Lab 1 Part 1 - Foundational Excercises

Team03: Savita Chari, Tymon Silva, Denny Lehman

Professional Magic

Your aunt (who is a professional magician), claims to have created a pair of magical coins that share a connection to each other that makes them land in the same way. The coins are always flipped at the same time. For a given flip $i \in \{1, 2, 3...\}$, let X_i be a Bernoulli random variable representing the outcome of the first coin, and let Y_i be a Bernoulli random variable representing the outcome of the second coin. You assume that each flip of the pair is independent of all other flips of the pair. You also assume that

$$P(X_i = 0) = P(X_i = 1) = P(Y_i = 0) = P(Y_i = 1) = 1/2,$$

and write,

$$P(X_i = Y_i) = p.$$

Your aunt claims that $p > 1/2$.

You design a test to evaluate your aunt's claim. You flip the coins 3 times and define your test statistic to be the sum $X_1 + Y_1 + X_2 + Y_2 + X_3 + Y_3$

Your null hypothesis is that $p = 1/2$. You plan to reject the null if your test statistic is 0 or 6.

1. What is the type 1 error rate of your test?
2. What is the power of your test for the alternate hypothesis that $p = 3/4$?

1. What is the type 1 error rate of the test?

The type 1 error occurs when we reject the null hypothesis when it is actually true. Type I error, a false positive, is typically denoted as alpha

$$\alpha = P(rejectH_o|H_o)$$

We are given that Ho is when p = 1/2. p is the probability that Xi = Yi

$$H_o : p = \frac{1}{2}, p = P(X_i = Y_i)$$

Let's define a trial as flipping both coins at the same time, where coin X has outcomes 0 and 1 and coin Y has 0 and 1

$$X \in \{0, 1\}$$
$$Y \in \{0, 1\}$$
$$\theta_1 = \text{outcome of the event, which is two coin flips}$$

Event space of θ_1

$$\theta_1 \in \{\{0, 0\}, \{0, 1\}, \{1, 0\}, \{1, 1\}\}$$
$$P(X = 0) = \frac{1}{2}$$
$$P(Y = 0) = \frac{1}{2}$$
$$P(X = 1) = \frac{1}{2}$$
$$P(Y = 1) = \frac{1}{2}$$
$$P(X = x, Y = y) = P(X = x) * P(Y = y)$$

We know that each flip of the pair has the following potential outcomes:

$$P(X_i = Y_i) = P(X_i = 0, Y_i = 0) + P(X_i = 1, Y_i = 1)$$
$$P(X_i = Y_i|H_o) = 0.5 = P(X_i = 0, Y_i = 0) + P(X_i = 1, Y_i = 1)$$
$$P(X_i = 0, Y_i = 0) = \frac{1}{4}$$
$$P(X_i = 1, Y_i = 1) = \frac{1}{4}$$

We are also given that our test statistic, theta_hat, will be equal to

$$\hat{\theta_n} = \sum_{i=1}^{n} X_i + Y_i$$
$$\hat{\theta_3} = X_1 + Y_1 + X_2 + Y_2 + X_3 + Y_3$$

From the problem definition, we reject the null when our test statistic theta is 0 or 6 after 3 flips of pairs of coins. Type I error is thus

$$\alpha = P(rejectH_o|H_o)$$
$$\alpha = P(\hat{\theta_3} \in \{0, 6\}|p = \frac{1}{2})$$
$$\alpha = P(\hat{\theta_3} = 0|p = \frac{1}{2}) + P(\hat{\theta_3} = 6|p = \frac{1}{2})$$

Solve for each term using the binomial distribution formula

binomial distribution formula

$$P_x = nCxq^x(1 - q)^{n-x}$$

q was chosen to avoid confusion with p from the problem statement

solve for $P(\hat{\theta_3} = 0|p = \frac{1}{2})$where q is $P(X_i = 0, Y_i = 0) = 1/4\hat{\theta_3} = 0$ means that after 3 three trials, the test statistic has valu

$$solve for P(\hat{\theta_3} = 6|p = \frac{1}{2})$$

where q is $P(X_i = 1, Y_i = 1) = 1/4$

$\hat{\theta_3} = 6$ means after all three trials, the test statistic has the value of 6 (all 1's)

