1. A model often used to predict the height in inches of an adult son is
\[ \hat{\text{height}} = \frac{1}{2} \text{father} + \frac{1}{2} \text{mother} + 2.5 \]

A different model suggests that the height of the father is more important in predicting the height of sons. That model is
\[ \hat{\text{height}} + \frac{2}{3} \text{father} + \frac{1}{3} \text{mother} + 2.5 \]

Of course neither model is “right.” Suppose that you have a lot of data (like Galton). How would you decide which of these two models is “better”? Be as precise as you can.

2. We might write Einstein’s law as \( E = mc^2 \) but we use the equation as if it is an exact equality (unlike the models in the first problem). Explain why you would not expect exact equality in any real application of this law.

3. Create a vector \( x \) in \( \mathbb{R} \) containing the following numbers:
\[ 1, 4, 7, 9, 13, 19, 21, 25 \]

For each of the following \( \mathbb{R} \) commands, try to predict what the output will be and then write down (using \( \mathbb{R} \) of course), what the output actually is. In each case where an explanation is asked for, say how \( \mathbb{R} \) is computing the answer that you get.

(a) \( x \)
(b) \( x+1 \)
(c) \( \text{sum}(x) \)
(d) \( x>10 \)
(e) \( x[x>10] \) Explain.
(f) \( \text{sum}(x>10) \) Explain.
(g) \( \text{sum}(x[x>10]) \) Explain.
(h) \( x[-(1:3)] \) Explain.
(i) \( x^2 \)

4. Data on all the counties in the United States are in the \text{counties} dataframe in the \text{Stob} package. Each row in
the dataframe is a county.

(a) How many counties are there?
(b) How many variables are in the dataframe?
(c) The variable \text{Population} has the population of each county as of the 2010 census. What was the total population of the United States in 2010?
(d) Why is a histogram of the populations of all the counties not very informative?
(e) Two of the variables are not really necessary – they can be computed from other variables in the dataframe. Name one of these and show how it can be computed easily from other variables.

5. The dataframe \text{Chile} in the \text{car} package (be sure to load this package) has data on a survey of voters in Chile conducted in the spring of 1988 before the election that unseated Augusto Pinochet. Use what you know about data frames and the \text{tally()} and \text{favstats()} functions to answer the following questions.

(a) What is a case and how many cases are there?
(b) What are the variables and which are quantitative and categorical?
(c) How many of each gender participated in the survey?
(d) What was the average income of those participating in the survey? What caveat would you want to mention when reporting this statistic?

(e) What percentage of the voters surveyed expected to vote against Pinochet? (Read the help document to determine which variable and value you should be looking at.)

(f) Voters with incomes greater than 40,000 pesos could be considered rich. What percentage of the rich voters expected to vote against Pinochet? (Hint: first make a dataframe with only rich voters.)

6. Suppose that we are studying baseball statistics. We might use two different units of analysis: the player or the team.

(a) Suppose that we are studying baseball with the unit of analysis being the player. Give an example of an appropriate categorical variable for this situation and an appropriate quantitative variable.

(b) Do the same in the case that we are studying baseball with the unit of analysis being the team.

7. In the parts below, we list some convenience samples of Calvin students. For each of these methods for sampling Calvin students, indicate one way in which the sample is likely not to be representative of the population of all Calvin students.

(a) The students in Mathematics 243A.

(b) The students in Nursing 329.

(c) The first 30 students who walk into the FAC west door after 12:30 PM today.

(d) The first 30 students you meet on the sidewalk outside Hiemenga after 12:30 PM today.

(e) The first 30 students named in the Calvin directory.

(f) The men’s basketball team.

8. Suppose that we were attempting to estimate the average height of a Calvin student. For this purpose, which of the convenience samples in the previous problem would you suppose to be most representative of the Calvin population? Which would you suppose to be least representative? Why?

9. Consider the set of natural numbers \( P = \{1, 2, \ldots, 50\} \) to be a population.

(a) How many prime numbers are there in the population?

(b) If a sample of size 10 is representative of the population, how many prime numbers would we expect to be in the sample? How many even numbers would we expect to be in the sample?

