MA 354: Data Analysis I – Fall 2019
Exam 2:

Instructions:

- You have 45 minutes to complete the conceptual part of this exam.
- The data analysis is take home and due 12/06 by 11:59p.
- Take a deep breath. You're going to do well and the worst case is that it will be productive.

R/LaTeX Sweave notes – this should be all that you need.

- To run R and print the output.

```
<<>>=
        #Rcode goes here
        #Output is automatically printed in the .pdf
@
```

  Remark: All R chunks must have no spaces preceding the <<>>= or @ syntax.

- Provide R code for plot and place the plot into our document.

```
<<plotName,eval=FALSE>>=
        #Rcode for plot
        #We will call this later so make sure it has a unique name
@
\begin{figure}[H]
        \centering
        <<fig=TRUE,echo=FALSE>>=
        library("graphics")
        <<plotName>>
        @
        \caption{Some information about our plot} \label{Fig:plot1}
\end{figure}
```

  You can then reference a graph in latex using `\ref{Fig:plot1}`.
  Remark: All R chunks must have no spaces preceding the <<>>= or @ syntax.

- If you wanted a one line equation that is centered like this,

$$\widehat{y_i} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon$$

  you can use this LaTeX.

```
\[\widehat{y_i} = \beta_0 + \beta_1 x_{1i}+ \beta_2 x_{2i} + \epsilon\]
```

- If you wanted a multiple line equation that is centered like this,

$$\begin{aligned} f_X(x) &= 90x^8(1-x) \\ &= 90x^8 - 90x^9 \end{aligned}$$

  you can use this LaTeX.

```
\begin{align*}
        f_X(x) &= 90 x^8(1-x)\\
                 &= 90x^8 - 90x^9\\
\end{align*}
```

Help: You can ask for information about any of the following functions that we've used by asking `R`. For example, if I wanted help with the lm() function I would run ?lm() in the `R` console. Note that if you're asking a question about a function, its library must be loaded.

- Stock R functions
  - which()
  - subset()
  - summary()
  - names()
  - cumsum()
  - apply()
  - lapply()
  - sapply()
  - tapply()
  - table()
  - prop.table()
  - pie()
  - barplot()
  - hist()
  - density()
  - boxplot()
  - lines()
  - points()
  - jitter()
  - legend()
  - optim()
  - prop.test()
  - t.test()
  - var.test()
  - aov()
  - lm()
  - anova()
  - tukeyHSD()
  - p.adjust()
  - fisher.test()
  - chisq.test()
  - cor()
  - cor.test()
- stringr Package
  - str_split()
- extraDistr Package
  - dmnom()
- nleqslv Package
  - nleqslv()
- ggplot2 Package Plotting
  - ggplot()
  - geom_bar()
  - coord_polar()
  - geom_hline()
  - geom_text()
  - geom_histogram()
  - geom_density()
  - geom_freqpoly()
  - geom_boxplot()
  - geom_jitter()
  - geom_violin()
  - geom_point()
  - geom_line()
  - facet_grid()
  - coord_flip()
  - theme_bw()
  - xlab()
  - ylab()
  - ggtitle()
- Probability Distribution
  - dbinom()
  - dhyper()
  - dnbinom()
  - dpois()
  - dunif()
  - dnorm()
  - dlnorm()
  - dchisq()
  - dt()
  - df()
- gridExtra Package
  - grid.arrange()
- qqplotr Package
  - stat_qq_band()
  - stat_qq_line()
  - stat_qq_point()
- boot Package
  - boot()
  - boot.ci()
- BSDA Package
  - SIGN.test()
- simpleboot Package
  - two.boot()
- RVAideMemoire Package
  - mood.medtest()
  - cramer.test()
- rcompanion Package
  - pairwiseMedianTest()
  - cldList()
  - phi()
  - cramerV()
- multcomp Package
  - glht()
  - cld()
- FSA Package
  - dunnTest()
- DescTools Package
  - StuartTauC()

- Bernoulli Distribution

$$f_X(x|p) = p^x(1-p)^{1-x} I(x \in \{0,1\}) \quad \textbf{[PMF]}$$
$$E(X) = p \quad \textbf{[Expected Value]}$$
$$var(X) = p(1-p) \quad \textbf{[Variance]}$$

