# CollegeDistance Dataset

## The Dataset

The *CollegeDistance* dataset—stored in the *AER* library—looks at the relationship between several variables for 4,739 high school Seniors from over 1,100 schools. Including in the dataset are the following variables.

- *gender*: A factor indicating gender.
- *ethnicity*: A factor indicating ethnicity with three levels: **afam** (African-American), *hispanic* (Hispanic), and *other*.
- *score*: A composite test score on an achievement test.
- *fcollege*: A factor indicating if the father of the student was a college graduate.
- *mcollege*: A factor indicating if the mother of the student was a college graduate.
- *home*: A factor indicating if the family of the student owns their home.
- *urban*: A factor indicating if the student's school in an urban area.
- *unemp*: The student's county unemployment rate in 1980 (The year of the initial survey).
- *wage*: The student's state hourly wage in manufacturing in 1980.
- *distance*: The student's distance from a four-year college in tens of miles.
- *tuition*: The student's average state four-year college tuition in thousands of U.S. dollars.
- *education*: The student's ultimate years of education, determined from a follow-up survey in 1986.
- *income*: A factor indicating if the student's family income was greater than $25,000.
- *region*: A factor indicating student's region, with levels **West** and **other**.

# Potential Questions

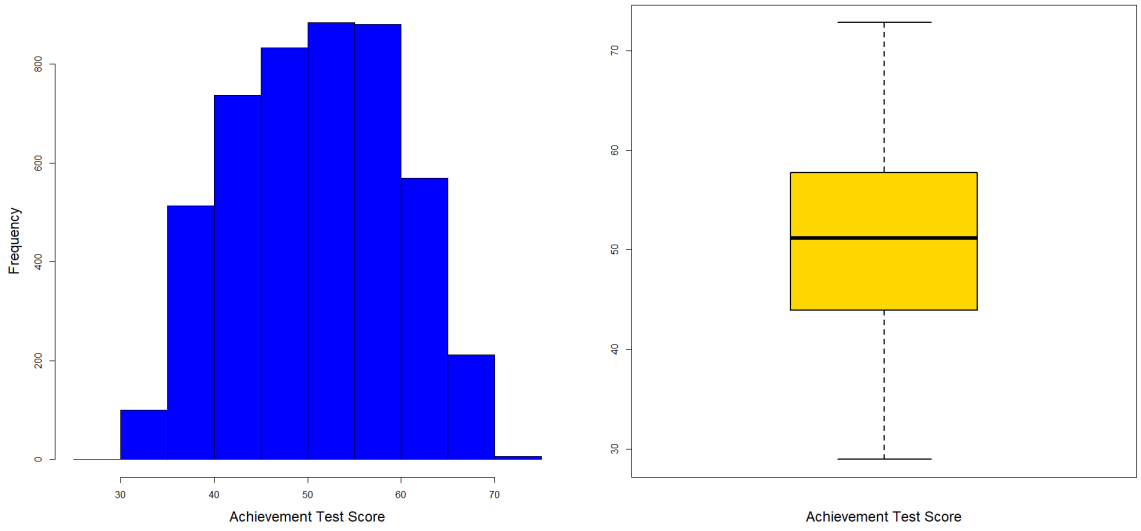
With a wide variety of quantitative and categorical characters, there are many different questions that students can answer. A short list of possible questions is given below, with the technique used in parentheses and the questions we will walk through marked in bold.

