Statistical Modeling: A Fresh Approach

Second Edition

End-of-Chapter Exercises

Chapter One Reading Questions

1. How can a model be useful even if it is not exactly correct?

2. Give an example of a model used for classification.

3. Often we describe personalities as "patient," "kind," "vengeful," etc. How can these descriptions be used as models for prediction?

4. Give three examples of models that you use in everyday life. For each, say what is the purpose of the model and in what ways the representation differs from the real thing.

5. Make a sensible statement about how precisely these quantities are typically measured:
   - The speed of a car.
   - Your weight.
   - The national unemployment rate.
   - A person's intelligence.

6. Give an example of a controlled experiment. What quantity or quantities have been varied and what has been held constant?

7. Using one of your textbooks from another field, pick an illustration or diagram. Briefly describe the illustration and explain how this is a model, in what ways it is faithful to the system being described and in what ways it fails to reflect that system.

Prob 1.01

Many fields of natural and social science have principles that are identified by name. Sometimes these are called "laws," sometimes "principles", "theories," etc. Some examples:

Kepler's Law	Newton's Laws of Motion	
Ohm's Law	Grimm's Law	Nernst equation
Raoult's Law	Nash equilibrium	Boyle's Law
Zipf's Law	Law of diminishing marginal utility	
Pareto principle	Snell's Law	Hooke's Law
Fitt's Law	Laws of supply and demand	
Ideal gas law	Newton's law of cooling	
Le Chatelier's principle		Poiseuille's law

These laws and principles can be thought of as models. Each is a description of a relationship. For instance, Hooke's law relates the extension and stiffness of a spring to the force exerted by the spring. The laws of supply and demand relate the quantity of a good to the price and postulates that the market price is established at the equilibrium of supply and demand.

Pick a law or principle from an area of interest to you — chemistry, linguistics, sociology, physics, ... whatever. Describe the law, what quantities or qualities it relates to one another, and the ways in which the law is a model, that is, a representation that is suitable for some purposes or situations and not others.

Enter your answer here:

An example is given below.

> *EXAMPLE: As described in the text, Hooke's Law, $f = -kx$, relates the force (f), the stiffness (k) and the extension past resting length (x) for a spring. It is a useful and accurate approximation for small extensions. For large extensions, however, springs are permanently distorted or break. Springs involve friction, which is not included in the law. Some springs, such as passive muscle, are really composites and show a different pattern, e.g., $f = -k\ x^3/|x|$ for moderate sized extensions.*

Prob 1.02

NOTE: Remember to load the `mosaic` package before starting:

```
> load(mosaic)
```

Each of the following statements has a syntax mistake. Write the statements properly and give a sentence saying what was wrong. (Cut and paste the correct statement from R, along with any output that R gives and your sentence saying what was wrong in the original.)

Here's an example:

> QUESTION: What wrong with this statement?
>
> ```
> > a = fetchData(myfile.csv)
> ```
>
> ANSWER: *It should be*
>
> ```
> > a = fetchData("myfile.csv")
> ```
>
> *The file name is a character string and therefore should be in quotes. Otherwise it's treated as an object name, and there is no object called myfile.csv.*

Now for the real thing. Say what's wrong with each of these statements for the purpose given:

1. `> seq(5;8)` to give `[1] 5 6 7 8`
2. `> cos 1.5` to calculate the cosine of 1.5 radians
3. `> 3 + 5 = x` to make x take the value 3+5
4. `> sqrt[4*98]` to find the square root of 392
5. `> 10,000 + 4,000` adding two numbers to get 14,000
6. `> sqrt(c(4,16,25,36))=4` intended to give

   `[1] FALSE TRUE FALSE FALSE`
7. `> fruit = c(apple, berry, cherry)` to create a collection of names

   `[1] "apple" "berry" "cherry"`
8. `> x = 4(3+2)` where x is intended to be assigned the value 20
9. `> x/4 = 3+2` where x is intended to be assigned the value 20

Prob 1.03

The operator `seq` generates sequences. Use `seq` to make the following sequences:

1. the integers from 1 to 10
2. the integers from 5 to 15
3. the integers from 1 to 10, skipping the even ones
4. 10 evenly spaced numbers between 0 and 1

Prob 1.04

According to legend, when the famous mathematician Johann Carl Friedrich Gauss (1777-1855) was a child, one of his teachers tried to distract him with busy work: add up the numbers 1 to 100. Gauss did this easily and immediately without a computer. But using the computer, which of the following commands will do the job?

A sum(1,100)
B seq(1,100)
C sum of seq(1,100)
D sum(seq(1,100))
E seq(sum(1,100))
F sum(1,seq(100))

Prob 1.05

Try the following command:

```
> seq(1,5,by=.5,length=2)
```

Why do you think the computer responded the way it did?

Prob 1.08

`1e6` and `10e5` are two different ways of writing one-million. Write 5 more different forms of one-million using scientific notation.

Prob 1.09

The following statement gives a result that some people are surprised by

```
> 10e3 == 1000
[1] FALSE
```

Explain why the result was FALSE rather than TRUE.

A 10e3 is 100, not 1000
B 10e3 is 10000, not 1000
C 10e3 is not a valid number
D It should be true; there's a bug in the software.

Chapter Two Reading Questions

1. What are the two major different kinds of variables?
2. How are variables and cases arranged in a data frame?
3. How is the relationship between these things: population, sampling frame, sample, census?
4. What's the difference between a longitudinal and cross-sectional sample?
5. Describe some types of sampling that lead to the sample potentially being unrepresentative of the population?

Prob 2.01

Read in the file `kidsfeet.csv`. For each of the following, hand in the R statements to find what is asked for. Provide both the command you give and the output of the command.

1 The names of the variables.

2 The mean of the foot width variable.

3 Which of the cases are girls.

4 The mean foot width for the subset of data for the girls.

Of no particular statistical value, but to review the use of logical operators:

5 The mean foot width for the subset of data for people whose bigger foot is left and dominant hand is also left.

And, for some extra practice in using logical operators:

6 The mean foot width for the subset of data for people who are either male or whose bigger foot matches the dominant hand.

7 The mean foot width for the subset of data for people whose bigger foot does NOT match the dominant hand.

Prob 2.02

Using the `table` operator and the comparison operators (such as `>` or `==`), answer the following questions about the CO2 data. You can read in the CO2 data with the statement

```
CO2 = fetchData("CO2")

Called from: fetchData("CO2")
```

You can see the data set itself by giving the command

```
CO2
```

In this exercise, you will use R commands to count how many of the cases satisfy various criteria:

1. How many of the plants in CO2 are Mc1 for Plant?
   7 12 14 21 28 34
2. How many of the plants in CO2 are either Mc1 or Mn1?
   8 12 14 16 23 54 92
3. How many are Quebec for Type and nonchilled for Treatment?
   8 12 14 16 21 23 54 92
4. How many have a concentration (conc) of 300 or bigger?
   12 24 36 48 60
5. How many have a concentration between 300 and 450 (inclusive)?
   12 24 36 48 60
6. How many have a concentration between 300 and 450 (inclusive) and are nonchilled?
   6 8 10 12 14 16
7. How many have an uptake that is less than 1/10 of the concentration (in the units reported)?
   17 33 34 51 68

Prob 2.04

Here is a small data frame about automobiles.

Make and model	Vehicle type	Trans. type	# of cyl.	City MPG	Hwy MPG
Kia Optima	compact	Man.	4	21	31
Kia Optima	compact	Auto.	6	20	28
Saab 9-7X AWD	SUV	Auto.	6	14	20
Saab 9-7X AWD	SUV	Auto.	8	12	16
Ford Focus	compact	Man.	4	24	35
Ford Focus	compact	Auto.	4	24	33
Ford F150 2WD	pickup	Auto.	8	13	17

(a) What are the cases in the data frame?

- [A] Individual car companies
- [B] Individual makes and models of cars
- [C] Individual configurations of cars
- [D] Different sizes of cars

(b) For each case, what variables are given? Are they categorical or quantitative?

- Kia Optima: not.a.variable categorical quantitative
- City MPG: not.a.variable categorical quantitative
- Vehicle type: not.a.variable categorical quantitative
- SUV: not.a.variable categorical quantitative
- Trans. type: not.a.variable categorical quantitative
- # of cyl.: not.a.variable categorical quantitative

(c) Why are some cars listed twice? Is this a mistake in the table?

- [A] Yes, it's a mistake.
- [B] A car brand might be listed more than once, but the cases have different attributes on other variables.
- [C] Some cars are more in demand than others.

Prob 2.09

Here is a data set from an experiment about how reaction times change after drinking alcohol.[**?**] The measurements give how long it took for a person to catch a dropped ruler. One measurement was made before drinking any alcohol. Successive measurements were made after one standard drink, two standard drinks, and so on. Measurements are in seconds.

Before	After 1	After 2	After 3
0.68	0.73	0.80	1.38
0.54	0.64	0.92	1.44
0.71	0.66	0.83	1.46
0.82	0.92	0.97	1.51
0.58	0.68	0.70	1.49
0.80	0.87	0.92	1.54
and so on ...			

(a) What are the rows in the above data set?

- [A] Individual measurements of reaction time.
- [B] An individual person.
- [C] The number of drinks.

(b) How many variables are there?

- [A] One — the reaction times.
- [B] Two — the reaction times with and without alcohol.
- [C] Four — the reaction times at four different levels of alcohol.

The format used for these data has several limitations:

- It leaves no room for multiple measurements of an individual at one level of alcohol, for example, two or three baseline measurements, or two or three measurements after one standard drink.

- It provides no flexibility for different levels of alcohol, for example 1.5 standard drinks, or for taking into account how long the measurement was made after the drink.

Another format, which would be better, is this:

SubjectID	ReactionTime	Drinks
S1	0.68	0
S1	0.73	1
S1	0.80	2
S1	1.38	3
S2	0.54	0
S2	0.64	1
S2	0.92	2
and so on ...		

What are the cases in the reformatted data frame?

A Individual measurements of reaction time.
B An individual person.
C The number of drinks.

How many variables are there?

A The same as in the original version. It's the same data!
B Three — the subject, the reaction time, the alcohol level.
C Four — the reaction times at four different levels of alcohol.

The lack of flexibility in the original data format indicates a more profound problem. The response to alcohol is not just a matter of quantity, but of timing. Drinks spread out over time have less effect than drinks consumed rapidly, and the physiological response to a drink changes over time as the alcohol is first absorbed into the blood and then gradually removed by the liver. Nothing in this data set indicates how long after the drinks the measurements were taken. The small change in reaction time after a single drink might reflect that there was little time for the alcohol to be absorbed before the measurement was taken; the large change after three drinks might actually be the response to the first drink finally kicking in. Perhaps it would have been better to make a measurement of the blook alcohol level at each reaction-time trial.

It's important to think carefully about how to measure your variables effectively, and what you should measure in order to capture the effects you are interested in.

Prob 2.14

Sometimes categorical information is represented numerically. In the early days of computing, it was very common to represent everything with a number. For instance the categorical variable for sex, with levels male or female, might be stored as 0 or 1. Even categorical variables like race or language, with many different levels, can be represented as a number. The codebook provides the interpretation of each number (hence the word "codebook").

Here is a very small part of a dataset from the 1960s used to study the influence of smoking and other factors on the weights of babies at birth.[?] `gestation.csv`

gest.	wt	race	ed	wt.1	inc	smoke	number
284	120	8	5	100	1	0	0
282	113	0	5	135	4	0	0
279	128	0	2	115	2	1	1
244	138	7	2	178	98	0	0
245	132	7	1	140	2	0	0
351	140	0	5	120	99	3	2
282	144	0	2	124	2	1	1
279	141	0	1	128	2	1	1
281	110	8	5	99	2	1	2
273	114	7	2	154	1	0	0
285	115	7	2	130	1	0	0
255	92	4	7	125	1	1	5
261	115	3	2	125	4	1	5
261	144	0	2	170	7	0	0

At first glance, all of the data seems quantitative. But read the codebook:

```
gest. - length of gestation in days

wt -  birth weight in ounces (999 unknown)

race - mother's race
   0-5=white 6=mex 7=black 8=asian
   9=mixed 99=unknown

ed - mother's education
   0= less than 8th grade,
   1 = 8th -12th grade - did not graduate,
   2= HS graduate--no other schooling ,
   3= HS+trade,
   4=HS+some college
   5= College graduate,
   6&7 Trade school HS unclear,
   9=unknown

marital 1=married, 2= legally separated, 3= divorced,
  4=widowed, 5=never married

inc - family yearly income in $2500 increments
  0 = under 2500, 1=2500-4999, ...,
  8= 12,500-14,999, 9=15000+,
  98=unknown, 99=not asked

smoke - does mother smoke? 0=never, 1= smokes now,
    2=until current pregnancy, 3=once did, not now,
    9=unknown

number  - number of cigarettes smoked per day
   0=never, 1=1-4, 2=5-9, 3=10-14, 4=15-19,
   5=20-29, 6=30-39, 7=40-60,
   8=60+, 9=smoke but don't know, 98=unknown, 99=not asked
```

Taking into account the codebook, what kind of data is each variable? If the data have a natural order, but are not genuinely quantitative, say "ordinal." You can ignore the "unknown" or "not asked" codes when giving your answer.

(a) Gestation categorical ordinal quantitative

(b) Race categorical ordinal quantitative

(c) Marital categorical ordinal quantitative

(d) Inc categorical ordinal quantitative

(e) Smoke categorical ordinal quantitative

(f) Number categorical ordinal quantitative

The disadvantage of storing categorical information as numbers is that it's easy to get confused and mistake one level for another. Modern software makes it easy to use text strings to label the different levels of categorical variables. Still, you are likely to encounter data with categorical data stored numerically, so be alert.

A good modern practice is to code missing data in a consistent way that can be automatically recognized by software as meaning missing. Often, NA is used for this purpose. Notice that in the number variable, there is a clear order to the categories until one gets to level 9, which means "smoke but don't know." This is an unfortunate choice. It would be better to store number as a quantitative variable telling the number of cigarettes smoked per day. Another variable could be used to indicate whether missing data was "smoke but don't know," "unknown", or "not asked."

Chapter Three Reading Questions

1. What is the disadvantage of using a 100% coverage interval to describe variation?

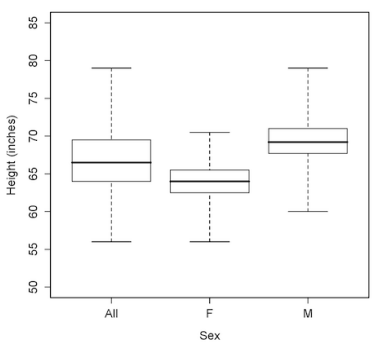
2. In describing a sample of a variable, what is the relationship between the variance and the standard deviation?

3. What is a residual?

4. What's the difference between "density" and "frequency" in displaying a variable with a histogram?

5. What's a normal distribution?

6. Here is the graph showing boxplots of height broken down according to sex as well as for both males and females together.



Which components of the boxplot for "All" match up exactly with the boxplot for "M" or "F"? Explain why.

7. Variables typically have units. For example, in Galton's height data, the height variable has units of inches. Suppose you are working with a variable in units of degrees celsius. What would be the units of the standard deviation of a variable? Of the variance? Why are they different?

Prob 3.01

Here is a small table of percentiles of typical daily calorie consumption of college students.

Percentile	Calories
0	1400
5	1800
10	2000
25	2400
50	2600
75	2900
90	3100
95	3300
100	3700

(a) What is the 50%-coverage interval?

Lower Boundary 1800 1900 2000 2200 2400 2500 2600
Upper Boundary 2600 2750 2900 3000 3100 3200 3500

(b) What percentage of cases lie between 2900 and 3300?
10 20 25 30 40 50 60 70 80 90 95

(c) What is the percentile that marks the upper end of the 95%-coverage interval? 75 90 92.5 95 97.5 100
Estimate the corresponding calorie value from the table.
2900 3000 3100 3300 3500 3700

(d) Using the 1.5 IQR rule-of-thumb for identifying an outlier, what would be the threshold for identifying a low calorie consumption as an outlier?
1450 1500 1650 1750 1800 2000

Prob 3.02

Here are some useful operators for taking a quick look at data frames:

`names`	Lists the names of the components.
`ncol`	Tells how many components there are.
`nrow`	Tells how many lines of data there are.
`head`	Prints the first several lines of the data frame.

Here are some examples of these commands applied to the `CO2` data frame:

```
  CO2 = fetchData("CO2")
Called from: fetchData("CO2")
  names(CO2)
[1] "Plant"     "Type"      "Treatment" "conc"      "uptak
  ncol(CO2)
[1] 5
```

```
nrow(CO2)
```

```
[1] 84
```

```
head(CO2)
```

```
  Plant   Type  Treatment conc uptake
1   Qn1 Quebec nonchilled   95   16.0
2   Qn1 Quebec nonchilled  175   30.4
3   Qn1 Quebec nonchilled  250   34.8
4   Qn1 Quebec nonchilled  350   37.2
5   Qn1 Quebec nonchilled  500   35.3
6   Qn1 Quebec nonchilled  675   39.2
```

- The data frame `iris` records measurements on flowers. You can read in with

```
iris = fetchData("iris")
```

```
Called from: fetchData("iris")
```

creating an object named `iris`.

Use the above operators to answer the following questions.

1. Which of the following is the name of a column in `iris`?

   flower Color Species Length
2. How many rows are there in `iris`?
   1 50 100 150 200
3. How many columns are there in `iris`?
   2 3 4 5 6 7 8 10
4. What is the `Sepal.Length` in the third row?
   1.2 3.6 4.2 4.7 5.9

- The data frame `mtcars` has data on cars from the 1970s. You can read it in with

```
cars = fetchData("mtcars")
```

```
Called from: fetchData("mtcars")
```

creating an object named `cars`.

Use the above operators to answer the following questions.

