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

1. Succinctly explain the significance of Central Limit Theorem.
   Solution:
   The Central Limit Theorem allows us to estimate population statistics using a sample distribution. The central limit theorem operates off of the assumption that as sample size increases, the sample distribution becomes more gaussian regardless of the underlying distribution.The second CLT states that a nonnormal distribution with a known population mean and variance will have gaussian distributed sample statistics. This is important because sample statistics can be gaussian distributed as long as the sample is representative.

2. Succinctly explain why one might choose to conduct a sign test instead of a t test.
   Solution:
   One may use a sign test instead of a t test if the underlying data is skewed. When the data is severely skewed, we may choose to compare medians of the data instead of means, since means are averages and medians are the middle point and the medians may help us better identify the long-run average we are interested in.

3. Succinctly describe the difference between a population distribution, a sampling distribution.
   Solution:
   A population distribution is a distribution that has the true population statistics, or the information we are actually interested in. This distribution is what we want to be able to model with the limited resources we may have. We use a sampling distribution, or a representative (randomly selected) subset of the data, in order to estimate the population distribution. It is important to note that these calculations are based off of the idea that as the size of observations in our sampling distributions increases, the expected value and variance of our sampling distribution approach the true population value and variance of the population distribution.

4. Succinctly describe what 95% confidence means with respect to a construced confidence interval.
   Solution:
   95% confidence with respect to a constructed confidence interval reflects the idea that if one was to sample and create a confidence interval for a certain population R amount of times, the true population value would lie in (.95)R cases, or 95% of all confidence intervals.

5. Succinctly describe what a 0.05 significance level ($\alpha = 0.05$) means with respect to a hypothesis test.
   Solution:
   Hypothesis tests have a null hypothesis where a test statistic equals some value and the alternative hypothesis poses that the test statistic is different from some value. A 0.05 significance levels means that if we want to be able to reject our null hypothesis, our observation/test statistic has to be so rare, that it would only appear in 100(1-α) percent of cases, or 5 percent of the time. This rarity gives us sufficient evidence to reject the null hypothesis.

6. Succinctly describe why post-hoc testing is necessary for the ANOVA, Kruskal Wallis, or Mood's median test.
   Solution:
   Post-hoc testing is necessary for the ANOVA, Kruskal Wallis, or Mood's median test because these procedures provide us with adjusted p-values. When we perform omnibus tests, we are comparing variances between many groups, which means our p values are being miscalculated and specifically are overestimating confidence. For example, if we compare across 3 groups, a 95% confidence level is actually a confidence level of (95%)(95%)(95%) which is about .7. The post-hoc tests adjust p-values so we can properly analyze p values across groups and the entire test. ANOVA substitutes the t distribution with the Tukey-HSD distribution, Kruskal Wallis uses the Dunn's test, and Mood's Median uses the Pair-wise median test.

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:

   The Bracht et al. analysis is trying to figure out if MFAP4 levels differ across different stages of disease, meaning that they are essentially comparing across "treatments". For our case, we would need to assume before moving forward that the sample is representative. When comparing across treatmeants, the researchers should be using an omnibus test. Since we don't know the skew of the data, it is hard to know whether the ANOVA, Kruskal Wallis, or Mood's median test should be used. If the data is skewed, we may use the Mood's Median test to compare variance across groups using medians. If the data is not skewed, we may use ANOVA or Kruskal Wallis. Rank may be important in this case in terms of ranking the severity of the stages of disease, and therefore we may want to use Kruskal Wallis. Accompanying the Kruskal Wallis test, the Dunn's test may be used to adjust P-values and identify which ranked groups are different and to what extent. If we are using a continuous distribution with a discrete distribution, we also would then want to implement a continuity correction to construct a confidence interval. This would be used to solve for the population proportion and we could therefore use a Wilson interval to find the true population proportion or MFAP4 within the different stages of the disease.

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.