Part 1:

1. Professional Magic:
    a. What is the type 1 error rate for this test?

W203 Lab 1: Hypothesis Testing

Professional Magician

1. $X_i \sim \begin{cases} \frac{1}{2} & X_i = 1 \\ \frac{1}{2} & X_i = 0 \end{cases} = P_{mf}\, X_i = P_{mf}\, Y_i = \begin{cases} \frac{1}{2} & Y_i = 1 \\ \frac{1}{2} & Y_i = 0 \end{cases}$

- Each pair of flips is $\perp$
- Consider $P(X_i = Y_i) = p$
- $P(X_i = Y_i) = P(X, Y=0) + P(X, Y=1) = P(X=0)P(Y=0 \mid X=0) + P(X=1)P(Y=1 \mid X=1) = p$
- Null $H_0: P = 1/2$; Alternate $H_a: p > 1/2$
- Procedure: 3 consecutive flips, Test Statistic $T = \sum_{i=1}^{3} X_i + \sum_{i=1}^{3} Y_i = \sum_{i=1}^{3}(X_i + Y_i)$
- Rejection Region: $T=0$, $T=6$

a) P(Type I error). Type I error = rejecting Ho when it is true

$P(\text{Type I error}) = P(\text{rejection region} \mid H_0) = P(T=0 \mid H_0) + P(T=6 \mid H_0)$

- $P(T=0 \mid H_0) = \prod_{i=1}^{3} P(X_i, Y_i = 0) = \prod_{i=1}^{3} P(X_i=0)P(Y_i=0 \mid X_i=0) = \prod_{i=1}^{3} \frac{1}{2} P(Y_i=0 \mid X_i=0)$

$= \frac{1}{8}\left(P(Y_i=0 \mid X_i=0)\right)^3$

- $P(T=1 \mid H_0) = \prod_{i=1}^{3} P(X_i, Y_i = 1) = \prod_{i=1}^{3} P(X_i=1)P(Y_i=1 \mid X_i=1) = \prod_{i=1}^{3} \frac{1}{2} P(Y_i=1 \mid X_i=1)$

$= \frac{1}{8}\left(P(Y_i=1 \mid X_i=1)\right)^3$

- $P(\text{rejection region} \mid H_0) = \frac{1}{8}\left(\overbrace{P(Y_i=0 \mid X_i=0)}^{1-a}{}^3 + \overbrace{P(Y_i=1 \mid X_i=1)}^{b}{}^3\right)$

- $P_{Y_i \mid X_i=0} = \begin{cases} a & Y_i = 1 \\ 1-a & Y_i = 0 \end{cases}$, $P_{Y_i \mid X_i=1} = \begin{cases} b & Y_i = 1 \\ 1-b & Y_i = 0 \end{cases}$

- $P(Y=1) = \frac{1}{2} = P(X=1)P(Y=1 \mid X=1) + P(X=0)P(Y=1 \mid X=0) = \frac{1}{2}(b+a) = \frac{1}{2} \Rightarrow a+b=1$
- Assuming null is true, $\frac{1}{2}(1-a+b) = \frac{1}{2} \Rightarrow 1-a+b=1 \Rightarrow -a+b=0 \Rightarrow a=b$

$\Rightarrow a+a=1 \Rightarrow a=b=\frac{1}{2}$

- $P(\text{rejection region} \mid H_0) = \frac{1}{8}\left(\left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^3\right) = \frac{1}{8}\left(\frac{1}{4}\right) = \frac{1}{32} = P(\text{Type I error})$

$= .03125$

b. What is the power of your test for the alternate hypothesis that p = ¾:

b) Statistical Power = $1-\beta = 1 - P(\text{Type II error})$, $H_a = \frac{3}{4} = p$

- Type II error = Fail to reject Ho when it is false
- $P(\text{Type II error}) = P(T \in \{1,2,3,4,5\} \mid H_a) = 1 - P(T \in \{0,6\} \mid H_a)$

$= 1 - P(T=0 \mid H_a) - P(T=6 \mid H_a)$

- $P(T=0 \mid H_a) = \prod_{i=1}^{3} P(x_i, Y_i = 0) = \frac{1}{8}\left(P(Y_i=0 \mid X_i=0)\right)^3$
- $P(T=6 \mid H_a) = \prod_{i=1}^{3} P(x_i, Y_i=1) = \frac{1}{8}\left(P(Y_i=1 \mid X_i=1)\right)^3$

by part a

$P_{Y_i \mid X_i=0} = \begin{cases} a & Y_i = 1 \\ 1-a & Y_i = 0 \end{cases}$ $\quad P_{Y_i \mid X_i=1} = \begin{cases} b & Y_i=1 \\ 1-b & Y_i = 0 \end{cases}$