1. Which of the following is the name of a column in `cars`?
   carb color size weight wheels
2. How many rows are there in `cars`?
   30 31 32 33 34 35
3. How many columns are there in `cars`?
   7 8 9 10 11
4. What is the `wt` in the second row?
   2.125 2.225 2.620 2.875 3.215

Prob 3.03

Which one of these commands will give the 95th percentile of the children's heights in Galton's data? `galton.csv`

A `quantile(galton$height,95)`
B `quantile(galton$height,0.95)`
C `quantile(galton$father,95)`
D `quantile(galton$father,0.95)`

Which of these command will give the 90-percent coverage interval of the children's heights in Galton's data?

A `quantile(galton$height,c(0.05,0.95))`
B `quantile(galton$height,c(0.025,0.975))`
C `quantile(galton$height,0.90)`
D `quantile(galton$height,90)`

Find the 50-percent coverage interval of the following variables in Galton's height data:

- Father's heights
  A 59 to 73 inches
  B 68 to 71 inches
  C 63 to 65.5 inches
  D 68 to 74 inches
- Mother's heights
  A 59 to 73 inches
  B 68 to 71 inches
  C 63 to 65.5 inches
  D 68 to 74 inches

Find the 95-percent coverage interval of

- Father's heights
  A 65 to 73 inches
  B 65 to 74 inches
  C 68 to 73 inches
  D 59 to 69 inches
- Mother's heights
  A 62.5 to 68.5 inches
  B 65 to 69 inches
  C 63 to 68.5 inches
  D 59 to 69 inches

Prob 3.04

In Galton's data, are the sons typically taller than their fathers? Create a variable that is the difference between the son's height and the father's height. (Arrange it so that a positive number refers to a son who is taller than his father.)

1. What's the mean height difference (in inches)?
   -2.47 -0.31 0.06 66.76 69.23
2. What's the standard deviation (in inches)?
   1.32 2.63 2.74 3.58 3.75

3. What is the 95-percent coverage interval (in inches)?

   A -3.7 to 4.8
   B -4.6 to 4.9
   C -5.2 to 5.6
   D -9.5 to 4.5

Prob 3.05

Use R to generate the sequence of 101 numbers: $0, 1, 2, 3, \cdots, 100$.

1. What's the mean value?
   25 50 75 100
2. What's the median value?
   25 50 75 100
3. What's the standard deviation?
   10.7 29.3 41.2 53.8
4. What's the sum of squares?
   5050 20251 103450 338350 585200

Now generate the sequence of perfect squares $0, 1, 4, 9, \cdots, 10000$, or, written another way, $0^2, 1^2, 2^2, 3^2, \cdots, 100^2$. (Hint: Make a simple sequence 0 to 100 and square it.)

1. What's the mean value?
   50 2500 3350 4750 7860
2. What's the median value?
   50 2500 3350 4750 7860
3. What's the standard deviation?
   29.3 456.2 3028 4505 6108
4. What's the sum of squares?
   5050 20251 338350 585200 2050333330

Prob 3.06

Using Galton's height data (`galton.csv`), compute the answers to these questions about outliers using the 1.5 IQR rule of thumb and the `outlier` function.

1. Which of these statements will compute the number of cases that are outliers with respect to height? (Assume that the data frame is named `galton`.)

   A `outlier(galton$height)`
   B `table(outlier(galton$height))`
   C `outlier(table(galton$height))`
   D `subset(galton, outlier(galton$height))`
   E `outlier(subset(galton, galton$height))`

2. How many of the cases are outliers in height ?
   0 1 2 3 5 10 15 20
3. How many of the cases are outliers in mother?
   0 11 22 33 44 55 66
4. How many of the cases are outliers in father?
   0 4 9 14 19 24 29
5. Looking just at the cases where mother is an outlier, how many of the children involved (variable sex) are female?
   0 5 10 15 20 25 30 35

Prob 3.07

To exercise your ability to extract subsets of data, take each of the following subsets of the swimming records in `swim100m.csv` and calculate the mean and minimum swimming time for the subset. (Answers have been rounded to one decimal place.)

- All records between 1920 and 1940 (including 1920 and 1940).

  Mean: 54.6 60.2 64.7 69.6 71.3
  Min: 56.4 60.2 64.7 69.6 71.3

- Female records in the 1970s and 1980s

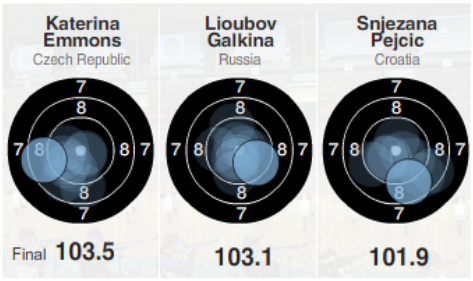
  Mean: 54.1 54.7 56.2 60.2 64.7 69.6 71.3
  Min: 54.1 54.7 56.2 60.2 64.7 69.6 71.3

- All records that are **slower** than 60 seconds. (Hint: Think what "slower" means in terms of the swimming times.)

  Mean: 56.2 60.2 64.7 69.6 71.3 73.2 75.8
  Min: 56.2 60.2 61.5 64.7 69.6 71.3

Prob 3.08

The figure shows the results from the medal winners in the women's 10m air-rifle competition in the 2008 Olympics. (Figure from the New York Times, Aug. 10, 2008)



The location of each of 10 shots is shown as transluscent light circles in each target. The objective is to hit the bright target dot in the center. There is random scatter (variance) as well as steady deviations (bias) from the target.

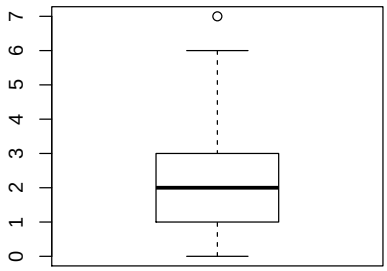
What is the direction of the apparent bias in Katerina Emmons's results? (Directions are indicated as compas directions, E=east, NE=north east, etc.)
NE NW SW SE

To measure the size of the bias, find the center of the shots and measure how far that is from the target dot. Take the distance between the concentric circles as one unit.

What is the size of the bias in Katerina Emmon's results?
0 1 3 4 6 10

Prob 3.09

Here is a boxplot:



Reading from the graph, answer the following:

(a) What is the median?

0 1 2 3 6 Can't estimate from this graph

(b) What is the 75th percentile?

0 1 2 3 6 Can't estimate from this graph

(c) What is the IQR?

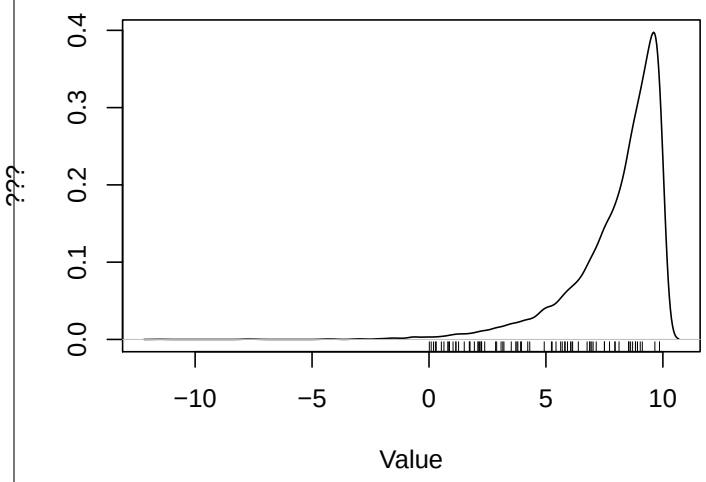
0 1 2 3 4 6 Can't estimate from this graph

(d) What is the 40th percentile?

A between 0 and 1
B between 1 and 2
C between 2 and 3
D between 3 and 4
E between 4 and 6
F Can't estimate from this graph.

Prob 3.10a

The plot shows two different displays of density. The displays might be from the same distribution or two different distributions.



(a) What are the two displays?

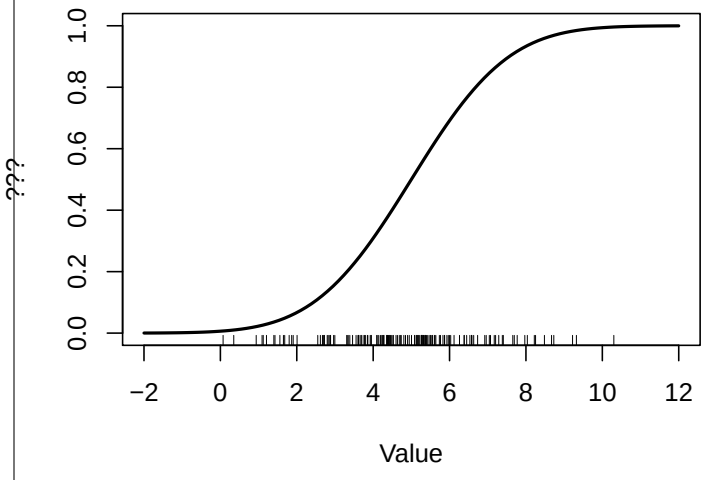
A Density and cumulative
B Rug and cumulative
C Cumulative and box plot
D Density and rug plot
E Rug and box plot

(b) The two displays show the same distribution. TRUE or FALSE

(c) Describe briefly any sign of mismatch or what features convince you that the two displays are equivalent.

Prob 3.10b

The plot shows two different displays of density. The displays might be from the same distribution or two different distributions.



(a) What are the two displays?

- [A] Density and cumulative
- [B] Rug and cumulative
- [C] Cumulative and box plot
- [D] Density and rug plot
- [E] Rug and box plot

(b) The two displays show the same distribution. TRUE or FALSE

(c) Describe briefly any sign of mismatch or what features convince you that the two displays are equivalent.

Prob 3.11

By hand, calculate the mean, the range, the variance, and the standard deviation of each of the following sets of numbers:

(A) $1, 0, -1$

(B) $1, 3$

(C) $1, 2, 3$.

1. Which of the 3 sets of numbers — A, B, or C — is the most spread out according to the range?
   - [A] A
   - [B] B
   - [C] C
   - [D] No way to know
   - [E] All the same

2. Which of the 3 sets of numbers — A, B, or C — is the most spread out according to the standard deviation?
   - [A] A
   - [B] B
   - [C] C
   - [D] No way to know
   - [E] All the same

Prob 3.12

A standard deviation contest. For (a) and (b) below, you can choose numbers from the set $0, 1, 2, 3, 4, 5, 6, 7, 8$, and 9. Repeats are allowed.

(a) Which list of 4 numbers has the largest standard deviation such a list can possibly have?

- [A] 0,3,6,9
- [B] 0,0,0,9
- [C] 0,0,9,9
- [D] 0,9,9,9

(b) Which list of 4 numbers has the smallest standard deviation such a list can possibly have?

- [A] 0,3,6,9
- [B] 0,1,2,3
- [C] 5,5,6,6
- [D] 9,9,9,9

Prob 3.13

(a) From what kinds of variables would side-by-side boxplots be generated?

- [A] categorical only
- [B] quantitative only
- [C] one categorical and one quantitative
- [D] varies according to situation

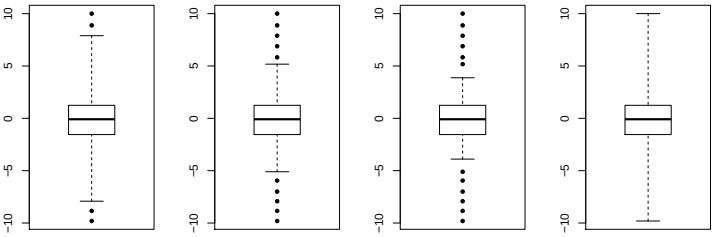
(b) From what kinds of variables would a histogram be generated?

- [A] categorical only
- [B] quantitative only
- [C] one categorical and one quantitative
- [D] varies according to situation

Prob 3.14

The boxplots below are all made from exactly the same data. One of them is made correctly, according to the "1.5 IQR" convention for drawing the whiskers. The others are drawn differently.

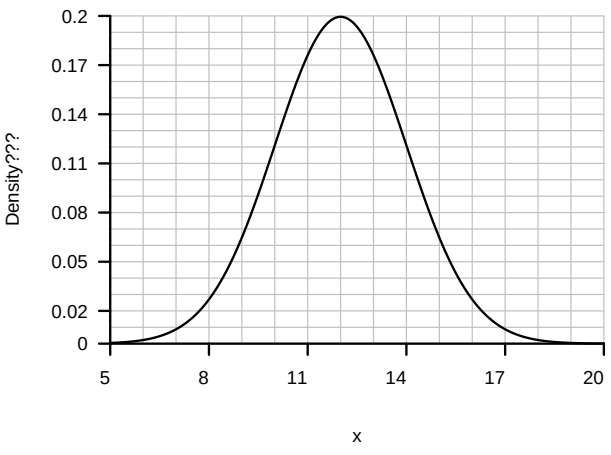
Plot 1 **Plot 2** **Plot 3** **Plot 4**



- Which of the plots is correct? 1 2 3 4

Prob 3.15

The plot purports to show the density of a distribution of data. If this is true, the fraction of the data that falls between any two values on the x axis should be the area under the curve between those two values.



Answer the following questions. In doing so, keep in mind that the area of each little box on the graph paper has been arranged to be 0.01, so you can calculate the area by counting boxes. You don't need to be too fanatical about dealing with boxes where only a portion in under the curve; just eyeball things and estimate.

(a) The total area under a density curve should be 1. Assuming that the density curve has height zero outside of the area of the plot, is the area under the entire curve consistent with this? yes no

(b) What fraction of the data falls in the range $12 \leq x \leq 14$?

A 14%
B 22%
C 34%
D 56%
E Can't tell from this graph.

(c) What fraction of the data falls in the range $14 \leq x \leq 16$?

A 14%
B 22%
C 34%
D 56%
E Can't tell from this graph.

(d) What fraction of the data has $x \geq 16$?

A 1%
B 2%
C 5%
D 10%
E Can't tell from this graph.

(e) What is the width of the 95% coverage interval. (Note: The coverage interval itself has top and bottom ends. This problem asks for the spacing between the two ends.)

A 2
B 4
C 8
D 12
E Can't tell from this graph.

Prob 3.16

If two distributions have the same five-number summary, must their density plots have the same shape? Explain.

Prob 3.17

As the name suggests, the Old Faithful geyser in Yellowstone National Park has eruptions that come at fairly predictable intervals, making it particularly attractive to tourists.

(a) You are a busy tourist and have only 10 minutes to sit around and watch the geyser. But you can choose when to arrive. If the last eruption occurred at noon, what time should you arrive at the geyser to maximize your chances of seeing an eruption?

A 12:50
B 1:00
C 1:05
D 1:15
E 1:25

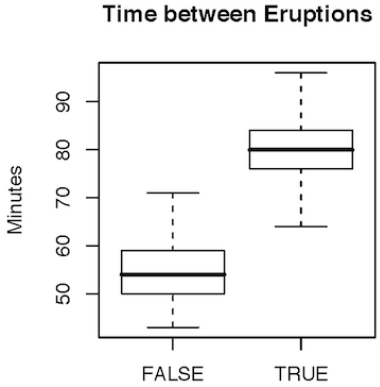
(b) Roughly, what is the probability that in the best 10-minute interval, you will actually see the eruption:

A 5%
B 10%
C 20%
D 30%
E 50%
F 75%

(c) A simple measure of how faithful is Old Faithful is the interquartile range. What is the interquartile range, according to the boxplot above?

A 10 minutes
B 15 minutes
C 25 minutes
D 35 minutes
E 50 minutes
F 75 minutes

(d) Not only are you a busy tourist, you are a smart tourist. Having read about Old Faithful, you understand that the time between eruptions depends on how long the previous eruption lasted. Here's a box plot indicating the distribution of inter-eruption times when the previous eruption duration was less than three minutes. (That is, "TRUE" means the previous eruption lasted less than three minutes.)



You can easily ask the ranger what was the duration of the previous eruption.

What is the best 10-minute interval to return (after a noon eruption) so that you will be most likely to see the next eruption, given that the previous eruption was less than three minutes in duration (the "TRUE" category).

A 1:00 to 1:10
B 1:05 to 1:15
C 1:10 to 1:20
D 1:15 to 1:25
E 1:20 to 1:30
F 1:25 to 1:35

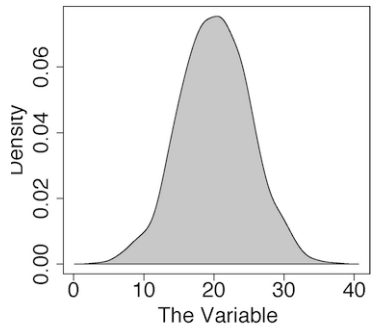
(e) How likely are you to see an eruption if you return for the most likely 10-minute interval?

A About 5%
B About 10%
C About 20%
D About 30%
E About 50%
F About 75%

Prob 3.18

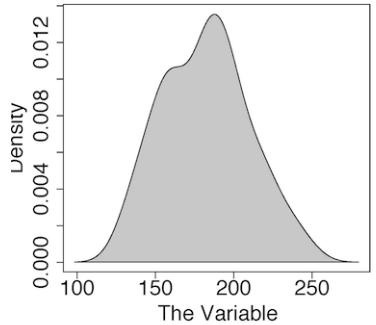
For each of the following distributions, estimate by eye the mean, standard deviation, and 95% coverage interval. Also, calculate the variance.

Part 1.



- Mean. 10 15 20 25 30
- Std. Dev. 2 5 12 15 20
- 95% coverage interval.
  - Lower end: 1 3 10 15 20
  - Upper end : 20 25 30 35 40
- Variance. 2 7 10 20 25 70 140 300

Part 2.



- Mean. 0.004 150 180 250
- Std. Dev. 10 30 60 80 120
- 95% coverage interval.
  - Lower end: 50 80 100 135 150 200 230
  - Upper end: 50 80 100 180 200 230
- Variance. 30 80 500 900 1600 23000

Prob 3.19

Consider a large company where the average wage of workers is \$15 per hour, but there is a spread of wages from minimum wage to \$35 per hour.

