STA130 Term Test – Section LEC0101 - **Solutions**

October 26, 2018

Last Name: ______________________ Last Name: ______________________

Student number: ______________________

Tutorial section or teaching assistant's name: ______________________

Instructions:

- Total marks: 50
- There are 15 pages including this page.
- Backs of pages will not be marked. Write your answers only in the space provided.
- The test is 90 minutes long.
- You are permitted a 1-sided, handwritten, 8.5 x 11 inch aid sheet. You must hand in your aid sheet with your test.
- Calculators are not permitted.
- Questions begin on the next page

Introduction

Throughout this test, we will work with two data sets: (1) 2014 United Nations data on Human Development `HDIndex`, and The 2018 World Happiness Report `WHRData`. The United Nations data contains the Human Development Index (HDI), a summary measure (i.e., one number) of average achievement in key dimensions of human development. The 2018 Happiness Report contains a measure of happiness for countries from 2005 through 2017 along with several other variables.

Here is a look at the data sets:

```
glimpse(HDIndex)
```

```
## Observations: 188
## Variables: 6
## $ Country         <chr> "Norway", "Australia", "Switzerland", "Denmark...
## $ HDI             <dbl> 0.9438773, 0.9349583, 0.9296131, 0.9233279, 0....
## $ Life_expect     <dbl> 81.6, 82.4, 83.0, 80.2, 81.6, 80.9, 80.9, 79.1...
## $ Expected_school <dbl> 17.49259, 20.22107, 15.79043, 18.68933, 17.923...
## $ Mean_school     <dbl> 12.63100, 12.96338, 12.82348, 12.72939, 11.889...
## $ GNI             <dbl> 64992.34, 42260.61, 56431.07, 44025.48, 45435....
```

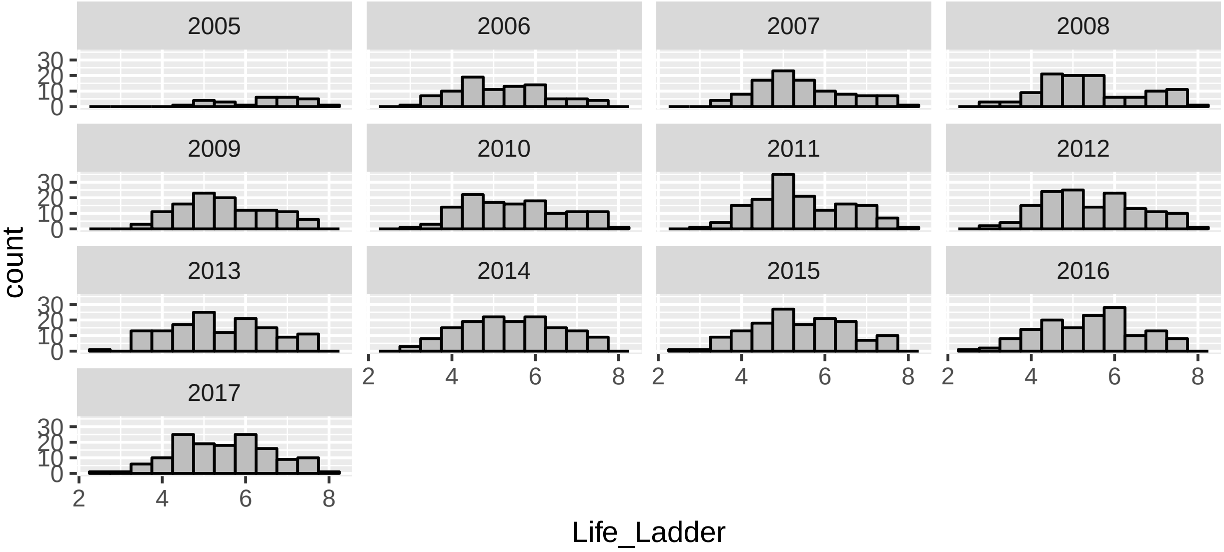
```
glimpse(WHRData)
```

```
## Observations: 1,562
## Variables: 6
## $ country                <chr> "Afghanistan", "Afghanistan", "Afghanis...
## $ year                   <dbl> 2008, 2009, 2010, 2011, 2012, 2013, 201...
## $ Life_Ladder            <dbl> 3.723590, 4.401778, 4.758381, 3.831719,...
## $ Social_support         <dbl> 0.4506623, 0.5523084, 0.5390752, 0.5211...
## $ Generosity             <dbl> 0.18181947, 0.20361446, 0.13763019, 0.1...
## $ Perceptions_corruption <dbl> 0.8816863, 0.8500354, 0.7067661, 0.7311...
```

