Lesson 2 Afghan Schools Solution

MA376: LTC James K. Starling

In 2007-2008 researchers tested the effect of village based schools vs. traditional government schools on test scores in northwestern Afghanistan. An initial sample of 34 villages were grouped into 12 equally sized village groups based on political and cultural alliances. Five village groups were randomly selected to establish a village based school. Unfortunately, due to an inter-village conflict, one group could never be surveyed and had to be dropped from the sample. As a result, the final sample consisted of 11 village groups (5 treatment and 6 control) and 31 villages (13 treatment and 18 control).

Step 1: Ask a research question

1) What was the research question for conducting this study?

Do village based schools have higher test scores than traditional government schools?

Step 2: Design a study and collect data

First we will bring in our data set:

```r
library(tidyverse)
afghan = read_csv("https://raw.githubusercontent.com/jkstarling/MA376/main/afghan_school.csv")
```

The data set contains the following variables:

Variable	Description	Units
girl	indicator variable for the sex of the child	1 if female, 0 if male
age	age of the child	years
num_people	number of people living in the household	people
jerib_cnt	number of jeribs of land owned by the household	jeribs
sheep_cnt	number of sheep owned by the household	sheep
treatment	indicator variable for if a child is in the treatment group (village school)	1 if in treatment group, 0 if in control group
test_score	final test score in the spring of 2008	points

2) Identify the response variable. Is this variable quantitative or categorical? (If categorical, note the number of categories. If quantitative, note the measurement units.)

The response variable is the number of points on the final test. This is a quantitative variable with units of "number of points."

3) Was this an observational study or an experiment? How are you deciding? If it was an experiement was it double blind, single blind, or not blinded?

This study is an experiment, each village was randomly assigned to one of two groups.

4) Identify the experimental units. How many are there?

The experimental units are children; n = 1394 children.

```
afghan %>% count()
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1  1394
```

5) Identify the treatment variable and it's levels.

The treatment variable is the group the each child was assigned to. the levels of the variable are 1 (village school) vs. 0 (traditional government school)

```
table(afghan$treatment)
```

```
##
##   0   1
## 697 697
```

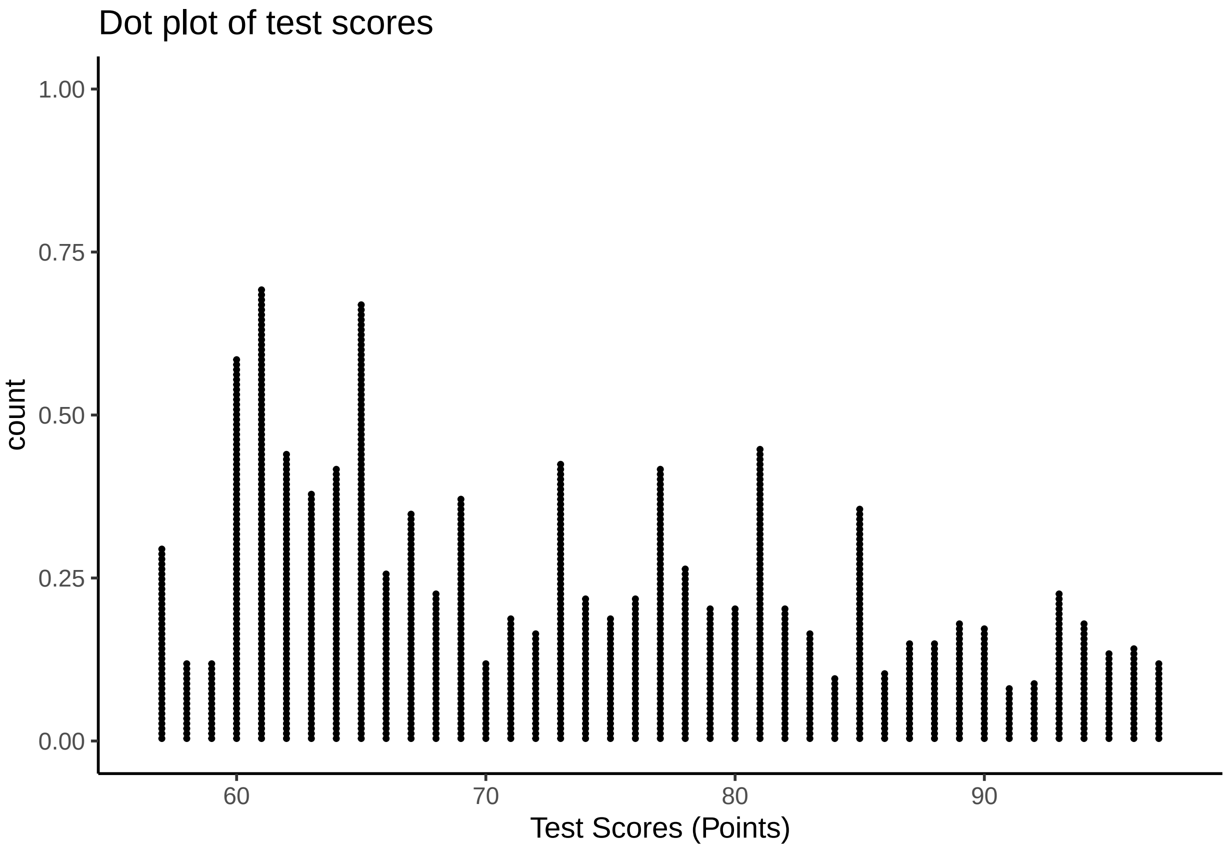
6) If planning the study, how would you determine who gets which treatment? What would you try to accomplish?