1. **Does achievement test score differ from 50? (1-sample mean)**
2. Does the proportion of mothers who graduated from college in 1980 differ from the proportion of women who graduate from college today?
3. Does achievement test score differ across genders? (2-sample mean)
4. Does achievement test score differ depending on if the student's father went to college? (2-sample mean)
5. **Does achievement test score differ depending on if the student's family owns their home? (2-sample mean)**
6. Does the proportion of mothers who graduate from college differ from the proportion of fathers who graduate from college? (2-sample proportion)
7. **Does the proportion of mothers who graduate from college differ depending on region? (2-sample proportion)**
8. Is there an association between county unemployment and achievement test score? (Correlation)
9. **Is there an association between distance from a four-year college and achievement score? (Correlation)**
10. **What is the relationship between parental and child education?***
11. How does family socioeconomic status affect a child's education?*

** I would denote this a challenge question. Challenge questions are ones that, due to the open-ended nature of the question or coding that can stretch the student or may not even be in book, challenges students to a certain degree.*

# Question 1

The achievement test given to the high school Seniors in this dataset ranges from 0 to 100. While not explicitly stated, it seems reasonable that the scores would be centered around 50. A quick look at either a histogram or a boxplot of the dataset bears this thought out.




The *score* variable seems to be unimodal and symmetric, maybe even Normally distributed, centered around 50 with a range from 30 to 70 (A guess at the standard deviation would be that it's near 10). If we were to set up a hypothesis test to see if the average test score is 50 versus not 50 tested at the $\alpha = 0.05$ level, we would have

$$H_0 : \mu_{Score} = 50$$
$$H_A : \mu_{Score} \neq 50$$

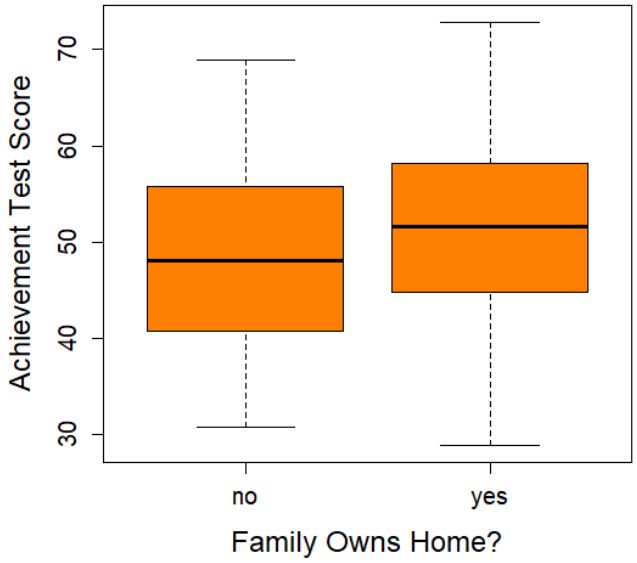
By using our *t.test* function, we find that our test statistic is $t = 7.0331$ with a p-value of $2 \times P\big(t_{4738} > |7.0331|\big) = 2.309 \times 10^{-12}$. With this in mind, we would reject the null hypothesis and conclude that the true average test score is not equal to 50. It's worth noting that the sample average for score is $\bar{x}_{Score} = 50.88903$, which is really close to the null hypothesis value of 50. This in part illustrates the difference between statistical significance and practical significance. If the true average was $\mu = 50.01$ you would eventually collect enough data to reject the null hypothesis given above, but for most practical purposes those values are identical. The answers to your questions have to be considered in the context of the data and practically.

```
library(AER)
data(CollegeDistance)
hist(CollegeDistance$score, xlab=''Achievement Test Score'',
main='''')
boxplot(CollegeDistance$score, xlab=''Achievement Test Score'')
t.test(CollegeDistance$score, mu=50)
```

# Question 5

Unfortunately, socioeconomic status can be a predictor of educational achievement. In this dataset, we have one particular indicator of socioeconomic status: home ownership. It then becomes a reasonable question if home ownership—as a proxy of socioeconomic status—results in different achievement test scores.

Looking at a side-by-side boxplot of achievement test score split out by home ownership, we do see that the median score for students whose family owns their home is higher than for those who don't. The gap is somewhat small, but still may be statistically significant.