- By part a, $P(Y=1) = \frac{1}{2} \Rightarrow a+b=1$

- Assuming $H_a: p = \frac{3}{4}$ is true,

$P(X_i=0)P(Y_i=0 \mid X_i=0) + P(X_i=1)P(Y_i=1 \mid X_i=1) = \frac{1}{2}(1-a+b) = \frac{3}{4}$

$\Rightarrow 1-a+b = \frac{3}{2} \Rightarrow -a+b = \frac{1}{2} \Rightarrow b = \frac{1}{2}+a$

$\Rightarrow a + \frac{1}{2} + a = 1 \Rightarrow 2a = \frac{1}{2} \Rightarrow a = \frac{1}{4}$, $b = \frac{1}{2}+\frac{1}{4} = \frac{3}{4}$

$P(\text{Rejection Region} \mid H_a) = \frac{1}{8}\left((1-a)^3 + (b)^3\right) = \frac{1}{8}\left(\left(\frac{3}{4}\right)^3 + \left(\frac{3}{4}\right)^3\right)$

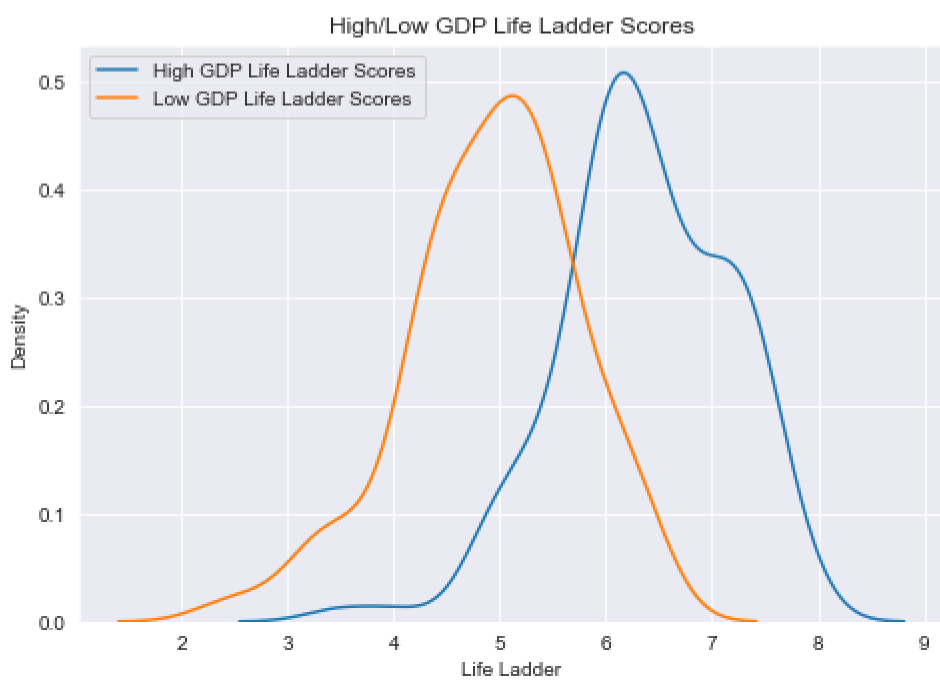
$= \frac{1}{8}\left(\frac{27}{64} + \frac{27}{64}\right) = \frac{1}{8}\left(\frac{54}{64}\right) = \frac{54}{512}$

- $P(\text{Type II error}) = 1 - \frac{54}{512}$
- Power $= 1 - \left(1 - \frac{54}{512}\right) = 1 - 1 + \frac{54}{512} = \frac{54}{512} = \frac{27}{256} \approx .105$

2. Wrong Test, Right Data. Imagine that your organization surveys a set of customers to see how much they like your regular website, and how much they like your mobile website, both measured on 5-point Likert scales. If you were to run a paired t-test, what consequences would the violation of the metric scale assumption have for your interpretation of the test results?
    a. Given we would be using a statistical test not designed for our data, the test results would be unreliable and likely would not accurately portray the actual relationship between how much customers like the regular vs. the mobile website. The differences between scores cannot be accurately measured therefore we can assume it does not meet a sufficiently normal distribution, even with a larger sample size because data cannot be aggregated.

