Introduction to Statistical Inference

STAT/BIOS 823

Homework 7

Directions

Using `RMarkdown` in `RStudio`, complete the following questions. Launch RStudio and open a new RMarkdown file or use the class RMarkdown template provided and save it on your working directory as a `.Rmd` file. At the end of the activity, save your **pdf** generated from `RMarkdown+Knitr` and submit your homework on the Blackboard.

If you have questions, please post them on the lesson discussion board.

All questions are mandatory except the `One-Way ANOVA question. This is not required`. Partial `R-code` and **output** from the code has been provided for you in some of the question.

`R code` and output must be clearly shown.

Homework submitted after the due date will attract a penalty of 10 points per day after the due date.

---

1 Analyzing Data with a Categorical Outcome

Binge drinking is defined by the National Institute on Alcohol Abuse and Alcoholism (NIAAA) as a pattern of drinking that brings blood alcohol concentration (BAC) levels to 0.08 g/dL. This typically occurs after 4 drinks for women and 5 drinks for men in about 2 hours. The Substance Abuse and Mental Health Services Administration (SAMHSA) conducts an annual National Survey on Drug Use and Health (NSDUH) and defines binge drinking as drinking 5 or more alcoholic drinks on the same occasion on at least 1 day in the past 30 days. In 2014, 24.7 percent of people ages 18 or older reported that they engaged in binge drinking in the past month. To assess the level of binge drinking on campus, University A conducted a survey of alcohol use of undergraduate students. A random sample of 1300 undergraduate students were classified as to whether they reported engaging in binge drinking in the last month:

Table 1: Frequency data.

Binge Drink	Count
Yes	630
No	670
Total	1300

1.1 Enter the data into R and re-produce the barchart (Figure 1).

Table 1 shows a summary of the data (it's frequency table).

(a) Enter the data into R in an expanded form (long form) such that you will have a dataset with 1300 rows, of which, 630 will represent the Yes category and 670 represent the No category. **Hint:** Use existing R functions such as `rep()`, `factor()`, `data.frame()` or write your own function.

(b) Explore the data and produce a frequency table, by, for example typing `tab1(binge)` or `summ()` using the `epiDisplay` package. Reproduce the bar-chart below. Hint: `tab1()` will produces a bar chart for you with frequencies on the Y axis default. You can have percentages instead of frequencies by typing `tab1(x, bar.values ="percent"`.