In setting up our hypothesis test, we'll test if the difference in *score* for non-owners minus owners is equal to or less than 0. Written out, this will be

$$H_0 : \mu_{No} - \mu_{Yes} = 0$$
$$H_A : \mu_{No} - \mu_{Yes} < 0$$

tested at the $\alpha = 0.05$. Before going to *R*, the last thing we'll need to decide is whether to test this under the equal variances assumption or not. Doing a little calculation, we find that our variances are $s^2_{No} = 80.827$ and $s^2_{Yes} = 73.162$. This implies—by the rule of thumb—that the variances are equal. Thus we can use the *t.test* function, taking care with how we define the *alternative* argument.

The function returns that our test statistic is $t = -8.7328$, following a t-distribution with $\nu = 4737$ degrees of freedom. Also from the function comes of p-value of $P\big(t_{4737} < -8.7328\big) < 2.2 \times 10^{-16}$. Thus, we'd reject the null hypothesis and conclude that the mean score for students whose families do not own their home is lower than that for those who do.

Checking the assumptions using the *table* function, we can see that the assumptions check out as 3,887 families in the sample own their home while 852 do not. These are well above our thresholds for the Central Limit Theorem to kick in.

```
library(AER)
data(CollegeDistance)
boxplot(score~home, data=CollegeDistance,
ylab=''Achievement Test Score'',
xlab=''Family Owns Home?'', col=''darkorange1'',
cex.lab=1.5,cex.axis=1.25)
var(CollegeDistance$score[CollegeDistance$home==''no''])
var(CollegeDistance$score[CollegeDistance$home==''yes''])
t.test(score[home==''no''],score[home==''yes''],
data=CollegeDistance, alternative=''less'',
var.equal=TRUE)
table(CollegeDistance$home)
```

# Question 7

Education varies greatly by region in the United States. Statewide percentage of bachelors degree attainment ranges from the low twenties to the low forties. In this case, we may be interested in if education attainment of mothers in our dataset differs across the various regions—here limited to the West region versus all others. A quick look at the contingency table—the variable *mcollege* on the rows and *region* on the columns—for the two variables shows the following;

| | West | Other |
|---|---|---|
| Yes | 122 | 529 |
| No | 821 | 3267 |

We can do a little basic calculation to find our sample proportions of mothers who graduated from college split out by region: $\hat{p}_{West} = \dfrac{122}{122+821} = 0.1294$ and $\hat{p}_{Other} = \dfrac{529}{529+3267} = 0.1394$. While these two proportions are very similar, in large sample sizes its possible to pick up on small differences and reject null hypotheses. Specifically the hypotheses that we'll look at—at the $\alpha = 0.1$ level—are

$$H_0 : p_{West} - p_{Other} = 0$$
$$H_A : p_{West} - p_{Other} \neq 0$$

The *prop.test* function gives us our results, although with a slight adjustment. The function gives us that the test statistic will be $t = \sqrt{0.63525} = -0.7970$, with the negative known

because $p_{West} < p_{Other}$. We should note that the reason why the function returns the test statistic squared is that the *prop.test* function is doing a chi-squared test of independence on the contingency table, with the results being identical to the two-sample test for proportions.

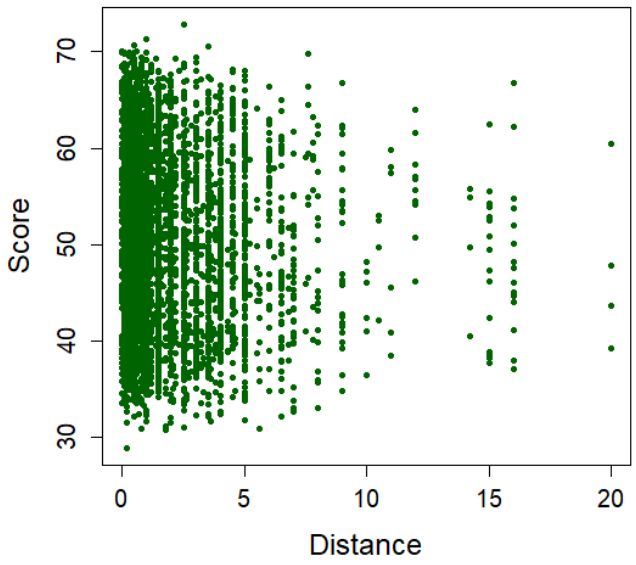
From our test statistic and alternative hypothesis, we can get that our p-value is $2 \times P\big(N(0,1) > |-0.797|\big) = 0.4254$. Because of this, we would fail to reject the null hypothesis and conclude that it's plausible that there's no difference between the proportions of mothers who graduated college for the West and Other regions.