- Binomial Distribution

$$f_X(x|n,p) = \binom{n}{x} p^x(1-p)^{n-x} I(x \in \{0,1,\dots n\}) \quad \textbf{[PMF]}$$
$$E(X) = np \quad \textbf{[Expected Value]}$$
$$var(X) np(1-p) \quad \textbf{[Variance]}$$

- Hypergeometric Distribution

$$f_X(x|N,n,m,k) = \frac{\binom{m}{x}\binom{n}{(k-x)}}{\binom{N}{k}} I(x \in \mathcal{X}) \quad \textbf{[PMF]}$$
$$E(X) = \frac{km}{m+n} \quad \textbf{[Expected Value]}$$
$$var(X) = \frac{km}{m+n}\ \frac{-n}{m+n}\ \frac{m+n-k}{m+n-1} \quad \textbf{[Variance]}$$

- Negative Binomial Distribution

$$f_X(x|n,p) = \binom{n+x-1}{x} p^n(1-p)^x I(x \in \{0,1,\dots\}) \quad \textbf{[PMF]}$$
$$E(X) = \frac{n(1-p)}{p} \quad \textbf{[Expected Value]}$$
$$var(X) = \frac{n(1-p)}{p^2} \quad \textbf{[Variance]}$$

- Poisson Distribution

$$f_X(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}\ I(x \in \{0,1,\dots\}) \quad \textbf{[PMF]}$$
$$E(X) = \lambda$$
$$var(X) = \lambda$$

- Uniform Distribution

$$f_X(x|a,b) = \frac{1}{b-a}\ I(x \in [a,b]) \quad \textbf{[PDF]}$$
$$E(X) = \frac{a+b}{2} \quad \textbf{[Expected Value]}$$
$$var(X) = \frac{(b-a)^2}{12} \quad \textbf{[Variance]}$$

- Gaussian Distribution

$$f_X(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}\ I(x \in \mathbb{R}) \quad \textbf{[PDF]}$$
$$E(X) = \mu \quad \textbf{[Expected Value]}$$
$$var(X) = \sigma^2 \quad \textbf{[Variance]}$$

- Log-Normal Distribution

$$f_X(x|\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{\frac{(ln(x)-\mu)^2}{2\sigma^2}}\ I(x \in (0,\infty)) \quad \textbf{[PDF]}$$
$$E(X) = e^{\mu+\sigma^2/2} \quad \textbf{[Expected Value]}$$
$$var(X) = e^{2\mu+\sigma^2} e^{\sigma^2-1} \quad \textbf{[Variance]}$$

- Chi-squared Distribution

$$f_X(x) = \frac{1}{\Gamma\left(\frac{v}{2}\right) 2^{v/2}} x^{\frac{v}{2}-1} e^{\frac{-x}{2}} \quad \textbf{[PDF]}$$
$$E(X) = v \quad \textbf{[Expected Value]}$$
$$var(X) = 2v \quad \textbf{[Variance]}$$

- Student T distribution

$$f_T(t) = \frac{\Gamma(\frac{v+1}{2})}{\sqrt{\pi}\ \Gamma(v/2)} \left(1 + \frac{t^2}{2}\right)^{-(v+1)/2} \quad \textbf{[PDF]}$$
$$E(X) = 0 \quad \textbf{[Expected Value for } v > 1\textbf{]}$$
$$var(X) = \frac{v}{v-2} \quad \textbf{[Variance for } v > 2\textbf{]}$$

- F distribution

$$f_W(w) = \frac{\Gamma(\frac{u+v}{2})}{\Gamma(\frac{u}{2})\Gamma(\frac{v}{2})} \left(\frac{u}{v}\right)^{u/2} \frac{w^{\frac{u}{2}-1}}{[1+(\frac{u}{v})w]^{(u+v)/2}}\ I(w > 0) \quad \textbf{[PDF]}$$
$$E(W) = \frac{v}{v-2} \quad (\textbf{[Expected Value for } v > 2\textbf{]})$$
$$var(W) = \left(\frac{u-2}{u}\right)\left(\frac{v}{v+2}\right) \quad (\textbf{[Variance]})$$

1 In-exam Portion:

Part I (30 points)
In Part I, I'm simply evaluating your engagement with the material. If you've worked through the material, there should be clear distinctions to make. I have provided as much room as I think is necessary to answer these questions. Take a minute to think or do some scratch work – your answer should fit in the space provided, only keep the important distinctions. I do not expect you to recite the formulas but to explain the procedures, their hypotheses, conclusions and/or their differences.