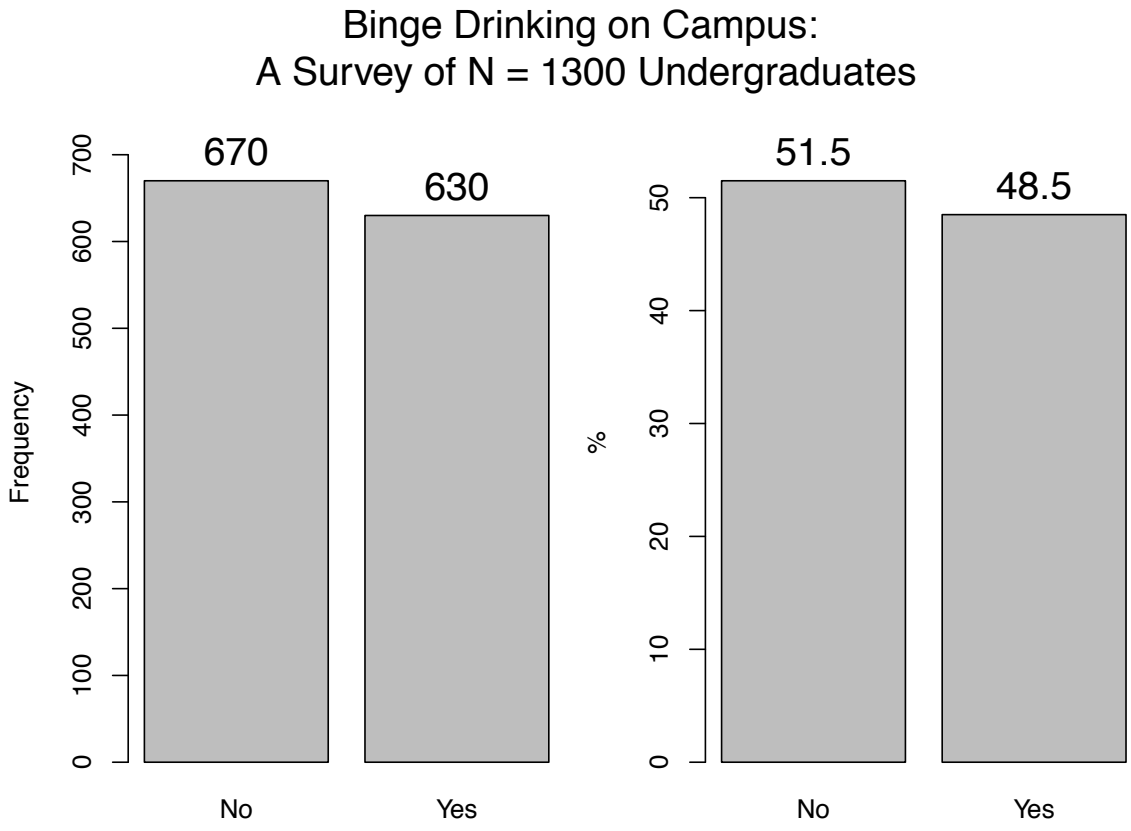


Figure 1: Barplots for drinking Binge

1.2 Chi-square Goodness of Fit Test

(a) To test if the proportion of Yes's is significantly different from 0.5, we can type the following code: **prop.test(x=630, n=1300)** or **binom.test()**. *What is your conclusion based on the output?*

(b) Perform a chi-square goodness of fit test to investigate the hypothesis that:

(i) there is no difference in proportion between the students who binge drink and those that do not. That is, test that $H_0 : p_1 = p_2$ versus $H_1 : p_1 \neq p_2$. *What is your conclusion?*

(ii) the proportion of undergraduates on campus who binge drink $\pi = P(Y = yes)$ is different than the average reported by NSDUH: $H_1 : \pi \neq 0.247$. The chi-square goodness of fit is testing whether the sample in hand appears to have been drawn from a population where the true rate of binge drinking is 0.247, or 24.7%. Note: we can verify the assumption that is based on sample size: $\pi \times n > 5$, that is, $0.247 \times 1300 = 321.1$ and $(1 - \pi) \times n > 5$, that is $(1 - 0.247) \times 1300 = 978.9$. *What is your conclusion?*

(c) Since the administration of University A would only be concerned if the rate of binge drinking is high, we should really be focusing on evidence supporting rates that are higher than 24.7%. That corresponds to investigating the hypothesis $H_1 : \pi > 0 : 247$. The **binomial** test is an exact test for one proportion that can be used when you your sample is too small to full fill the requirements of the chi-square goodness of fit test. Perform both the **Chi-square** test and the **binomial** test. *What is your conclusion?*

```
prop.test(x = 630, n = 1300, p = 0.247, alternative = "g")
binom.test(x = 630, n = 1300, p = 0.247, alternative = "g")
```

1.3 Sales Data

The CEO of Company Z is interested in learning more about sales patterns for the chain's retail outlets across the United States. Tomorrow you have to let the CEO know whether the type of gear sold stores is associated with geographic location. For each store, you have information on the store's geographic region (A, B, C, D) and its most popular type of sports gear sold (winter sports, summer sports, all-season sports; based on total sales volume) in the last three calendar years. The `sales` data is used to answer the questions that follow.

Table 2: Summary of Sales data.

Region	Winter	Sport-Summer	All-Season	Total
A	22	6	9	37
B	7	31	13	51
C	0	15	7	22
D	13	13	14	40
Total	42	65	43	150

(a) Read in and examine the sales data.

The summary table can be reproduced by typing

```
library(MASS)
(tabs <- xtabs(~Region + Sport, data = sales))
```

```
##        Sport
## Region  A  S  W
##      A  9  6 22
##      B 13 31  7
##      C  7 15  0
##      D 14 13 13
```

```
# or tabpct(Region, Sport, graph=FALSE)
```

(b) What is the distribution of most popular gear type within each region? To answer this question, use prop.table() to generate a table of proportions and use the mosaic() to generate a mosaic plot. Then describe what you find.