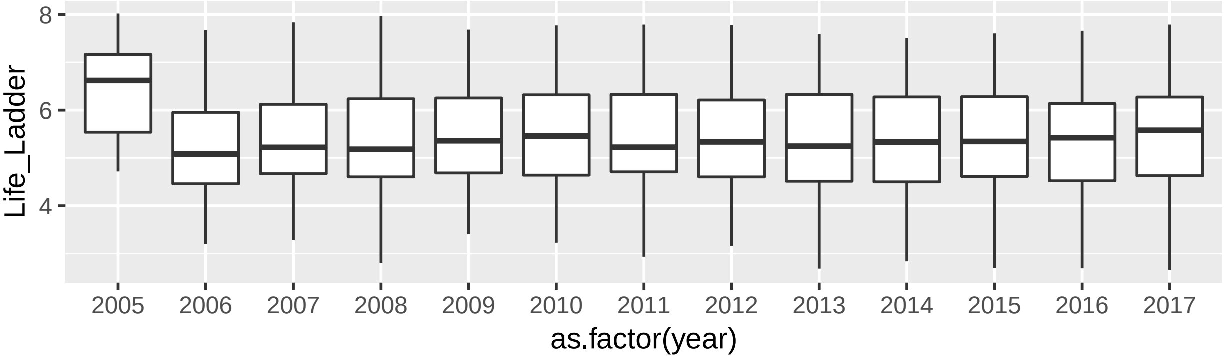
Question 1

The following two vizualizations labelled **Plot A** and **Plot B** show the distributions of the variable `Life_Ladder` for each `year`.

Plot A



Plot B



Answer the following questions based on the scenario described in Question 1. [12 marks, 2 marks each part]

(a) Which vizualization is more appropriate for comparing the median `Life_Ladder` between 2005 and 2017? **Plot A** or **Plot B**? Briefly explain your choice in one sentence.

Plot B, displays median values of `Life_Ladder` for each year.

(b) Which vizualization is more appropriate for comparing the number of observations and shape of the distribution of `Life_Ladder` for each year? **Plot A** or **Plot B**? Briefly explain your choice in one sentence.

Plot A, show histograms which display the number of observations in a histogram bin and show the shape of a ditsribution.

In parts (c) and (d) circle the correct answer.

(c) The variable `year` in the `WHRData` data is an R atomic vector that is stored as a

Double vector ~~**Factor vector**~~.

(d) The variable `Country` in the `HDIndex` data is a

~~**Quantitative variable**~~ **Categorical variable**.

Question 1 continued

(e) `Perceptions_corruption` is the proportion of people in a country that perceive widespread corruption in business or government in their country. The mean of `Perceptions_corruption` was calculated two different ways, in the R code below, resulting in **Table A** and **Table B**.

Table A

```
WHRData %>% group_by(year) %>%
  filter(year >= 2015) %>%
  summarise(n = n(), mean = mean(Perceptions_corruption))
```

```
## # A tibble: 3 x 3
##    year     n  mean
##   <dbl> <int> <dbl>
## 1  2015   143    NA
## 2  2016   142    NA
## 3  2017   141    NA
```

Table B

```
WHRData %>% group_by(year) %>%
  filter(year >= 2015 & is.na(Perceptions_corruption) == FALSE) %>%
  summarise(n = n(), mean = mean(Perceptions_corruption))
```

```
## # A tibble: 3 x 3
##    year     n  mean
##   <dbl> <int> <dbl>
## 1  2015   133 0.737
## 2  2016   131 0.747
## 3  2017   129 0.735
```

(i) In the context of this question, state what `n` represents in **Table A** and **Table B**.

In **Table A** `n` represents the number of countries in `WHRData` from 2015 to 2017.

In **Table B** `n` represents the number of countries in `WHRData` from 2015 to 2017 with a non-missing value of `Perceptions_corruption`.

(ii) In one sentence explain why the mean of `Perceptions_corruption` is `NA` in each row of **Table A**, but not `NA` in each row of **Table B**.