$$P(\hat{\theta_3} = 6|p = \frac{1}{2}) = 3C3q^3 * (1-q)^0$$

$$P(\hat{\theta_3} = 6|p = \frac{1}{2}) = (1)(1/4)^3(1)$$

$$P(\hat{\theta_3} = 6|p = \frac{1}{2}) = (1/4)^3$$

now substitute back in for alpha

$$\alpha = P(\hat{\theta_3} = 0|H_o) + P(\hat{\theta_3} = 6|H_o)$$

$$\alpha = (\frac{1}{4})^3 + (\frac{1}{4})^3$$

$$\alpha = \frac{2}{64} = \frac{1}{32}$$

2. What is the power of your test for the alternate hypothesis that $p = 3/4$?

The power of our test is the Probability that we have successfully rejected the null hypothesis when the alternative is true. $Power = 1 - \beta$ Where β is the type 2 error, or the false negative rate. Type 2 error occurs when we fail to reject the null hypothesis when we should.

$$H_a : p = \frac{3}{4}, p = P(X_i = Y_i) Power = P(rejectnull|H_a)$$

We assume that the likelihood for heads and tails remains the equal, but the likelihood of getting $X_i = Y_i$ changes to $\frac{3}{4}$. Therefore,

$P(X_i = Y_i|H_a) = 3/4 P(X_i = Y_i) = P(X_i = 0, Y_i = 0) + P(X_i = 1, Y_i = 1) P(X_i = 0, Y_i = 0) = P(X_i = 1, Y_i = 1) \therefore P(X_i = 0$

When fail to reject

$\hat{\theta_3} = X_1 + Y_1 + X_2 + Y_2 + X_3 + Y_3$ reject Ho when $\hat{\theta_3} = 0 or \hat{\theta_3} = 6 Power = P(rejectnull|H_a) Power = P(\hat{\theta_3} \in \{0, 6\}|p = \frac{3}{4}) Powe$

Now we solve for the two terms

$P(\hat{\theta_3} = 0|p = 3/4) = 3C0 * (3/8)^3 * (1-3/8)^0 = (1)(3/8)^3(1) = \frac{27}{512} P(\hat{\theta_3} = 6|p = 3/4) = 3C3 * (3/8)^3 * (1-3/8)^0 = (1)(3/8)^3(1)$

$$Power = P(\hat{\theta_3} = 0|p = 3/4) + P(\hat{\theta_3} = 6|p = 3/4) Power = (\frac{3}{8})^3 + (\frac{3}{8})^3 Power = 2 * (\frac{3}{8})^3 Power \approx 0.105$$

Wrong Test, Right Data

Our Analysis :

The response categories in Likert scales creates rank order/ordinal data, but the intervals between their values cannot be compared or presumed equal. Therefore, standard statistical analysis such as the mean, standard deviation etc. are inappropriate for ordinal data. Hence, this data is not suitable for a paired t-test. Paired t-test is parametric which means it expects a normal distribution of metric scaled data in its test population. It may be used for ordinal data such as True/False but for our use case a non-parametric test such as Wilcoxon signed-rank test or Sign test would be a better choice.

Test Assumptions

For the four following questions, your task is to evaluate the assumptions for the given test using your background knowledge and examining the data.

World Happiness

You would like to know whether people in countries with high GDP per capita (higher than the mean) are more happy or less happy than people in countries with low GDP (lower than the mean). List all assumptions for a two-sample t-test and evaluate them.

Two-sample t-test assumptions

1. Metric variables
   The Life Ladder variable is a metric variable. We can justify this by reviewing the FAQ for this dataset that states, "... it asks respondents to think of a ladder, with the best possible life for them being a 10, and the worst possible life being a 0. They are then asked to rate their own current lives on that 0 to 10 scale." From this description, we know the responses fall on a 0 to 10 scale, where the distances or gaps between each point on the scale are the same, since each rung on a ladder is the same distance for the next.
2. IID
   Here, we assume IID is upheld since Gallup [1] is a reputable research firm, and we expect their sampling and data collection processes maintain the independence and identically distributed assumption.
3. Data is normally distributed (check sample size)
   Since we have sufficient sample sizes (greater than 30) in each of the Low (n=105) and High (n=121) groups for GDP, we can assume normality by the Central Limit Theorem.

Thus, all the assumptions for a two-sample t-test are valid and the test would be appropriate.

Study question 3: Legislator age

You would like to test whether Democratic or Republican senators are older. List all assumptions for a Wilcoxon rank-sum test and evaluate them.