Randomly assign village groups to receive village schools. The goal is to create groups that are as similar as possible with regard to other variables, i.e., we want to remove confounding variables which allows us to make a cause-effect conclusion.

Step 3: Explore the data

**7) Create a dot plot of the outcomes of the response variable. If you can not see the full distribution add `dotsize = .1` inside `geom_dotplot()`. Note the axis labels on the y axis do not really have any meaning in dot plots made with ggplot. What does this plots tell you about your data? Summarize your observations in context of the problem.

```
afghan %>% ggplot(aes(x = test_score)) +
  geom_dotplot(dotsize = 0.2,binwidth = 1) +
  labs(x = "Test Scores (Points)", title = "Dot plot of test scores") +
  theme_classic()
```



The dot plot of test scores shows a relatively spread out distribution from 58 points to 95 points. The greatest number of observations are between 59 and 68 points and the number of observations taper off as the point values increase.

8) Calculate the mean and standard deviation for all test scores in the sample.

```
afghan %>% summarise(Mean = mean(test_score), SD = sd(test_score))
```

```
## # A tibble: 1 x 2
##    Mean    SD
##   <dbl> <dbl>
## 1  73.2  11.1
```

The average test score is 73.2 points and the standard deviation of test scores is 11.1 points

9) Create and specify a statistical model for predicting future results using the overall mean score for your sample and specifying the standard error of the residuals. Also, plot a histogram of the residuals of the model.

```
#Single Mean Model
single.model = lm(test_score~1, data = afghan)
summary(single.model)
```