(c) Perform a Chi-square test of independence to investigate the hypothesis that the sales of sports gear at a retail outlet is dependent on geographic region of the retail outlet. The only assumption we need to verify is based on sample size. None of the expected cell counts may be less than 5. The expected cell count for the cell in the ith row and jth column of the table can be calculated as: $(C_i \times R_j)/N$ where R_i is the total count for the ith row, C_j is the total count for the jth column, and N is the overall sample size. Based on the output of the `chisq.test` function, we should be able to check that we have sufficient numbers to use the chi-square test. *Do we? What is your conclusion about the dependence of sales of sports gear on geographic region?*

```
# chisq.test(tabs) # Chi-square test
```

2 Analysis of Continuous Outcome Data

2.1 One-Sample Tests

Fuel economy is measured under controlled conditions in a laboratory using a series of tests specified by federal law. Manufacturers test their own vehicles-usually pre-production prototypes-and report the results to the Environmental Protection Agency (EPA). EPA reviews the results and confirms about 15% to 20% of them through their own tests at the National Vehicles and Fuel Emissions Laboratory. Car Company A submitted to the EPA an estimate of 25 mpg (city and highway combined) for the fuel economy of their 2018 Sedan. The EPA tested a random sample of 30 2018 Sedans to confirm the estimate submitted by Company A:

```
car <- c(19, 26, 24, 21, 24, 23, 26, 24, 23, 20, 21,
    24, 18, 21, 20, 23, 24, 26, 25, 19, 24, 23, 27,
    24, 26, 25, 20, 21, 19, 23)
```

(a) Perform a one-sample `t-test` to investigate the hypothesis that the average fuel economy of the 2018 Sedan is not equal to 25 mpg (i.e., $H_1 : \mu \neq 25$). *What is your conclusion? Produce a reasonable plot to visualize this single continuous outcome. Generate the mean and standard deviation of mpg.*

```
# t.test(car, mu=25)
```

(b) The assumption of the `t-test` is approximate normality of the outcome. This assumption is not necessary for large samples, but since we're dealing with 30 observations we need to check. Use the "eyeball" method with a normal Q-Q plot. A normal **QQ plot** is a scatterplot of the observed quantiles (percentiles) of the data

against its expected quantiles assuming it follows a normal distribution. If normality is a reasonable assumption, the actual quantiles and the expected quantiles should be similar and thus follow a straight line. Use the **Shapiro-Wilk** test to assess normality. If the p-value is small (e.g., less than 0.05), it's likely that the data have violated the normality assumption. Perform the **Shapiro-Wilk** test. *Does the data appear to follow the straight line?*

(c) In general, we're not concerned if our cars are getting better fuel economy than what is advertised. It makes sense in this situation then to focus our efforts on identifying whether Company A has overestimated the mpg of the 2018 Sedan (i.e., $H_1 : \mu < 25$). Perform this one-sided test. Does it appear the cars are a random sample from a population of vehicles whose average mpg is less than 25? You should word your conclusion something like this: "There is sufficient evidence to conclude that the average fuel economy of 2018 Sedans is less than the 25 mpg reported by Company A ($p = 0.03, 95\%CI : UpperLimit$)", OR "There is insufficient evidence to conclude that the average fuel economy of 2018 Sedans is less than the 25 mpg reported by Company A ($p = 0.3, 95\%CI : UpperLimit$)"