Submit your exam by emailing the following to wcipolli@colgate.edu

1. A LASTNAME_FIRSTNAME.pdf file just containing your answers (pages 5-6)

2 Out-of-exam Portion:

Part II (70 points)
In Part II, you're completing a data analysis. In this analysis you should provide numerical and graphical summaries that provide information for the researcher related to their research question.

Submit your exam by emailing the following to wcipolli@colgate.edu

1. A LASTNAME_FIRSTNAME.pdf of your final draft data analysis
2. Your .Rnw file.

Part III (Optional with likely increased score)
Shortly after the exam, you will receive an email to anonymously review two exams. You should review their data analysis for completeness, correctness, and communication. You will type up **constructive** notes to make the response better. The idea is to provide guidance for what's needed for the full data analysis to be effectively communicated to where you can understand the logic and the conclusions made about the data analysis. The format is discussed below.

- Write a paragraph about the general pros and cons of the paper you're reviewing. There is something good about every paper – find it and discuss that part. Also provide, in broad strokes a **constructive** critique of the response.
- Provide a list of major issues.
- Provide a list of minor issue.
- Provide a list of typographical errors you've found while reading.
- Ensure to provide specific line item comments where applicable.

Part VI (Optional with likely increased score)
After you receive comments about your work, revisit your analysis from the exam. Write a final draft of your analysis and provide responses to reviewer comments.

- Write a revision of your original solution which incorporates comments made in the reviews you've received.
- Provide a list of responses to specific line item comments; e.g.,
  - On page 1, line 2, you appear to interpret the statistics incorrectly.
    Response: This was actually done correctly because I was treating the predictor as categorical and not continuous. I've added a sentence to make this distinction clear when fitting the model.
  - On page 2, line 4, you're missing a period at the end of the sentence.
    Response: Thank you for pointing this out; I've added the missing period to the end of the sentence.

Part I – Use only the space provided to answer the following.