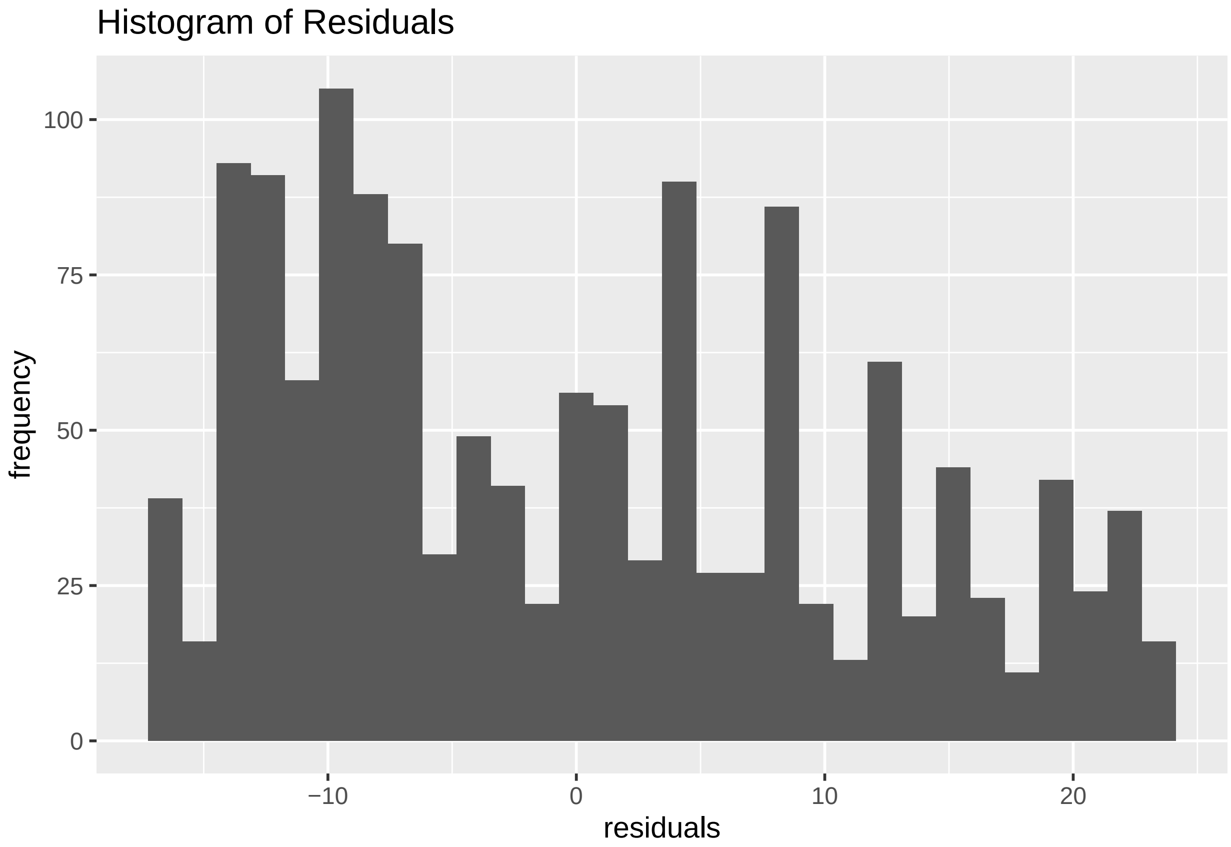
```
##
## Call:
## lm(formula = test_score ~ 1, data = afghan)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.244  -9.244  -1.244   7.756  23.756
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  73.2439     0.2984   245.4   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 11.14 on 1393 degrees of freedom
```

Predicted score = 73.24 points; SE of residuals = 11.14 points.

Plot residuals:

```
ggplot(data=afghan, aes(x = single.model$residuals)) +
  geom_histogram() +
  labs(title = "Histogram of Residuals", x = "residuals", y = "frequency")
```

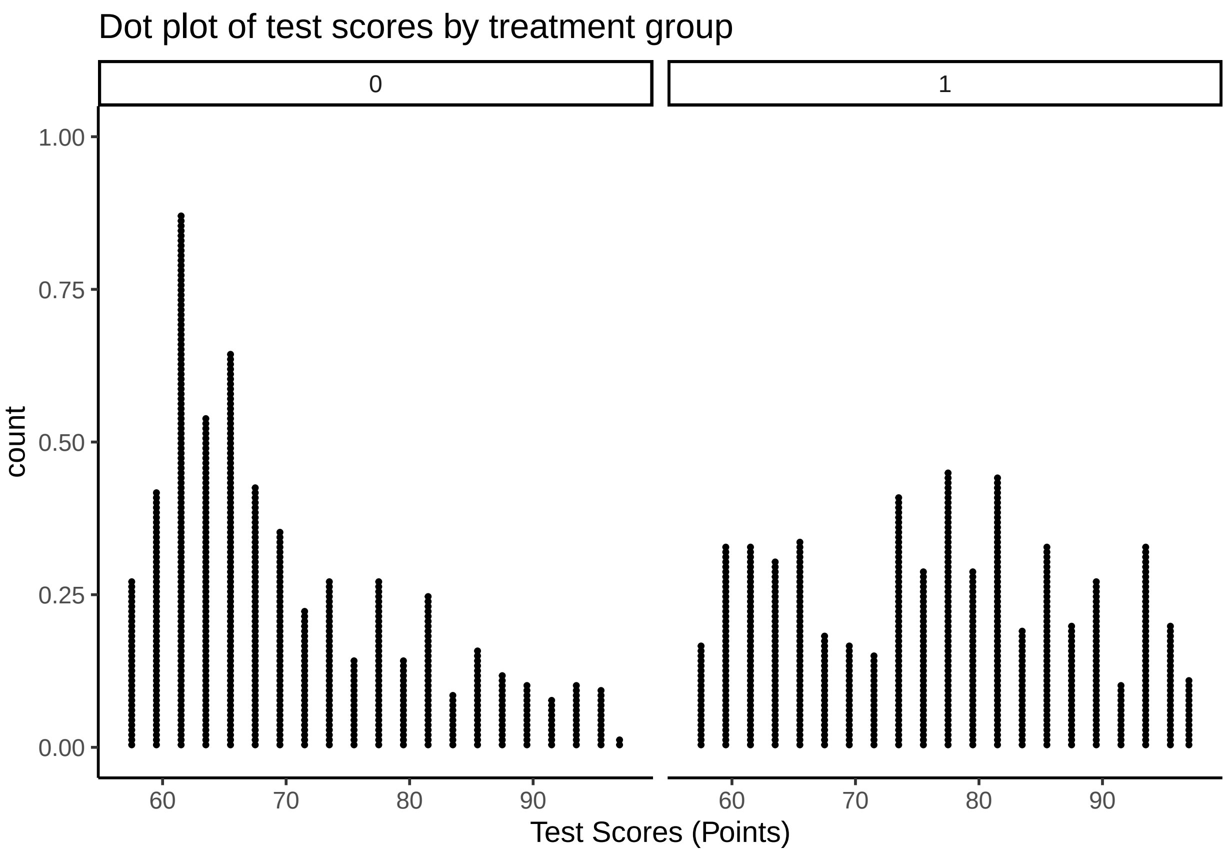
```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```



10) Recreate your dot plot from question 7 stratifying on the treatment variable. Hint: you can separate the plots by adding `facet_wrap(~treatment)` to your ggplot code.