After a contract negotiation, all workers receive a \$2 per hour raise. What happens to the standard deviation of hourly wages?

A No change
B It goes up by \$2 per hour
C It goes up by \$4 per hour
D It goes up by 4 dollars-square per hour
E It goes up by \$4 per hour-square
F Can't tell from the information given.

The annual cost-of-living adjustment is 3%. After the cost-of-living adjustment, what happens to the standard deviation of hourly wages?

A No change
B It goes up by 3%
C It goes up by 9%
D Can't tell from the information given.

Prob 3.20

Construct a data set of 10 hypothetical exam scores (use integers between 0 and 100) so that the inter-quartile range equals zero and the mean is greater than the median.

Give your set of scores here:

Prob 3.23

Here are some familiar quantities. For each of them, indicate what is a typical value, how far a typical case is from this typical value, and what is an extreme but not impossible case.

Example: Adult height. Typical value, 1.7 meters (68 inches). Typical case is about 7cm (3 inches) from the typical value. An extreme height is 2.2 meters (87 inches).

- An adult's weight.
- Income of a full-time employed person.
- Speed of cars on a highway in good conditions.
- Systolic blood pressure in adults. [You might need to look this up on the Internet.]
- Blood cholesterol LDL levels. [Again, you might need the Internet.]
- Fuel economy among different models of cars.

- Wind speed on a summer day.
- Hours of sleep per night for college students.

Prob 3.24

Data on the distribution of economic variables, such as income, is often presented in quintiles: divisions of the group into five equal-sized parts.

Here is a table from the US Census Bureau (Historical Income Tables from March 21, 2002) giving the distribution of income across US households in year 2000.

Quintile	Upper Boundary	Mean Value
Lowest	$17,955	$10,190
Second	$33,006	$25,334
Third	$52,272	$42,361
Fourth	$81,960	$65,729
Fifth	—	$141,260

Based on this table, calculate:

(a) The 20th percentile of family income.

10190 17955 33006 25334 52272 42361 81960 141260

(b) The 80th percentile of family income.

10190 17955 33006 25334 52272 42361 81960 141260

(c) The table doesn't specify the median family income but you can make a reasonable estimate of it. Pick the closest one.

10000 18000 25500 42500 53000 65700

(d) Note that there is no upper boundary reported for the fifth quintile, and no lower boundary reported for the first quintile. Why?

(e) From this table, what evidence is there that family income has a skew rather than "normal" distribution?

Prob 3.25

Use the Internet to find "normal" ranges for some measurements used in clinical medicine. Pick one of the following or choose one of particular interest to you: blood pressure (systolic, diastolic, pulse), hematocrit, blood sodium and potassium levels, HDL and LDL cholesterol, white blood cell counts, clotting times, blood sugar levels, vital respiratory capacity, urine production, and so on. In addition to the normal range, find out what "normal" means, e.g., a 95% coverage interval on the population or a range inconsistent with proper physiological function. You may find out that there are differing views of what "normal" means — try to indicate the range of such views. You may also find out that "normal" ranges can be different depending on age, sex, and other demographic variables.

Prob 3.28

An advertisement for "America's premier weight loss destination" states that "a typical two week stay results in a loss of 7-14 lbs." (The *New Yorker*, 7 April 2008, p 38.)

The advertisement gives no details about the meaning of "typical." Give two or three plausible interpretations of the quoted 7-14 pound figure in terms of "typical." What interpretation would be most useful to a person trying to predict how much weight he or she might lose?

Prob 3.29

A seemingly straightforward statistic to describe the health of a population is **average age at death**. In 1842, the *Report on the Sanitary Conditions of the Labouring Population of Great Britain* gave these averages: "gentlemen and persons engaged in the professions, 45 years; tradesmen and their families, 26 years; mechanics, servants and laborers, and their families, 16 years."

A student questioned the accuracy of the 1842 report with this observation: "The mechanics, servants and laborer population wouldn't be able to renew itself with an average age at death of 16 years. Mothers would be dying so early in life that they couldn't possibly raise their kids."

Explain how an average age of death of 16 years could be quite consistent with a "normal" family structure in which parents raise their children through the child's adolescence in the teenage years. What other information about ages at death would give a more complete picture of the situation?

Prob 3.30

The identification of a case as an outlier does not always mean that the case is invalid or abnormal or the result of a mistake. One situation where perfectly normal cases can look like outliers is when there is a mechanism of proportionality at work. Imagine, for instance, that there is a typical rate of production of a substance, and the normal variability is proportional in nature, say from 1/10 of that typical rate to 10 times the rate. This leads to a situation where some normal cases are 100 times as large as others.

To illustrate, look at the `alder.csv` data set, which contains field data from a study of nitrogen fixation in alder plants. The SNF variable records the amount of nitrogen fixed in soil by bacteria that reside in root nodules of the plants. Make a box plot and a histogram and describe the distribution. Which of the following descriptions is most appropriate:

- [A] The distribution is skewed to the left, with outliers at very low values of SNF.
- [B] The distribution is skewed to the right, with outliers at very high values of SNF.
- [C] The distribution is roughly symmetrical, although there are a few outliers.

In working with a variable like this, it can help to convert the variable in a way that respects the idea of a proportional change. For instance, consider the three numbers 0.1, 1.0, and 10.0, which are evenly spaced in proportionate terms — each number is 10 times bigger than the preceding number. But

as absolute differences, 0.1 and 1.0 are much closer to each other than 1.0 and 10.0.

The *logarithm* function transforms numbers to a scale where even proportions are equally spaced. For instance, taking the logarithm of the numbers 0.1, 1.0, and 10.0 gives the sequence −1, 0, 1 — exactly evenly spaced.

The logSNF variable gives the logarithm of SNF. Plot out the distribution of logSNF. Which of the following descriptions is most apt?

A The distribution is skewed to the left.
B The distribution is skewed to the right.
C The distribution is roughly symmetrical.

You can compute logarithms directly in R, using the functions `log`, `log2`, or `log10`. Which of these functions was used to compute the quantity logSNF from SNF. (Hint: Try them out!)

log log2 log10

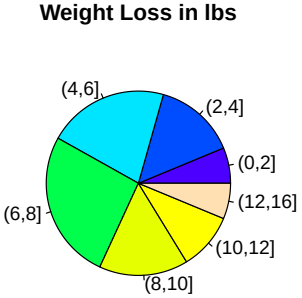
The *base* of the logarithm gives the size of the proportional change that corresponds to a 1-unit increase on the logarithmic scale. For example, `log2` calculates the base-2 logarithm. On the base-2 logarithmic scale, a doubling in size corresponds to a 1-unit increase. In contrast, on the base-10 scale, a ten-fold increase in size gives a 1-unit increase.

Logarithmic transformations are often used to deal with variables that are positive and strongly skewed. In economics, price, income and production variables are often this way. In general, any variable where it is sensible to describe changes in terms of proportion might be better displayed on a logarithmic scale. For example, price inflation rates are usually given as percent (e.g., "The inflation rate was 4% last year.") and so in dealing with prices over time, the logarithmic transformation can be appropriate.

Prob 3.31

This exercise deals with data on weight loss achieved by clients who stayed two weeks at a weight-loss resort. The same data using three different sorts of graphical displays: a pie chart, a histogram, and a box-and-whiskers plot. The point of the exercise is to help you decide which display is the most effective at presenting information to you.

In many fields, pie charts are used as "statistical graphics." Here's a pie chart of the weight loss:



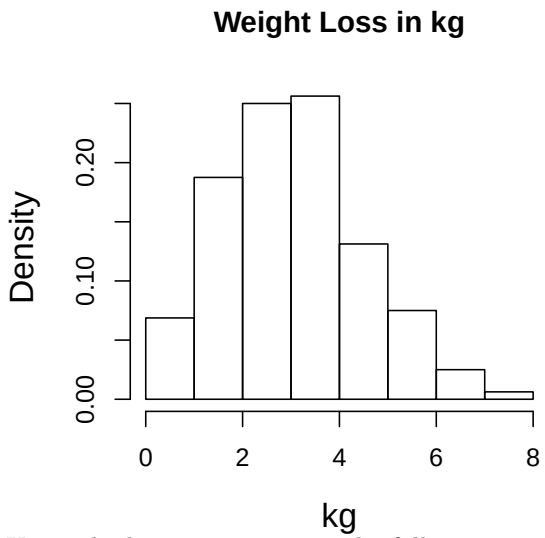
Using the pie graph, answer the following:

(a) What's the "typical" (median or mean) weight loss?
3.7 4.2 5.5 6.8 8.3 10.1 12.4

(b) What is the central 50% coverage interval?
2.3to6.8 4.2to10.7 4.4to8.7 6.1 to 9.3 5.2to12.1

(c) What is an upper extreme value? 10 13 16 18 20

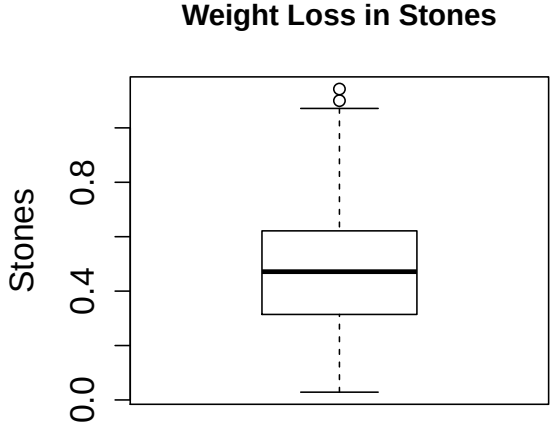
Now to display the data as a histogram. So that you can't just re-use your answers from the pie chart, the weights have been rescaled into kilograms.



Using the histogram, answer the following:

1. What's the "typical" (median or mean) weight loss?
   1.9 2.1 3.1 3.7 4.6 5.6
2. What is the central 50% coverage interval?
   1.1to3.3 2.0to4.8 2.0to3.9 2.8 to 4.4 2.5to5.4
3. What is an upper extreme value? 6 8 10 12 14

Finally, here is a boxplot of the same data. It's been rescaled into a traditional unit of weight: stones.



Using the boxplot, answer the following:

1. What's the "typical" (median or mean) weight loss?
   0.20 0.35 0.50 0.68 0.83 1.2

2. What is the central 50% coverage interval?
   0.2to0.5 0.3to0.8 0.4to0.8 0.5to0.7 0.3to0.6

3. What is an upper extreme value? 0.7 0.9 1.0 1.1 1.3

Which style of graphic made it easiest to answer the questions?
pie.chart histogram box.plot

Prob 3.36

Elevators typically have a close-door button. Some people claim that this button has no mechanical function; it's there just to give impatient people some sense of control over the elevator.

Design and conduct an experiment to test whether the button does cause the elevator door to close. Pick an elevator with such a button and record some details about the elevator itself: place installed, year installed, model number, etc.

Describe your experiment along with the measurements you made and your conclusions. You may want to do the experiment in small teams and use a stopwatch in order to make accurate measurements. Presumably, you will want to measure the time between when the button is pressed and when the door closes, but you might want to measure other quantities as well, for instance the time from when the door first opened to when you press the button.

Store the data from your experiment in a spreadsheet in Google Docs. Set the permissions for the spreadsheet so that anyone with the link can read your data. Make sure to **paste the link** into the textbox so that your data can be accessed.

Please don't inconvenience other elevator users with the experiment.

Prob 3.50

What's a "normal" body temperature? Depending on whether you use the Celsius or Fahrenheit scale, you are probably used to the numbers 37° (C) or 98.6° (F). These numbers come from the work of Carl Wunderlich, published in *Das Verhalten der Eigenwarme in Krankenheiten* in 1868 based on more than a million measurements made under the armpit. According to Wunderlich, "When the organism (man) is in a normal condition, the general temperature of the body maintains itself at the physiologic point: 37°C= 98.6°F."

Since 1868, not only have the techniques for measuring temperatures improved, but so has the understanding that "normal" is not a single temperature but a range of temperatures.

A 1992 article in the Journal of the American Medical Association (PA Mackowiak et al., "A Critical Appraisal of 98.6°F ..." JAMA v. 268(12) pp. 1578-1580) examined temperature measurements made orally with an electronic thermometer. The subjects were 148 healthy volunteers between age 18 and 40.

The figure shows the distribution of temperatures, separately for males and females. Note that the horizontal scale is given in both C and F — this problem will use F.

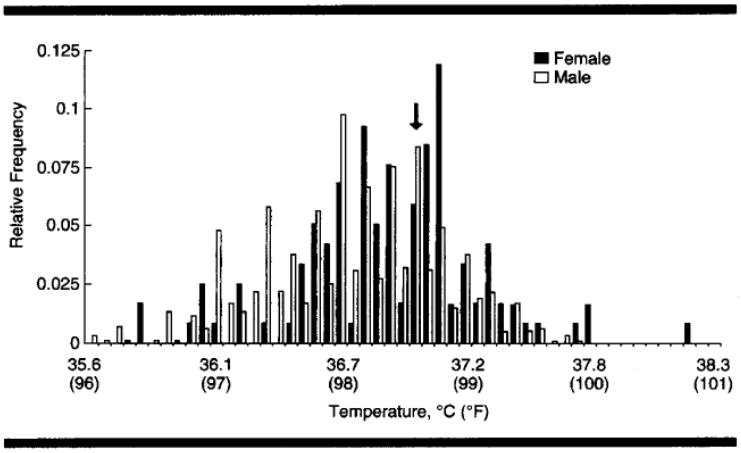


Fig 1.—Frequency distribution of 700 baseline oral temperatures obtained during two consecutive days of observation in 148 healthy young male and female volunteers. Arrow indicates location of 37.0°C (98.6°F).

What's the absolute range for females?

- Minimum: 96.1 96.3 97.1 98.6 99.9 100.8
- Maximum: 96.1 96.3 97.1 98.6 99.9 100.8

And for males?

- Minimum: 96.1 96.3 97.1 98.6 99.9 100.8
- Maximum: 96.1 96.3 97.1 98.6 99.9 100.8

Notice that there is an outlier for the females' temperature, as evidenced by a big gap in temperature between that bar and the next closest bar. How big is the gap?

A About 0.01° F.
B About 0.1° F.
C Almost 1° F.

Give a 95% coverage interval for females. Hint: The interval will exclude the most extreme 2.5% of cases on each of the left and right sides of the distribution. You can find the left endpoint of the 95% interval by scanning in from the left, adding up the heights of the bars until they total 0.025. Similarly, the right endpoint can be marked by scanning in from the right until the bars total 0.025.

A About 96.2°F to about 99.0°F
B About 96.8°F to about 100.0°F
C About 97.6°F to about 99.2°F

And for males?

A About 96.2°F to about 99.2°F
B About 96.7 to about 99.4°F.
C About 97.5°F to about 99.6°F

Prob 3.53

There are many different numerical descriptions of distributions: mean, median, standard deviation, variance, IQR, coverage interval, ... And these are just the ones we have touched on so far. We'll also encounter "standard error," "margin of error," "confidence interval." There are so many that it becomes a significant challenge to students to keep them straight. Eventually, statistical workers learn the subtleties of the different descriptions and when each is appropriate.

Then, like using near synonyms in English, it becomes second nature.

As an example, consider the verb "spread.". Here are some synonyms from the thesaurus, each of which is appropriate in a particular context: broadcast, scatter, propagate, sprawl, extend, stretch, cover, daub, ... If you were talking to a farmer about sewing seeds, the words "broadcast" or "scatter" would be appropriate, but it would be silly to say the seeds are being "daubbed" or "sprawled". On the other hand, to an urbanite concerned with congestion in traffic, the growth of the city might well be summarized with "sprawl." You have to know the context and the intent to choose the correct term.

To help to understand the different context and intents, here are two important ways of categorizing what a particular description captures:

(1) Location and scatter

- What is a typical value? ("center")
- What are the top and bottom range of the values? ("range")
- How far are the values scattered? ("scatter")
- What is high? or What is low? ("non-central")

(2) Including the "extremes"

- All inclusive, and sensitive to outliers. ("not-robust")
- All inclusive, but not sensitive to outliers. ("robust")
- Leaves out the very extremes. ("plausible"')
- Focuses on the middle. ("mainstream")

Note that descriptors of both the "plausible" and the "mainstream" type are necessarily robust, since they leave out the outliers.

(3) Individual versus whole sample.

- Description relevant to individual cases
- Description or summary of entire samples, combining many cases.

You won't have to deal with this until later, where it explains terms that you haven't yet encountered like like "standard error", "margin of error", "confidence interval."

Example: The **mean** describes the *center* of a distribution. It is calculated from all the data and *not-robust* against outliers.

For each of the following descriptors of a distribution , choose the items that best characterize the descriptor.

1. Median
   (a) center range scatter non-central
   (b) robust not-robust plausible mainstream
2. Standard Deviation
   (a) center range scatter non-central
   (b) robust not-robust plausible mainstream
3. IQR
   (a) center range scatter non-central
   (b) robust not-robust plausible mainstream
4. Variance
   (a) center range scatter non-central
   (b) robust not-robust plausible mainstream
5. 95% coverage interval
   (a) center range scatter non-central
   (b) robust not-robust plausible mainstream
6. 50% coverage interval
   (a) center range scatter non-central
   (b) robust not-robust plausible mainstream
7. 50th percentile
   (a) center range scatter non-central
8. 80th percentile
   (a) center range scatter non-central
9. 99th percentile
   (a) center range scatter non-central
10. 10th percentile
    (a) center range scatter non-central

One of the reasons why there are so many descriptive terms is that they have different roles in theory. For example, the variance turns out to have simple theoretical properties that make it useful when describing sums of variables. It's much simpler than, say, the IQR.

Prob 3.54

There are two kinds of questions that are often asked relating to percentiles:

- What is the value that falls at a given percentage? For instance, in the `ten-mile-race.csv` running data, how fast are the fastest 10% of runners? In R, you would ask in this way:

```
> run = fetchData("ten-mile-race.csv")
> qdata(0.10, run$net)
 10%
4409
```

  The answers is in the units of the variable, in this case seconds. So 10% of the runners have net times faster than or equal to 4409 seconds.

