Part 1: Review of statistical concepts

In this part of the problem set you will consider the relationship between spending per pupil, per capita income, and student test scores in a sample of school districts. Use chapters 2 and 3, the lecture notes, and the formulas given in the appendix to this problem set.

Basic summary statistics for test scores, spending per pupil, and per capita income:

```
. summ totscore8 regday percap if totscore8~=. & regday~=. & percap~=.;

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------
   totscore8 |     184    698.7554   21.09904        641        747
      regday |     184     4.72487   .8686434      3.023      8.759
      percap |     184    18.70823   5.598836      9.686     46.855
```

1. Using this table, report the following statistics:

a) Mean test score $(\bar{X})$.

b) Variance of test score (s_X^2)

c) Variance of mean test score $(s_{\bar{X}}^2)$

d) Standard error of mean test score $(s_{\bar{X}})$

2. Estimate a 95% confidence interval for the mean test score in this sample.

3. Consider the average test score in high-spending and low-spending districts:

```
-> high = 0 (low-spending districts)

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------
   totscore8 |     156    693.9551   18.59426        634        736
```

```
-> high = 1 (high-spending districts)

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------
   totscore8 |      94    706.0106   20.89374        641        747
```

Indicate the value of:

a) Mean difference in test scores between high-spending districts and low-spending ones

b) Standard error of mean difference

4. Test the null hypothesis that the two means are the same, against the alternative that they are different. Be explicit about the null hypothesis, the alternative hypothesis, and test statistics used.

a) State your conclusion in both statistical terms and policy terms.

b) Does this analysis mean that we should spend more per pupil if our objective is to increase test scores?

Part 2: Using Stata

Using the "PS1 Data.dta" dataset, create a do-file to give you the information necessary to answer the following questions. Please submit your do-file and you log-file with your problem set.

You need to show your work in answering questions (1)-(3), i.e. write down the formulas you are using, and use the numbers from the Stata output to compute the relevant confidence interval or test statistic.

1. Use the information that you get from using either the tabulate or the summarize command and construct a 95% confidence interval for the proportion of women in the sample. Verify your calculations by using the ci command in Stata.

2. Test the null hypothesis that the average wage in the population is equal to $11/hour (against the alternative hypothesis that it is not equal to $11) using a 5% significance level. (Note that you will need to create the wage variable). In performing this test, indicate:

a) Null hypothesis.

b) Alternative hypothesis.

c) Test statistic used.

d) Interpretation of the test statistic and conclusion of the statistical test.

Verify your result by using the command ttest.

3. What is the difference in the average wage between men and women? Is this difference statistically significant at the 5% significance level? In testing this, indicate

a) Null hypothesis.

b) Alternative hypothesis.

c) Test statistic used.

d) Interpretation of the test statistic and conclusion of the statistical test.

Verify your results by using the command ttest.

4. Does the result of question 3 provide evidence to conclude that there is gender discrimination in wage, i.e. that women earn less simply because they are women? Explain.

Appendix: Some useful formulas

Sample variance:

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Sample standard deviation:

$$s_X = \sqrt{s_X^2}$$

We can also estimate the variance of the sample mean $(\bar{X})$:

$$s_{\bar{X}}^2 = \frac{s_X^2}{n}$$

Standard error of the sample mean $(\bar{X})$:

$$SE(\bar{X}) = \frac{s_X}{\sqrt{n}}$$

Standard error of the difference between two sample means (assuming independent samples and not restricting the variances in the two samples to be equal):

$$SE(\bar{X} - \bar{Z}) = \sqrt{\frac{s_X^2}{n_X} + \frac{s_Z^2}{n_Z}}$$

Formulating the hypothesis (H_0) that the mean is equal to a certain value against an alternative hypothesis (H_1):

$$H_0: E(X) = \mu_{X,0}$$

$$H_1: E(X) \neq \mu_{X,0}$$

t-statistic for this test:

$$t = \frac{\bar{X} - \mu_{X,0}}{SE(\bar{X})}$$

Formulating the hypothesis that two means are equal to each other against the alternative that they are not:

$$H_0: E(X) - E(Z) = 0$$

$$H_1: E(X) - E(Z) \neq 0$$

t-statistic for this test:

$$t = \frac{\bar{X} - \bar{Z}}{SE(\bar{X} - \bar{Z})}$$

A 95 percent confidence interval for the sample mean $\bar{X}$ is given by:

$$\bar{X} \mp 1.96 \cdot SE(\bar{X})$$