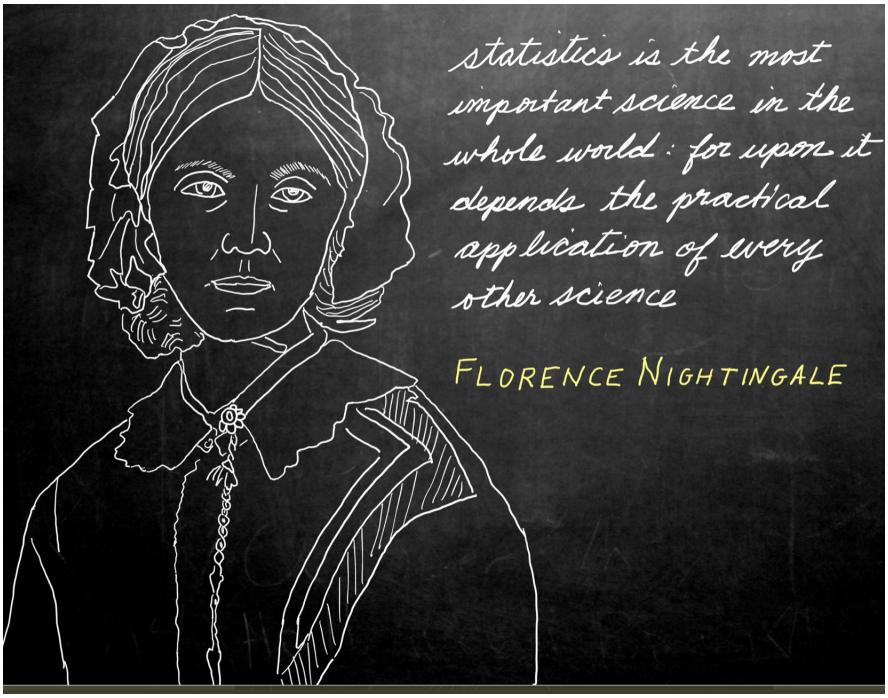
**Instructions:**

- You may use your one page of notes, R, StatKey, and a calculator. You may NOT a book or the rest of the internet.
- Show your work. Organize your work in a reasonably neat and coherent way.
- If you used R, StatKey, or a calculator, write down what you typed!
- You are not permitted to talk to anybody who has not yet taken the test about the test.



1. Short Answer ($a$, $b$, and $c$ are not related to each other)
   (a) You want to test if the average height of Portland trees is taller than the average height of Seattle trees.
      i. State the null-hypothesis in words AND with symbols [4]

      ii. State the alternative-hypothesis in words AND with symbols [4]

      iii. You gather data on 100 trees from Portland and 80 trees from Seattle. What test statistic (or method) would you use? [3]

   (b) You want to test to see if the average height of trees is the same in Portland, Seattle, and Eugene. What test statistic would you use? [3]

   (c) It is estimated that 8% of men (and 0.5% of women) have some degree of 'color vision deficiency' or CVD (colloquially called being color blind). Your friend invents a new type of surgery that they claim can cure CVD. Unfortunately, like all surgeries there are potential side-effects. In this case, a very low chance of major and permanent vision damage. You are hired to test whether this new surgery is more effective than standard treatments. You gather data on 100 individuals, 50 of whom are given the surgery and the other 50 are given a fake surgery (placebo). [10]
      i. What type of study do you perform (the study, not the statistical technique)?

      ii. Describe what a Type I error would be in this case

      iii. Describe what a Type II error would be in this case

      iv. Which type of error do you consider to be worse? Explain. (there are different possible interpretations here)

      v. You calculate a p-value of .18, what can you conclude?

2. Use StatKey and/or R for these questions. Each letter is unrelated

   (a) You want to test if the same proportion of United and Delta flights are late. You randomly sample 1000 flights from each airline and find that for United 114/1000 were late and for Delta it was 44/1000. (In StatKey this is labelled as '*Late Arrivals -3e*')

      i. Compute the p-value for this statistical test. [3]

      ii. Use a normal approximation. Find $Z$. Compute the p-value. [3]

      iii. What can we conclude (in context)? [2]

   (b) You want to test if Systolic Blood Pressure has a negative correlation with Heart Rate. You randomly sample 200 patients in the intensive care unit. You then perform linear regression and find the sample correlation of your regression line is $r = -.057$ with a standard error of 0.071. (In StatKey this is labelled as '*ICU Admissions*')

      i. Compute the p-value for this statistical test. [3]

      ii. What can we conclude (in context)? [3]

      iii. Does this method of data collection allow us to make any conclusions about the relationship between blood pressure and heart rate in the general population? [2]

3. At his previous college, your professor gave grades with the following percentage of students getting each grade:

| Grade | A | B | C | D | F |
|---|---|---|---|---|---|
| % of students | 28% | 36% | 20% | 8% | 8% |

You suspect that your professor will have a different grade distribution here at UP.