- What percentage falls at a given value? For instance, what fraction of runners are faster than 4000 seconds?

```
> pdata(4000, run$net)
[1] 0.04029643
```

The answer includes those whose net time is exactly **equal to** or **less than** 4000 seconds.

It's important to pay attention to the p and q in the statement. pdata and qdata ask related but different questions.

Use pdata and qdata to answer the following questions about the running data.

1. Below (or equal to) what age are the youngest 35% of runners?
   - Which statement will do the correct calculation?
     A `pdata(0.35,run$age)`
     B `qdata(0.35,run$age)`
     C `pdata(35,run$age)`
     D `qdata(35,run$age)`
   - What will the answer be?
     28 29 30 31 32 33 34 35

2. What's the net time that divides the slowest 20% of runners from the rest of the runners?
   - Which statement will do the correct calculation?
     A `pdata(0.20,run$net)`
     B `qdata(0.20,run$net)`
     C `pdata(0.80,run$net)`
     D `qdata(0.80,run$net)`
   - What will the answer be?
     4921 5318 5988 6346 7123 7431 seconds

3. What is the 95% coverage interval on age?
   - Which statement will do the correct calculation?
     A `pdata(c(0.025,0.975),run$age)`
     B `qdata(c(0.025,0.975),run$age)`
     C `pdata(c(0.050,0.950),run$age)`
     D `qdata(c(0.050,0.950),run$age)`
   - What will the answer be?
     A 22 to 60
     B 20 to 65
     C 25 to 59
     D 20 to 60

4. What fraction of runners are 30 or younger?
   - Which statement will do the correct calculation?
     A `pdata(30,run$age)`
     B `qdata(30,run$age)`
     C `pdata(30.1,run$age)`
     D `qdata(30.1,run$age)`
   - What will the answer be?
     In percent: 29.3 30.1 33.7 35.9 38.0 39.3

5. What fraction of runners are 65 or older? (Caution: This isn't yet in the form of a BELOW question.)
   - Which statement will do the correct calculation?
     A `pdata(65,run$age)`
     B `pdata(64.99,run$age)`
     C `pdata(65.01,run$age)`
     D `1-pdata(65,run$age)`
     E `1-pdata(64.99,run$age)`
     F `1-pdata(65.01,run$age)`
   - What will the answer be?
     In percent: 0.5 1.1 1.7 2.3 2.9

6. The time it takes for a runner to get to the start line after the starting gun is fired is the difference between the time and net.

   ```
   run$to.start = run$time - run$net
   ```

   - How long is it before 75% of runners get to the start line?
     In seconds: 164 192 213 294 324 351
   - What fraction of runners get to the start line before one minute? (Caution: the times are measured in seconds.)
     In percent: 10 15 19 22 25 31 34

7. What is the 95% coverage interval on the ages of female runners?
   A 19 to 61 years
   B 22 to 61 years
   C 19 to 56 years
   D 22 to 56 years

8. What fraction of runners have a net time BELOW 4000 seconds? (That is, don't include those who are at exactly 4000 seconds.)
   In percent: 3.72 4.00 4.03 4.07 5.21

Chapter Four Reading Questions

1. Which is larger: variance of residuals, variance of the model values, or the variance of the actual values?
2. How can a difference in group means clearly shown by your data nonetheless be misleading?
3. What does it mean to partition variation? What's special about the variance — the square of the standard deviation — as a way to measure variation?

Prob 4.05

Create a spreadsheet with the three variables distance, team, and position, in the following way:

distance	team	position
5	Eagles	center
12	Eagles	forward
11	Eagles	end
2	Doves	center
18	Doves	end
12	Penguins	forward
15	Penguins	end
19	Eagles	back
5	Penguins	center
12	Penguins	back

(a) After entering the data, you can calculate the mean distance in various ways.

- What is the grand mean distance?
  4 9.25 10 11 11.1 11.75 12 14.67 15.5
- What is the group mean distance for the three teams?
  - Eagles 4 9.25 10 11 11.1 11.75 12 14.67 15.5
  - Doves 4 9.25 10 11 11.1 11.75 12 14.67 15.5
  - Penguins 4 9.25 10 11 11.1 11.75 12 14.67 15.5
- What is the group mean distance for the following positions?
  - back 4 9.25 10 11 11.1 11.75 12 14.67 15.5
  - center 4 9.25 10 11 11.1 11.75 12 14.67 15.5
  - end 4 9.25 10 11 11.1 11.75 12 14.67 15.5

(b) Now, just for the sake of developing an understanding of group means, you are going to change the dist data. Make up values for dist so that the mean dist for Eagles is 14, for Penguins is 13, and for Doves is 15.

Cut and paste the output from R showing the means for these groups and then the means taken group-wise according to position.

(c) Now arrange things so that the means are as stated in (b) but **every** case has a residual of either 1 or −1.

Prob 4.06

Read in the Current Population Survey wage data:

```
> w = fetchData("CPS.csv")
```

(a) What is the grand mean of wage?
7.68 7.88 8.26 8.31 9.02 9.40 10.88

(b) What is the group-wise mean of wage for females?
7.68 7.88 8.26 8.31 9.02 9.40 10.88

(c) What is the group-wise mean of wage for married people?
7.68 7.88 8.26 8.31 9.02 9.40 10.88

(d) What is the group-wise mean of wage for married females?
7.68 7.88 8.26 8.31 9.02 9.40 10.88

Prob 4.07

Read in the Galton height data

```
> g = fetchData("galton.csv")
```

(a) What is the standard deviation of the height?

(b) Calculate the grand mean and, from that, the residuals of the actual heights from the grand mean.

```
> mean(height, data=g)
[1] 66.76069
> res0 = g$height - 66.76069
> sd(res0)
```

What is the standard deviation of the residuals from this "grand mean" model? 2.51 2.58 2.92 3.58 3.82

(c) Calculate the group-wise mean for the different sexes and, from that, the residuals of the actual heights from this group-wise model.

```
> mod1 = mean( height ~ sex, data=g)
> res1 = g$height - fitted( mod1, data=g)
> sd(res1)
```

What is the standard deviation of the residuals from this group-wise model? 2.51 2.58 2.92 3.58 3.82

(d) Explain why the standard deviation of the residuals of the group-wise model less than the standard deviation of the residuals of the "grand mean" model.

Chapter Five Reading Questions

- What is a sampling distribution? What sort of variation does it reflect?
- What is resampling and bootstrapping?
- What is the difference between a "confidence interval" and a "coverage interval?"

Prob 5.04

After a month's hard work in the laboratory, you have measured a growth hormone from each of 40 plants and computed a confidence interval on the grand mean hormone concentration of 36 ± 8 ng/ml. Your advisor asks you to collect more samples until the margin of error is 4 ng/ml. Assuming the typical $1/\sqrt{n}$ relationship between the number of cases in the sample and the size of the margin of error, how many plants, including the 40 you have already processed, will you need to measure?
40 80 160 320 640

Prob 5.08

You are calculating the mean of a variable B and you want to know the standard error, that is, the standard deviation of the sampling distribution of the mean. Which of the following statements will estimate the standard error by bootstrapping?

A `sd(do(500)*resample(mean(B)))`
B `resample(do(500)*mean(sd(B)))`
C `mean(do(500)*mean(resample(B)))`
D `sd(do(500)*mean(resample(B)))`
E `resample(sd(do(500)*mean(B)))`

Prob 5.09

A student writes the following on a homework paper:

> "The 95% confidence interval is (9.6, 11.4). I'm very confident that this is accurate, because my sample mean of 10.5 lies within this interval."

Comment on the student's reasoning.

Source: Prof. Rob Carver, Stonehill College

Prob 5.10

A perennial problem when writing scientific reports is figuring out how many significant digits to report. It's naïve to copy all the digits from one's calculator or computer output; the data generally do not justify such precision.

Once you have a confidence interval, however, you do not need to guess how many significant digits are appropriate. The standard error provides good guidance. Here is a rule of thumb: keep two significant digits of the margin of error and round the point estimate to the same precision.

For example, suppose you have a confidence interval of 1.7862 ± 0.3624 with 95% confidence. Keeping the first two significant digits of the margin of error gives 0.36. We'll keep the point estimate to the same number of digits, giving altogether 1.79 ± 0.36.

Another example: suppose the confidence interval is 6548.23 ± 1321. Keeping the first two digits of the margin of error gives 1300, with a corresponding margin of error of 6500 ± 1300.

(a) Suppose the computer output is $0.03234232 \pm 0.01837232$.

Using this rule of thumb, what should you report in as the confidence interval?

A 0.3234 ± 0.01837
B 0.3234 ± 0.0183
C 0.0323 ± 0.0184
D 0.0323 ± 0.018
E 0.032 ± 0.018
F 0.032 ± 0.01
G 0.03 ± 0.01

(b) Now suppose the computer output were $99.63742573 \pm 1.48924367$.

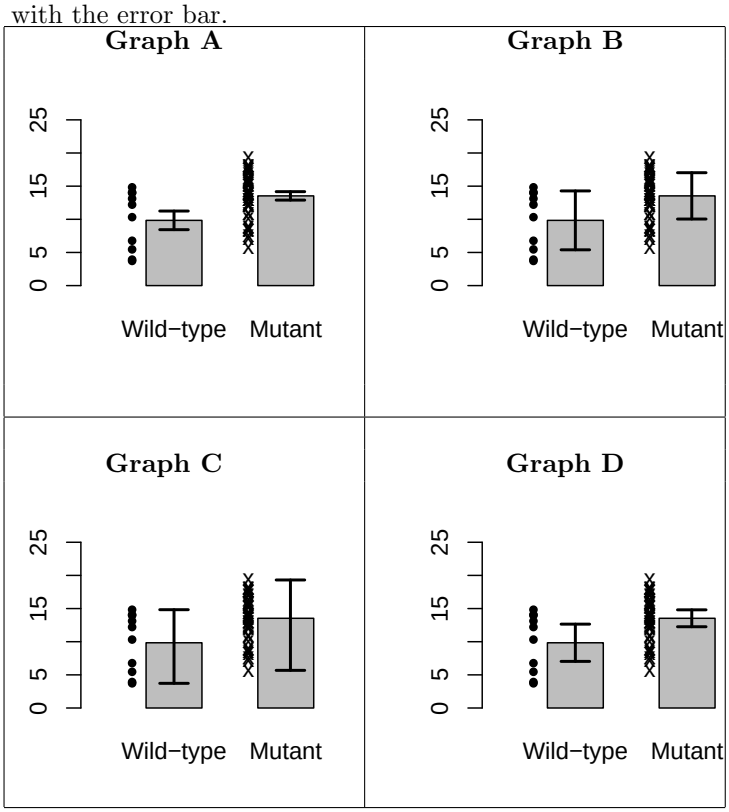
What should you report as the confidence interval?

A 100 ± 1
B 99 ± 1.5
C 99.6 ± 1.5
D 99.64 ± 1.49
E 99.647 ± 1.489

Prob 5.12

Scientific papers very often contain graphics with "error bars." Unfortunately, there is little standardization of what such error bars mean so it is important for the reader to pay careful attention in interpreting the graphs.

The following four graphs — A through D — each show a distribution of data along with error bars. The meaning of the bars varies from graph to graph according to different conventions used in different areas of the scientific literature. In each graph, the height of the filled bar is the mean of the data. Your job is to associate each error bar with its meaning. You can do this by comparing the actual data (shown as dots) with the error bar.



- Range of the data
  Graph A Graph B Graph C Graph D
- Standard deviation of the data
  Graph A Graph B Graph C Graph D
- Standard error of the mean
  Graph A Graph B Graph C Graph D
- 95% confidence interval on the mean
  Graph A Graph B Graph C Graph D

This problem is based on G. Cumming, F. Fidler, and DL Vaux (2007), "Error bars in experimental biology", *J. Cell Biology* **177**(1):7-11

Prob 5.13

An advertisement for "America's premier weight loss destination" states that "a typical two week stay results in a loss of 7-14 lbs." (The *New Yorker*, 7 April 2008, p 38.)

The advertisement gives no details about the meaning of "typical," but here are some possibilities:

- The 95% coverage interval of the weight loss of the individual clients.
- The 50% coverage interval of the weight loss of the individual clients.
- The 95% confidence interval on the mean weight loss of all the clients.
- The 50% confidence interval on the mean weight loss of all the clients.

Explain what would be valid and what misleading about advertising a confidence interval on the mean weight loss.

Why might it be reasonable to give a 50% coverage interval of the weight loss of individual clients, but not appropriate to give a 50% confidence interval on the mean weight loss.

Prob 5.17

Standard errors and confidence interval apply not just to model coefficients, but to any numerical description of a variable. Consider, for instance, the median or IQR or standard deviation, and so on.

A quick and effective way to find a standard error is a method called **bootstrapping**, which involves repeatedly resampling the variable and calculating the description on each resample. This gives the sampling distribution of the description. From the sampling distribution, the standard error — which is just the standard deviation of the sampling distribution — can be computed.

Here's an example, based on the inter-quartile range of the kids' foot length measurements.

First, compute the desired sample statistic on the actual data

```
IQR(kids$length)
```

```
[1] 1.6
```

Next, modify the statement to incorporate resampling of the data:

```
IQR(resample(kids$length))
```

```
[1] 1.4
```

Finally, run this statement many times to generate the sampling distribution and find the standard error of this distribution:

```
samps = do(1000) * IQR(resample(kids$length))
sd(samps)
```

```
[1] 0.3586868
```

Use the bootstrapping method to find an estimate of the standard error of each of these sample statistics on the kids' foot length data:

1. The sample median. (Pick the closest answer.)
   0.01 0.07 0.14 0.24 0.34 0.71 1.29 1.32 24.6
2. The sample standard deviation. (Pick the closest answer.)
   0.01 0.07 0.14 0.24 0.34 0.71 1.29 1.32 24.6
3. The sample 75th percentile.
   0.01 0.07 0.14 0.24 0.34 0.71 1.29 1.32 24.6

Bootstrapping works well in a broad set of circumstances, but if you have a very small sample, say less than a dozen cases, you should be skeptical of the result.

Prob 5.19

In this activity, you are going to look at the sampling distribution and how it depends on the size of the sample. This will be done by simulating a sample drawn from a population with known properties. In particular you'll be looking at a variable that is more or less like the distribution of human adult heights — normally distributed with a mean of 68 inches and a standard deviation of 3 inches.

Here's one random sample of size $n = 10$ from this simulated population:

```
rnorm(10, mean=68, sd=3)
```

```
[1] 62.842 71.095 62.357 68.896 67.494
[6] 67.233 69.865 71.664 69.241 70.581
```

These are the heights of a random sample of $n = 10$. The sampling distribution refers to some numerical description of such data, for example, the *sample mean*. Consider this sample mean the output of a single trial.

```
mean( rnorm(10, mean=68, sd=3) )
```

```
[1] 67.977
```

If you gave exactly this statement, it's very likely that your result was different. That's because you have a different random sample — `rnorm` generates random numbers. And if you repeat the statement, you'll likely get a different value again, for instance:

```
mean( rnorm(10, mean=68, sd=3) )
```

```
[1] 66.098
```

Note that both of the sample means above differ somewhat from the population mean of 68.

The point of examining a sampling distribution is to be able to see the reliability of a random sample. Do to this, you generate many trials — say, 1000 — and look at the distribution of the trials.

For example, here's how to look at the sampling distribution for the mean of 10 random cases from the population:

```
s = do(1000)*mean( rnorm(10, mean=68, sd=3) )
```

By examining the distribution of the values stored in `s`, you can see what the sampling distribution looks like.

Generate your own sample

1 What is the mean of this distribution?

2 What is the standard deviation of this distribution?

3 What is the shape of this distribution?

Now modify your simulation to look at the sampling distribution for $n = 1000$.

4 What is the mean of this distribution?

5 What is the standard deviation of this distribution?

6 What is the shape of this distribution?

Which of these two sample sizes, $n = 10$ or $n = 1000$, gave a sampling distribution that was more reliable? How might you measure the reliability?

The idea of a sampling distribution applies not just to means, but to any numerical description of a variable, to the coefficients on models, etc.

Now modify your computer statements to examine the **sampling distribution of the standard deviation** rather than the mean. Use a sample size of $n = 10$. (Note: Read the previous sentence again. The statistic you are asked to calculate is the **sample standard deviation**, not the sample mean.)

7 What is the mean of this distribution?

8 What is the standard deviation of this distribution?

9 What is the shape of this distribution?

Repeat the above calculation of the distribution of the sample standard deviation with $n = 1000$.

10 What is the mean of this distribution?

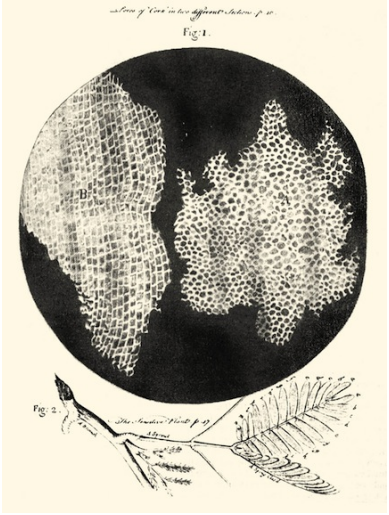
11 What is the standard deviation of this distribution?

12 What is the shape of this distribution?

For this simulation of heights, the population standard deviation was set to 3. You expect the result from a random sample to be close to the population parameter. Which of the two sample sizes, $n = 10$ or $n = 1000$ gives results that are closer to the population value?