An important part of these tests is making sure that the assumptions of these tests are met. For this test, we just need to make sure that the sample size for both samples is greater than 30—which is true—and that we observe at least a bare minimum observed "successes" and "failures"—which is also true.

```
library(AER)
data(CollegeDistance)
table(CollegeDistance$mcollege,
CollegeDistance$region)
prop.test(prop.test(x=c(529,122),
n=c(529+3267,122+821),
correct=F)
sqrt(0.63525)
```

# Question 9

This dataset is named *CollegeDistance*, so it seems reasonable to do some analysis on the title variable. So, let's look at the relationship between distance to the nearest four-year college and achievement test score. Both of these are quantitative variables, so this question comes down to one of correlation. Looking at the scatterplot of the two variables, we would at most expect a weak correlation.

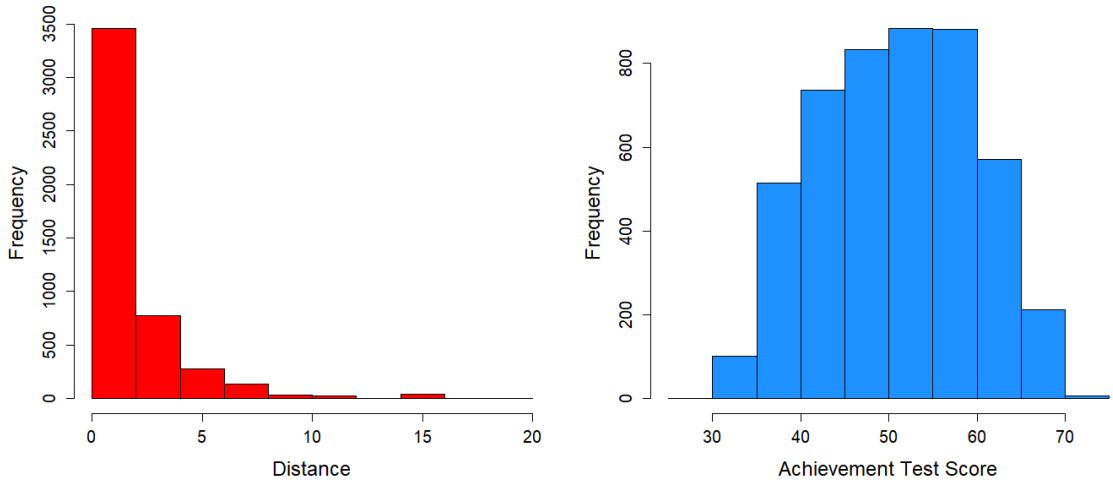



The actual sample correlation bears this out, with $r = -0.06798$. However, we should note that in large sample sizes statistical significance is easier to find. Our hypotheses to investigate if there's an association between *distance* and *score* would be

$$H_0 : \rho = 0$$
$$H_A : \rho \neq 0.$$

Let's say that we test this at the $\alpha = 0.01$ level. The *cor.test* function gives us a test statistic of $t = -4.6896$ with a p-value of $2 \times P\big(t_{4737} > |-4.6896|\big) = 2.815 \times 10^{-6}$. Thus, we would reject the null hypothesis and conclude that the correlation between distance from a four-year college and achievement test score is non-zero. This is an instance where practical significance and statistical significance may be at odds.