```
afghan %>% ggplot(aes(x = test_score))+
  geom_dotplot(dotsize = 0.3)+
  facet_wrap(~treatment)+
  labs(x = "Test Scores (Points)",
       title = "Dot plot of test scores by treatment group")+
  theme_classic()
```

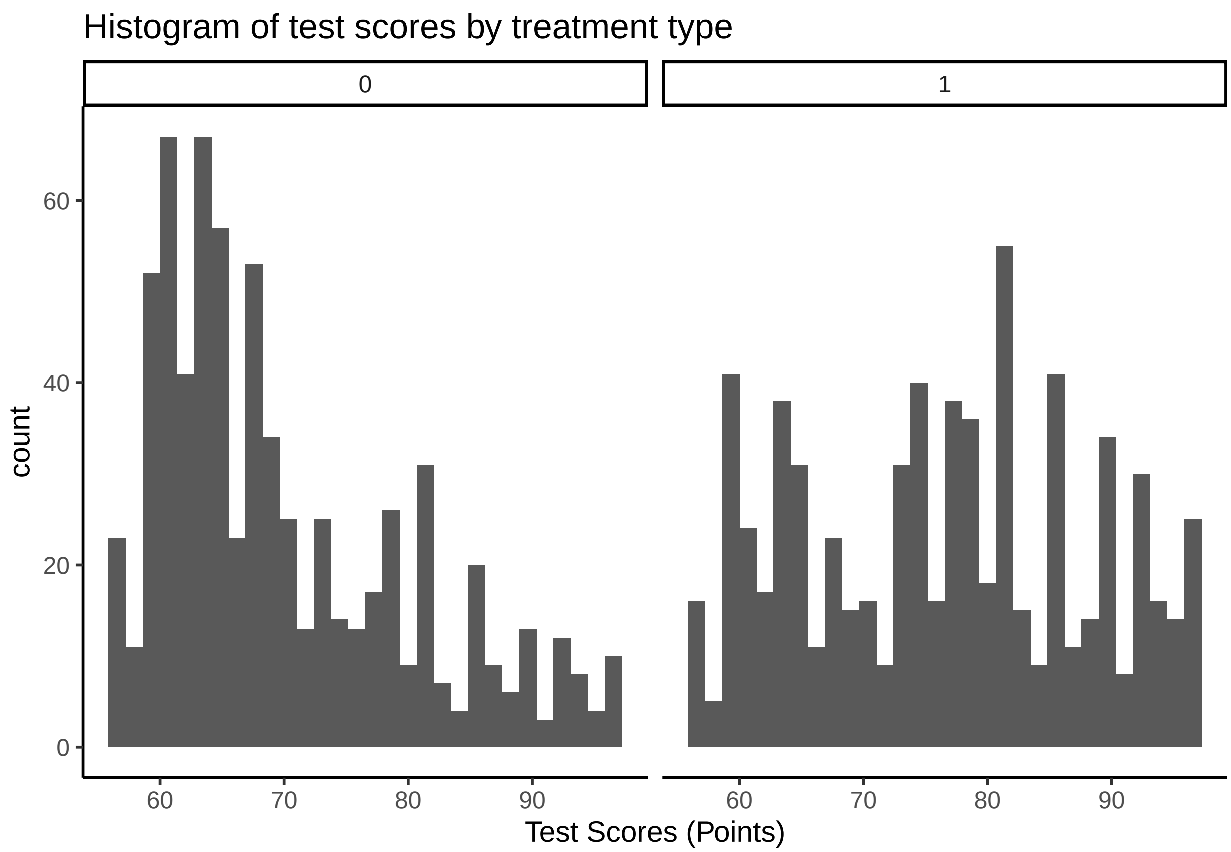
```
## Bin width defaults to 1/30 of the range of the data. Pick better value with 'binwidth'.
```



11) Create the same plot as a histogram. What do these two plots tell you about your data? Summarize your observations in context of the problem.