Using Wilcoxon rank-sum test we have been asked to run a statistical test using the data provided and maintained by GovTrack, ProPublica, MapLight, FiveThirtyEight etc. The correctness of id's which relate the record to other databases, name information (first, last, etc.), biographical information (birthday, gender), and terms served in Congress guaranteed. [2]

Test Assumption:

The database maintains information about all the members of the United States Congress, Senators, Presidents and Vice Presidents of the United States. The study question focuses our test on Democratic and Republican *senators*, therefore we will limit the study only to the senators of USA. The field $type = 'sen'$ filters only desired rows of data from this dataset.

[1] https://www.gallup.com/corporate/212381/who-we-are.aspx
[2] https://github.com/unitedstates/congress-legislators

Hypothesis

H_0: Probability that the Democrat Senators are older than Republicans is same as the probability of Republican Senators being older than their Democrat counterpart.

What is Wilcoxon rank-sum test?

Wilcoxon rank-sum test is a non-parametric, distribution-free test for two independent samples. Though it involves assumptions, but those assumptions are less restrictive than the assumptions for parametric tests. It considers the ranks instead of the metric value of the variable. It uses the order of variables to construct statistics that can be used to test hypothesis.

Advantage : The population distribution doesn't have to be normal, so it's easier to justify a rank-based test. It is a good choice of test for smaller sample size.

Disadvantage : Since these tests do not use the metric information they loose statistical power.

The Wilcox rank-sum test works on following principals:

1. The samples are interval scale and list the score from lowest to highest. Higher rank gets higher score
2. Only considers rank instead of looking at metric value of the variable
3. Uses order of variables to construct statistics that is used for hypothesis

The Wilcox rank-sum test makes the following assumptions:

1. Test involves independent, unpaired samples
2. The data is IID
3. The Data is Ordinal Variable

Data and Methodology

In order to perform Wilcoxon test we need 2 types of columns from our dataset

1. [X] A categorical column with 2 distinct groups : We chose the 'party' column for this requirement but the data set includes more than two categories. We filtered only the Republican and Democrat data in order to perform Wilcoxon rank-sum test as we are interested in only those categories. Thus we fulfill the requirement of having only 2 categories.
2. [Y] A statistical column with numeric outcome: We chose the 'birthday' field for this requirement. Though this field is of class 'Date', with some manipulation we can make it a numeric outcome

Data Transformation

As mentioned above, the birthday field is of 'Date' class and is in YYYY/MM/DD format. In order to find out the age of the Senator we calculate the current age of the Senator by doing a datediff between current date and the birthday

```
## # A tibble: 3 x 5
##   full_name          birthday   party    type  AgeInYears
##   <chr>              <date>     <chr>    <chr>      <dbl>
## 1 Sherrod Brown      1952-11-09 Democrat sen           69
## 2 Maria Cantwell     1958-10-13 Democrat sen           63
## 3 Benjamin L. Cardin 1943-10-05 Democrat sen           78
```