(c) Using R choose 5 different random samples of size 10 from the population \( P \). Record how many prime numbers and how many even numbers are in each sample. Make any comments about the results that strike you as relevant.

10. Sometimes we can improve on simple random sampling by incorporating randomness but in a more complicated way. As of Day 10 in the fall, the breakdown of the Calvin student body by level was given by the following table

<table>
<thead>
<tr>
<th>Class Level</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>First-year</td>
<td>1,078</td>
</tr>
<tr>
<td>Sophomore</td>
<td>975</td>
</tr>
<tr>
<td>Junior</td>
<td>821</td>
</tr>
<tr>
<td>Senior</td>
<td>1,054</td>
</tr>
<tr>
<td>Other</td>
<td>106</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>4,034</strong></td>
</tr>
</tbody>
</table>

If we were to choose a simple random sample, we might get too many juniors or not enough first-year students. But since we know the true proportion of each class, we might separately sample from each class to get an overall sample that is representative of the whole. If a sample of size 200 is desired, explain how many of each class we should choose. (Of course withing each class, we should choose a simple random sample. This sampling design is called **stratified random sampling**).
11. The dataframe Allegan (in the Stob package) has data on 57 years of weather in Allegan, MI. Each case is a day. The variable TMAX is the maximum temperature recorded on the given day.

(a) What is a 50% coverage interval for the maximum daily temperature?
(b) What is a 95% coverage interval for the maximum daily temperature?
(c) What is an interesting feature of the distribution of this variable? (Hint: draw a density plot.)

12. The following histograms and boxplots are of five different datasets. Match the boxplot to the histogram of the same data. Justify your choices with a sentence or two.

![Boxplots and histograms](image)

13. Sometimes it is useful to change the units of a variable. For instance, we might change from inches to feet or from degrees centigrade to fahrenheit. Obviously, such statistics as the mean, median, variance and standard deviation will change if we do that.

(a) A new variable $Y$ is created from a variable $X$ by multiplying each case of $X$ by a constant $c$ (i.e., $Y = cX$). How are the mean, median, variance and standard deviation of $Y$ related to those of $X$?
(b) A new variable $Y$ is created from a variable $X$ by adding a constant $d$ to each case of $X$. (i.e., $Y = X + d$). How are the mean, median, variance and standard deviation of $Y$ related to those of $X$?

14. Boxplots are fairly popular. But boxplots cannot show some of the more interesting, and often important, features of a variable. The data frame faithful has data on over 270 eruptions of the Old Faithful geyser in Yellowstone National Park. The variable eruptions gives the length in minutes of each of the eruptions. What interesting feature of the distribution of this variable is shown by a density plot but which cannot be seen in a box plot?

15. The mean of a variable is one choice for a simple model for that variable. But it is not the only one. For example, we might choose the median instead.

(a) Use the mean as a model for the GPA of Calvin seniors (using the data in the sr dataframe). Compute the sum of squares of residuals of this model.
(b) Use instead the median as a model for the GPA of Calvin seniors using the very same data. Compute the sums of squares of residuals for this model.
(c) Compare the sums of squares of residuals for these two models. Which is smaller? (It turns out the mean is the value that minimizes the sums of squares of residuals.)

16. The Current Population Survey data (CPS85) has data on both the wage of the respondent (wage) and the employment sector in which the respondent worked (sector). Suppose that we try to explain the variation in wages by using sector as an explanatory variable.
(a) What are the model values (i.e., $\hat{y}_i$) for each of the eight sectors?
(b) What is the variance of the wage variable?
(c) What is the variance of the residuals of the model?
(d) Which observation has the largest (in absolute value) residual?
(e) Construct a side-by-side boxplot of the wages by sector. Notice that the wage distributions for any two sectors overlap by quite a bit. For which sectors is it true that at least 75% of the respondents in that sector made more money than half of those respondents in the manufacturing sector.

17. This made-up data frame is so small that you can answer all the questions below without the use of a computer (or even a calculator). Do that.