The mean is `NA` in each row of **Table A** since `mean` returns `NA` if the input values include `NA` values. Tables A and B show that there are `NA` values of `Perceptions_corruption` in each year so the mean in each row of Table A is `NA`.

Question 2

In order to compare the health of different countries using the `HDIndex` data it is important to transform Life Expectancy (`Life_expect`) into a value that can be compared across countries. The United Nations does this by setting minimum and maximum values of Life Expectancy (`Life_expect`) that act as "natural zeros" and "aspirational targets", respectively, to standardize Life expectancy.

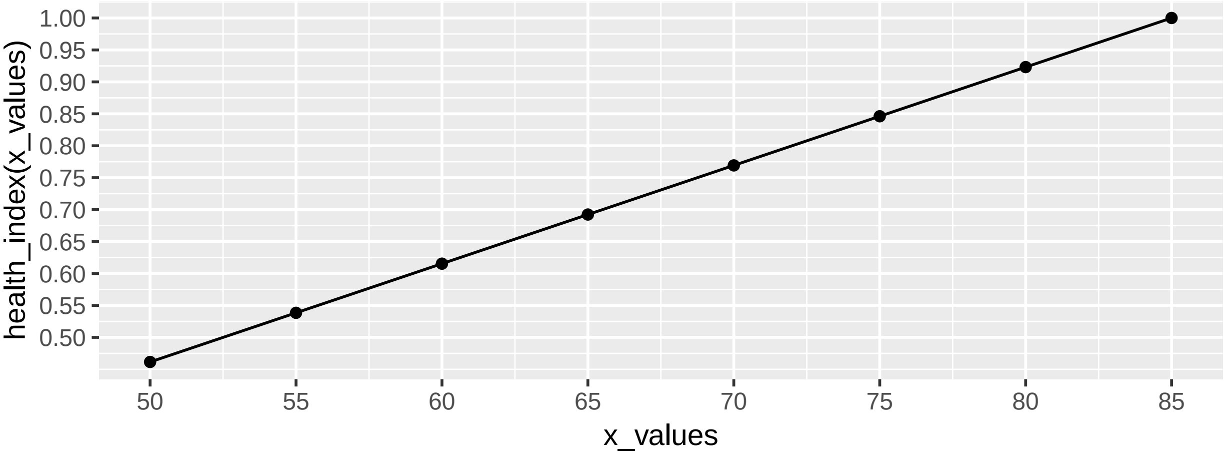
	Minimum	Maximum
Life Expectancy (years)	20	85

The next step in creating a health index is to calculate:

$$\text{Health index} = \frac{\text{actual value} - \text{minimum value}}{\text{maximum value} - \text{minimum value}}.$$

Health index is implemented in R in the function below, and plotted for values 50, 55, 60, 65, 70, 75, 80, 85.

```r
health_index <- function(actual_value){
  (actual_value - 20)/(85 - 20)
}
```



The function `health_index()` is used in the function `class_hi()` below.

```r
class_hi <- function(actual_value) {
  ifelse(health_index(actual_value) >= 0.9, "High", "Otherwise")
}
```

Answer the following questions based on the scenario described in Question 2. [7 marks]

(a) A random sample of three countries are selected from `HDIndex`. Fill in the values of `class_hi`, in the table below, using the `class_hi()` function. [3 marks]

Country	Life_expect	class_hi
Denmark	80.2	High
Bahamas	75.4	Otherwise
Haiti	62.8	Otherwise

(b) Which statement provides the *best explanation* of the values in the table produced by the R code below. **Circle the correct statement below.** [2 marks]

```r
HDIndex %>%
  group_by(class_hi(Life_expect)) %>%
  filter(Country == "Singapore" | Country == "Finland" |
  Country == "Cuba" | Country == "Gambia") %>%
  summarise(n = n())
```

class_hi(Life_expect)	n
High	3
Otherwise	1

(I) Singapore, Finland, and Gambia all have a Health Index value that is at least 0.9, and Cuba has a Health Index less than 0.9.

(II) At most one of Singapore, Finland, Cuba, and Gambia have a Health Index that is at least 0.9, and one of these countries has a Health Index less than 0.9.