Prob 5.23

Robert Hooke (1635-1703) was a contemporary of Isaac Newton. He is famous for his law of elasticity (Hooke's Law) and is considered the father of microscopy. He was the first to use the word "cell" to name the components of plant tissue; the structures he observed during his observations through a microscope reminded him of monks' cells in a monastery. He drew this picture of cork cells under the microscope:



Regarding these observations of cork, Hooke wrote:

> *I could exceedingly plainly perceive it to be all perforated and porous, much like a Honey-comb, but that the pores of it were not regular. . . . these pores, or cells, . . . were indeed the first* microscopical *pores I ever saw, and perhaps, that were ever seen, for I had not met with any Writer or Person, that had made any mention of them before this ....*

He went on to measure the cell size.

> *But, to return to our Observation, I told several lines of these pores, and found that there were usually about threescore of these small Cells placed end-ways in the eighteenth part of an Inch in length, whence i concluded that there must be neer eleven hundred of them, or somewhat more then a thousand in the length of an Inch, and therefore in a square Inch above a Million, or 1166400. and in a Cubick Inch, above twelve hundred Millions, or 1259712000. a thing almost incredible, did not our* Microscope *assure us of it by ocular demonstration ....* — from Robert Hooke, *Micrographia*, 1665

There are several aspects of Hooke's statement that reflect its origins at the start of modern science. Some are quaint, such as the spelling and obsolete use of Capitalization and the hyperbolic language ("a thing almost incredible," which, to be honest, is true enough, but not a style accepted today in scientific writing). Hooke worked before the development of the modern notion of precision. The seeming exactness of the number 1,259,712,000 for the count of cork cells in a cubic inch leaves a modern reader to wonder: did he really count over a billion cells?

It's easy enough to trace through Hooke's calculation. The observation at the base of the calculation is threescore cells — that's 60 cells — in 1/18 of an inch. This comes out to $60 \times 18 = 1080$ cells per linear inch. Modeling each cell as a little cube allows this to be translated into the number of cells covering a square inch: 1080^2 or 1,116,400. To estimate the number of cells in a cubic inch of cork material, the calculation is 1080^3 or 1,259,712,000.

To find the precision of these estimates, you need to go back to the precision of the basic observation: 60 cells in 1/18th of an inch. Hooke didn't specify the precision of this, but it seems reasonable to think it might be something like 60 ± 5 or so, at a confidence level of 95%.

1. When you change the units of a measurement (say, miles into kilometers), both the point estimate and the margin of error are multiplied by the conversion factor.

   Translate Hooke's count of the number of cells in 1/18 inch, 60 ± 5 into a confidence interval on the number of cells per linear inch.

   [A] 1080 ± 5
   [B] 1080 ± 90
   [C] 1080 ± 180

2. In calculating the number of cells to cover a square inch, Hooke simply squared the number of cells per inch. That's a reasonable approximation.

   To carry this calculation through a confidence interval, you can't just square the point estimate and the margin of error separately. Instead, a reasonable way to proceed is to take the endpoints of the interval (e.g., 55 to 65 for the count of cells in 1/18 inch), and square those endpoints. Then convert back to $\pm$ format.

   What is a reasonable confidence interval for the number of cells covering a square inch?

   [A] $1,200,000 \pm 500,000$
   [B] $1,170,000 \pm 190,000$
   [C] $1,166,000 \pm 19,000$
   [D] $1,166,400 \pm 1,900$

3. What is a reasonable confidence interval for the number of cork cells that fit into a cubic inch?

   [A] $1,300,000,000 \pm 160,000,000$
   [B] $1,260,000,000 \pm 16,000,000$
   [C] $1,260,000,000 \pm 1,600,000$
   [D] $1,259,700,000 \pm 160,000$
   [E] $1,259,710,000 \pm 16,000$
   [F] $1,259,712,000 \pm 1,600$

   It's usually better to write such numbers in scientific notation, so that the reader doesn't have to count digits to make sense of them. For example, $1,260,000,000 \pm 16,000,000$ might be more clearly written as $1260 \pm 16 \times 10^6$.

Chapter 6 Reading Questions

- What is an "explanatory variable" and how does it differ from a "response variable?"
- What is the difference between a "model term" and a "variable?"
- Why are there sometimes multiple model terms in a model?
- What is an interaction term and how does it differ from a variable?
- In graphs of the model response value versus an explanatory variable, quantitative explanatory variables are associated with slopes and categorical explanatory variables are associated with step-like differences. Explain why.
- How can models be useful that fail to represent the actual causal connections (if any) between variables? Give an example.

Prob 6.01

In McClesky vs Georgia, lawyers presented data showing that for convicted murderers, a death sentence was more likely if the victim was white than if the victim was black. For each case, they tabulated the race of the victim and the sentence (death or life in prison). Which of the following best describe the variables their models?

[A] Response is quantitative; explanatory variable is quantitative.
[B] Response is quantitative; explanatory variable is categorical.
[C] Response is categorical; explanatory variable is quantitative.
[D] Response is categorical; explanatory variable is categorical.
[E] There is no explanatory variable.

[Note: Based on an example from George Cobb.]

Prob 6.02

In studies of employment discrimination, several attributes of employees are often relevant:

> age, sex, race, years of experience, salary, whether promoted, whether laid off

For each of the following questions, indicate which is the response variable and which is the explanatory variable.

1. Are men paid more than women?

   Response Variable:
   age sex race years.experience salary promoted laid.off

   Explanatory Variable:
   age sex race years.experience salary promoted laid.off

2. On average, how much extra salary is a year of experience worth?

   Response Variable:
   age sex race years.experience salary promoted laid.off

   Explanatory Variable:
   age sex race years.experience salary promoted laid.off

3. Are whites more likely than blacks to be promoted?

   Response Variable:
   age sex race years.experience salary promoted laid.off

   Explanatory Variable:
   age sex race years.experience salary promoted laid.off

4. Are older employees more likely to be laid off than younger ones?

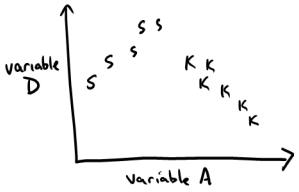
   Response Variable:
   age sex race years.experience salary promoted laid.off
   Explanatory Variable:
   age sex race years.experience salary promoted laid.off

[Note: Thanks to George Cobb.]

Prob 6.03

The drawings show some data involving three variables:

- D — a quantitative variable
- A — a quantitative variable
- G — a categorical variable with two levels: S & K



Sketch the graph on a piece of paper. Over that sketch, draw a function that shows the pattern of the fitted model values for each of the following models:

(a) D ~ A+G

(b) D ~ A-1

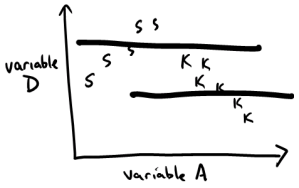
(c) D ~ A

(d) D ~ A*G

(e) D ~ 1

(f) D ~ poly(A,2)

Example: D ~ G.



Prob 6.04

Using your general knowledge about the world, think about the relationship between these variables:

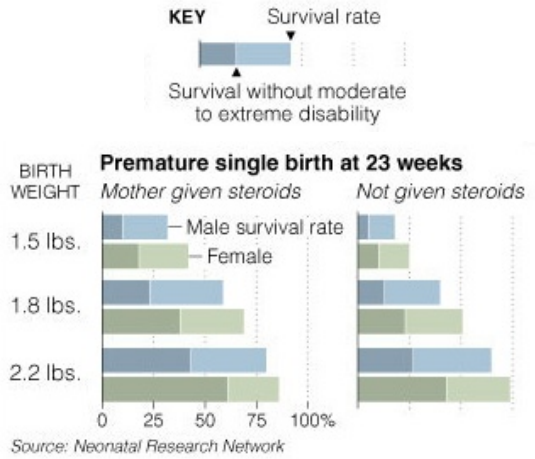
- speed of a bicyclist.
- steepness of the road, a quantitative variable measured by the grade (rise over run). 0 means flat, + means uphill, − means downhill.
- fitness of the rider, a categorical variable with three levels: unfit, average, athletic.

On a piece of paper, sketch out a graph of speed versus steepness for reasonable models of each of these forms:

1. Model 1: speed ~ 1 + steepness
2. Model 2: speed ~ 1 + fitness
3. Model 3: speed ~ 1 + steepness+fitness
4. Model 4: speed ~ 1 + steepness+fitness + steepness:fitness

Prob 6.05

The graphic (from the New York *Times*, April 17, 2008) shows the fitted values from a model of the survival of babies born extremely prematurely.



Caption: "A new study finds that doctors could better estimate an extremely premature baby's chance of survival by considering factors including birth weight, length of gestation, sex and whether the mother was given steroids to help develop the baby's lungs."

Two different response variables are plotted: (1) the probability of survival and (2) the probability of survival without moderate to severe disabilities. Remarkably for a statistical graphic, there are three explanatory variables:

1. Birth weight (measured in pounds (lb) in the graphic).
2. The sex of the baby.
3. Whether the mother took steroids intended to help the fetus's lungs develop.

Focus on the survival rates without disabilities — the darker bars in the graphic.

(a) Estimate the effect of giving steroids, that is, how much extra survival probability is associated with giving steroids?

   A No extra survival probability with steroids.
   B About 1-5 percentage points
   C About 10 to 15 percentage points
   D About 50 percentage points
   E About 75 percentage points

(b) For the babies where the mother was given steroids, how does the survival probability depend on the birth weight of the baby:

- [A] No dependence.
- [B] Increases by about 25 percentage points.
- [C] Increases by about 50 percentage points.
- [D] Increases by about 25 percentage points per pound.
- [E] Increases by about 50 percentage points per pound.

(c) For the babies where the mother was given steroids, how does the survival probability depend on the sex of the baby?

- [A] No dependence.
- [B] Higher for girls by about 15 percentage points.
- [C] Higher for boys by about 20 percentage points.
- [D] Higher for girls by about 40 percentage points.
- [E] Higher for boys by about 40 percentage points.

(d) How would you look for an interaction between birth weight and baby's sex in accounting for survival?

- [A] Compare survival of males to females at a given weight.
- [B] Compare survival of males across different weights.
- [C] Compare survival of females across different weights.
- [D] Compare the difference in survival between males and females across different weights.

Do you see signs of a substantial interaction between birth weight and sex in accounting for survival? (Take substantial to mean "greater than 10 percentage points.") Yes No

(e) How would you look for a substantial interaction between steroid use and baby's sex in accounting for survival.

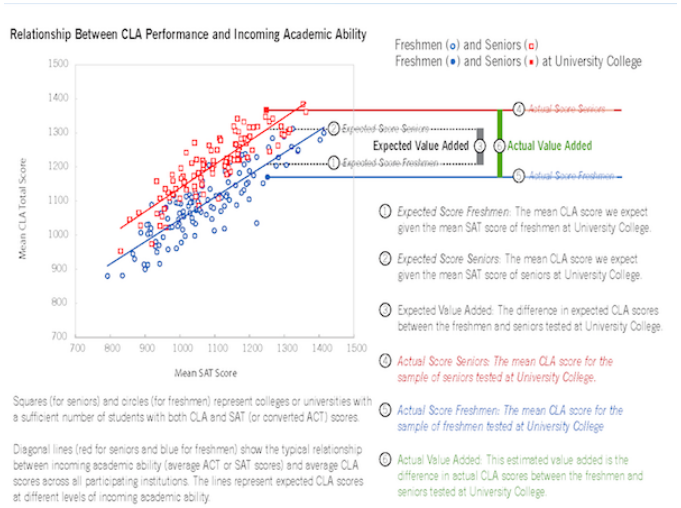
- [A] Compare survival of males to females when the mother was given steroids.
- [B] Compare survival of males between steroid given and steroid not given.
- [C] Compare survival of females between steroid given and steroid not given.
- [D] Compare the difference in survival between males and females between steroid given and steroid not given.

Do you see signs of a substantial interaction between steroid use and sex in accounting for survival? (Take substantial to mean "greater than 10 percentage points.") Yes No

Prob 6.06

The graphic in the Figure is part of a report describing a standardized test for college graduates, the Collegiate Learning Assessment (CLA). The test consists of several essay questions which probe students' critical thinking skills.

Although individual students take the test and receive a score, the purpose of the test is not to evaluate the students individually. Instead, the test is intended to evaluate the effect that the institution has on its students as indicated by the difference in test scores between 1st- and 4th-year students (freshmen and seniors). The cases in the graph are institutions, not individual students.



Council for Aid to Education, "Collegiate Learning Assessment: Draft Institutional Report, 2005-6" `http://www.cae.org`

There are three variables involved in the graphic:

cla The CLA test score (averaged over each institution) shown on the vertical axis

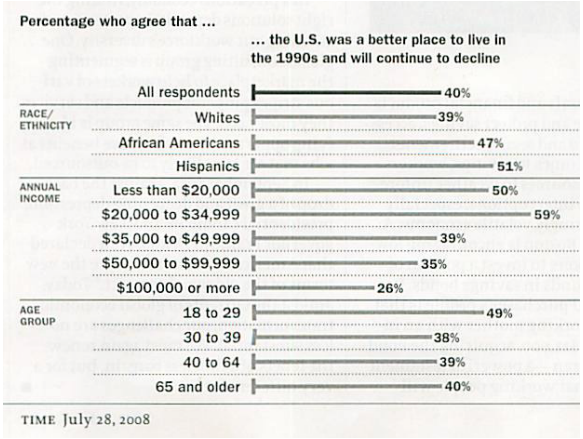
sat The SAT test score of entering students (averaged over each institution) shown on the horizontal axis

class Whether the CLA test was taken by freshmen or seniors. (In the graph: blue for freshmen, red for seniors)

What model is being depicted by the straight lines in the graph? Give your answer in the standard modeling notation (e.g, A $\sim$ B+C) using the variable names above. Make sure to indicate what interaction term, if any, has been included in the model and explain how you can tell whether the interaction is or is not there.

Prob 6.07

Time Magazine reported the results of a poll of people's opinions about the U.S. economy in July 2008. The results are summarized in the graph.



[Source: *Time*, July 28, 2008, p. 41]

The variables depicted in the graph are:

- Pessimism, as indicated by agreeing with the statement that the U.S. was a better place to live in the 1990s and will continue to decline.
- Ethnicity, with three levels: White, African American, Hispanic.
- Income, with five levels.
- Age, with four levels.

Judging from the information in the graph, which of these statements best describes the model pessimism ~ income?

A Pessimism declines as incomes get higher.
B Pessimism increases as incomes get higher.
C Pessimism is unrelated to income.

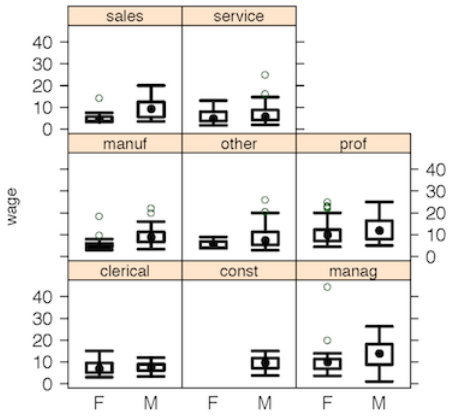
Again, judging from the information in the graph, which of these statements best describes the model pessimism ~ age?

A Pessimism is highest in the 18-29 age group.
B Pessimism is highest in the 64 and older group.
C Pessimism is lowest among whites.
D Pessimism is unrelated to age.

Poll results such as this are almost always reported using just one explanatory variable at a time, as in this graphic. However, it can be more informative to know the effect of one variable while *adjusting for* other variables. For example, in looking at the connection between pessimism and age, it would be useful to be able to untangle the influence of income.

Prob 6.08

Here is a display constructed using the Current Population Survey wage data:



Which of the following commands will make this? Each of the possibilities is a working command, so try them out and see which one makes the matching plot. Before you start, make sure to read in the data with

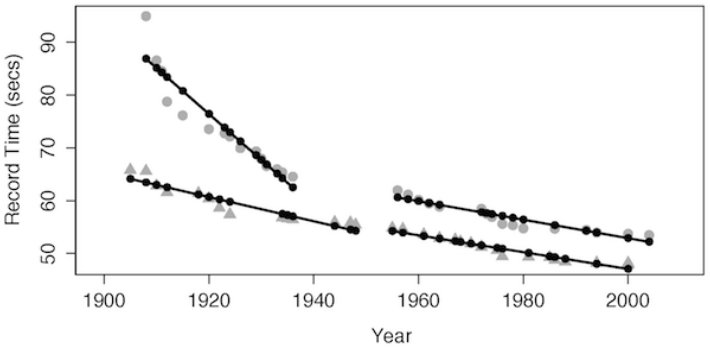
```
> cps = fetchData("cps.csv")
```

A `bwplot(wage~sex,groups=sector,data=cps)`
B `bwplot(wage~sex|sector,data=cps)`
C `bwplot(wage~cross(sex,sector),data=cps)`

Prob 6.10

It's possible to have interaction terms that involve more than two variables. For instance, one model of the swimming record data displayed in Chapter 4 was time ~ year*sex. This model design includes an interaction term between year and sex. This interaction term produces fitted model values that fall on lines with two different slopes, one for men and one for women. Now consider a third possible term to add to the model, the transform term that is "yes" when the year is larger than 1948 and "no" when the year is 1948 or earlier. Call this variable "post-war," since World War II ended in 1945 and the Olympic games resumed in 1948. This can be interpreted to represent the systematic changes that occurred after the war.

Here is the model of the swimming record data that includes an intercept term, main terms for year, sex, and post-war, and interaction terms among all of those: a three-way interaction.



A two-way interaction term between sex and year allows there to be differently sloping lines for men and women. A

three-way interaction term among sex, year, and post-war allows even more flexibility; the difference between slopes for men and women can be different before and after the war. You can see this from the graph. Before 1948, men's and women's slopes are very different. After the war the slopes are almost the same.

Explain how this graph gives support for the following interpreation: Before the war, women's participation in competitive sports was rapidly increasing. As more women became involved in swimming, records were rapidly beaten. After the war, both women and men had high levels of participation and so new records were the result of better methods of training. Those methods apply equally to men and women and so records are improving at about the same rate for both sexes.