Test Assumptions:

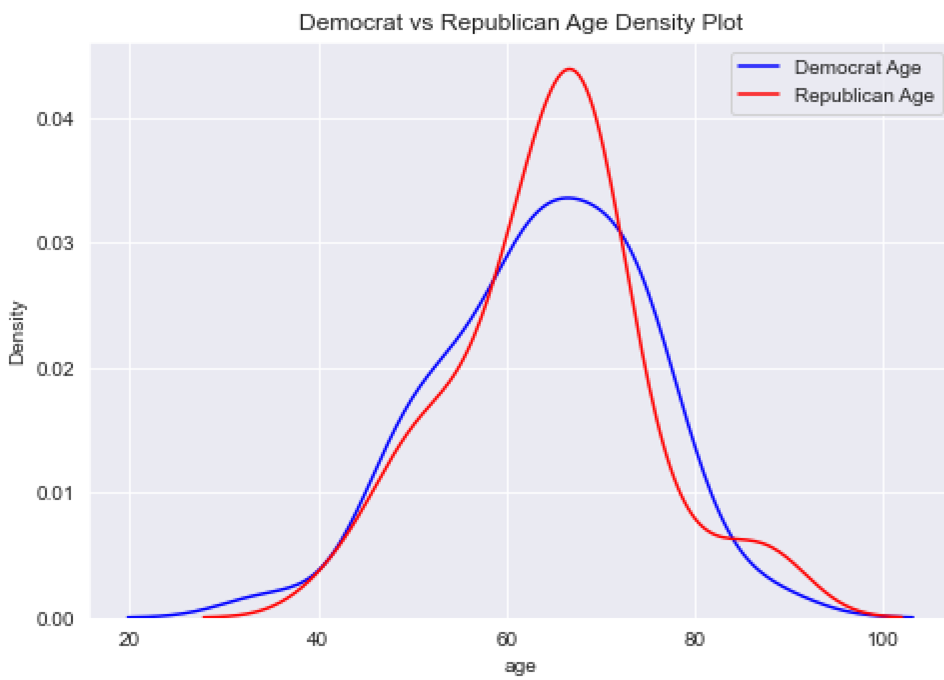
1. **World Happiness** - Assumptions for a two sample t test:
    a. Data is independent
        i. The classification between GDP groups would not meet independent criteria because GDP is interrelated. GDP is related to the trade between countries and this is dependent. We have no reason to believe the happiness of one country is dependent on the happiness of another country.
    b. Data is normally distributed
        i. While both samples indicate deviations from normality, we have large enough sample sizes (121 and 105 for the high/low gdp groups respectively) that the CLT should imply relative normality



ii.

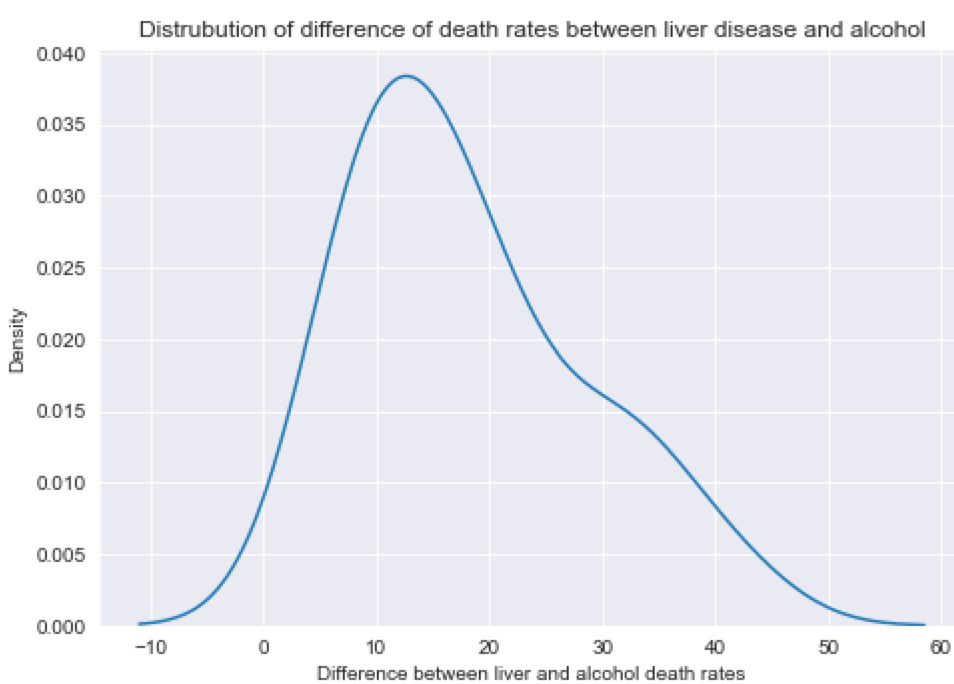
c. Equal variances
    i. The variances in both groups are pretty small (.64 for the high group and .67 for the low group). While not exact, these variances are close enough and small enough that we believe the equal variance assumption is substantively met
d. Metric Scale - Data is continuous
    1. Yes, we are making the assumption that the intervals between Cantril Ladder scores are equal. The study provided the mean scores for each country, so the data would be continuous in (0,10). However, happiness as a variable alone would typically be ordinal so the approach by the study of providing a mean could be flawed. If that were the case, normality and variance would also be flawed.