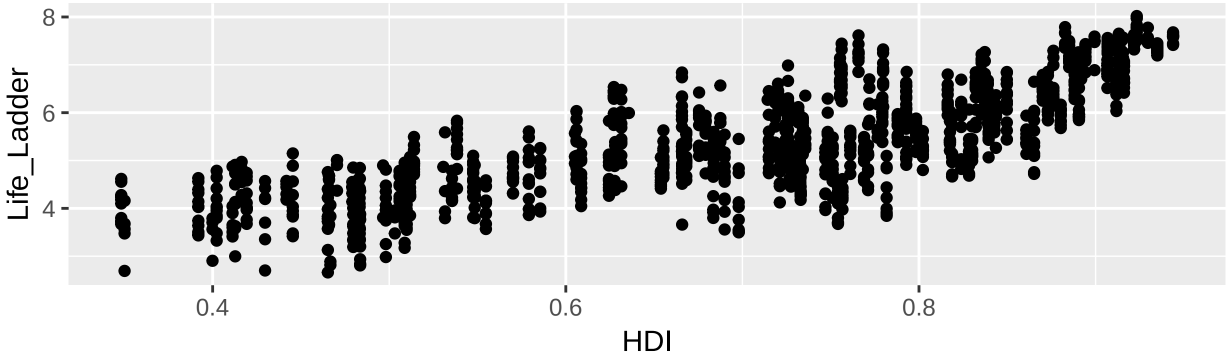
(III) Three countries in `HDIndex` have a Health Index that is at least 0.9, and one of the countries has a Health Index less than 0.9.

(IV) Three countries among, Singapore, Finland, Cuba, and Gambia, have Health Index values that are at least 0.9, and one of these countries has a Health Index less than 0.9.

(V) Only one country in `HDIndex` has a Health Index less than 0.9.

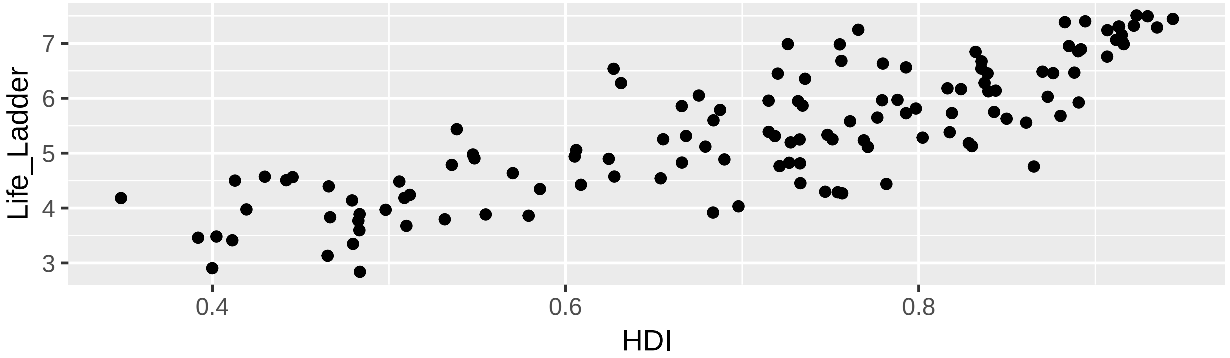
Plot A

```
WHRData %>%
  rename( Country = country) %>%
  select(Country, Life_Ladder) %>%
  inner_join(HDIndex, by = "Country") %>%
  ggplot(aes(x = HDI, y = Life_Ladder)) + geom_point()
```



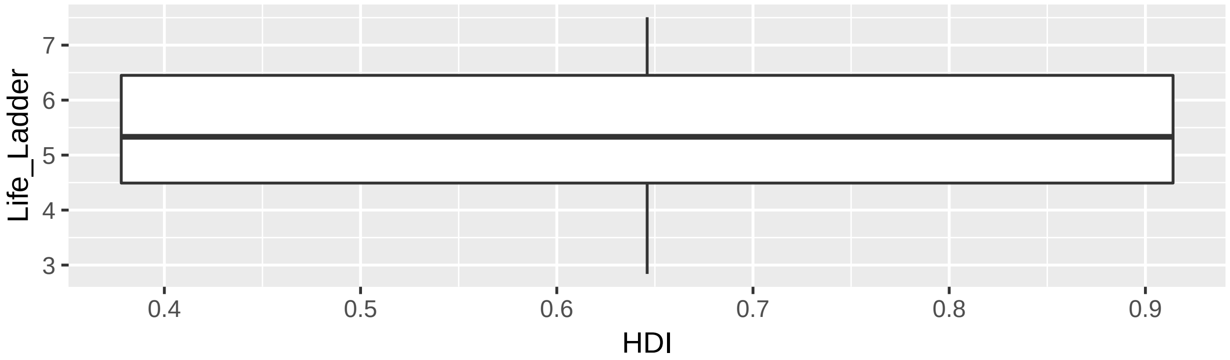
Plot B

```
WHRData %>%
  filter(year == 2014) %>%
  rename( Country = country) %>%
  inner_join(HDIndex, by = "Country") %>%
  ggplot(aes(x = HDI, y = Life_Ladder)) + geom_point()
```