1. Succinctly explain the significance of Central Limit Theorem.
   Solution: CLT has many formulations, but it basically states that regardless of the population distribution, the sampling distribution of $\bar{X}$, for any random variable X, would be approximately Gaussian distributed with $\mu_{\bar{x}} = \mu_x$ and $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$. It is often invoked to fulfil distributional assumptions for various hypothesis tests and confidence intervals. For example, one-sample and two-sample t tests require normal population distributions as an assumption to conduct the test, and the CLT is invoked for non-normal population data to fulfil the test's distribution requirement. It also allows approximating a discrete ditribution using a continuous distribution, as we do in Z-hypothesis test to check evidence for or against a population proportion.
2. Succinctly explain why one might choose to conduct a sign test instead of a t test.
   Solution: First, it may be that the sample data is extremely skewed. So, t test as a hypothesis test to check evidence for or against a pre-specified value of μ in H_0 may not be the best for making statistical inference, because checking evidence for or against a pre-specified value of median in H_0 using sign test would be better for skewed data. Secondly, if our assumptions for t test are not met. Even though the t-test is robust to some normality departures, having grossly non-normal data may warrant us to use sign test to make inference about the population, as long as the observations are independent and the sample is representative of the population.
3. Succinctly describe the difference between a population distribution, a sampling distribution.
   Solution: A population distribution is pre-defined for a given set of data and has $\mu_x = E(X)$ and $\sigma_x = var(X)$ and may be unknown. The sampling distribution, is the distribution of any sample statistic if we take repeated random samples from the population or use resampling with a sample that is reasonably large and is representative of the population. Note that the sampling distribution may be different from the population distribution, example CLT may be invoked to show that the distribution of $\bar{X}$ is Approximately Gaussian regardless of population distribution. The population distribution is characterized by population parameters that we don't always know, and we use sampling distributions to use observed data and construct C.I. and conduct hypothesis test to estimate these population parameters.
4. Succinctly describe what 95% confidence means with respect to a construced confidence interval.
   Solution: It means that theoretically, if we were to take repeated random samples from the population and construct 95% confidence intervals, 95% of those confidence intervals are believed to contain the true value of the population parameter. Note that a 95% confidence interval allows us to have an interval of values to estimate the true value of a population parameter using the given sample data, which allows taking into account variability of observed data which is different than a point estimate which doesn't consider variability. The 95% confidence interval, then, simply refers to the **long run** behavior of construction of confidence intervals using repeated random sampling from the population.
5. Succinctly describe what a 0.05 significance level ($\alpha = 0.05$) means with respect to a hypothesis test.
   Solution: In a hypothesis test, we begin by assuming a sharp null hypothesis, meaning we assume 1 value for any population parameter, and also H_a related to the population parameter which may be one-sided or two sided. So, α really tells us how unusual our observed data has to be, considering H_0 is true, for us to have evidence to reject H_0 and gain evidence in support of H_a. So, we calculate a test statistic value using the observed data, after checking our assumptions for the test, and find the probability of observed data assuming H_0 is true which is p-value. If p-value$<\alpha = 0.05$, we say that the probability of observed data or more exterme with respect to H_a is less than $\alpha = 0.05$, so we have evidence to reject H_0.
6. Succinctly describe why post-hoc testing is necessary for ANOVA, Kruskal Wallis, or Mood's median test.
   Solution: The reason is all of them are ominbus test, or they merely check evidence for against H_0 that the value of a particular population parameter is equal across all treatment groups. This parameter is the mean (μ) for ANOVA, median(M) for Mood's Median test and the mean population rank (μ^R) for Kruskal Wallis test. If we reject H_0 in either of the omnibus tests, we simply know that at least one of the population parameters in not equal to the value assumed under H_0, but we don't know parameters for which specific treatment groups are different and in what direction (greater or less). So, we use TukeyHSD, median pairwise intervals and Dunn's test for μ, M and μ^R respectively, to construct $100(1-\alpha)\%$ C.I. for each interval and check which treatment groups have statistically different value of parameters and in what direction.

Longer (but still succinct) Answer

7. The analysis of Bracht et al. (2016) aimed to consider the ability of MFAP4 (a continuous variable) to differentiate between stages of the disease (ordinal) – fibrosis stages (0-2) and cirrhosis (3-4) based on the Scheuer scoring system. What analysis should they use? Ensure to include any questions that need to be answered to make the correct decision.
   Solution: Note that this is an observational study since none of the values for the variables in the dataset are influenced by the experimental design. We simply observe the value of MFAP4 with the stages of the disease in the dataset. Based on the research question, it is clear that we have the different **ordinal** stages of the disease as 5 distinct treatment groups. Firstly, we need to be sure that observations in the samples are independent, and that the sampling procedure and experimental design is such that it doesn't introduce bias. If these assumptions are not met, we should not proceed with any analysis since our conclusions would be biased. Our first instinct, then, may be to check if μ_x, where x represents the MFAP4 levels in a sample in any treatment group, is equal across the five distinct treatment groups. This involves the use of ANOVA and Tukey's HSD as a post-hoc test subsequently if H_0 is rejected in ANOVA. However, we need to ask if the population variances (σ^2) are equal for all four treatment groups which is a critical assumption. We also need to ask if the populations for treatment groups are normally distributed, but ANOVA is robust to normality departures. Using Levene's test, we may check evidence against the population variances being equal to each other. If we find evidence against population variances being equal to each other, we may instead rely on Mood's Median Test or Kruskal Wallis test to check if M_x or μ_x^R are all equal to each other, and then median pairwise intervals and Dunn's test as post hoc testing respectively if we reject the H_0 under ominbus tests to see which treatment groups have different parameter values, and the nature of such difference.

References

Bracht, T., Molleken, C., Ahrens, M., Poschmann, G., Schlosser, A., Eisenacher, M., Stuhler, K., Meyer, H. E., Schmiegel, W. H., Holmskov, U., Sorensen, G. L., and Sitek, B. (2016). Evaluation of the biomarker candidate mfap4 for non-invasive assessment of hepatic fibrosis in hepatitis C patients. *Journal of Translational Medicine*, 14.