Prob 6.11

Consider the following situation. In order to encourage schools to perform well, a school district hires an external evaluator to give a rating to each school in the district. The rating is a single number based on the quality of the teachers, absenteeism among the teachers, the amount and quality of homeworks the the teachers assign, and so on.

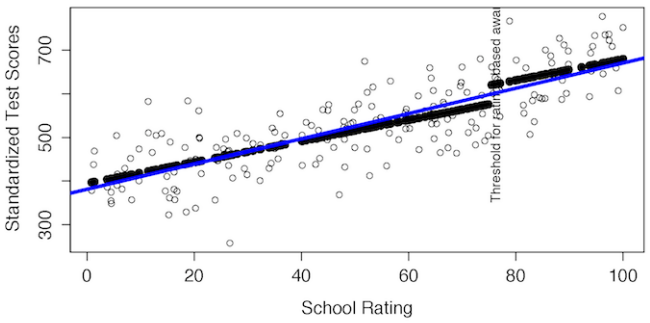
To reward those schools that do well, the district gives a moderate salary bonus to each teacher and a fairly large budget increase to the school itself.

The next year, the school district publishes data showing that the students in schools that received the budget increases had done much better on standarized test scores than the schools that hadn't gotten the increases. The school district argues that this means that increasing budgets overall will improve performance on standardized tests.

The teacher's union supports the idea of higher budgets, but objects to the rating system, saying that it is meaningless and that teacher pay should not be linked to it. The Taxpayers League argues that there is no real evidence that higher spending produces better results. They interpret the school district's data as indicating only that the higher ranked schools are better and, of course, better schools give better results. Those schools were better before they won the ratings-based budget increase.

This is a serious problem. Because of the way the school district collected its data, being a high-rated school is confounded with getting a higher budget.

A modeling technique for dealing with situations like this is called **threshold regression**. Threshold regression models student test scores at each school as a function of the school rating, but includes another variable that indicates whether the school got a budget increase. The budget increase variable is almost the same thing as the school rating: because of the way the school district awarded the increases, it is a threshold transformation of the school rating.



The graph shows some data (plotted as circles) from a simulation of this situation in which the budget increase had a genuine impact of 50 points in the standardized test. The solid line shows the model of test score as a function of school rating, with only the main effect. This model corresponds to the claim that the threshold has no effect. The solid dots are the model values from another model, with rating as a main effect and a threshold transformation of rating that corresponds to which schools got the budget increase.

Explain how to interpret the models as indicating the effect of the budget increase. In addition to your explanation, make sure to give a numerical estimate of how big the effect is, according to the model as indicated in the graph.

An important statistical question is whether the data provide good support for the claim that the threshold makes a difference. (Techniques for answering this question are discussed later in the book). The answer depends both on the size of the effect, and how much data is used for constructing the model. For the simulation here, it turns out that the threshold model has successfully detected the effect of the budget increase.

Graphics Technique 6.1

Given data and a model design, the computer will find the model function and model values for you. As an example, consider the Current Population Survey data `cps.csv` . Suppose you want to build a model with wage as a response variable and age and sex as explanatory variables incorporated as main terms. Also include the intercept term, as usual.

Using the model design language, this model is wage $\sim$ 1 + age + sex.

You first need to read in the data frame.

```
w = fetchData("cps.csv")
```

Next, use the `lm` operator to find the model function:

```
mod1 = lm( wage ~ 1 + age + sex, data=w)
```

The two arguments are:

the model design wage $\sim$ 1 + age + sex.

the data to be used This always looks like `data=w` where the name of the data frame is used.

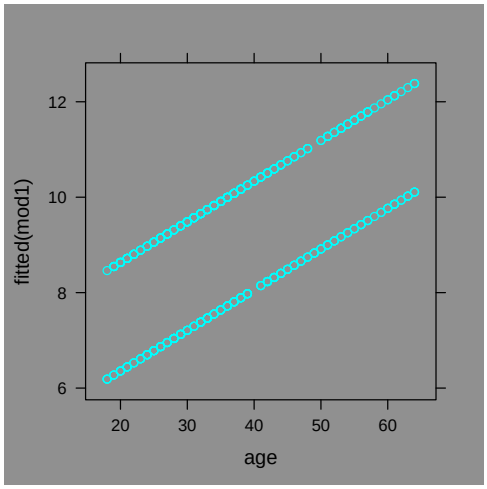
The `mod1 =` ... part of the command simply gives the model a name so that you can use it later on. If you construct more than one model, it makes sense to give them different

names. But don't re-use the name of the data frame; a name can be used for only one thing at a time.

In making a graph of the function, the model values will always be plotted on the vertical axis. But you have a choice of what to put on the horizontal axis. Here are the commands which use the `xyplot` operator:

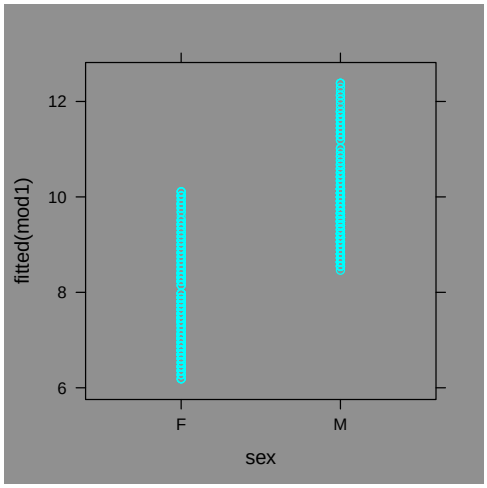
Model values versus age .

```
xyplot( fitted(mod1) ~ age, data=w )
```



Model values versus sex .

```
print(xyplot( fitted(mod1) ~ sex, data=w ))
```



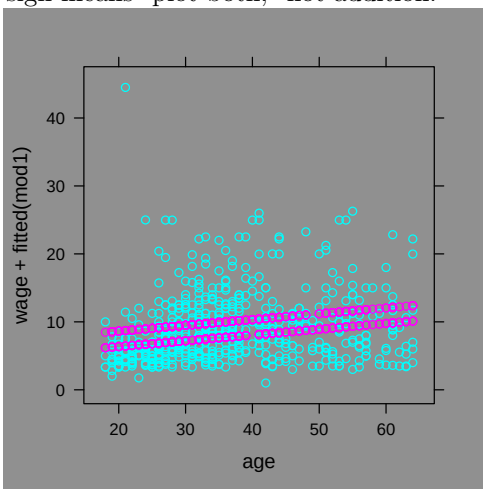
The fitted model values can be accessed with `fitted(mod1)`. The choice of which explanatory variable to plot on the horizontal axis is specified by the name following the ~ sign in the plotting command. Remember to include the name of the data frame in the last argument: `data=w`.

Some elaborations are possible.

Show the response variable in addition to the model values. This involves putting the response variable name to the left of the ~ sign:

```
print(xyplot( wage + fitted(mod1) ~ age, data=w ))
```

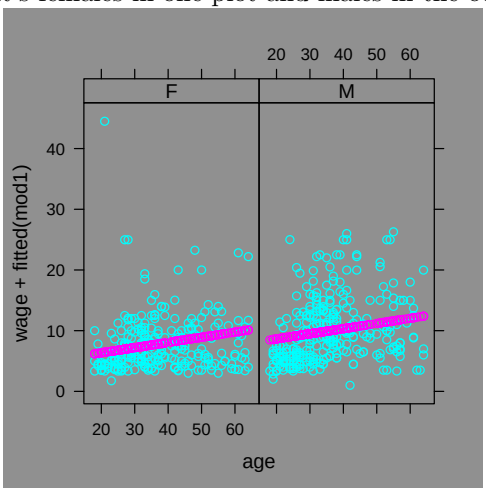
The + sign means "plot both," not addition.



Break up the display according to one or more categorical variables.

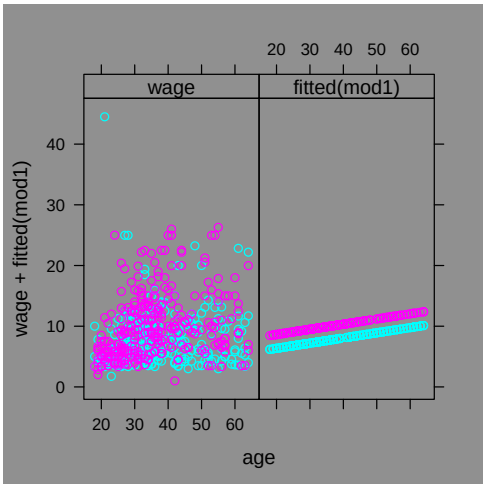
```
xyplot( wage + fitted(mod1) ~ age | sex, data=w )
```

This uses a vertical bar followed by the name of the categorical variable. Each of the levels of the variable to the right of the vertical bar | will have a separate plot, so in the graphics it's females in one plot and males in the other.



A different way to break up the display makes it easier to compare the model values for the different groups:

```
xyplot(wage+fitted(mod1) ~ age,groups=sex,data=w)
```



Graphics Technique 6.2

Sometimes when making a graphic, you want to change some aspect of it or add new elements to it. The lattice graphics system provides ways to do this. To illustrate, make a simple density plot in the ordinary way.

```
galton = fetchData("galton.csv")
densityplot( galton$height )
```

In looking at the plot, you realize that you want a better label for the horizontal axis. One option is to redo the plot from scratch:

```
densityplot( galton$height, xlab="Children")
```

Another option is to instruct the graphics system just to change the specific parameters you want, for example, to change the x-label and to delete the points plotted at the bottom:

```
trellis.last.object( xlab="Height (inches)",
  plot.points=FALSE)
```

Now suppose you decide to superimpose density plots of the mothers' and fathers' heights. To do this, you need to tell the lattice system that you want to *add on* to the plot. This is done with the `trellis.focus` command:

```
trellis.focus()
```

You'll notice that the original graphic is surrounded by a red line. Once this is done, you can add on to the plot.

To start, use `density` to compute the density curve for the mothers and the fathers. This will not plot out those curves, yet.

```
mother = density(galton$mother)
father = density(galton$father)
```

Now you can add those curves to the plot, using the `llines` command. This takes a series of x- and y- points, but those are already contained in the output of the `density` program:

```
llines(mother, col='red', lwd=2)
llines(father, col='black', lwd=2)
```

Finally, tell the lattice system that you are done adding to the original plot:

```
trellis.unfocus()
```

Use these commands to construct the graph comparing the distribution of the mothers' and fathers' heights to the children's heights. Describe how they are different and explain why this is.

Other useful commands for adding elements to lattice plots are `llines`, `lpoints`, and `ltext`.

Chapter Seven Reading Questions

- What is the role of the response variable in a model formula?
- What is the purpose of constructing indicator variables from categorical variables?
- How can model coefficients be used describe relationships? What are the relationships between?
- What is Simpson's paradox?
- Given an example of how the meaning of a coefficient of a particular term can depend on what other model terms are included in the model?

Prob 7.01

There is a correspondence between the model formula and the coefficients found when fitting a model.

For each of the following model formulas, tell what the coefficient is:

(a) $3 - 7x + 4y + 17z$

- Intercept: -7 3 4 17
- z coef: -7 3 4 17
- y coef: -7 3 4 17
- x coef: -7 3 4 17

(b) $1.22 + 0.12\mathsf{age} + 0.27\mathsf{educ} - 0.04\mathsf{age} : \mathsf{educ}$

- Intercept: -0.04 0.12 0.27 1.22
- educ coef: -0.04 0.12 0.27 1.22
- age coef: -0.04 0.12 0.27 1.22
- age:educ coef: -0.04 0.12 0.27 1.22

(c) $8 + 3\mathsf{color}\mathbf{Red} - 4\mathsf{color}\mathbf{Blue}$

- Intercept: -4 3 8
- color**Red** coef: -4 3 8
- color**Blue** coef: -4 3 8

Prob 7.02

For each of the following coefficient reports, tell what the corresponding model formula is:

(a)

term	coef
Intercept	10
x	3
y	5

[A] $x + y$
[B] $1 + x + y$
[C] $10 + 3 + 5$
[D] $10 + 3x + 5y$
[E] $10x + 5y + 3$

(b)

term	coef
Intercept	4.15
age	-0.13
educ	0.55

[A] age
[B] age + educ
[C] $4.15 - 0.13 + 0.55$
[D] $4.15\mathsf{age} - 0.13\mathsf{educ} + 0.55$
[E] $4.15 - 0.13\mathsf{age} + 0.55\mathsf{educ}$

Prob 7.03

For some simple models, the coefficients can be interpreted as grand means, group-wise means, or differences between group-wise means. In each of the following, A, B, and C are quantitative variables and color is a categorical variable with levels "red," "blue," and "green."

(a) The model A ~ color gave these coefficients:

term	coefficient
Intercept	10
color**Blue**	5
color**Green**	12

- What is the mean of A for those cases that are Blue:
  5 10 12 15 17 22 27 unknown
- What is the mean of A for those cases that are Green:
  5 10 12 15 17 22 27 unknown
- What is the mean of A for those cases that are Red:
  5 10 12 15 17 22 27 unknown
- What is the grand mean of A for all cases:
  5 10 12 15 17 22 27 unknown

(b) The model B ~ color - 1 gave these coefficients:

term	coefficient
color**Red**	100
color**Blue**	-40
color**Green**	35

- What is the group mean of B for those cases that are Blue:
  -40 -5 0 35 60 65 100 135 unknown
- What is the group mean of B for those cases that are Red:
  -40 -5 0 35 60 65 100 135 unknown
- What is the group mean of B for those cases that are Green:
  -40 -5 0 35 60 65 100 135 unknown
- What is the grand mean of B for all cases:
  -40 -5 0 35 60 65 100 135 unknown

(c) The model C ~ 1 gave this coefficient:

term	coefficient
Intercept	4.7

- What is the group mean of C for those cases that are Blue:
  0.0 4.7 unknown
- What is the grand mean of C for all cases:
  0.0 4.7 unknown

Prob 7.04

Using the appropriate data set and correct modeling statements, compute each of these quantities and give the model statement you used (e.g., age ~ sex)

(a) From the `cps.csv` data, what is the mean age of single people? (Pick the closest answer.)
28 31 32 35 39 years.
What was your model expression?

(b) From the `cps.csv` data, what is the difference between the mean ages of married and single people? (Pick the closest answer.)

A Single people are, on average, 5 years younger.
B Single people are, on average, 5 years older.
C Single people are, on average, 7 years younger.
D Single people are, on average, 7 years older.

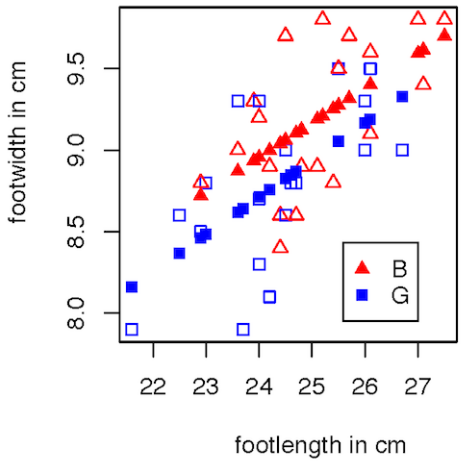
What was your model expression?

(c) From the `swim100m.csv` data, what is the mean swimming time for women? (Pick the closest.)
55 60 65 70 75 80 seconds.
What is your model expression?

(d) From the `utilities.csv` data, what is the mean CCF for November? (Pick the closest.) (Hint: use `as.factor(month)` to convert the month number to a categorical variable.)
-150 -93 42 150 192
What is your model expression?

Prob 7.05

Here is a graph of the kids feet data showing a model of footwidth as a function of footlength and sex. Both the length and width variables are measured in cm.



The model values are solid symbols, the measured data are hollow symbols.

Judging from the graph, what is the model value for a boy with a footlength of 22 cm?

A 8.0cm
B 8.5cm
C 9.0cm
D 9.5cm
E Can't tell from this graph.

According to the model, after adjusting for the difference in foot length, what is the typical difference between the width of a boy's foot and a girl's foot?

A no difference
B 0.25cm
C 0.50cm
D 0.75cm
E 1.00cm
F Can't tell from this graph.

Judging from the graph, what is a typical size of a residual from the model?

A 0.10cm
B 0.50cm
C 1.00cm
D 1.50cm
E Can't tell from this graph.

Prob 7.06

In the `swim100m.csv` data, the variables are

- time: World record time (in seconds)
- year: The year in which the record was set
- sex: Whether the record is for men or women.

Here are the coefficients from several different fitted models.

```
> lm( time ~ year, data=swim)
Coefficients:
(Intercept)      year
   567.2420   -0.2599

> lm( time ~ year+sex, data=swim)
Coefficients:
(Intercept)      year      sexM
   555.7168   -0.2515   -9.7980

> lm( time ~ year*sex, data=swim)
Coefficients:
(Intercept)      year      sexM year:sexM
   697.3012   -0.3240 -302.4638    0.1499

> lm( time ~ sex, data=swim)
Coefficients:
(Intercept)      sexM
      65.19    -10.54
```

For each of the following, pick the appropriate model from the set above and use its coefficients to answer the question.

(a) How does the world record time typically change from one year to the next for both men and women taken together?

-302.4 -10.54 -9.79 -0.2599 -0.2515 -0.324 -0.174

(b) How does the world record time change from one year to the next for women only?

-302.4 -10.54 -9.79 -0.2599 -0.2515 -0.324 -0.174

(c) How does the world record time change from one year to the next for men only?

-302.4 -10.54 -9.79 -0.2599 -0.2515 -0.324 -0.174

Prob 7.07

In the SAT data `sat.csv` , the variables have these units:

- sat has units of "points."
- expend has units of "dollars."
- ratio has units of "students."
- frac has units of "percentage points."

Consider the model formula

$$\text{sat} = 994 + 12.29\ \text{expend} - 2.85\ \text{frac}$$

(a) What are the units of the coefficient 994?

A points
B dollars
C students
D percentage points
E points per dollar
F students per point
G points per student
H points per percentage points

(b) What are the units of the coefficient 12.29?