<table>
<thead>
<tr>
<th>Sex</th>
<th>Job Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>5</td>
</tr>
<tr>
<td>F</td>
<td>7</td>
</tr>
<tr>
<td>F</td>
<td>9</td>
</tr>
<tr>
<td>M</td>
<td>4</td>
</tr>
<tr>
<td>M</td>
<td>8</td>
</tr>
</tbody>
</table>

(a) If we use a means model to predict job rating from sex, what are the five values of $\hat{y}$ for each of the five cases?
(b) What are the five residual values for this model?
(c) How much do we gain by using gender to predict job rating? In other words, quantify the difference between this model and a model that doesn’t use any explanatory variable at all.

18. The data frame deathpenalty in the Stob package has data on whether the death penalty was given in 326 homicide cases in Florida in 1976 and 1977.

(a) Use an appropriate measure to “prove” that black defendants are less likely to get the death penalty than white defendants – perhaps a counterintuitive result.
(b) The R function cross makes a new variable from two categorical variables. Use the following syntax:

```r
tally(Penalty ~ cross(Defendant, Victim), data = deathpenalty)
```

Explain how the result of this computation suggests that a more nuanced conclusion than that of part (a) might be warranted.

19. For each of the following stories, identify the explanatory and response variables and indicate whether you think an observational study or experiment would be used. (Explain your choice in each case.)

(a) Calvin wants to compare the GPAs of women living in Heyns to those living in Noordewier.
(b) A doctor wants to compare the effectiveness of atenolol and quinapril for his patients in controlling high blood pressure.
(c) A child psychologist wants to find out whether the amount of video game playing by fifth graders is related to the frequency with which they bully children at school.

20. Suppose that we want to decide who is the better Mathematics 171 instructor of two instructors A and B who are teaching two different sections.

(a) Give several reasons why we cannot simply compare the instructors by comparing the grades of the students in class.
(b) There is a uniform final exam in Mathematics 171. Suppose we compare the instructors by comparing the scores of their students on the uniform final exam. How does that help? What problems still remain?
Suppose that we also give a pretest to all the students in Mathematics 171 and compare the two instructors by comparing how much their students improved over the course of the semester. How does that help? What problems still remain?

21. Ultramarathoners often develop respiratory infections after running an ultra. Researchers were interested in whether vitamin C was useful in reducing these infections. The researchers studied a group of 100 ultra runners. They gave 50 a daily regimen of 600 mg vitamin C while the other 50 received a placebo. They also studied 100 nonrunners over the same period, also giving 50 of them the vitamin C treatment and 50 of them the placebo. All 200 subjects were watched for 14 days after the race to determine if infections developed.

(a) What are the explanatory and response variables?
(b) What is the name of this experimental design?
(c) Why did the researchers use nonrunners at all if the purpose was to determine whether vitamin C helped prevent infections in ultramarathoners?

22. (AP Statistics Free Response, 2007, Problem 2) As dogs age, diminished joint and hip health may lead to joint pain and thus reduce a dog’s activity level. Such a reduction in activity can lead to other health concerns such as weight gain and lethargy due to lack of exercise. A study is to be conducted to see which of dietary supplements, glucosamine or chondroitin, is more effective in promoting joint and hip health and reducing the onset of canine osteoarthritis. Researchers will randomly select a total of 300 dogs from ten different large veterinary practices around the country. All of the dogs are more than 6 years old, and their owners have given consent to participate in the study. Changes in joint and hip health will be evaluated after 6 months of treatment.

(a) What would be an advantage to adding a control group in the design of this study?
(b) Assuming a control group is added to the other two groups in the study, explain how you would assign the 300 digs to these three groups for a completely randomized design.
(c) Rather than using a completely randomized design, one group of researchers proposes blocking on clinics, and another group of researchers proposes blocking on breed of dog. How would you decide which one of these two variables to use as a blocking variable?

23. The dataframe `normaltemp` has data on the temperatures of 130 Calvin students (taken in Psychology 151 in 1998).

(a) What is the sample mean temperature for this group of students?
(b) Use the bootstrap to compute a standard error for your statistic.
(c) It is folklore that the “normal” body temperature is 98.6 degrees. Explain how your analysis in (b) gives strong evidence that this folklore is wrong.