Even further, checking the assumptions of this test bring the results into further doubt. While the sample size is sufficient and the scatterplot shows a linear relationship, we can easily see that distance from a four-year college is heavily skewed to the right. This violates the assumption that the two variables are normally distributed.




I have little doubt that were we to know the exact degrees of freedom for the correct t-distribution, or were we to use bootstrapping to get an estimated p-value, we would still reject the null hypothesis. However, the violation of the assumptions still needs to be acknowledged in the analysis.

```
library(AER)
data(CollegeDistance)
plot(CollegeDistance$distance, CollegeDistance$score,
pch=20,xlab=''Distance'', ylab=''Score'',
col=''darkgreen'')
cor(CollegeDistance$distance, CollegeDistance$score)
cor.test(CollegeDistance$distance, CollegeDistance$score)
hist(CollegeDistance$distance, main='''',
xlab=''Distance'', col=''red'')
hist(CollegeDistance$score, main='''',
xlab=''Achievement Test Score'',
col=''dodgerblue'')
```

# Question 10

As was mentioned, this is a challenge question. There are many routes that one can go in answering the question "What is the relationship between parental and child education?" But we need to start with another question: how do we define child education?

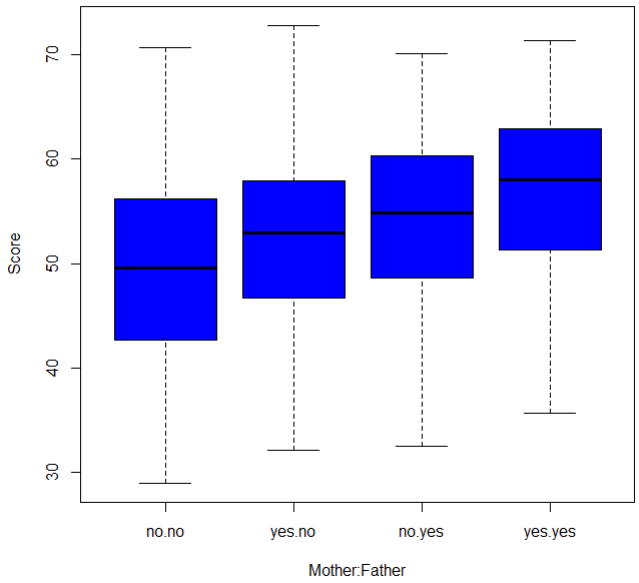
This isn't a trivial question, as our answer will affect the analyses that are possible as well as the variables we will use. In my mind, there are two options: educational achievement as measured by the achievement test administered (*score*) and educational achievement as measure by the followup years of education (*education*).

We already know a fair bit about the *score* variable, so let's look at *education*. Looking at the variable in *R*, it appears that education is a numeric variable ranging from 12 to 18. However, a little more digging shows that there's a little more to the variable. Looking at the description of the dataset online, we see that *education* is better described as a categorical variable, with each number in the dataset corresponding to an educational level—12 is high school, 13 is a vocational degree, etc. Treating this variable as a quantitative variable would be incorrect, so we'll need to adjust things a bit. Since the dataset has variables about whether the student's mother and father graduated from college, let's create a variable *college* that describes if the student in question graduated with at least a bachelor's degree.

So we now have two variables we can consider to define student education. Now, we need to define parental education. This seems trivial, but it brings some challenges as well. Namely, we have two variables that make up parental education: *mcollege* and *fcollege*. If we combine them into a single variable, we'll have four groups.

| | Father No | Father Yes |
|---|---|---|
| Mother No | Group 1 | Group 2 |
| Mother Yes | Group3 | Group 4 |

There are few ways to look at this. We could run two hypothesis tests, comparing the response based on mother's education and father's education. This does leave out some information, namely, does mother's and father's education interact? We can partially investigate this looking at the side-by-side boxplot of the *score* for these four groups.