```
afghan %>% ggplot(aes(x = test_score))+
  geom_histogram()+
  facet_wrap(~treatment)+
  labs(x = "Test Scores (Points)",
       title = "Histogram of test scores by treatment type")+
  theme_classic()
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



The dot plot and histogram of the control group (traditional government schools) has the bulk of the observations between 60 and 70. The distribution for the treatment group (village Schools) appear to be pretty evenly spread from 60 to 95. This indicates that on average test scores are higher in the treatment group.

12) Calculate the mean and standard deviation of test scores for each group of the treatment variable. Hint: you can use `group_by()` to group by treatment prior to any calculating values. Based on the group means, did one of the groups tend to score higher than the other? By a lot or just a little? Which group had more variable results?

```
afghan %>%
  group_by(treatment) %>%
  summarise(Mean = mean(test_score), SD = sd(test_score))
```

```
## # A tibble: 2 x 3
##   treatment  Mean    SD
##       <dbl> <dbl> <dbl>
## 1         0  70.0  10.0
## 2         1  76.5  11.3
```

The control group has a mean 70 points and a standard deviation of 10 points while the treatment group has a mean of 76.5 points and a standard deviation of 11.3 points. The treatment group had a higher average score by 6.5 points and had slightly more variability than the control group (11.3 points vs 10 points)

13) Create and specify a statistical model for predicting outcomes depending on which treatment condition someone is assigned to, using the treatment-specific mean scores ("school grouping model"). Hint: you will need to ensure that the treatment variable is a factor variable.

```r
# Separate Means Model
multi.model = lm(test_score ~ 0 + as.factor(treatment), data = afghan)
summary(multi.model)
```

```
##
## Call:
## lm(formula = test_score ~ 0 + as.factor(treatment), data = afghan)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.485  -9.003  -1.485   7.997  26.997
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## as.factor(treatment)0   70.003      0.404   173.3   <2e-16 ***
## as.factor(treatment)1   76.485      0.404   189.3   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 10.66 on 1392 degrees of freedom
## Multiple R-squared:  0.9793, Adjusted R-squared:  0.9793
## F-statistic: 3.294e+04 on 2 and 1392 DF,  p-value: < 2.2e-16
```

$$\text{Predicted } testscore = \begin{cases} 70.003 \text{ points}, & \text{if traditional school} \\ 76.485 \text{ points}, & \text{if village school} \end{cases}$$

SE of Residuals = 10.66 points

14) Is the standard error of the residuals for the school grouping model smaller than the standard deviation of the residuals in the single mean model?

Yes, the SE of the residuals is less in the multiple means model than in the single mean model.

15) Does knowing which treatment group each person was assigned to explain all the variation in responses? How are you deciding?

No, there is still variation in the scores within each treatment group.

Step 4: Draw inferences beyond the data

We will discuss this in section 1.3.

Step 5: Formulate conclusions.