24. The `normaltemp` data also records the gender of each individual. (Unfortunately gender has been coded as a quantitative rather than categorical variable. To convert this variable to categorical, use the function `factor`. For example `bwplot(Temp~factor(Gender),data=normaltemp)` draws a boxplot that you will want to look at.)

(a) What is the mean temperature for each gender group in this sample?
(b) Use the bootstrap to compute confidence intervals for the mean temperature of each of men and women.
(c) Does your analysis in (b) suggests that there is a difference in the population in the mean temperature of men and women?

25. The bootstrap can help us estimate parameters other than the mean of the population. Using the `normaltemp` data, compute an estimate of the third quartile of the temperature distribution of the population in question and its standard error. (Note that you will have to look at the dataframe that results from your `do` function to see where the various sample statistics are.)

26. The `tips` dataframe in the `reshape2` package has data on the tips received by a waiter in a certain restaraunt.
(a) Fit the model $\text{tips} \sim 1$. What is the coefficient? What is the sum of squares of residuals?

(b) Fit the model $\text{tips} \sim 1 + \text{total\_bill}$. What are the coefficients? What is the sum of squares of residuals?

(c) For each of the two models above, what is the predicted tip for a $20$ bill?

(d) From the results of the preceding two parts, do you think it important to know the total bill in predicting the tip? Give a statistical answer.

27. The SAT dataframe in the mosaic package has data on state by state educational inputs and outputs (for high schools). The variable sat has average SAT scores of high school students in the state and the variable expend has data on per pupil expenditures on education.

(a) Fit the model $\text{sat} \sim 1 + \text{expend}$.

(b) Interpret the coefficient of expend in the model. What seems odd about this?

(c) Examine the other variables in the datafram. Do you have an explanation for this “odd” result?

28. Suppose that response is a quantitative response variable and color is a categorical variable with the levels Red, Green, and Blue. Suppose that the fitted model from R gave the following coefficients:

(Intercept) 15
ColorBlue 8
ColorGreen 13

(a) What is the mean of response for all those cases that are Blue?

(b) What is the mean of response for all those cases that are Red?

(c) We cannot determine the mean of response for all cases. Why not?

(d) What are the possible values could the mean of response for all cases (using only the information given)?

29. Using the CPS85 data from the mosaic package, write a model and then use lm to answer the following questions.

(a) What is the average age of single people in the sample?

(b) What is the average age difference between single and married people in the sample?

(c) Write a good statistical statement that describes how wages vary with age.

(d) Write a good statistical statement that describes how wages vary by gender with age being held fixed.

30. In the SAT dataset of the mosaic package are variables on education in the states.

(a) Fit the model $\text{sat} \sim \text{expend} + \text{frac}$.

(b) What are the units of the three coefficients in the model?

(c) Compare this model to problem 27. How does the model of this problem help understand the “strangeness” of problem 27?
31. A certain drug has a side effect of increasing hemoglobin levels in individuals. The effect depends on the dosage strength and is somewhat different for males and females. Consider the model

\[
\text{Hemoglobin} \sim 1 + \text{dose} + \text{sex} + \text{sex:dose}
\]

Suppose that Males are the reference group for sex (so that the model formula has a variable \text{sexF}). In each part below, choose the correct response and give a short explanation for your choice.

(a) The coefficient of the intercept in the model formula will be: (choose one) positive negative 0
(b) The coefficient of \text{dose} in the model formula will be: (choose one) positive negative 0
(c) The coefficient of \text{sexF} in the model formula will be: (choose one) positive negative 0
(d) The coefficient of \text{dose:sexF} in the model formula will be: (choose one) positive negative 0

32. The data frame \text{mammals} in the \text{MASS} package has data on the brain and body weight of various mammals.

(a) The model \text{brain} \sim 1 + \text{body} doesn’t seem like a very good model. Why not? (Look at a graph.)

(b) An appropriate transformation of both brain and body size results in a much better model. Find such a transformation. (Hint: you might notice that these data extend over several orders of magnitude.)