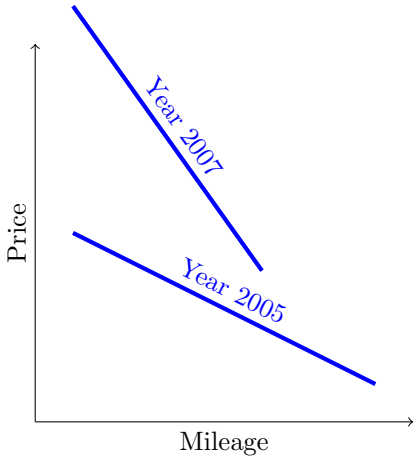
A points
B dollars
C students
D dollars per student
E points per dollar
F students per point
G points per student

(c) What are the units of the coefficient 2.85?

A points
B dollars
C percentage points
D points per dollar
E students per point
F points per student
G points per percentage points

Prob 7.08

The graph shows schematically a possible relationship between used car price, mileage, and the car model year.



Consider the model price ~ mileage*year.

In your answers, treat year as a simple categorical variable, and use year 2005 as the reference group when thinking about coefficients.

(a) What will be the sign of the coefficient on mileage?

- [A] Negative
- [B] Zero
- [C] Positive
- [D] No way to tell from the information given

(b) What will be the sign of the coefficient on model year?

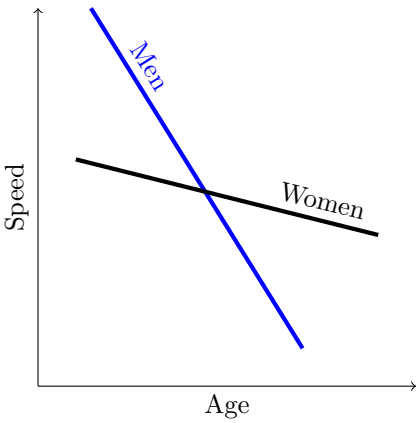
- [A] Negative
- [B] Zero
- [C] Positive
- [D] No way to tell from the information given

(c) What will be the sign of the interaction coefficient?

- [A] Negative
- [B] Zero
- [C] Positive
- [D] There is no interaction coefficient.
- [E] No way to tell from the information given

Prob 7.09

The graph shows schematically a hypothesized relationship between how fast a person runs and the person's age and sex.



Consider the model speed ~ age*sex.

(a) What will be the sign of the coefficient on age?

- [A] Negative
- [B] Zero
- [C] Positive
- [D] No way to tell, even roughly, from the information given

(b) What will be the sign of the coefficient on sex? (Assume that the sex variable is an indicator for women.)

- [A] Negative
- [B] Zero
- [C] Positive

(c) What will be the sign of the interaction coefficient? (Again, assume that the sex variable is an indicator for women.)

- [A] Negative
- [B] Zero
- [C] Positive
- [D] There is no interaction coefficient.
- [E] No way to tell, even roughly, from the information given

Prob 7.10

Consider this model of a child's height as a function of the father's height, the mother's height, and the sex of the child.

height ~ father*sex + mother*sex

Use the Galton data `galton.csv` to fit the model and examine the coefficients. Based on the coefficients, answer the following:

(a) There are two boys, Bill and Charley. Bill's father is 1 inch taller than Charley's father. According to the model, and assuming that their mothers are the same height, how much taller should Bill be than Charley?

- [A] They should be the same height.
- [B] 0.01 inches
- [C] 0.03 inches
- [D] 0.31 inches
- [E] 0.33 inches
- [F] 0.40 inches
- [G] 0.41 inches

(b) Now imagine that Bill and Charley's fathers are the same height, but that Charley's mother is 1 inch taller than Bill's mother. According to the model, how much taller should Charley be than Bill?

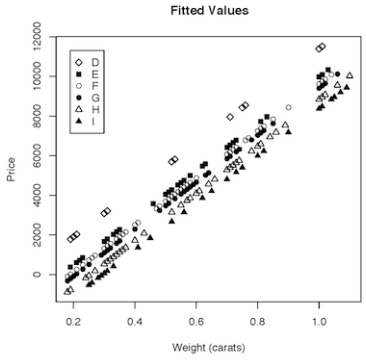
- [A] They should be the same height.
- [B] 0.01 inches
- [C] 0.03 inches
- [D] 0.31 inches
- [E] 0.33 inches
- [F] 0.40 inches
- [G] 0.41 inches

(c) Now put the two parts together. Bill's father is one inch taller than Charley's, but Charley's mother is one inch taller than Bill's. How much taller is Bill than Charley?

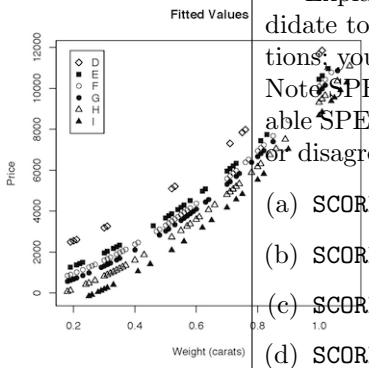
- [A] They should be the same height.
- [B] 0.03 inches
- [C] 0.08 inches
- [D] 0.13 inches
- [E] 0.25 inches

Prob 7.11

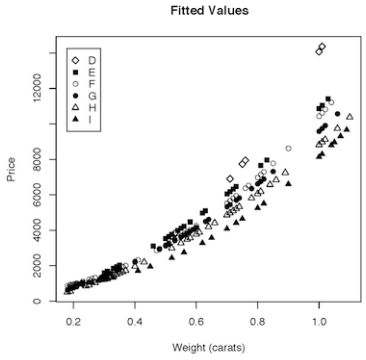
The file `diamonds.csv` contains several variables relating to diamonds: their price, their weight (in carats), their color (which falls into several classes — D, E, F, G, H, I), and so on. The following several graphs show different models fitted to the data: price is the response variable and weight and color are the explanatory variables.



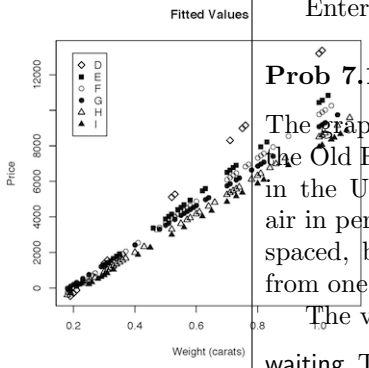
Graph 1



Graph 2



Graph 3



Graph 4

Which model corresponds to which graph?

(a) `lm( price~carat + color, data=diamonds)`

Which graph? Graph 1 Graph 2 Graph 3 Graph 4

(b) `lm( price~carat * color, data=diamonds)`

Which graph? Graph 1 Graph 2 Graph 3 Graph 4

(c) `lm( price~poly(carat,2) + color, data=diamonds)`

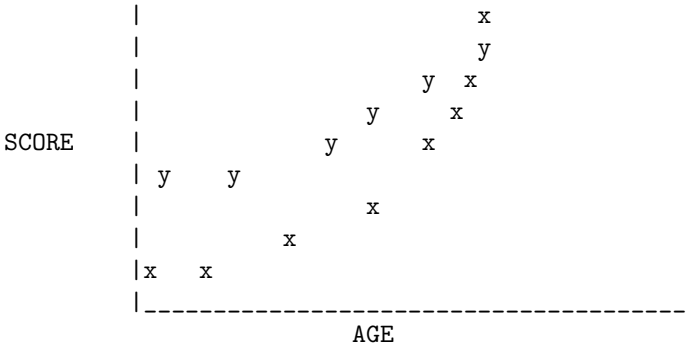
Which graph? Graph 1 Graph 2 Graph 3 Graph 4

(d) `lm( price~poly(carat,2) * color, data=diamonds)`

Which graph? Graph 1 Graph 2 Graph 3 Graph 4

Prob 7.12

The graph shows data on three variables, SCORE, AGE, and SPECIES. The SCORE and AGE are quantitative. SPECIES is categorical with levels `x` and `y`.



Explain which of the following models is plausibly a candidate to describe the data. (Don't do any detailed calculuations, you can't because the axes aren't marked with a scale.) Note SPECIESx means that the case has a level of x for variable SPECIES. For each model explain in what ways it agrees or disagrees with the graphed data.

(a) `SCORE = 10 - 2.7 AGE + 1.3 SPECIESx`

(b) `SCORE = 10 + 5.0 AGE - 2 AGE^2 - 1.3 SPECIESx`

(c) `SCORE = 10 + 5.0 AGE + 2 AGE^2 - 1.3 SPECIESx`

(d) `SCORE = 10 + 2.7 AGE + 2 AGE^2 - 1.3 SPECIESx + 0.7 AGE * SPECIESx`

Enter your answers for all four models here:

Prob 7.13

The graphs below show models values for different models of the Old Faithful geyser, located in Yellowstone National Park in the US. The geyser blows water and steam high in the air in periodic eruptions. These eruptions are fairly regularly spaced, but there is still variation in the time that elapses from one eruption to the next.

The variables are

waiting The time from the previous eruption to the current one

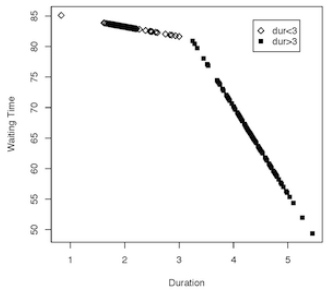
duration The duration of the previous eruption

biggerThan3 A categorical variable constructed from duration, which depicts simply whether the duration was greater or less than 3 minutes.

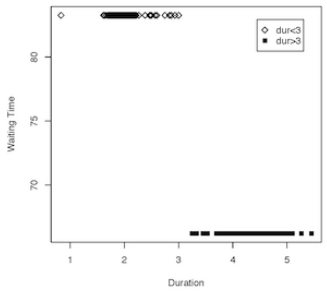
In each case, judge from the shape of the graph which model is being presented.

- (A) waiting ~ duration
- (B) waiting ~ duration + biggerThan3
- (C) waiting ~ duration*biggerThan3
- (D) waiting ~ biggerThan3
- (E) waiting ~ poly(duration,2)

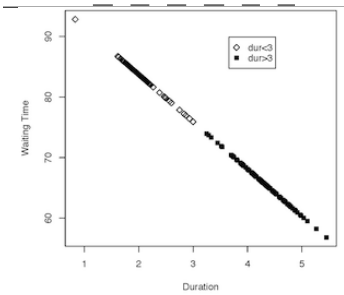
- (F) waiting ∼ poly(duration,2)*biggerThan3



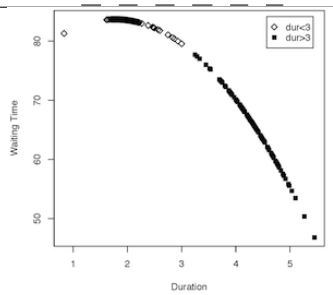
1. A B C D E F



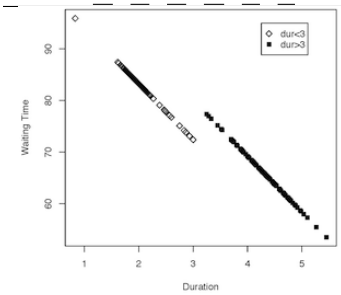
2. A B C D E F



3. A B C D E F



4. A B C D E F



5. A B C D E F

Prob 7.14

Here is a report from the New York *Times*:

> It has long been said that regular physical activity and better sleep go hand in hand. But only recently have scientists sought to find out precisely to what extent. One extensive study published this year looked for answers by having healthy children wear actigraphs — devices that measure movement — and then seeing whether more movement and activity during the day meant improved sleep at night.
>
> The study found that sleep onset latency — the time it takes to fall asleep once in bed — ranged from as little as roughly 10 minutes for some children to more than 40 minutes for others. But physical activity during the day and sleep onset at night were closely linked: every hour of sedentary activity during the day resulted in an additional three minutes in the time it took to fall asleep at night. And the children who fell asleep faster ultimately slept longer, getting an extra hour of sleep for every 10-minute reduction in the time it took them to drift off. (Anahad O'Connor, Dec. 1, 2009 — the complete article is at `http://www.nytimes.com/2009/12/01/health/01really.html`.)

There are two models described here with two different response variables: sleep onset latency and duration of sleep.

(a) In the model with sleep onset latency as the response variable, what is the explanatory variable?

- A Time to fall asleep.
- B Hours of sedentary activity.
- C Duration of sleep.

(b) In the model with duration of sleep as the response variable, what is the explanatory variable?

- A Time to fall asleep.
- B Hours of sedentary activity.
- C Duration of sleep.

(c) Suppose you are comparing two groups of children. Group A has 3 hour of sedentary activity each day, Group B has 8 hours of sedentary activity. Which of these statements is best supported by the article?

- A The children in Group A will take, on average, 3 minutes less time to fall asleep.
- B The children in Group B will have, on average, 10 minutes less sleep each night.
- C The children in Group A will take, on average, 15 minutes less time to fall asleep.
- D The children in Group B will have, on average, 45 minutes less sleep each night.

(d) Again comparing the two groups of children, which of these statements is supported by the article?

- A The children in Group A will get, on average, about an hour and a half hours of extra sleep compared to the Group B children.
- B The children in Group A will get, on average, about 15 minutes more sleep than the Group B children.
- C The two groups will get about the same amount of sleep.

Prob 7.15

Car prices vary. They vary according to the model of car, the optional features in the car, the geographical location, and the respective bargaining abilities of the buyer and the seller.

In this project, you are going to investigate the influence of at least three variables on the asking price for used cars:

- Model year
- Mileage
- Geographical location

These variables are relatively easy to measure and code. There are web sites that allow us quickly to collect a lot of cases. One site that seems easy to use is `www.cars.com`. Pick a particular model of car that is of interest to you. Also, pick a few scattered geographical locations. (At `www.cars.com` you can specify a zip code, and restrict your search to cars within a certain distance of that zip code.)

For each location, use the web site to find prices and the other variables for 50-100 cars. Record these in a spreadsheet with five variables: price, model year, mileage, location, model name. (The model name will be the same for all your data. Recording it in the spreadsheet will help in combining data for different types of cars.) You may also choose to record some other variables of interest to you.

Using your data, build models make a series of claims about the patterns seen in used-car prices. Some basic claims that you should make are in this form:

- Looking just at price versus mileage, the price of car model XXX falls by 12 cents per mile driven.
- Looking just at price versus age, the price of car model XXX falls by 1000 dollars per year of age driven.
- Considering both age and mileage, the price of car model XXX falls by ...
- Looking at price versus location, the price differs ...

You may also want to look at interaction terms, for example whether the effect of mileage is modulated by age or location.

Note whether there are outliers in your data and indicate whether these are having a strong influence on the coefficients you find.

Year	Vehicle	Price	Mileage
2006	Mazda Miata MX-5	$23,995	969
2004	Mazda Miata MX-5	$19,977	10,198
2004	Mazda Miata MX-5	$16,999	20,560
2002	Mazda Miata MX-5 SE	$15,995	18,826
2003	Mazda Miata MX-5	$15,995	15,947
2002	Mazda Miata MX-5	$14,995	44,215
2002	Mazda Miata MX-5 LS	$14,500	18,000
2002	Mazda Miata MX-5	$14,500	1,800
2001	Mazda Miata MX-5	$13,995	70,136
2002	Mazda Miata MX-5	$12,995	43,978
1999	Mazda Miata MX-5	$9,500	63,000
1999	Mazda Miata MX-5	$9,500	63,000
1991	Mazda Miata MX-5	$4,200	90,000

Price and other information about used Mazda Miatas in the Saint Paul, Minnesota area from `www.cars.com`.

Prob 7.16

Here is a news article summarizing a research study by Bingham *et al.*, "Drinking Behavior from High School to Young Adulthood: Differences by College Education," Alcoholism: Clinical & Experimental Research; Dec. 2005; vol. 29; No. 12

After reading the article, answer these questions:

1. The article headline is about "drinking behavior." Specifically, how are they measuring drinking behavior?
2. What explanatory variables are being studied?
3. Are any interactions reported?
4. Imagine that the study was done using a single numerical indicator of drinking behavior, a number that would be low for people who drink little and don't binge drink, and would be high for heavy and binge drinkers. For a model with this numerical index of drinking behavior as the output, what structure of model is implied by the article?
5. For the model you wrote down, indicate which coefficients are positive and which negative.

Binge Drinking Is Age-Related Phenomenon

By Katrina Woznicki, MedPage Today Staff Writer December 14, 2005

ANN ARBOR, Mich., Dec. 14 - Animal House notwithstanding, going to college isn't an excess risk factor for binge drinking any more than being 18 to 24 years old, according to researchers here.

The risks of college drinking may get more publicity, but the college students are just late starters, Raymond Bingham, Ph.D., of the University of Michigan and colleagues reported in the December issue of Alcoholism: Clinical & Experimental Research.

Young adults in the work force or in technical schools are more likely to have started binge drinking in high school and kept it up, they said.

The investigators said the findings indicated that it's incorrect to assume, as some do, that young adults who don't attend college are at a lower risk for alcohol misuse than college students.

"The ones who don't go on to a college education don't change their at-risk alcohol consumption," Dr. Bingham said. "They don't change their binge-drinking and rates of drunkenness."

In their study comparing young adults who went to college with those who did not, they found that men with only a high school education were 91% more likely to have greater alcohol consumption than college students in high school. Men with only a postsecondary education (such as technical school) were 49% more likely to binge drink compared with college students.

There were similar results with females. Women with only a high school education were 88% more likely to have greater alcohol consumption than college students.

The quantity and frequency of alcohol consumption increased significantly from the time of high school graduation

at the 12th grade to age 24 ($p < 0.001$), investigators reported in the December issue of Alcoholism: Clinical & Experimental Research

College students drank, too, but their alcohol use peaked later than their non-college peers. By age 24, there was little difference between the two groups, the research team reported.

"In essence," said Dr. Bingham, "men and women who did not complete more than a high-school education had high alcohol-related risk, as measured by drunkenness and heavy episodic drinking while in the 12th grade, and remained at the same level into young adulthood, while levels for the other groups increased."