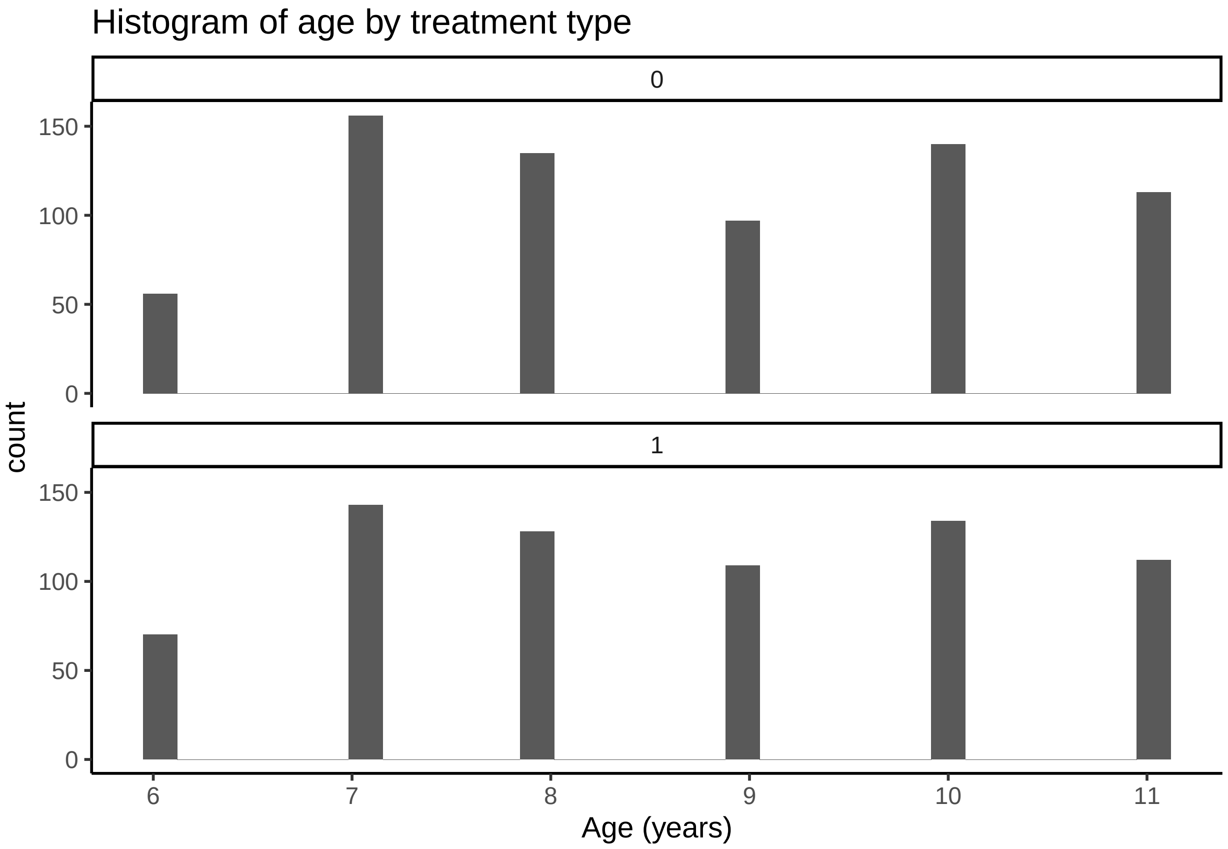
16) Summarize your "school grouping model" with a Sources of Variation Diagram:

Observed variation in:	Sources of explained variation:	Sources of unexplained variation:
Test score (in points)	Treatment Group: traditional school vs. village school	Sex
		Age
Inclusion criteria: Students enrolled in school in selected villages of northwestern Afghanistan (6-11 y.o. in Ghor Province)		Land ownership
Design: Students were enrolled in the same school during 2007 and 2008		Unknown

17) Examine the distributions of age for the two treatment groups. Does age appear to be a confounding variable in this study? How are you deciding?

```r
afghan %>% ggplot(aes(x = age))+
  geom_histogram()+
  facet_wrap(~treatment, ncol = 1)+
  labs(x = "Age (years)",
       title = "Histogram of age by treatment type")+
  theme_classic()
```

