Homework policies reminder: Late homework is not accepted, with one exception: you may hand in an assignment late (by the beginning of the next class meeting) *once* during the semester. If you are planning to hand in a homework late, email me by the beginning of class on the original due date.

Please ***type*** your assignments as much as possible, except for equations, symbols, and graphs, which may be handwritten. Please integrate graphics into the flow of your text; do not put all your graphics and tables at the end of your paper.

I encourage you to discuss the homework with other students and with me. Your final answer should be written in your own words.

1. Install R (0 points)

We will be using the software R. R is a free program widely used by statisticians (and many others) and runs on MacOS, Windows, and Linux. You can download and install R from this web page: https://cran.r-project.org. Use the link for your platform from "Download and Install R" (not "Source code for all Platforms", unless you want to compile R yourself, which I don't recommend).

You may also want to install RStudio, which is a free front-end interface to R that some people find more user-friendly. You can download RStudio here: https://www.rstudio.com/products/rstudio/download/ (be sure to install R first). Use the link for your platform from "Installers for Supported Platforms".

2. R Tutorial (0 points)

Read the handout, "A Brief Introduction to R", posted on Moodle. This is the handout I used in Stat 11 last year. I was intending to update the handout for Stat 21 this semester, but my TeX software (the typesetting software used to create this handout) seems to have stopped working after I upgraded my operating system. However, most of the handout should still be relevant for us this semester. The main difference is that we used the web version of RStudio in Stat 11, whereas we are using the desktop version of R or RStudio in Stat 21. So ignore the sections of the handout that pertain to the web version of RStudio.

You should work through the entire document by actually typing in (or copying and pasting) the commands and running them in R to see what happens, rather than just reading the text.

3. Blood Pressure (12 points)

Many years ago, I worked on a study on women with polycystic ovarian syndrome (PCOS), an endocrine disease related to diabetes (Legro, Bentley-Lewis, Driscoll, Wang, and Dunaif (2002): *J. Clinical Endocrinology and Metabolism* 87:5). As part of the study, 371 women with PCOS were selected. Do women with this condition tend to differ in blood pressure from the general population?

The 371 women had a mean systolic blood pressure of 121.07, with an SD of 16.59. Let μ denote the mean blood pressure of the population from which these 371 subjects were selected. (You may assume that these subjects can be considered to be representative of some larger population.)

(a) A "normal" systolic blood pressure is considered to be 120. Carry out a hypothesis test of the following hypotheses using an alpha level of .05: $H_0: \mu = 120$ vs. $H_a: \mu \neq 120$

Be sure to label the test statistic and the p-value. Do these control subjects significantly differ from a "normal" population in systolic blood pressure?

(b) If the sample size had been $n = 3710$ instead of 371, how would your conclusions change? How does this illustrate the concept of *statistical significance* as distinct from *effect size*— that is, the size of the effect in a medical setting?
(c) Calculate a 95% confidence interval for μ. Does this interval contain 120? Is your confidence interval consistent with your conclusion in part (a)?

4. Normal Probability Plots (10 points)

(a) Use R to generate a random sample having 10 observations from a standard Normal distribution. Then make a histogram of the dataset. You can use the following commands:

```
data <- rnorm(10, 0,1)    # generate sample from Normal population with n = 10, mean = 0, SD = 1
hist(data)                # make histogram
```

Repeat this a total of twelve times. To plot all twelve histograms on one page, give this command before creating the histograms:

```
par(mfrow=c(4,3))         # plot 12 plots per page in a 4-by-3 array in row order
```

Do the samples appear to be normally distributed? Explain briefly. (Hand in the page with the twelve histograms.)
(b) Now repeat the above using normal probability plots (NPP). You don't need to plot the same twelve datasets; just create twelve new datasets as above, and use these command to make the NPP:

```
qqnorm(data)              # make NPP
abline(0,1, col=gray(.7))  # add gray line with intercept = 0, slope = 1
```

Do the samples appear to be normally distributed? Explain briefly. (Hand in the page with the twelve NPPs.)
(c) Repeat parts (a) and (b) using twelve samples of 100 observations each. How does the appearance of the histograms and NPPs change with the increased sample size?
(d) You can think of an entire NPP as being a statistic, since it is calculated from a sample. As such, it must have a sampling distribution, like any other statistic. Make a rough sketch (by hand) of the range in which you think 95% of NPPs would fall if sampled under these conditions, for both $n = 10$ and $n = 100$. Hand in these two pictures. (Think carefully about the shape of these ranges.)

5. Randomization Test (10 points)

Here we analyze data from the same study as in question 3. Women afflicted with PCOS often experience weight gain. Here we want to see if the PCOS subjects in our dataset have, on average, higher body mass index (BMI, a measure of weight that accounts for height) than control subjects. We will compare the two groups using a randomization test, as opposed to the traditional parametric analysis you did in question 3.

(a) Download the dataset `bmi.csv` from Moodle. Then read the dataset into R using the following command:

```
data <- read.csv("~/Desktop/bmi.csv", header=T)
```

(You may have to change the path to the file, depending on where you saved the downloaded file.) Verify that your dataset has 243 subjects with BMI in the first column and PCOS status in the second column (1 = PCOS, 99 = control). Define variables for PCOS and status as follows:

```
bmi <- data[,"bmi"]
status <- data[,"status"]
```

(b) Calculate the difference in the mean BMI for PCOS subjects and the control subjects using the following command:

```
mean(bmi[status==1]) - mean(bmi[status==99])          # difference of group means
```

(c) Now we will conduct a randomization test of the null hypotheses that PCOS subjects have the same average BMI as control subjects, against the one-sided alternative hypothesis that the PCOS subjects are higher. We will randomly permute the status labels 1000 times and observe the distribution of the test statistic (namely, the differences in the means of the two groups). First, define a vector diffs to hold the 1000 differences:

```
numshuffles <- 1000                  # number of shuffles to run
diffs <- rep(NA, numshuffles)        # create empty vector to hold shuffled differences
```

Now use a for loop to set up the 1000 randomizations:

```
for(i in 1:numshuffles)  {


}
```

Inside the for loop, you will need to randomly shuffle the status labels, calculate the difference in means using the shuffled group labels, and save the difference. To do the random shuffling, use the sample command:

```
newstatus <- sample(status)
```

The other commands are left for you to figure out.

(d) Finally, once the loop has executed, make a histogram of the 1000 differences. Add a red vertical line showing the magnitude of the difference in the real dataset; you can do this using the abline command. Calculate what proportion of the differences from the shuffled datasets exceed the difference seen in the real dataset; this is your empirical p-value. What do you conclude about the null hypothesis?

WHAT TO HAND IN: Hand in all of your R code, and your histogram showing the 1000 shuffled differences with the size of the real difference shown. (You can save R graphics windows as pdfs and then copy and paste them into a word processor document.) Also include your empirical p-value and your conclusion.

6. NEW YORK CITY MARATHON (5 POINTS)

The dataset nycmarathon.csv has finishing times of 3000 New York City marathon runners. Download the dataset from Moodle and read it into R using the following command:

```
marathon <- scan("~/Desktop/nycmarathon.csv")
```

Make a histogram using the following command:

```
hist(marathon, col="darkgray", border="white", nclass=20)
```

It may be useful to change the nclass option to get more or fewer bars. Describe any interesting features of the dataset you discover.