33. (There is no collaboration allowed on this problem.) A dataset available from Danny’s website \text{Diamonds} has data on 308 diamonds. You can get the dataset by

\begin{verbatim}
Diamonds = fetchData("Diamonds.csv")

Retrieving from http://www.mosaic-web.org/go/datasets/Diamonds.csv
\end{verbatim}

\text{head}(Diamonds)

\begin{verbatim}
carat color clarity certification price
1 0.30 D VS2 GIA 1302
2 0.30 E VS1 GIA 1510
3 0.30 G VVS1 GIA 1510
4 0.30 G VS1 GIA 1260
5 0.31 D VS1 GIA 1641
6 0.31 E VS1 GIA 1555
\end{verbatim}

The variables include
carat  weight (one carat is 200 mg)
color  on a scale from D (colorless) to Z (yellow)
clarirty  a categorical variable, see http://www.diamondinfo.org
certification  third party certification organization
price  price in dollars

(a) Anyone who knows diamonds knows that diamonds of greater carat weight are more highly valued. Fit a model \( \text{price} \sim 1 + \text{carat} \).

(b) From the graph, you might suppose that a second degree polynomial fits the data somewhat better. That is, you might want to fit the model \( \text{price} \sim 1 + \text{carat} + \text{carat}^2 \). R provides a function that makes this easy to do:

\[
1 \leftarrow \text{lm}(\text{price} \sim \text{poly(carat, 2)}, \text{data} = \text{Diamonds})
\]

Fit this model and write the model equation in more conventional notation than that returned by R.

(c) Give an argument that the quadratic model fits the data better.

34. The dataframe KidsFeet in the mosaic package has data on a number of children including measurements of their feet.

(a) Fit the following model for the width of the feet: \( \text{width} \sim 1 + \text{length} \). List the coefficients and find the value of \( R^2 \) for this model.

(b) Now fit the model \( \text{width} \sim 1 + \text{length} + \text{sex} \) and find the value of \( R^2 \). Would you say that knowing the gender of the student is of significant help in predicting the width of a foot given the length?

35. Suppose that \( y \) is a response variable, \( x \) and \( z \) are quantitative explanatory variables, and \( g \) and \( h \) are categorical explanatory variables. Consider the following models:

A  \( y \sim 1 \)
B  \( y \sim 1 + x \)
C  \( y \sim 1 + x + g \)
D  \( y \sim 1 + z + h \)
E  \( y \sim 1 + x + g + h \)
F  \( y \sim 1 + x + g + x:g \)
G  \( y \sim 1 + x + z + g + h \)
H  \( y \sim 1 + z + g + h \)

For some pairs of the above models, the \( R^2 \) for one is certainly less than or equal to the \( R^2 \) of the other. For other pairs, it depends on what the variables are as to which \( R^2 \) is greater. List all pairs of models for which it is possible to determine the relationship of \( R^2 \) and say what that relationship is. (For example, the \( R^2 \) for model A is less than or equal to that of model B, and so forth.)

36. The dataframe trees has data on the Volume, Girth, and Height of some felled cherry trees. Using \( R^2 \) as a guide, give a good reason for using the model \( \text{Volume} \sim 1 + \text{Girth} + \text{Girth}^2 \) instead of either a first degree or third degree polynomial.

37. The Duncan dataframe of the car package has data (from 1950) on 45 different professions. Each of the 45 professions was rated on prestige and the typical characteristics of persons in that profession.

prestige  percent of raters rating occupation as excellent or good
income  percent of males earning $3,500 or more
education  percent of males who were high school graduates

(a) Fit a model \( \text{prestige} \sim 1 + \text{income} \).

(b) Fit a model \( \text{prestige} \sim 1 + \text{income} + \text{education} \).

(c) Explain why the coefficient of \( \text{income} \) is different in the two models. Give an explanation that explains the size of the difference.
38. The story in the following link
   http://news.sciencemag.org/social-sciences/2014/02/scienceshot-why-you-should-talk-your-baby
   describes the result of a study done on verbal abilities in babies.

   (a) The headline claims that parents should talk to their babies. The headline is the summary of a conclusion of a study. What are the explanatory and response variables of that study?

   (b) One of the variables in the study is socioeconomic status of the parents. What is the role of that variable in this study?

39. The following study
   makes claims about diets that include brown rice.

   (a) What are the explanatory and response variables referred to in the headline?