2. **Legislator Age** - Assumptions for a Wilcoxon rank-sum test:
    a. Observations from each group are independent of each other:
        i. The observations are independent. One person's age is not affected by another's.
    b. Responses are metric (hypothesis by means) :
        i. Yes, the age data is metric - we can rank one person as older than another and measure the difference.
    c. Normal distribution or large sample size:
        i. The distributions are a bit abnormal, but since we have a suitably large sample size (48 for democrats, 50 for republicans) we should meet this assumption due to the CLT



ii.

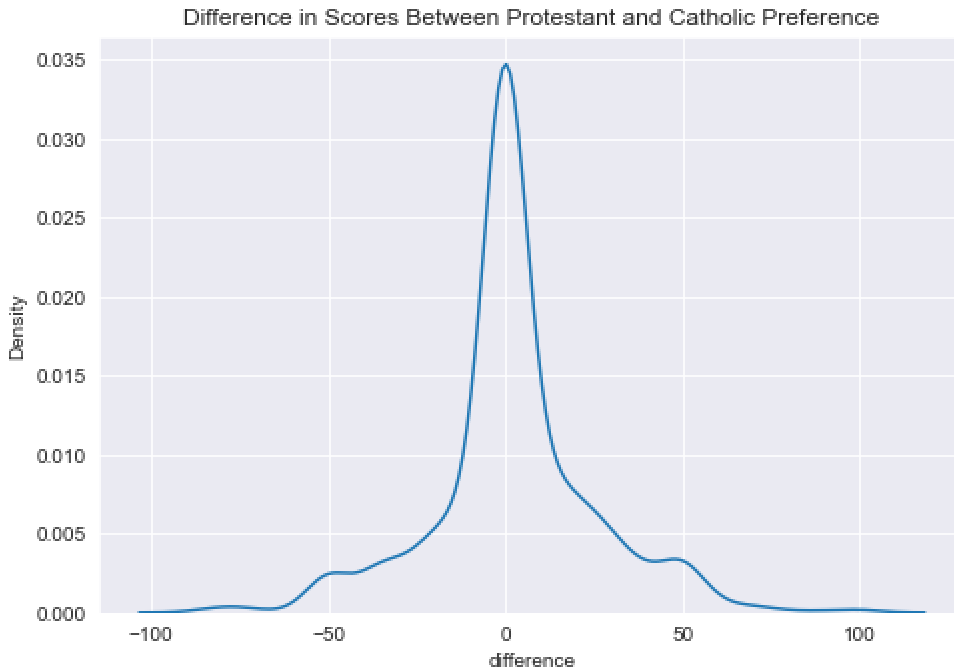
3. **It's For Your Health! -** Assumptions for a signed-rank test:
   a. Symmetric distribution:
      i. The distribution of data is not symmetric



   b. Response variables are independent:

        i. There could be a dependency on geographical clustering by country. However, if the data collection process could avoid this, then the samples are independent, meaning the difference in death rates from one country does not inform on the difference in death rates from another
    c. The measurement scale of the differences is metric:
        i. Our data meets this assumption - death rates are metric scale

4. **Positive Vibes -** Assumptions for a paired t-test:
    a. The dependent variable is continuous:
        i. The samples appear to be classifying people's sentiment into categories. Although these are numerical values, the response is subjective and cannot be measured between two different respondents. In this example, we are looking at the difference in sentiment between Catholic and Protestant from one respondent. One respondent may have a +30 response in the direction of Protestants which could be subjectively different than a respondent with the same measurement.
    b. Observations are independent:
        i. The excerpt doesn't explain how the data collection process occurred. Were these randomly selected people or were they groups that showed up together at a survey location. The first would support IID but the second could allude to some dependency between groups. For example, if the respondents came from the same household and responded together.
    c. Data is paired:
        i. Yes, we can pair one person's response for each religious group.
    d. Dependent variable is normally distributed:
        i. We defined this data as ordinal, which would not meet the normally distributed assumption. If the data was metric then the differences between the two variables is pretty abnormal - essentially a large peak at 0 and not a lot of values elsewhere. With that said, since we have a very large sample size (n=802) the CLT would allow for the distribution of the sample mean to be approximately normal



e. Cannot be highly skewed:
   i. The distribution does not appear to be highly skewed