Looking at these boxplot, I don't see much evidence for an interactive effect, at least based on achievement test score. With this in mind, we'll limit our scope to testing score

based on parental education separately, with our alternative hypotheses both being of the form $H_A : \mu_{No} - \mu_{Yes} < 0$—implying that parental education is positively associated with student achievement—and tested at the $\alpha = 0.01$ level. These tests result in p-values of $6.91 \times 10^{-37}$ for mother's education and $2.77 \times 10^{-69}$ for father's education, thus we'd reject both null hypotheses and conclude that the average test score for student's whose mother graduated from college was greater than those whose mother did not graduate from college, with a similar result for fathers.

It should be noted that although the probability of a Type I error for each individual test is $\alpha = 0.01$, the probability of making at least one Type I error across the two tests is $1 - 0.99^2 = 0.0199$. This is due to the classical multiple testing problem: the fact that as you do more hypothesis tests to answer a question, the more likely you are to make at least one Type I Error. Really, this isn't the best technique to answer this question. Rather, we should be doing an Analysis of Variance (ANOVA), which better attacks this question without having problems with multiple testing.

Of course, achievement was only one way to look at the question of parental and child education. We can also investigate the *college* variable we created earlier. Again, if we only concentrate on the main effects of mother's and father's education, we can do a pair of two-sample hypothesis tests. In this case, we'd test if the proportion of students who graduate from college is greater than for those whose mothers graduated from college versus those whose mothers did not. This would be

$$H_0 : p_{No} - p_{Yes} = 0$$
$$H_0 : p_{No} - p_{Yes} < 0$$

We then would test the same for fathers, testing both at the $\alpha = 0.01$ level. An initial look at the data suggests that we will likely reject the null hypotheses.

| | Student No | Student Yes |
|---|---|---|
| Mother No | 2713 | 1375 |
| Mother Yes | 250 | 401 |

| | Student No | Student Yes |
|---|---|---|
| Father No | 2581 | 1172 |
| Father Yes | 382 | 604 |

After using the *prop.test* function, we find that our p-values are $1.18 \times 10^{-42}$ and $2.55 \times 10^{-67}$ for mother's and father's education respectively. Thus, we would reject both null hypotheses again, while again noting that our probability of making a Type I error on these two tests combined is higher than the significance level. Just like with our first option, this is not the ideal way to answer this question. We could use the $\chi^2$ test for independence, or even better logistic regression where we would be able to quantify the effects of mother's and father's education on the probability that a student receives a college degree.

In any event, it does seem that parental education has a positive affect on student's education, whether quantified through achievement testing or degree attainment. The degree to

which it's important is hard to state, as is the question of whether the affect is cumulative or interactive in any way. Exploratory analyses seem to suggest that, yes, the affect is cumulative (Having both parents hold a college degree looks like it potentially results in even higher achievement test scores than either parent singularly) but not interactive. Overall, while there does seem to be an effect, more fully quantifying this effect would require techniques not covered in the book.

```
library(AER)
data(CollegeDistance)
CollegeDistance$college=ifelse(CollegeDistance$education>13,''yes'',''no'')
boxplot(score~mcollege+fcollege,data=CollegeDistance,xlab=''Mother:Father'',
color=''blue'')
var(CollegeDistance$score[CollegeDistance$mcollege==''yes''])
var(CollegeDistance$score[CollegeDistance$mcollege==''no''])
var(CollegeDistance$score[CollegeDistance$fcollege==''yes''])
var(CollegeDistance$score[CollegeDistance$fcollege==''no''])
t.test(score~mcollege,data=CollegeDistance,
alternative=''less'',var.equal=TRUE)
t.test(score~fcollege,data=CollegeDistance,
alternative=''less'',var.equal=TRUE)
table(CollegeDistance$college,CollegeDistance$mcollege)
table(CollegeDistance$college,CollegeDistance$fcollege)
prop.test(table(CollegeDistance$college,CollegeDistance$mcollege),
correct=FALSE)
prop.test(table(CollegeDistance$college,CollegeDistance$fcollege),
correct=FALSE)
```