Plot C

```
WHRData %>%
  filter(year == 2014) %>%
  rename( Country = country) %>%
  inner_join(HDIndex, by = "Country") %>%
  select(Country, HDI, Life_Ladder) %>%
  ggplot(aes(x = HDI, y = Life_Ladder)) + geom_boxplot()
```



(c) A data scientist would like to investigate the relationship in 2014 between the Human Development Index (`HDI`) and Happiness (`Life_ladder`). Which plot (A, B, or C) is the most appropriate vizualization of this relationship.? Explain your choice in one sentence. [2 marks]

A scatterplot is appropriate to vizualize the relationshiop between two quantitative variables. Plot B is the appropriate vizualization since it's a scatterplot that only includes data from 2014.

Question 3

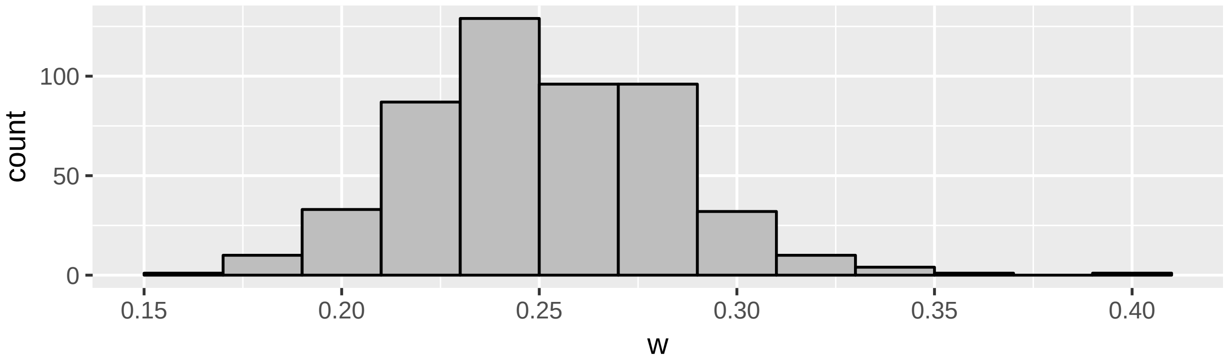
Countries with very high human development have an `HDI` value of at least 0.80. A recent article stated that 25% of countries have very high human development. Is this statement supported by the 2014 data on human development in `HDIndex`? [15 marks]

The R code below was used to investigate this question.

```r
HDIndex %>% select(Country, HDI) %>%
  mutate(vhigh = ifelse(HDI >= 0.80, "Yes", "No")) %>%
  group_by(vhigh) %>%
  summarise(n = n()) %>%
  mutate(prop = n / sum(n))
```

```
## # A tibble: 2 x 3
##   vhigh     n  prop
##   <chr> <int> <dbl>
## 1 No      139 0.739
## 2 Yes      49 0.261
```

```r
set.seed(101)
N <- 500
w <- rep(NA, N)
m <- 188

for (i in 1:N) {
  x <- sample( c("Yes", "No"),
               size = m,
               replace = TRUE,
               prob = c(0.25, 0.75))
  p <- sum(x == "Yes") / m
  w[i] <- p
}

dat <- data_frame(w)

dat %>%
  ggplot(aes(x = w)) +
  geom_histogram(binwidth = 0.02, colour = "black", fill = "grey")
```



```
y <- as.numeric(
  HDIndex %>% select(Country, HDI) %>%
  mutate(vhigh = ifelse(HDI >= 0.80, "Yes", "No")) %>%
  group_by(vhigh) %>%
  summarise(n = n()) %>%
  mutate(prop = n / sum(n)) %>% filter(vhigh == "Yes") %>% select(prop)
  )

y
```

```
## [1] 0.2606383
```

```
dat %>% filter(w >= y) %>% summarise(n = n())
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   200
```

```
dat %>% filter(w <= 0.25 - y) %>% summarise(n = n())
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1     0
```

```
dat %>% filter(w <= (0.25 - (y - 0.25))) %>% summarise(n = n())
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   160
```

Answer the following questions based on the scenario in Question 3.

