Lesson 2 Afghan Schools

CDT I. M. Smrt

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?

Step 2: Design a study and collect data

First we will bring in our data set:

```
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.)

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?

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

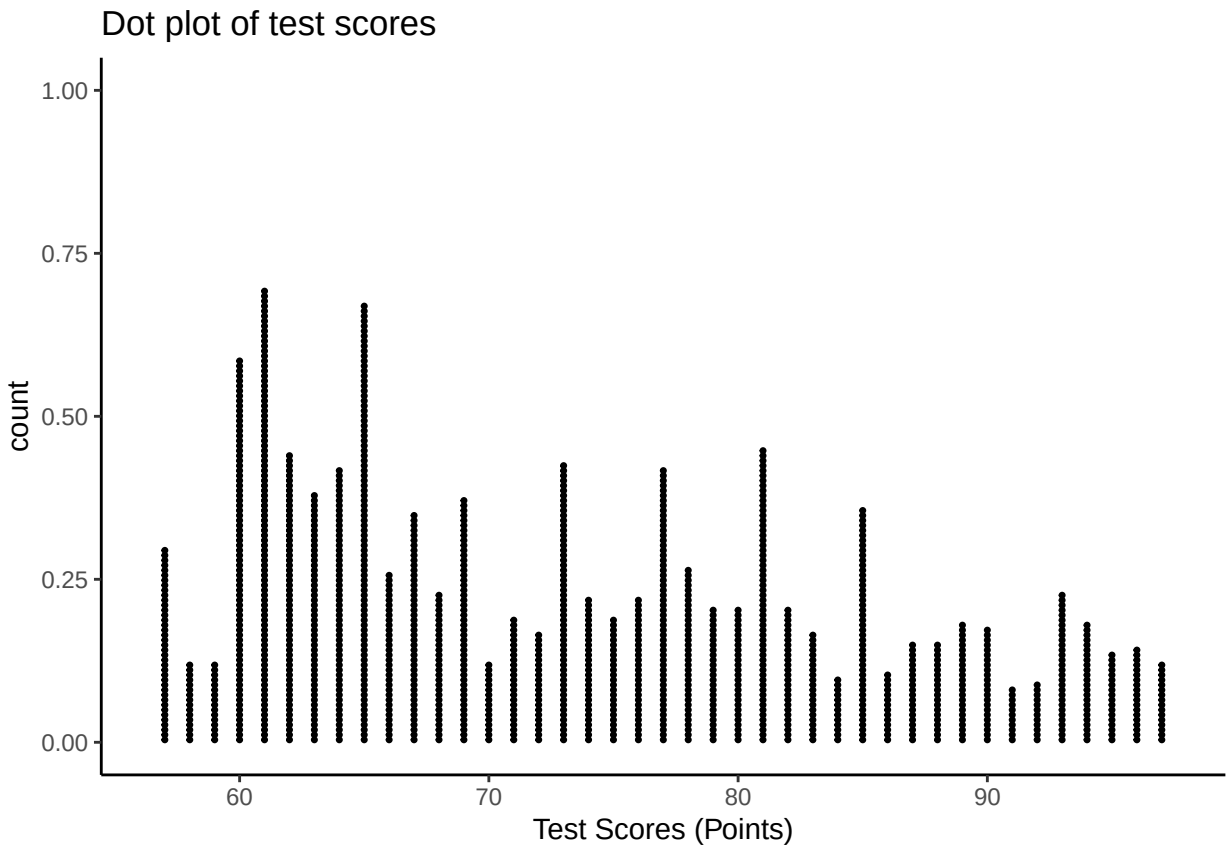
5) Identify the treatment variable and it's levels.

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

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 = .2` 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()
```



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
```

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.

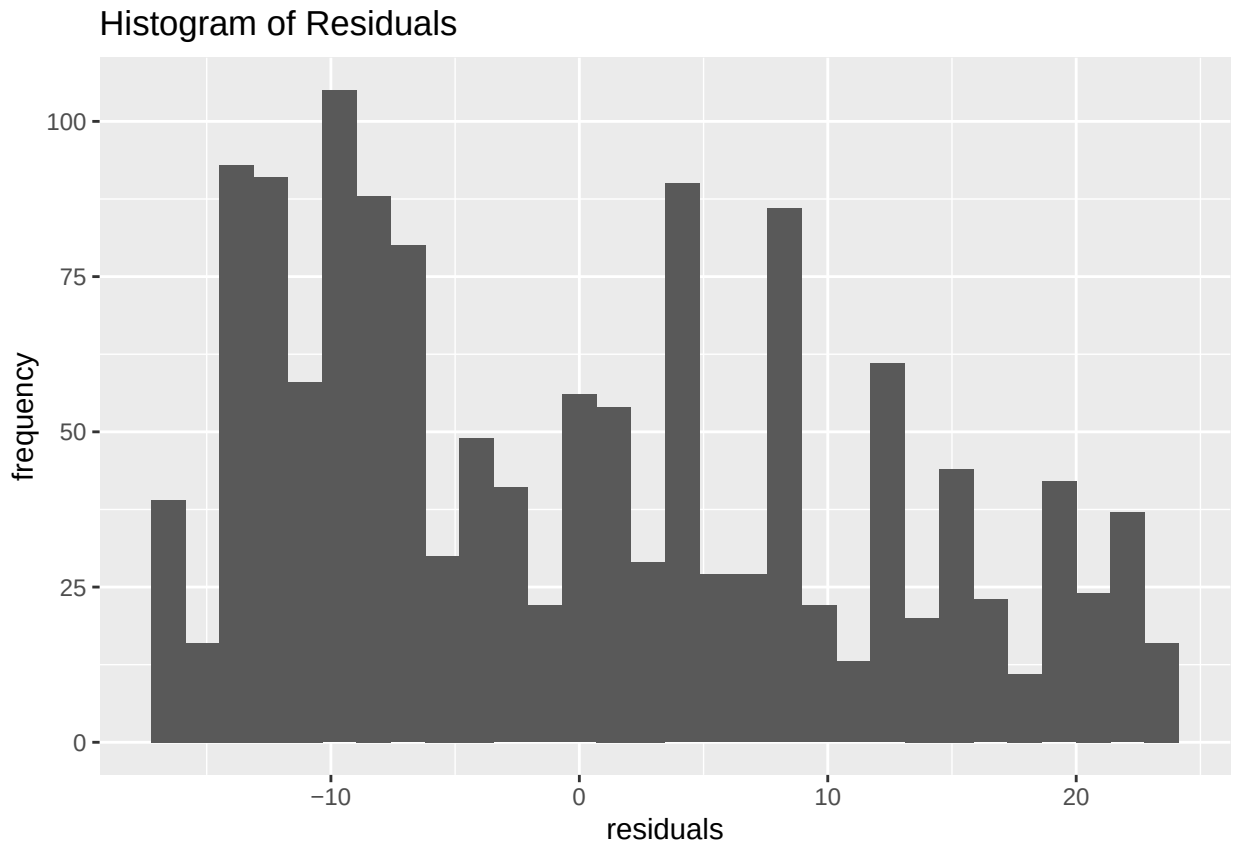
```r
#Single Mean Model
single.model = lm(test_score~1, data = afghan)
summary(single.model)
```

```
##
## 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
```

Plot residuals:

```r
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.

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.

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?

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.

```
# 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
```

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?

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

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:
Inclusion criteria: Design:		

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? Hint: Use a histogram.

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 (use a boxplot).

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.

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.

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.