=tab;
VBAR v1 / group=v2;
RUN;
```

Exercise: *A study at the University of Texas Southwestern examined whether the risk of hepatitis C was related to whether people had tattoos and to where they got their tattoos. The data can be found on the course website under the name "tat.csv". Read in the data with SAS, tabulate the data, plot it, and run a chi-squared test. Are the variables independent at the 0.05 significance level? Comment on the results.*
[Source: https://www2.stat.duke.edu/courses/Summer13/sta111.001-2/Lectures_site/lec21_site.pdf]

Exercise: *A news article reports that "Americans have differing views on two potentially inconvenient and invasive practices that airports could implement to uncover potential terrorist attacks." This news piece was based on a survey conducted among a random sample of 1137 adults nationwide, interviewed by telephone November 7-10, 2010, where one of the questions on the survey was "Some airports are now using full-body digital x-ray machines to electronically screen passengers in airport security lines. Do you think these new x-ray machines should or should not be used at airports?." The data can be found on the course website under the name "xray.csv". Is policy support independent of political affiliation? Tabulate the data, plot it, and run a chi-squared test. [Source: https://www.openintro.org/]*

3. ONE-WAY ANOVA AND KRUSKAL-WALLIS

Last time, we saw that we can use PROC ANOVA to do one-way ANOVAs. If your data is called anova, your outcome is called "outcome" and your grouping variable is called "group":

```
PROC ANOVA data=anova;
class group;
model outcome = group;
RUN;
```

There's a nonparametric analogue of ANOVA which looks at ranks. Its name is Kruskal-Wallis, and it is implemented in SAS in PROC NPAR1WAY. Again, if your data is called "anova", your outcome is called "outcome" and your grouping variable is called "group":

```
PROC NPAR1WAY data=anova;
VAR outcome;
CLASS group;
RUN;
```

Exercise: *At some university, an introductory math course is taught by 5 different instructors. The math department is interested in knowing if there are significant differences in scores in the different sections. You can find the data on the course website under the name "scores.csv". Plot and summarize the data and help the math department make a decision.*