(a) State the null and alternative hypotheses. Define all the parameters. [4 marks]

$H_0 : p = 0.25$,$H_A : p \neq 0.25$, where p is the proportion of countries that have a very high `HDI`.

(b) The observed value of the test statistic is **0.2606383**. [1 mark]

(c) Complete the sentences by filling in the missing words.

Sentence 1: In the R code above the object `m` represents the **sample size**, and `p` represents **the propoprtion of countries with high `HDI` in a simulation**. [2 mark]

Sentence 2: The y-axis of the histogram above, labelled `count`, corresponds to the number of **number of simulations** in a histogram bin. [1 mark]

Sentence 3: The values shown in the histogram were calculated under the assumption that the **null** hypothesis is **true**. [2 mark]

(d) Calculate the P-value of the hypothesis test. Show how you obtained the value. [1 marks]

The number of simulations more extreme than test statistic in the positive direction is 200 and the number more extreme in the negative direction is 160. Therefore, the P-value $= \frac{200+160}{500} = 0.72$

(e) What can you conclude about the proportion of countries with very high human development in 2014? Write 2-3 sentences explaining your conclusions. [4 marks]

The hypothesis test had a large p-value, 0.72, indicating 72% of proportions were as large as the one observed using data simulated under the null hypothesis. The large P-value suggests that there is no evidence to reject the null hypothesis that the proportion of countries with very high human development is 0.25. Thus, there is no evidence to support rejecting the null hypothesis.