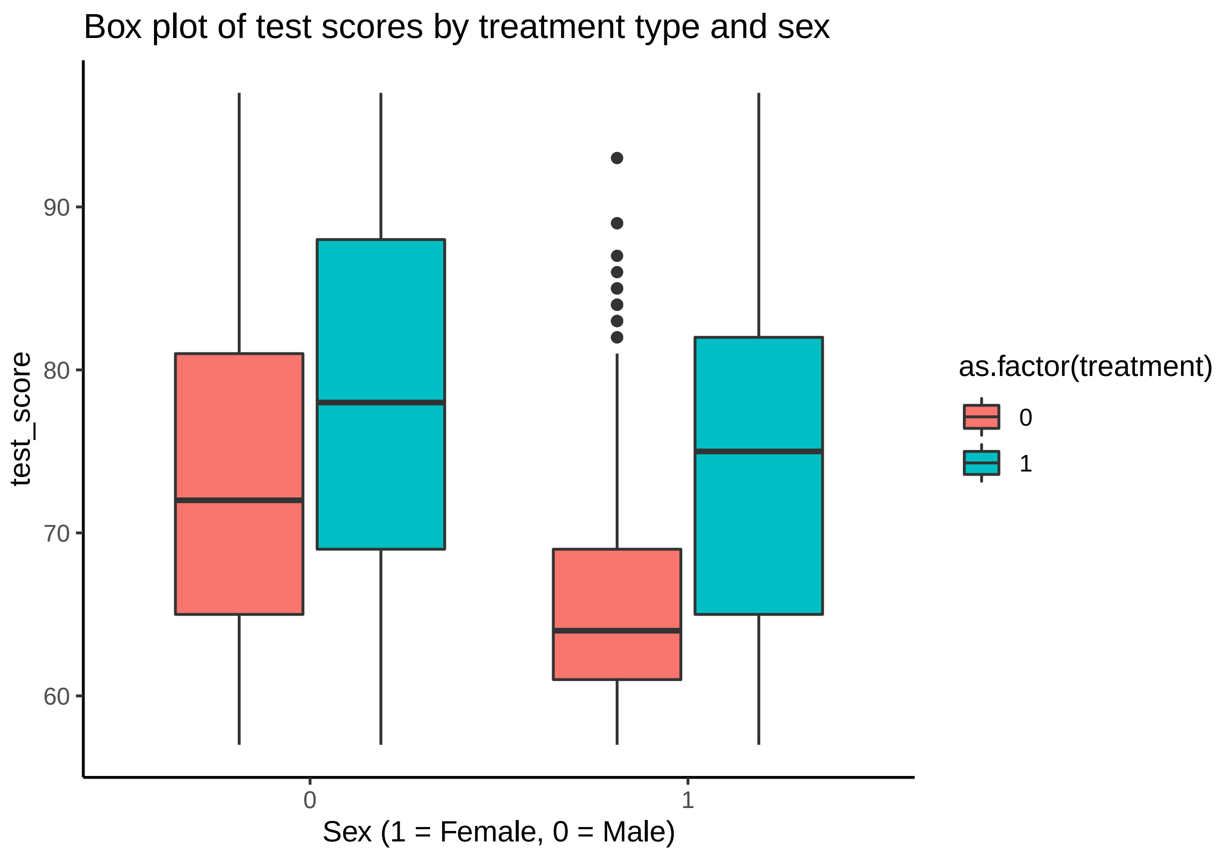
```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```



Age does not appear to be a confounding variable. The distribution of ages appears to be very similar in both groups.

18) Is Sex a confounding variable in this study? How are you deciding? Hint: if you are creating a plot, ensure that both the variable for sex and the variable for treatment are factors

```
afghan %>%
  ggplot(aes(x = as.factor(girl), y = test_score,
             fill = as.factor(treatment)))+
  geom_boxplot()+
  labs(x = "Sex (1 = Female, 0 = Male)",
       title = "Box plot of test scores by treatment type and sex")+
  theme_classic()
```



Sex does not appear to be a confounding variable. The group enrolled in village schools had higher test scores regardless of sex.

19) Could there be another explanation, apart from the type of school, that could explain the difference in the group means that we found? Explain.

Yes, see the sources of variation diagram (unexplained sources of variation). Ideally, random assignment eliminates confounding variables.

Step 6: Look back and ahead

20) Suggest at least one way you would improve this study if you were to carry it out yourself.

We could improve this study by using a larger sample. Also, we randomly assign school type at the individual student level.

References

Burde, Dana, and Leigh L Linden. “Bringing Education to Afghan Girls: A Randomized Controlled Trial of Village-Based Schools.” American Economic Journal: Applied Economics 5, no. 3 (July 2013): 27–40. https://doi.org/10.1257/app.5.3.27.