   (b) What covariates did the article imply were considered?

40. Do problem 10.04 from the collection of exercises from the book. (Link here or on the homework webpage.)

41. Suppose that a manufacturer of computer chips claims to produce no more than 5% defective chips. You inspect 100 chips produced.

   (a) Would you have a strong reason to doubt the claim of the manufacturer if you find that 6 of the chips are defective? Why or why not?

   (b) If in part (a) you said no, how many defective chips would you have to find before you think that you would have a strong reason to doubt the manufacturer? Defend your choice.

42. Breanna Verkaik made 72 out of 89 free throws this past basketball season.

   (a) Use an appropriate binomial model to estimate the probability that Breanna would make all 10 free throws if she shot 10.

   (b) What assumption of the binomial model might be questioned in this situation?

43. Do problem 11.23 from the textbook exercises. here

44. Do problem 11.33 from the textbook exercises. here

45. Do problem 11.34 from the textbook exercises. here

46. The whiteside data in the MASS package has data on gas usage in a certain house before and after the owner installed insulation.

   Gas    weekly gas consumption in 1000s of cubic feet
   Insul  before or after insulation was installed
   Temp   the average of outside temperature in degrees Celsius

   (a) Fit the model \( \text{Gas} \sim 1 + \text{Insul} + \text{Temp} \).

   (b) Compute a 95% confidence interval for \( \text{Insul} \) using the bootstrap method.

   (c) Compute a 95% confidence interval for \( \text{Insul} \) using the standard method. Comment on any difference you see.

   (d) Find fitted values of the model for an average temperature of 0 both before and after insulation.

   (e) Write a 95% prediction interval for each of your fitted values in part (d).

   (f) Compare the two intervals and write a nice sentence saying qualitatively what these two intervals tell us.
47. It might be expected that more successful baseball teams might tend to draw more fans to the ballpark. The dataframe `Baseball21` in the `Stob` package has data on the first 11 years of major league baseball seasons in the 21st century. Here are a few of the variables:

- `attendance` the attendance at home games for the team on the season
- `W` the number of games won by the team in the season
- `R` runs scored by the team
- `LG` the league, American or National, that the team is in

(a) Fit the model `attendance ∼ 1 + W`.
(b) Compute a confidence interval for the coefficient of wins in the model of part (a). Write a sentence that interprets this coefficient.
(c) Note the confidence interval for the intercept term. Why is it so wide?
(d) Now fit a model `attendance ∼ 1 + W + R`. One might consider this model by thinking that teams that score more runs, even if they lose, are more interesting to watch and so might attract more fans. Does the fitted model support this conjecture?
(e) You might instead think that National League teams are, in general, more interesting to watch and so might attract more fans. Fit a model and decide whether the data support this conjecture.

48. At this website, http://www.yale.edu/infantlab/socialevaluation/Helper-Hinderer.html are some videos that show some of the experiments done on the question of whether very young children already have a sense of social relationships. The first two videos were shown to several young children. In the first video, a triangle helps the circle up the hill. In the second video, the square hinders the circle from reaching the top of the hill. After a child is shown both sequences, they are tested to see whether they have a preference for the helper triangle or the hinderer square. The first video in the second row shows a pre-verbal 6 month old child clearly showing a preference for the helper triangle. In one particular run of this experiment, 12 children were tested and 9 preferred the helper in this way.

(a) What is the natural null hypothesis in this experiment?
(b) If the null hypothesis is true, how many children are expected to choose the helper rather than the hinderer?
(c) Is the result in this experiment, 9 out of 12 favoring the helper, strong evidence against the null hypothesis? Give a quantitative justification for your claim.
(d) Of course it was important to conduct this experiment very carefully in order to ensure that it was the helper-hinderer distinction that was important. Give two examples of possible confounding variables and explain how randomization of some aspect of the experiment would address each.
(e) In this experiment, where specifically should blinding have been used?