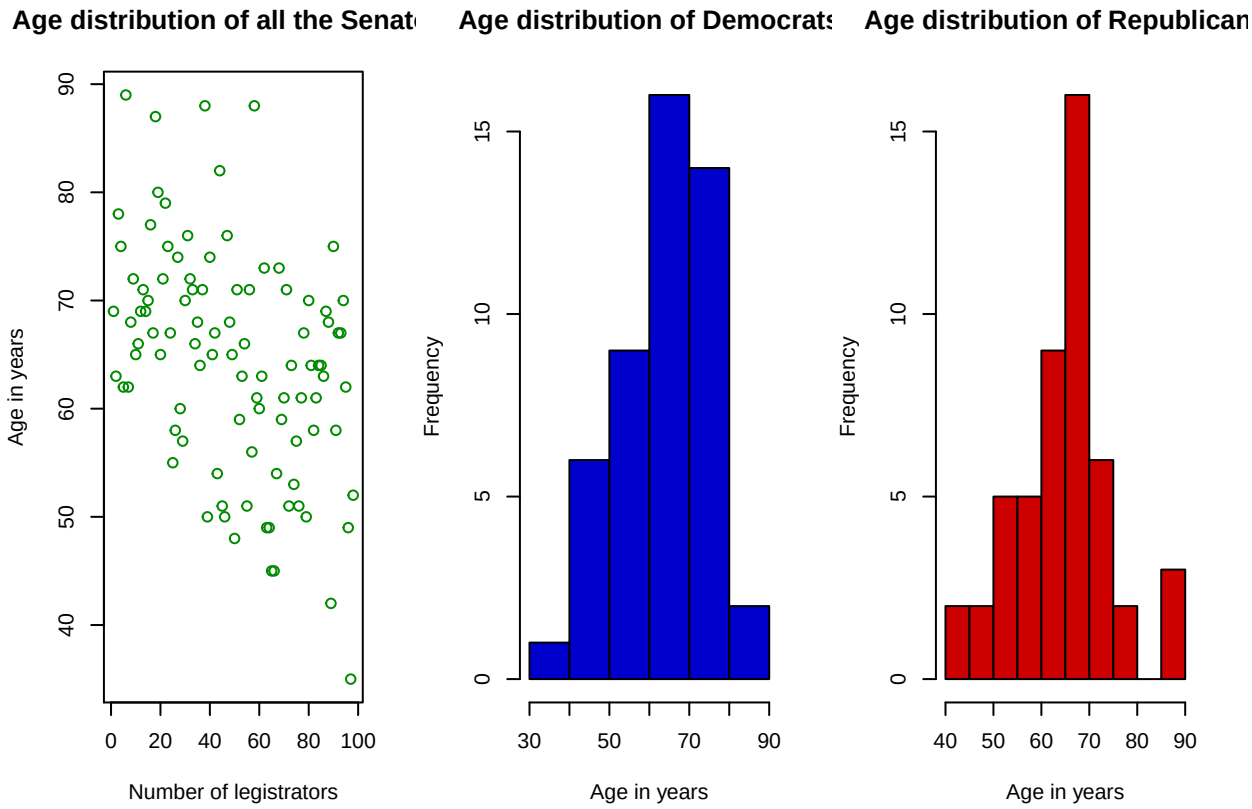
Exploring Assumption 1 : Are the data sets independent?

Well, yes, a Senator cannot belong to 2 parties simultaneously, so the data is independent. As mentioned in the introduction section, the source of the data is the congress-legislators project. Data such as information (first, last, etc.), biographical information (birthday, gender) of every members of the United States Congress (1789-Present) is maintained. Data of one Senator does not depend on the other. So we can safely assume the data is independent. In this test we are using the entire population of data available to us.

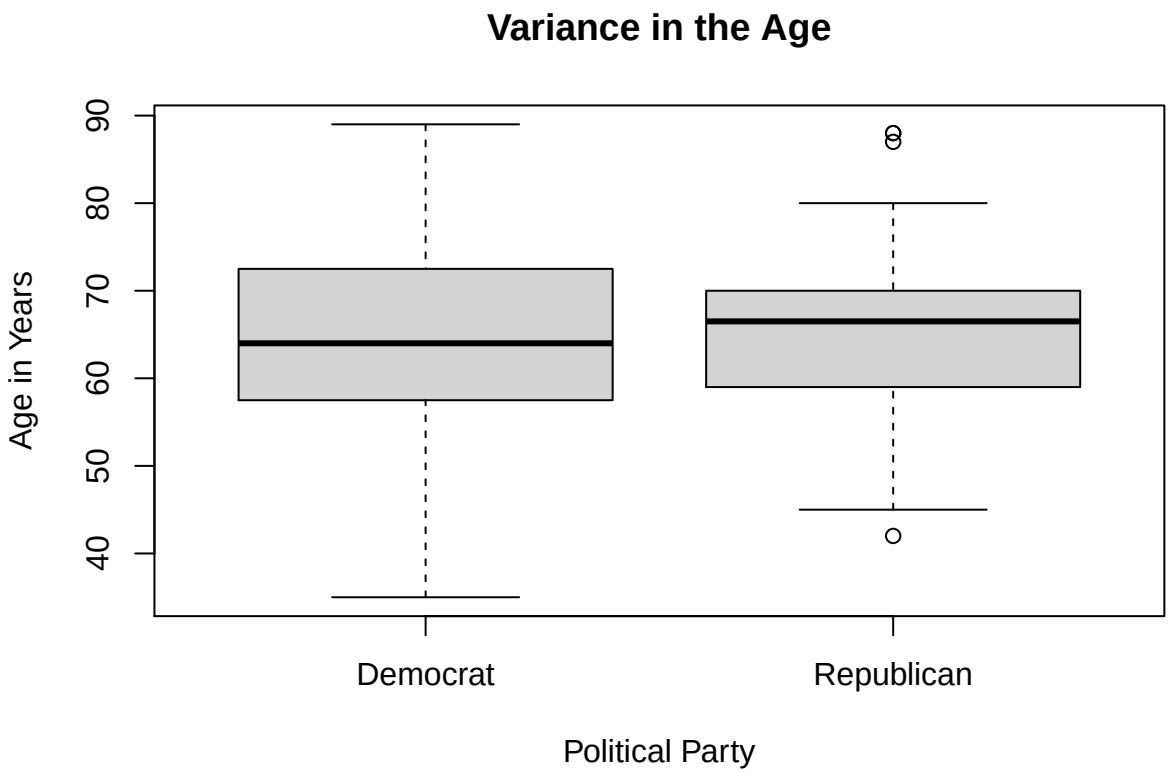
Exploring Assumption 2 : Is the population IID?

