# STA 100 Homework 5

Due 11:59 pm Wednesday, August 2 onto Gradescope

1. Men with prostate cancer were randomly assigned to undergo surgery, or "watchful waiting" (WW). The results over the next several years are found below:

| | Surgery | WW |
|---|---|---|
| Died | 83 | 106 |
| Alive | 264 | 242 |

Assume we want to test if death was independent of what group (Surgery, WW) they were in with $\alpha = 0.05$.

(a) State the null and alternative hypotheses.

(b) Calculate the test statistic and find the critical value.

(c) Find the range of $p$-value.

(d) Do you reject or fail to reject the null?

(e) State the conclusion in terms of the problem.

2. A randomized experiment was conducted in which patients with coronary artery disease either had angioplasty (A) or bypass surgery (B). The accompanying table shows the treatment type and if chest pain occurred over the next 5 years:

| | A | B |
|---|---|---|
| Pain | 111 | 74 |
| No pain | 402 | 441 |

Assume we want to test if chest pain was independent of treatment, with $\alpha = 0.01$.

(a) State the null and alternative hypotheses.

(b) Calculate the test statistic and find the critical value.

(c) Find the range of $p$-value.

(d) Do you reject or fail to reject the null?

(e) State the conclusion in terms of the problem.

3. The following data summarizes the incidence of Coronary Heart Disease (CHD) for people who smoked regularly and for people who did not smoke regularly:

| | CHD | No CHD |
|---|---|---|
| Smoked | 84 | 87 |
| Did not smoke | 2916 | 4913 |

(a) Find a 95% confidence interval for the difference in **proportion of CHD for smokers v.s. nonsmokers**.

(b) Interpret your confidence interval from (a).

(c) Does this interval suggest a dependence on CHD and smoking?

(d) Does this interval suggest the proportion for CHD for patients who smoke is 20% higher than for those who do not? Explain.

4. A sociology student wanted to test if the mean GPA of four sororities were significantly different. They found the following:

| Sorority | A | B | C | D |
|---|---|---|---|---|
| $\bar{Y}_i$ | 3.22 | 3.57 | 2.87 | 2.98 |
| $s_i$ | 0.54 | 0.35 | 0.21 | 0.23 |
| $n_i$ | 10 | 10 | 10 | 10 |

Denote the population means of Sorority A, B, C, D as $\mu_1, \mu_2, \mu_3, \mu_4$, separately and assume $\alpha = 0.01$.

(a) Fill out the ANOVA table

| **Source** | **df** | **SS** | **MS** |
|---|---|---|---|
| Between groups | | | |
| Within groups | | | |
| Total | | | |

(b) State the null and alternative hypotheses.

(c) Calculate the test statistic and find the critical value.

(d) Find the range of $p$-value.

(e) Do you reject or fail to reject the null?

(f) State the conclusion in terms of the problem.

(g) What type of error could we have made in this question?

(h) Calculate the family-wise (simultaneous, Bonferroni) 99% confidence intervals for $\mu_2 - \mu_1$, $\mu_2 - \mu_3$, and $\mu_2 - \mu_4$, where you make $k = 3$ total confidence intervals. Hint: R is needed to find the $t$ multiplier; use `qt(p = 1 -` $\alpha$`/(2*k), df = n - I)`.

(i) Which confidence intervals suggest a significant difference in the means?

5. A group of paramedics does not believe that the average number of calls in the morning, afternoon and night shifts are equal. They counted the number of calls over 7 days, and found the following:

| | Morning | Afternoon | Night |
|---|---|---|---|
| $\bar{Y}_i$ | 2.57 | 3.71 | 4.29 |
| $s_i$ | 0.98 | 1.11 | 1.38 |
| $n_i$ | 7 | 7 | 7 |

Denote the population mean number of calls in the morning, afternoon and night shifts as $\mu_1, \mu_2, \mu_3$, separately and assume $\alpha = 0.01$.

(a) Fill out the ANOVA table

| **Source** | **df** | **SS** | **MS** |
|---|---|---|---|
| Between groups | | | |
| Within groups | | | |
| Total | | | |

(b) State the null and alternative hypotheses.

(c) Calculate the test statistic and find the critical value.

(d) Find the range of $p$-value.

(e) Do you reject or fail to reject the null?

(f) State the conclusion in terms of the problem.