(a) State the null hypothesis [3]

(b) You gather data from the grades he gave out last semester and see that he gave out 75 grades. What test statistic should you use? [1]

(c) Calculate the p-value (may be helpful to use R or StatKey) given that he gave the following grades last semester: [5]

| Grade | A | B | C | D | F |
|---|---|---|---|---|---|
| Number | 15 | 22 | 23 | 10 | 5 |

(d) What can you conclude in context? [3]

(e) Calculate the contribution from the "C" grade to the test statistic [3]

4. You are interested in determining if different companies have different amounts of sugar in their cereals. Consider the following ANOVA table generated using R. The label 'sugar' measures the amount of sugar in each cereal and the label 'company' refers to the company that makes that particular brand of cereal (Kellogs, Quakers, General Mills)

```
> summary(aov(sugar~company))
            Df Sum Sq Mean Sq F value
company      2    1.7   0.825   0.027
Residuals   27  822.8  30.473
```

(a) State the null hypothesis and alternative hypothesis in english [4]

(b) Notice that there is usually one more column in ANOVA tables. Use the table and a computational tool (R/StatKey) to compute the p-value. [3]

(c) What can we conclude (in context)? [3]

(d) Notice that the Sum of Squares and Mean Squared values are much higher for 'Residuals' than they are for 'company'. Interpret what these mean in context. [3]

5. More details on the cereal data can be found in the table below which shows the number of cereals randomly chosen from each of the three companies, the mean amount of sugar, and standard deviation.

| Company | $n$ | mean | standard deviation |
|---|---|---|---|
| General Mills | 13 | 10.43 | 5.46 |
| Quaker | 6 | 10.83 | 4.92 |
| Kellogs | 11 | 10.18 | 5.86 |

(a) Using just the Kellogs data, find a 90% confidence interval for the mean amount of sugar in Kellogs cereal. [5]

(b) Using this table and the ANOVA table from the previous problem, find a narrower 90% confidence interval for the mean amount of sugar in Kellogs cereal. [5]

(c) Compare these two intervals. Explain why part (b) is narrower than part (a) [2]

6. We want to perform linear regression to predict 'GDP' from the 'BirthRate', 'Density', and 'Population'. We download the 'All Countries' data set from StatKey and type the following command into R, and we get the following output. Note that GDP is measure in $ per person and Birth Rate is measured in Births per 1000 people.

```
> allc=lm(AllCountries$GDP~AllCountries$BirthRate+AllCountries$Density+AllCountries$Population)
> summary(allc)

Call:
lm(formula = AllCountries$GDP ~ AllCountries$BirthRate + AllCountries$Density +
    AllCountries$Population)

Residuals:
   Min     1Q Median     3Q    Max
-22907  -9670  -3350   3767  88457

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             36270.6435  2877.2184  12.606  < 2e-16 ***
AllCountries$BirthRate  -1074.2143   121.7845  -8.821 1.16e-15 ***
AllCountries$Density        3.1600     0.7083   4.461 1.46e-05 ***
AllCountries$Population    -9.2787     8.1753  -1.135    0.258
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16290 on 174 degrees of freedom
  (39 observations deleted due to missingness)
Multiple R-squared:  0.3894,    Adjusted R-squared:  0.3789
F-statistic: 36.99 on 3 and 174 DF,  p-value: < 2.2e-16
```

(a) Write the linear equation [3]

(b) What percentage of the variation in 'GDP' scores is explained by our linear model (round to 4 digits)? [2]

(c) Is the model effective? Why or why not? [3]

(d) If the Birth Rate increased by 1 unit (one birth per 1000 people), and everything else remained constant, how much would we expect GDP to increase? [3]

7. Continuing with the 'All Countries' multiple regression from the previous problem

   (a) Population was measured in millions of people. Consider the population row on the previous page. How would each of the 4 numbers in that row change if we instead measure the population in thousands of people? Be specific. If something didn't change, write 'unchanged'. [6]

      i. How would the 'Estimate' change?

      ii. How would the 't-value' change?

      iii. How would the '$\Pr(>|t|)$' change?

   (b) If we wanted to refine our linear model and get rid of one predictor, which predictor would we remove? Why?

   We remove that predictor and get the results:

```
Residual standard error: 16310 on 175 degrees of freedom
  (39 observations deleted due to missingness)
Multiple R-squared:  0.3849,    Adjusted R-squared:  0.3779
F-statistic: 54.75 on 2 and 175 DF,  p-value: < 2.2e-16
```

   (c) Your friend says removing that predictor was good and that this new model is better than the original. Give one reason (based on the above table) that supports your friend's statement.

   (d) A different friend says removing that predictor was bad and that the original model was better. Give one reason (based on the above table) that supports this friend's statement.

8. Easy/Fun Points [8]

(a) What was your favorite statistical test/technique? Why?

(b) We didn't get to the section on probability (Appendix P) or Bayes Theorem. Would you have preferred to skip some of the material we covered so that we could have gotten to that section?

(c) Give one (or more) example of you used statistics or probability outside of this class.

(d) Draw a picture of whatever you want