From the graph below we can tell that the age distribution between republicans and democrats is some what normal, and not too skewed. Moreover, as per Central Limit Theorem (CMT), when we have a sufficiently large sample size (98 in our case), the sampling distribution starts to approximate a normal distribution .

Given this information, Wilcoxon rank − sum test is not a good choice for this use case



Also, from the box plot below, it is very clear that the we can calculate variance on the metric data (AgeInYears). Democrats have a larger variance than the Republicans. In the Republican data set there are some outliers too.



Performing two sided Wilcoxon rank-sum test

In order to prove the null hypothesis we have been asked to perform the Wilcoxon rank-sum test. Our data satisfies the requirements for conducting this test.

```
wilcox.test(legislator_DemRep_Age_data$AgeInYears~legislator_DemRep_Age_data$party, mu=0,
            alt = "two.sided", conf.int = T, conf.level=0.95,  paired=F, exact=F, correct=T)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  legislator_DemRep_Age_data$AgeInYears by legislator_DemRep_Age_data$party
## W = 1193.5, p-value = 0.966
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -4.000097  4.000001
## sample estimates:
## difference in location
##          -7.858072e-05
```

With p-value being greater that 0.05, We cannot reject the null hypothesis. Which stated that the probability that the Democrat Senators are older than Republicans is same as the probability of Republican Senators being older than their Democrat counterpart.

Observation

It is our opinion that in this scenario where we have an option to run either parametric or non-parametric test, we should choose parametric t-test becasue it has more statistical power as compared to non-parametric test such as wilcoxon Rank-Sum test.

Our data fulfills all the requirements to run a t-test

1. The samples are independent of one another and are Metric scale.
2. The populations have equal variance or spread
3. The populations are normally distributed (IID)

```
t.test(legislator_DemRep_Age_data$AgeInYears~legislator_DemRep_Age_data$party)
```

```
##
##  Welch Two Sample t-test
##
## data:  legislator_DemRep_Age_data$AgeInYears by legislator_DemRep_Age_data$party
## t = -0.30936, df = 95.063, p-value = 0.7577
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.858245  3.548245
## sample estimates:
##   mean in group Democrat mean in group Republican
##                   64.125                   64.780
```

With p-value being greater that 0.05, We cannot reject the null hypothesis. Which stated that the probability that the Democrat Senators are older than Republicans is same as the probability of Republican Senators being older than their Democrat counterpart.

The t-test further observes that the mean age of Republican Senators is approximately 6 months higher than their Democratic counterpart.

Test Outcome and it's correctness

1. We used the mean to create our test statistics which is well accepted distributions for statistical testing.
2. We followed the decision rules of a hypothesis test which gives us the guarantee that our false positive rejection rate (the type 1 error rate) is bounded.

3. Our test sample size is >30 so it meets the standard t-test heuristic.

Conclusion for Study question 3: Legislator age

It is important to observe the data before choosing the right kind of test to make correct predictions and to leverage the full power of statistics.

It's for your health!