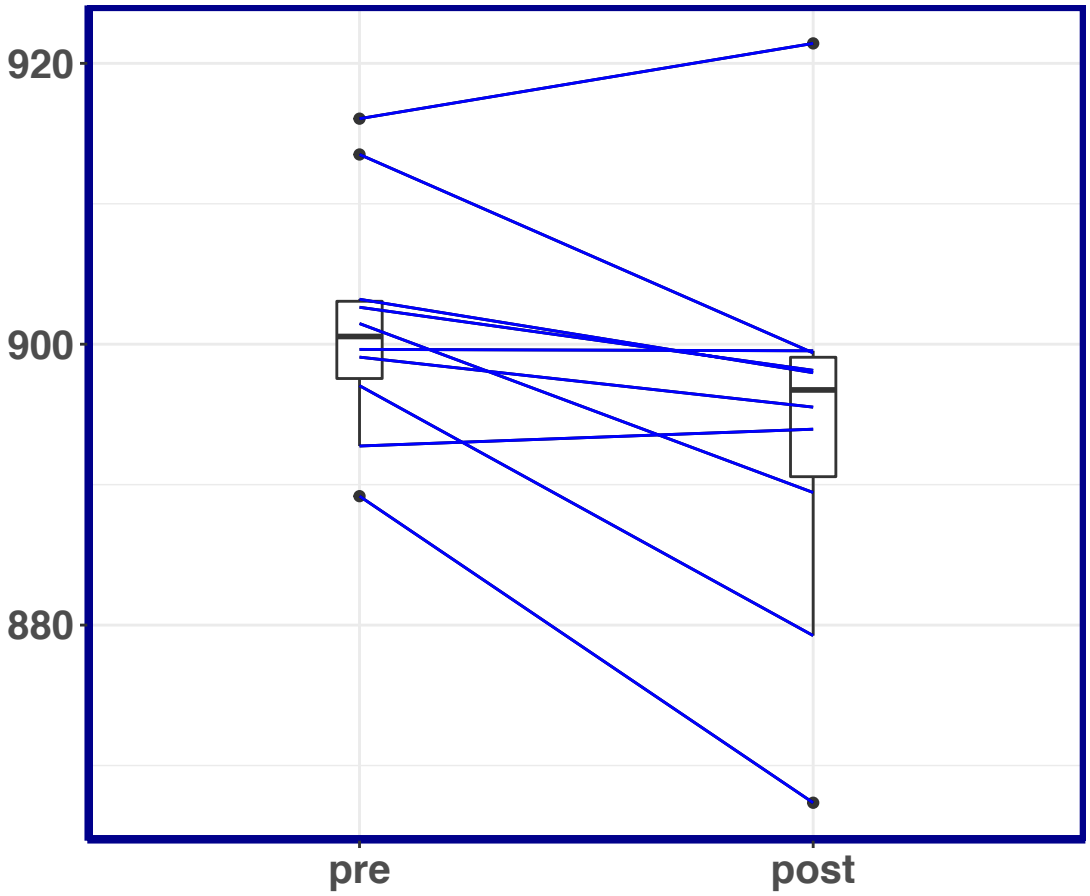
```
# t.test(car, alternative='less', mu=25)
```

2.2 Dependent Samples Tests

The coach for the USA Swim Team is trying to identify training programs to improve the performance of athletes preparing for the 2016 Summer Olympics. A potential training program has been identified but must be tested on the athletes. To assess its effectiveness, the coach will compare the times (in seconds) of 10 swimmers in the Men's 1500 free-style prior to the training and their times after training:

```
id <- c(1:10)
pre <- c(899.63, 913.51, 897.05, 889.18, 903.2, 916.06,
    899.08, 892.75, 901.47, 902.63)
post <- c(899.53, 899.38, 879.25, 867.35, 897.97, 921.42,
    895.52, 893.95, 889.44, 898.14)
datap <- data.frame(cbind(id, pre, post))
```

(a) Install and load the `PairedData` package and produce a profile plot showing the paired differences for each of the 10 swimmers.**Hint:** The first part of the code can be **paired.plotProfiles(datap, "pre", "post") + geom_line(color="blue")**



(b) Recall that the coach wants to investigate the hypothesis that the training program is effective in reducing swim times for male athletes in the 1500 freestyle; that is, whether or not swimmer's times decrease under the new program (i.e., $H_1 : \mu_{pre} - \mu_{post} > 0$).

(i) Generate the means and standard deviations of swim time pre- and post-training and produce a boxplot showing the pre-and post-training distributions.