49. In this problem, you will think about the relationship between confidence intervals and hypothesis tests. Recall the dataset from the last test on the survival time of Alzheimer’s patients.

```r
mydata <- fetchData("CSHA.csv")

Retrieving from http://www.mosaic-web.org/go/datasets/CSHA.csv
```

The variables are
- `Gender` gender
- `Education` number of years of education
- `AAO` age at onset of Alzheimer’s disease
- `Survival` number of days from onset of Alzheimer’s until death

(a) Suppose that you are investigating the model `Survival ∼ 1 + Gender`. For this model, what is the most reasonable null hypothesis?
(b) Construct a confidence interval for the coefficient of Gender in the model. Explain how that confidence interval might indicate that there is not enough evidence to say that the null hypothesis is false.

50. The males of stalk-eyed flies have long eye stalks. Is the male's long eye-stalk affected by the quality of its diet? Two groups of male flies were reared on different diets. One group was fed corn and the other cotton wool. The eye spans (distance between the eyes) were measured in mm. The data are in the package Stalkies2 in the abd package.

   eye.span   eye span in mm
   food       diet (Corn or Cotton)

   (a) Identify the null and alternate hypotheses.
   (b) If we fit the model eye.span ~ 1 + food, what is a reasonable test statistic?
   (c) Do a simulation to compute an approximate P-value for this test statistic.
   (d) Is there strong evidence against the null hypothesis?

51. The dataframe pulp in the faraway package has data on an experiment to test the brightness of paper produced by four different shift operators. (The units have been lost and these are not typical units. In standard units, paper brightness is usually between 80 and 90.)

   bright   brightness of the pulp
   operator  operator a-d

   Use a randomization test to determine if the shift operator makes a difference in the brightness of paper produced.

52. The dataframe corrosion in the faraway package has data on the corrosion of test bars with various percentages of added iron.

   Fe        percentage of added iron
   loss      material lost due to corrosion

   Fit a model loss ~ 1 + Fe.

   (a) Compute $R^2$ for this model.
   (b) What is the obvious null hypothesis in this situation?
   (c) What is the expected value of $R^2$ if the null hypothesis is true?
   (d) Compute the $F$-statistic. You could have $R$ do this, but computing it from the formula will make you feel really smart.

53. Do problem 14.22 of the problems from the text, found here.

54. The nels88 dataframe in the faraway package has data on a mathematics test taken as part of a national study. Some variables are

   math     the mathematics test score of the student
   ses      the socioeconomic status of the family of child
   paredu   the parents educational status

   For the following models, complete the following table

<table>
<thead>
<tr>
<th>Model</th>
<th>SSModel</th>
<th>df</th>
<th>SSResid</th>
<th>df</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>math ~ 1 + sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>math ~ 1 + sex + paredu</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>math ~ 1 + sex + paredu + ses</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
55. A study was done on a new filter that was supposed to reduce noise pollution in automobiles. Different sizes of cars were measured (size), two different types of filter (Type) and measurements were taken on each size of the car (Side). The model was Noise $\sim 1 + \text{Size} + \text{Type} + \text{Side}$ (with Size converted to a factor). The ANOVA table is below with some of the quantities replaced by letters.

**Analysis of Variance Table**

<table>
<thead>
<tr>
<th>Response: Noise</th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>factor(Size)</td>
<td>2</td>
<td>26051.4</td>
<td>145.9872</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td>1</td>
<td>1056.2</td>
<td>1056.2</td>
<td>B</td>
<td>0.001679 **</td>
</tr>
<tr>
<td>Side</td>
<td>1</td>
<td>0.7</td>
<td>0.7</td>
<td>0.0078</td>
<td>0.930268</td>
</tr>
<tr>
<td>Residuals</td>
<td>31</td>
<td>2766.0</td>
<td>89.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(a) Compute A.
(b) Compute B.
(c) We haven’t learned enough to compute C exactly. But you can say approximately what it is.
(d) How many cars were tested?
(e) Write a sentence that interprets correctly the value .930268 in the lower right of the table. Your sentence should interpret the number in the particular context of the problem.

56. Do problem 15.10 of the textbook problems. (The data are in TenMileRace in the mosaic package.)

57. The dataframe cathedral in the faraway package has data on the cathedral nave heights and lengths of several cathedrals in England.