You would like to use it [wine dataset] to test whether countries have more deaths from heart disease or from liver disease. List all assumptions for a signed-rank test and evaluate them.

```
##         country alcohol deaths heart liver
## 1     Australia     2.5    785   211  15.3
## 2       Austria     3.9    863   167  45.6
## 3      Belg/Lux     2.9    883   131  20.7
## 4        Canada     2.4    793   191  16.4
## 5       Denmark     2.9    971   220  23.9
## 6       Finland     0.8    970   297  19.0
## 7        France     9.1    751    71  37.9
## 8       Iceland     0.8    743   211  11.2
## 9       Ireland     0.7   1000   300   6.5
## 10       Israel     0.6    834   183  13.7
## 11        Italy     7.9    775   107  42.2
## 12        Japan     1.5    680    36  23.2
## 13  Netherlands     1.8    773   167   9.2
## 14  New Zealand     1.9    916   266   7.7
## 15       Norway     0.8    806   227  12.2
## 16        Spain     6.5    724    86  36.4
## 17       Sweden     1.6    743   207  11.2
## 18  Switzerland     5.8    693   115  20.3
## 19           UK     1.3    941   285  10.3
## 20           US     1.2    926   199  22.1
## 21 West Germany     2.7    861   172  36.7
```

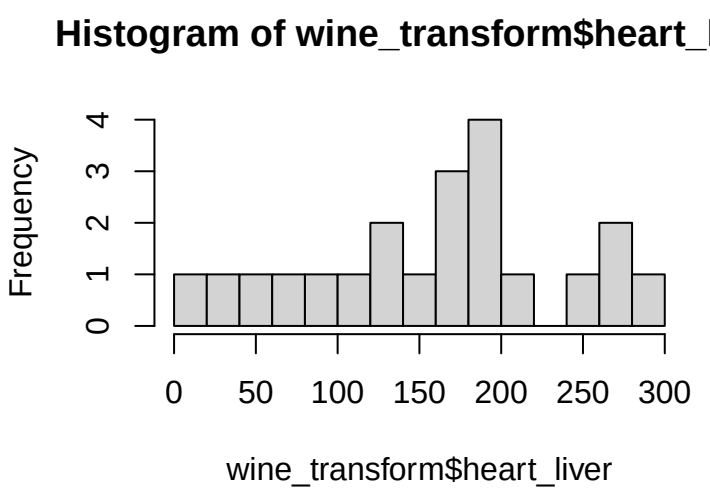
Wilcoxon signed-rank test assumptions

1. Metric variables We need metric variables for a paired Wilcoxon signed-rank test, and the heart and liver variables satisfy this assumption. The Heart and Liver variables represent the number of deaths from heart and liver disease in each country.

2. IID
   There are some concerns with IID for this data. Since there dataset only contains 21 countries, we do not know how these countries were drawn, particularly if they were randomly select. Furthermore, we do not know how the data was collected, which means the data draw could not have been identically distributed.

3. Difference is symmetric The last assumption requires the difference between pairs follows a symmetric distribution. If you look at the histogram below of the difference between heart and liver disease deaths, one could argue that the distribution is skewed left, which violates our difference is symmetric assumption.

Due to major concerns with assumptions 2 and 3, the Wilcoxon signed-rank test would not be a valid test to perform to answer the question.



Positive vibes

You would like to test whether the US population feels more positive towards Protestants or towards Catholics. List all assumption for a paired t-test and evaluate them.

```
## # A tibble: 6 x 5
##    ...1  year    id prottemp cathtemp
##   <dbl> <dbl> <dbl>    <dbl>    <dbl>
## 1     1  2004     4       50       50
## 2     2  2004     7        5       85
## 3     3  2004     9       50       50
## 4     4  2004    14       60       60
## 5     5  2004    21      100        0
## 6     6  2004    24        5        5
```

Paired two-sample t-test assumptions

1. Metric variables
   The Life Ladder variable is a metric variable. We can justify this by reviewing the FAQ for this dataset that states, "... it asks respondents to think of a ladder, with the best possible life for them being a 10, and the worst possible life being a 0. They are then asked to rate their own current lives on that 0 to 10 scale." From this description, we know the responses fall on a 0 to 10 scale, where the distances or gaps between each point on the scale are the same, since each rung on a ladder is the same distance for the next.

2. IID
   Here, we assume IID is upheld since Gallup[3] is a reputable research firm, and we expect their sampling and data collection processes maintain the independence and identically distributed assumption.

3. Data is normally distributed (check sample size)
   Since we have sufficient sample sizes (greater than 30) in each of the Low (n=105) and High (n=121) groups for GDP, we can assume normality by the Central Limit Theorem.

[3] https://www.gallup.com/corporate/212381/who-we-are.aspx

Thus, all the assumptions for a two-sample t-test are valid and the test would be appropriate.