(ii) Compute the paired differences and generate the mean and standard deviation of the differences

(iii) The assumption of the paired t-test is normality of the paired differences. Use a normal **Q-Q plot** and the **Shapiro-Wilk** test to verify this assumption.

(iv) Perform the `t-test` analysis for this paired data. What is your conclusion? Make sure you word it similar to: "There is sufficient evidence to conclude that the training program is effective at reducing swim times for Men's 1500 Freestyle ($p = 0.03$). The program, on average, decreased swim time by 7 seconds (95% CI on difference: $\mu_{pre} - \mu_{post} > 0$)."

```
# Starting code
datap$diff <- pre - post
t.test(pre, post, paired = TRUE, alternative = "greater")
```

2.3 Wilcoxon Signed-Rank Test

The Wilcoxon-Signed Rank test is a non-parametric test for comparing two dependent samples to assess whether their mean ranks differ. It's a less powerful alternative to the paired t-test that should be substituted when the normality assumption cannot be verified or met. Perform the `paired` one-sided test using **wilcox.test()** function, with `alternative = "greater"` option. Does your conclusion differ from the paired t-test above?

2.4 Optional Question: One-Way ANOVA

Six different insect sprays are in development to help combat infestation of crops. Each of the 6 sprays were applied on 12 different fields and the number of insects found dead in the field was recorded. Researchers are interested in finding any significant differences in effectiveness across the six sprays. Load the data:

```
data(InsectSprays)
```

Recall that researchers want to investigate whether **any difference** exists in the six insecticides under study. This hypothesis that we must nullify is called the **omnibus null**: $H_0 : \mu_1 = \mu_2 = \cdots \mu_6$ The alternative to this is that **at least one of the insecticides is different**: $H_1 : \mu_i \neq \mu_j$ for some $i \neq j$). That is, if we find a statistically significant result with ANOVA, it is interpreted as **at least one of the means is different**. It alone can't tell us how many or which are different.

(a) Perform the one-way ANOVA to investigate the omnibus hypothesis. Graphically visualize the data. What's the best graphic to visualize a continuous outcome across groups?

(b) Generate the mean and standard deviation of insect totals for each group using a function like **tapply()**, **ddply()** or **aggregate.numeric()**.

(c) The assumptions of ANOVA are the same as the two-sample t -test: (1) normality of the outcome within each group and (2) equal group variances. To Check (1), assess the residuals for normality after fitting the ANOVA model. To check (2), use **leveneTest()** from the **car** package. Is ANOVA appropriate for this data? ANOVA is robust to violations of normality (especially for large sample sizes) but a violation of equal variances can severely impact our ability to make accurate inferences. The **leveneTest** is a formal test investigating the hypothesis that at least one of the group variances are unequal (i.e., $\sigma_i^2 \neq \sigma_j^2$ for $i \neq j$). If you observe $p < 0.05$, you have sufficient evidence to conclude that at least one of the variances are different and the assumption is violated:

```
library(car)
data(InsectSprays)
leveneTest(InsectSprays$count ~ InsectSprays$spray)
```

(d) Use log transformation to transform the response variable `count`. Does it get better?

```
attach(InsectSprays)
countlog <- log(count + 1)
leveneTest(countlog ~ spray)
```

(e) Fit the ANOVA with the log-transformed variable. Check for normality of the residuals. Do the residuals appear normally distributed?

(f) Since we have a significant difference somewhere, it is important that we both identify it and describe it. Post-hoc comparisons can be performed using many methods, but an easy one to start with is Tukey's Highly Significant Differences (HSD) Test. It performs pairwise comparisons of all groups to find where any statistically significant differences exist between groups. The output includes a table of all pairwise comparisons (e.g., 'B-A'), an estimate of the difference between the groups, and a confidence interval on the difference. Which groups appear to be different?

```
TukeyHSD(fit)
```

(g) Perform a non-parametric Kruskal-Wallis Test. The Kruskal-Wallis test is analogous to the Mann-Whitney test for two groups. It is a robust (non-parametric) alternative to the one-way ANOVA that eliminates the need for normality and equal variances, but it also provides a less powerful test of differences than ANOVA. Do the results of this test agree with the ANOVA from above?