- **x** the length of the nave
- **y** the height of the nave
- **style** the style of the cathedral, romanesque or gothic

Fit a model to predict the height of the nave from the length of the nave and the style. ($y \sim 1 + x + \text{style}$)

(a) Write the equation of the fitted model.
(b) Use R to estimate the P-value for the null hypothesis that style does not matter in this model.
(c) Instead, use a simulation method (using shuffle) to compute the same P-value.
(d) Should style be included in this model?

58. The dataframe Ericksen in the car package has data on the undercount rate in the 1980 census. See the help document for a description of all the variables. The response variable is undercount which is the percentage of people undercounted in a given region.

(a) Make a big model for undercount by including all the other available variables as explanatory variables. Don’t include any interaction terms however.
(b) Make an argument that several of the variables are clearly not necessary in the model.
(c) Make a smaller model for undercount and give a statistical reason for believing that this smaller model is to be preferred to the larger model.

59. In this problem we are going to develop a hypothesis test and determine its power for a typical testing situation. Suppose a certain medical test is marketed with the statement that it detects a certain medical condition 90% of the time that the patient actually has the condition. (90% is called the sensitivity of the test.) To test this claim, a market research firm finds people with the condition and gives them the test.
(a) Suppose that the firm tests 20 persons with the disease. How many patients should test positive for the condition if the claim of the company is true?

(b) What is the probability that 16 or fewer of 20 persons would test positive? (Hint: you might want to think of the binomial model.)

(c) If the testing firm sets a significance level of 5%, when will they reject the null hypothesis if they test 20 patients?

(d) Suppose that the test only detects the disease in 80% of the patients that have it. What is the power of this hypothesis test at the 5% level of significance?

(e) Suppose that the testing firm desires 95% power and a 5% level of significance. How many people should the firm test if the true sensitivity is 80%?

60. Medical testing is even more complicated than the last problem suggests. Obviously the medical test should detect the disease in patients who have it but it should also not detect the disease in patients who don’t have it. The specificity of the test is the percentage of negative results in testees who don’t have the disease. So a specificity of 90% means that the test will produce 10% false positives. If a claim is made that a medical test is 90% accurate, it usually means that both the sensitivity and specificity are at least 90%. In this problem, we investigate the problem of going backward: if a test tells you that you have a certain disease, how likely is it that you have the disease? Seems like 90% right?

(a) Suppose that in a certain population, only 1% of the people have the disease. Now suppose that everyone is tested. What proportion of the population will test positive for the disease? (Remember that a person can test positive either by having the disease and the test is right or not having the disease and the test is wrong.)

(b) Based on your result in part (a), what proportion of those who test positive actually have the disease?

(c) If you did (a) and (b) right, you should be more than a little surprised. What do you think this result says about universal screening programs for various diseases?

61. In logistic regression, the values of the model \( y \) are “link” values that have to be transformed to get probabilities.

(a) For each of the following link values, compute the corresponding probability: \(-3, -2, -1, 0, 1, 2, 3\).

(b) For each of the following probabilities, compute the corresponding link value: \(0.1, 0.25, 0.5, 0.75, 0.9\).

62. The dataframe CAFE in the Stat2Data has data on a 2003 vote in the US Senate on an amendment to a bill sponsored by John Kerry and John McCain to mandate improved fuel economy on cars and light trucks. The amendment effectively killed the bill so was strongly supported by most car manufacturers. One might suppose that the political party of the Senator might be related to their vote (it usually is) and it also might be supposed that the contributions from car manufacturers had an affect. The variable LogContr is the logarithm of the contributions that a senator received in their lifetime while the variable Dem is 1 if the senator was a Democrat or Independent and 0 otherwise. The variable Vote is the vote of the senator (1 or 0). (Remember that for datasets in the Stat2Data package you need to load the package and use the command data(CAFE) to make the dataset accessible.)

(a) Fit a model for the Senator’s vote that uses LogContr and Dem as explanatory variables.

(b) The median of LogContr is approximately 4. For this value, what are the predicted probabilities of a YES vote from a Democrat and from a Republican?

(c) Write a good sentence interpreting the coefficient of LogContr in the model.

(d) Write a sentence interpreting the coefficient of Dem in the model.