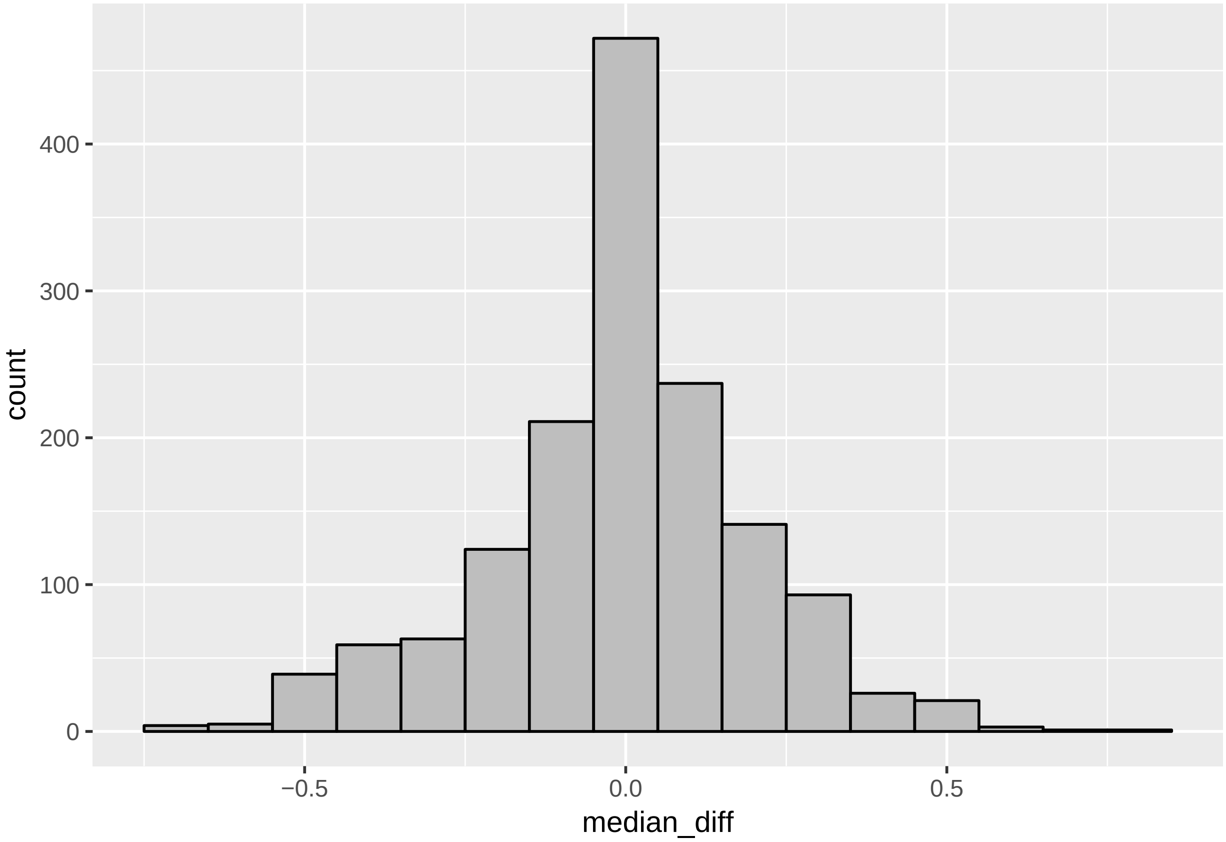
Question 4

Is there a difference in global happiness between 2017 and 2006? The variable `Life_Ladder` records the (weighted) average of happiness for a country. The R code below is a data analysis to investigate this question.

```r
WHRDat_1706 <- WHRData %>%
  filter(year == 2006 | year == 2017) %>%
  select(year, Life_Ladder)

WHRDat_1706 %>% group_by(year) %>% summarise(n = n(), median = median(Life_Ladder))
```

```
## # A tibble: 2 x 3
##    year     n median
##   <dbl> <int>  <dbl>
## 1  2006    89   5.08
## 2  2017   141   5.58
```

```r
set.seed(130)
K <- 1500
x <- rep(NA, K)

for (i in 1:K) {
  dat <- WHRDat_1706 %>% mutate(year = sample(year))

  y <- dat %>% group_by(year) %>%
    summarise(medians = median(Life_Ladder)) %>%
    summarise(sim_test_stat = diff(medians))

  x[i] <- as.numeric(y)
}

sim <- data_frame(median_diff = x)

sim %>% ggplot(aes(x = median_diff)) +
  geom_histogram(binwidth = 0.1, colour = "black", fill = "grey")
```



```
y <- as.numeric(WHRDat_1706 %>% group_by(year) %>%
                  summarise(medians = median(Life_Ladder)) %>%
                  summarise(x = diff(medians)))
y
```

```
## [1] 0.4947562
```

```
sim %>%
  filter(median_diff >= abs(y) | median_diff <= -abs(y)) %>%
  summarise(n() / K)
```

```
## # A tibble: 1 x 1
##   `n()/K`
##     <dbl>
## 1  0.0233
```

Answer the following questions based on the scenario in Question 4. [16 marks]

(a) State the null and alternative hypotheses. Make sure to define the all the statistical parameters explicitly. [4 marks]

$H_0 : M_{2017} = M_{2006}$, $H_A : M_{2017} \neq M_{2006}$, where M_{2017}, M_{2006} are the medians of `Life_Ladder` in 2017 and 2006 respectively.

(b) What test statistic is used in this hypothesis test? What is the observed value of the statistic? Make sure to define any symbols that you use. [2 marks]

The test statistic is the difference in the medians between 2017 and 2006. $\hat{M}_{2017} - \hat{M}_{2006} = 0.4947562$. $\hat{M}_{2017}, \hat{M}_{2006}$ are the sample medians in 2017 and 2006 respectively.

(c) How many of the simulations are more extreme than the observed value of the test statistic? Briefly explain your answer. [2 marks]

The P-value $= \frac{\text{\# simualtions more extreme than observed value of test statistic}}{\text{\# simulations}}$. Therefore, 0.0233*1500 = 35 are more extreme than the observed value of the test statistic.

(d) Is the following statement **TRUE** or **FALSE**? The P-value is the probability that the median of `Life_Ladder` is the same in 2006 and 2017. Circle the correct answer. [1 mark]

~~**TRUE**~~ **FALSE**

(e) Fill in the blanks. [3 marks]

A **type II** error would be made if there is **a real** difference between median happiness in 2006 and 2017, but the statistical test failed to reject the **null** hypothesis.

(f) Is there evidence of a change in happiness between 2006 and 2017? Write 2-3 sentences to justify your conclusion. [4 marks]

The hypothesis test had a low p-value, 0.0233, indicating only 2% of changes were as large as the one observed using data simulated under the null hypothesis. This is strong evidence to reject the null hypothesis. Therefore, there is evidence that global happiness is different in 2006 and 2017.