Design of Experiments: Lab 1

Exercise 1. [Adapted from Lloyd (1999)]. A surgery for a condition is performed in two hospitals: hospital A and hospital B. We are interested in comparing the survival rates of the hospitals. The code below will read in the data we will use for this exercise.

```
cond = matrix(c("A", "critical", "survived",
                "A", "critical", "died",
                "A", "noncritical", "survived",
                "A", "noncritical", "died",
                "B", "critical", "survived",
                "B", "critical", "died",
                "B", "noncritical", "survived",
                "B", "noncritical", "died"), ncol = 3, byrow = T)
colnames(cond) = c("Hospital", "Status", "Outcome")
df = as.data.frame(cond)
df$Total = c(17, 101, 100, 3, 2, 36, 175, 8)
```

a) Compare the (marginal) survival rates of hospitals A and B, ignoring the status of the patients. Comment on what you see.

b) Now, find the mortality rates by hospital for critical and noncritical patients, separately. Compare the results to what you found in part 1. Why is this happening?

Exercise 2. The goal of this exercise is reviewing the central limit theorem, which states that, under mild conditions, sample averages are approximately normal.

a) Draw a sample of size 10000 from an Exponential(1) distribution and create a histogram. Does it look approximately normal?

b) Draw 1000 samples of size 100 from Exponential(1). For each of the 1000 samples, find the sample mean. Create a histogram of the sample means. Do they look normal?

c) [Optional] Draw 1000 samples of size 100 from Exponential(1). For each of the 1000 samples, find a confidence interval for the population mean μ assuming normality. Check whether each interval contains the expected value of the Exponential(1). Comment on your results.

Exercise 3. In previous courses, you saw the two-sample t-test for comparing the population means of two groups. In class, we have seen a test (based on the F distribution) that lets us compare the populations means of k groups, where $k \geq 2$. The goal of this exercise is checking that when $k = 2$, the tests are equivalent. We will use `data(hsb2)` from `library(openintro)`. If you don't have `library(openintro)` installed, you can install it with the command `install.packages("openintro")`.

Social scientists are interested in checking whether average `math` scores depend on whether go to public or private schools.

a) Assume that the variances of the scores in public and private schools are the same. Find a 90% confidence interval for the difference in `math` averages between public and private schools.

b) Use the `t.test` function to test at the significance level $\alpha = 0.05$ whether average `math` scores are different in public and private schools.

c) Do the same with `aov`. Compare the p-values. Are they the same? Take the square of the observed T statistic you found in part a) and compare it to the F statistic you find with `aov`.

Exercise 4. The dataset `salary.csv` contains salaries (in USD), anxiety levels on a scale that goes from 0 (no anxiety) to 7 (very anxious), and education level. You can read in the dataset with the command:

```
salary = read.csv("http://vicpena.github.io/sta9750/salary.csv")
```

a) Create a plot that only shows the relationship between anxiety and salary. Comment on what you see.

b) Now create a plot that displays salary, anxiety, and education. Comment on what you see now. plot (using faceting or color-coding as needed).

c) An article claims that higher salaries come at the cost of higher anxiety levels. Do your findings agree with this claim? Explain why or why not keeping the technical considerations to the minimum.

Exercise 5. In this exercise, we will use `data(hsb2)` from `library(openintro)`.

a) Create a variable called `average` that has the final average scores the students got combining their results in `read`, `write`, `math`, `science`, and `socst`.

b) Find a 95% confidence interval for `average`.

c) Is there evidence at the $\alpha = 0.05$ significance level that the `average` scores are greater than 50 points at the population level?

Exercise 6. Let's keep on using `data(hsb2)`.

a) Find a 99% confidence interval for the difference in averages in `math` scores between `male` and `female` students.

b) Is there evidence at the $\alpha = 0.01$ significance level that `math` scores differ between `male` and `female` students?

Exercise 7. [From Montgomery (1986).] During cooking, doughnuts absorb fat in various amounts. A scientist wished to learn if the amount absorbed depends on the type of fat used. For each of four fats, six batches of doughnuts were prepared. The data in the table below are the grams of fat absorbed per batch, coded by deducting 100g to give simpler figures.

T1	T2	T3	T4
64	78	75	55
72	91	93	66
68	97	78	49
77	82	71	64
56	85	63	70
95	77	76	68

a) Read the data into `R` and plot it. By looking at your plot, do you think that there will be significant differences between the types of fat?

b) Check that the assumptions of the one-way ANOVA model are satisfied.

c) Fit a one-way ANOVA model with the sum-to-zero constraint. Find point estimates and confidence intervals for the grand mean and the treatment effects. Comment on the results.

d) Is there evidence to claim that there are differences between types of fat at the $\alpha = 0.01$ significance level?

e) Run `TukeyHSD` to perform pairwise comparisons and comment on the results.