(g) What type of error could we have made in this question?

(h) Calculate the family-wise (simultaneous, Bonferroni) 99% confidence intervals for $\mu_1 - \mu_2$, $\mu_1 - \mu_3$ and $\mu_2 - \mu_3$, where you make $k = 3$ total confidence intervals. Hint: R is needed to find the $t$ multiplier; use `qt(p = 1 - α/(2*k), df = n - I)`.

(i) Are your confidence intervals consistent with the conclusion in (f)?

6. The peak flow rate of a person is the fastest rate at which a person can expel air after taking a deep breath. Peak flow is measured in units of liters/min, and gives an indication of a persons respiratory health. 17 men were randomly sampled, and information on their peak flow and height (in cm) follows:

| | Height | Peak flow |
|---:|---:|---:|
| Sample mean | 180.4118 | 660 |
| Sample standard deviation | 8.5591 | 117.9952 |

The thought is that the height of a man should affect their peak flow. In addition, we are given that the sample correlation is $r = 0.32725$, and SSE $= 198909.3$.

(a) Identify the response variable $Y$ and the explanatory variable $X$.

(b) Calculate the slope and intercept of the fitted regression line.

(c) Interpret the slope and intercept of the fitted regression line in terms of the problem (if appropriate).

(d) Predict the peak flow of a man who is 174 cm tall.

(e) If we are comparing two men, and their difference in heights is 10 cm, what can we expect their average difference in peak flow to be?

7. Continue with the data from Problem 6. Assume the level of significance $\alpha = 0.05$.

(a) State the null and alternative hypotheses if one wants to test if height and peak flow are uncorrelated.

(b) Calculate the test statistic and find the critical value.

(c) Find the range of $p$-value.

(d) Do you reject or fail to reject the null?

(e) State the conclusion in terms of the problem.

(f) Find the 95% confidence interval for the slope. Is your confidence interval consistent with the conclusion in (e)?

(g) Interpret your confidence interval from (f) in terms of the problem.

R is necessary for the remaining questions. Attach source codes and any plots you produce to your homework submission. You may write down your numerical results.

8. On Canvas (Files → Data) you will find the dataset `blood.csv`, which has two columns: `Type` (blood type), and `Disease` (either yes or no). This dataset consists of a random sample from a particular area. Use this dataset and R and assume we want to test if having this particular disease is independent of blood type.

(a) Find the test statistic.

(b) Find the $p$-value.

(c) Do you reject or fail to reject the null if $\alpha = 0.01$? State your conclusion in terms of the problem.

(d) Were blood type A individuals more or less likely to have the disease than what we expected if the null was true?

(e) Were blood type O individuals more or less likely to have the disease than what we expected if the null was true?

(f) Which group contributed most to the value of the test statistic?

9. On Canvas (Files → Data) you will find the dataset `IQ.csv`. It has two columns, the first of which denotes what major a student is from (A, B, or C). The second is the IQ measured by the Stanford-Binet Intelligence Scales. The goal is to determine if this IQ measure differs on average between majors.

   (a) Provide the ANOVA table.

   (b) Find the test statistic and the $p$-value.

   (c) Do we fail to reject or reject the null if $\alpha = 0.05$?

   (d) State your conclusion in terms of the problem.

   (e) Plot a normal quantile plot of the IQ scores. Does this data appear to be approximately normally distributed?

   (f) Calculate the family-wise (simultaneous, Bonferroni) 95% confidence intervals for $\mu_A - \mu_B$, $\mu_A - \mu_C$ and $\mu_B - \mu_C$, where you make $k = 3$ total confidence intervals.

   (g) Which confidence intervals suggest a significant difference in the means?

10. On Canvas (Files → Data) you will find the dataset `fitness.csv`, which contains the following columns:

    Column 1: `Tread`: The typical amount of time training at high intensity on the treadmill ($X$).

    Column 2: `Run`: The time it took to complete a 10 kilometer run (in minutes) ($Y$).

    It is suspected that the amount of high-intensity training may affect the time it took to complete a 10 kilometer run.

    (a) Find the slope and intercept of the fitted regression line.

    (b) Find the 95% confidence interval for the slope.

    (c) Find the value of $s_e$.

    (d) Find the value of $r^2$.

    (e) Does your interval suggest a significant linear relationship?