The problem, Dr. Bingham said, is that while it's easier for clinicians to target college students, a homogenous population conveniently located on concentrated college campuses, providing interventions for at-risk young adults who don't go on to college is going to be trickier.

"The kids who don't complete college are everywhere," Dr. Bingham said. "They're in the work force, they're in the military, they're in technical schools."

Dr. Bingham and his team surveyed 1,987 young adults who were part of the 1996 Alcohol Misuse Prevention Study. All participants had attended six school districts in southeastern Michigan. They were interviewed when they were in 12th grade and then again at age 24. All were unmarried and had no children at the end of the study. Fifty-one percent were male and 84.3% were Caucasian.

The 1,987 participants were divided into one of three education status groups: high school or less; post-secondary education such as technical or trade school or community college, but not a four-year degree college; and college completion.

The investigators looked at several factors, including quantity and frequency of alcohol consumption, frequency of drunkenness, frequency of binge-drinking, alcohol use at young adulthood, cigarette smoking and marijuana use.

Overall, the men tend to drink more than the women regardless of education status. The study also showed while lesser-educated young adults may have started heavier drinking earlier on, college students quickly caught up.

For example, the frequency of drunkenness increased between 12th grade and age 24 for all groups except for men and women with only a high school education ($p < 0.001$).

"The general pattern of change was for lower-education groups to have higher levels of drunkenness in the 12th grade, and to remain at nearly the same level while college-completed men and women showed the greatest increases in drunkenness," the authors wrote.

Lesser-educated young adults also started binge-drinking earlier, but college students, again, caught up. High school-educated women were 27% more likely to binge drink than college women, for example. High-school-educated men were 25% more likely to binge drink than men with post-secondary education.

But binge-drinking frequency increased 21% more for college-educated men than post-secondary educated men. And college women were 48% more likely to have an increase in binge-drinking frequency than high school-educated women.

The study also found post-secondary educated men had the highest frequency of drunken-driving. High school educated men and women reported the highest frequencies of smoking in the 12th grade and at age 24 and also showed the greater increase in smoking prevalence over this period whereas college-educated men and women had the lowest levels of smoking.

Then at age 24, the investigators compared those who were students to those who were working and found those who were working were 1.5 times more likely to binge drink ($p < 0.003$), 1.3 times more likely to be in the high drunkenness group ($p < 0.018$), and were 1.5 times more likely to have a greater quantity and frequency of alcohol consumption ($p < 0.005$).

"The transition from being a student to working, and the transition from residing with one's family of origin to another location could both partially explain differences in patterns," the authors wrote.

Dr. Bingham said the findings reveal that non-college attending young adults "experience levels of risk that equal those of their college-graduating age mates."

Prob 7.17

For the simple model A ~ G. where G is a categorical variable, the coefficients will be group means. More precisely, there will be an intercept that is the mean of one of the groups and the other coefficients will show how the mean of the other groups each differ from the reference group.

Similarly, when there are two grouping variables, G and H, the model A ~ G + H + G:H (which can be abbreviated A ~ G*H) will have coefficients that are the group-wise means of the crossed groups. Perhaps "subgroup-wise means" is more appropriate, since there will be a separate mean for each subgroup of G divided along the lines of H. The interaction term G:H allows the model to account for the influence of H separately for each level of G.

However, the model A ~ G + H does **not** produce coefficients that are group means. Because no interaction term has been included, this model cannot reflect how the effect of H differs depending on the level of G. Instead, the model coefficients reflect the influence of H as if it were the same for all levels of G.

To illustrate these different models, consider some simple data.

Suppose that you found in the literature an article about the price of small pine trees (either Red Pine or White Pine) of different heights in standard case/variable format, which would look like this:

Case #	Color	Height	Price
1	Red	Short	11
2	Red	Short	13
3	White	Tall	37
4	White	Tall	35
and so on ...			

Commonly in published papers, the raw case-by-case data isn't reported. Rather some summary of the raw data is presented. For example, there might be a summary table like this:

SUMMARY TABLE

Mean Price			
	Color		
Height	Red	White	Both Colors
Short	$12	$18	$15
Tall	$20	$34	$27
Both Heights	$16	$26	$21

The table gives the mean price of a sample of 10 trees in each of the four overall categories (Tall and Red, Tall and White, Short and Red, Short and White). So, the ten Tall and Red pines averaged $20, the ten Short and White pines averaged $18, and so on. The margins show averages over larger groups. For instance, the 20 white pines, averaged $26, while the 20 short pines averaged $15.

The average price of all 40 trees in the sample was $21.

Based on the summary table, answer these questions:

1. In the model price $\sim$ color, which involves the coefficients "intercept" and "colorWhite", what will be the values of the coefficients?

   - Intercept 12 15 16 18 20 21 26 27 34
   - colorWhite -10 -8 0 5 8 10

2. In the model price $\sim$ height, which involves the coefficients "intercept" and "heightTall", what will be the values of the coefficients?

   - Intercept 0 4 8 12 15 16 18 20 21 26 27 34
   - heightTall 0 4 8 12 15 16 18 20 21 26 27 34

3. The model price $\sim$ height * color, with an interaction between height and color, has four coefficients and therefore can produce an exact match to the prices of the four different kinds of trees. But they are in a different format: not just one coefficient for each kind of tree. What are the values of these coefficients from the model? (Hint: Start with the kind of tree that corresponds to the intercept term.)

   - Intercept 0 4 6 8 10 12 16
   - heightTall 0 4 6 8 10 12 16
   - colorWhite 0 4 6 8 10 12 16
   - heightTall:colorWhite 0 4 6 8 10 12 16

4. The model price $\sim$ height + color gives these three coefficients:

   - Intercept : 10
   - heightTall : 12
   - colorWhite : 10

   It would be hard to figure out these coefficients by hand because they can't be read off from the summary table of Mean Price.

   According to the model, what are the fitted model values for these trees:

   - Short Red 10 12 15 16 20 22 32 34
   - Short White 10 12 15 16 20 22 32 34
   - Tall Red 10 12 15 16 20 22 32 34
   - Tall White 10 12 15 16 20 22 32 34

   Notice that the fitted model values aren't a perfect match to the numbers in the table. That's because a model with three coefficients can't exactly reproduce a set of four numbers.

Chapter Eight Reading Questions

- What is a residual? Why does it make sense to make them small when fitting a model?
- What is "least squares?"
- What does it mean to "partition variability" using a model?
- How can a model term be redundant? Why are redundant terms a problem?

Prob 8.01

Here are some (made-up) data from an experiment growing trees. The height was measured for trees in different locations that had been watered and fertilized in different ways.

height	water	light	compost	nitrogen
5	2	shady	none	little
4	1	bright	none	lot
5	1.5	bright	some	little
6	3	shady	rich	lot
7	3	bright	some	little
6	2	shady	rich	lot

(a) In the model expression height $\sim$ water, which is the explanatory variable?

A height
B water
C light
D compost
E Can't tell from this information.

(b) Ranger Alan proposes the specific model formula

$$\text{height} = 2 * \text{water} + 1.$$

Copy the table to a piece of paper and fill in the table showing the model values and the residuals.

height	water	model values	resids
5	2		
4	1		
5	1.5		
6	3		
7	3		
6	2		

(c) Ranger Bill proposes the specific model formula

$$\text{height} = \text{water} + 3.$$

Again, fill in the model values and residuals.

height	water	model values	resids
5	2		
4	1		
5	1.5		
6	3		
7	3		
6	2		

(d) Based on your answers to the previous to parts, which of the two models is better? Give a specific definition of "better" and explain your answer quantitatively.

(e) Write down the set of indicator variables that arise from the categorical variable compost.

(f) The fitted values are exactly the same for the two models water $\sim$ compost and water $\sim$ compost-1. This suggests that the **1** vector $(1, 1, 1, 1, 1, 1)$ is redundant with the set of indicator variables due to the variable compost. Explain why this redundancy occurs. Is it because of something special about the "compost" variable?

(g) Estimate, as best you can using only very simple calculations, the coefficients on the model water $\sim$ compost-1. (Note: there is no intercept term in this model.)

(h) Ranger Charley observes that the the following model is perfect because all of the residuals are zero.

$$\text{height} \sim 1+\text{water}+\text{light}+\text{compost}+\text{nitrogen}$$

Charley believes that using this model will enable him to make excellent predictions about the height of trees in the future. Ranger Donald, on the other hand, calls Charley's regression "ridiculous rot" and claims that Charley's explanatory terms could fit perfectly any set of 6 numbers. Donald says that the perfect fit of Charley's model does not give any evidence that the model is of any use whatsoever. Who do you think is right, Donald or Charley?

Prob 8.02

Which of these statements will compute the sum of square residuals of the model stored in the object `mod`?

- A `resid(mod)`
- B `sum(resid(mod))`
- C `sum(resid(mod))^2`
- D `sum(resid(mod)^2)`
- E `sum(resid(mod^2))`
- F None of the above.

Prob 8.03

Here is a simple model that relates foot width to length in children, fit to the data in `kidsfeet.csv`:

```
> kids = fetchData("kidsfeet.csv")
> mod = lm( width ~ length, data=kids)
> coef(mod)
(Intercept)    length
     2.8623    0.2479
```

(a) Using the coefficients, calculate the predicted foot width from this model for a child with foot length 27cm.

2.86 3.10 7.93 9.12 9.56 12.24 28.62

(b) The sum of squares of the residuals from the model provides a simple indication of how far typical values are from the model. In this sense, the standard deviation of the residuals tells us how much uncertainty there is in the prediction. (Later on, we'll see that another term needs to be added to this uncertainty.) What is the sum of squares of the residuals?

4.73 5.81 5.94 6.10 6.21

(c) What is the sum of squares of the fitted values for the kids in `kidsfeet.csv`?

42.5 286.3 3157.7 8492.0 15582.1

(d) What is the sum of squares of the foot widths for the kids in `kidsfeet.csv`.

3163.5 3167.2 3285.1 3314.8 3341.7

(e) There is a simple relationship between the sum of squares of the response variable, the residuals, and the fitted values. You can confirm this directly. Which of the following R statements is appropriate to do this:

- A
```
sum(kids$width)-(sum(resid(mod))+
sum(fitted(mod)))
```
- B
```
sum(kids$width^2)-(sum(resid(mod)
^2)+sum(fitted(mod)^2))
```
- C `sum(resid(mod))-sum(fitted(mod))`
- D
```
sum(resid(mod)^2)-sum(fitted(mod)
^2)
```

Note: It might seem natural to use the `==` operator to compare the equality of two values, for instance `A==B`. However, arithmetic on the computer is subject to small round-off errors, too small to be important when looking at the quantities themselves but sufficient to cause the `==` operator to say the quantities are different. So, it's usually better to compare numbers by subtracting one from the other and checking whether the result is very small.

Prob 8.04

Consider the data collected by Francis Galton in the 1880s, stored in a modern format in the `galton.csv` file. In this file, `heights` is the variable containing the child's heights, while the father's and mother's height is contained in the variables `father` and `mother`. The `family` variable is a numerical code identifying children in the same family; the number of kids in this family is in `nkids`.

```
> galton = fetchData("galton.csv")
> lm( height ~ father, data=galton)
```

```
Coefficients:
(Intercept)       father
    39.1104       0.3994
```

(a) What is the model's prediction for the height of a child whose father is 72 inches tall? 67.1 67.4 67.9 68.2

(b) Construct a model using both the father's and mother's heights, using just the main effect but not including their interaction. What is the model's prediction for the height of a child whose father is 72 inches tall and mother is 65 inches tall? 67.4 68.1 68.9 69.2

(c) Construct a model using mother and father's height, including the main effects as well as the interaction. What is the model's prediction for the height of a child whose father is 72 inches tall and mother is 65 inches tall? 67.4 68.1 68.9 69.2

Galton did not have our modern techniques for including multiple variables into a model. So, he tried an expedient, defining a single variable, "mid-parent," that reflected both the father's and mother's height. We can mimic this approach by defining the variable in the same way Galton did:

```
> midparent=(galton$father+1.08*galton$mother)/2
```

Galton used the multiplier of 1.08 to adjust for the fact that the mothers were, as a group, shorter than the fathers.

Fit a model to the Galton data using the mid-parent variable and child's sex, using both the main effects and the interaction. This will lead to a separate coefficient on mid-parent for male and female children.

(d) What is the predicted height for a girl whose father is 67 inches and mother 64 inches? 63.6 63.9 64.2 65.4 65.7

The following questions are about the size of the residuals from models.

(e) Without knowing anything about a randomly selected child except that he or she was in Galton's data set, we can say that the child's height is a random variable with a certain mean and standard deviation. What is this standard deviation? 2.51 2.73 2.95 3.44 3.58 3.67 3.72

(f) Now consider that we are promised to be told the sex of the child, but no other information. We are going to make a prediction of the child's height once we get this information, and we are asked to say, ahead of time, how good this prediction will be. A sensible way to do this is to give the standard deviation of the residuals from the best fitting model based on the child's sex. What is this standard deviation of residuals? 2.51 2.73 2.95 3.44 3.58 3.67 3.72

Prob 8.05

The "modern physics" course has a lab where students measure the speed of sound. The apparatus consists of an air-filled tube with a sound generator at one end and a microphone that can be set at any specified position within the tube. Using an oscilloscope, the transit time between the sound generator and microphone can be measured precisely. Knowing the position p and transit time t allows the speed of sound v to be calculated, based on the simple model:

$$\text{distance} = \text{velocity} \times \text{time} \qquad \text{or} \qquad p = vt.$$

Here are some data recorded by a student group calling themselves "CDT".

position (m)	transit time (millisec)
0.2	0.6839
0.4	1.252
0.6	1.852
0.8	2.458
1.0	3.097
1.2	3.619
1.4	4.181

Part 1.

Enter these data into a spreadsheet in the standard case-variable format. Then fit an appropriate model. Note that the relationship $p = vt$ between position, velocity, and time translates into a statistical model of the form p ~ t - 1 where the velocity will be the coefficient on the `t` term.

What are the units of the model coefficient corresponding to velocity, given the form of the data in the table above?

- A meters per second
- B miles per hour
- C millimeters per second
- D meters per millisecond
- E millimeters per millisecond
- F No units. It's a pure number.
- G No way to know from the information provided.

Compare the velocity you find from your model fit to the accepted velocity of sound (at room temperature, at sea level, in dry air): 343 m/s. There should be a reasonable match. If not, check whether your data were entered properly and whether you specified your model correctly.

Part 2.

The students who recorded the data wrote down the transit time to 4 digits of precision, but recorded the position to only 1 or 2 digits, although they might simply have left off the trailing zeros that would indicate a higher precision.

Use the data to find out how precise the position measurement is. To do this, make two assumptions that are very reasonable in this case:

1. The velocity model is highly accurate, that is, sound travels at a constant velocity through the tube.

2. The transit time measurements are correct. This assumption reflects current technology. Time measurements can be made very precisely, even with inexpensive equipment.

Given these assumptions, you should be able to calculate the position from the transit time and velocity. If the measured position differs from this model value — as reflected by the residuals — then the measured position is imprecise. So, a reasonable way to infer the precision of the position is by the typical size of residuals.

How big is a typical residual? One appropriate way to measure this is with the standard deviation of the residuals.

- Give a numerical value for this.
  0.001 0.006 0.010 0.017 0.084 0.128

Part 3.

The students' lab report doesn't indicate how they know for certain that the sound generator is at position zero. One way to figure this out is to measure the generator's position from the data themselves. Denoting the actual position of the sound generator as p_0, then the equation relating position and transit time is

$$p - p_0 = vt \qquad \text{or} \qquad p = p_0 + vt$$

This suggests fitting a model of the form $\texttt{p} \sim 1 + \texttt{t}$, where the coefficient on 1 will be p_0 and the coefficient on t will be v.

Fit this model to the data.

- What is the estimated value of p_0?
  -0.032 0.012 0.000 0.012 0.032

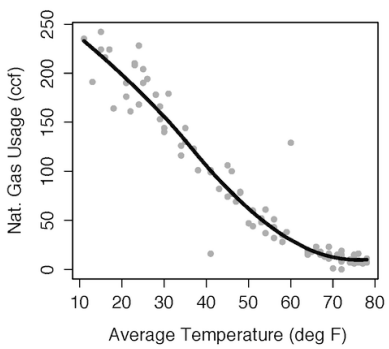
Notice that adding new terms to the model reduces the standard deviation of the residuals.

- What is the new value of the standard deviation of the residuals?
  0.001 0.006 0.010 0.017 0.084 0.128

Compare the estimated speed of sound found from the model p ~ t to the established value: 343 m/s . Notice that the estimate is better than the one from the model p ~ t - 1 that didn't take into account the position of the sound generator.

Prob 8.06

The graph shows some data on natural gas usage (in ccf) versus temperature (in deg. F) along with a model of the relationship.



(a) What are the units of the residuals from a model in which natural gas usage is the response variable?
ccf degF ccf.per.degF none

(b) Using the graph, estimate the magnitude of a typical residual, that is, approximately how far a typical case is from the model relationship. (Ignore whether the residual is positive or negative. Just consider how far the case is from the model, whether it be above or below the model curve.)
2ccf 20ccf 50ccf 100ccf

(c) There are two cases that are outliers with respect to the model relationship between the variables. Approximately how big are the residuals in these two cases?
2ccf 20ccf 50ccf 100ccf

Now ignore the model and focus just on those two outlier cases and their relationship to the other data points.

(d) Are the two cases outliers with respect to natural gas usage? True or False

(e) Are the two cases outliers with respect to temperature? True or False

Prob 8.07

It can be helpful to look closely at the residuals from a model. Here are some things you can easily do:

1. Look for outliers in the residuals. If they exist, it can be worthwhile to look into the cases involved more deeply. They might be anomalous or misleading in some way.

2. Plot the residuals versus the fitted model values. Ideally there should be no evident relationship between the two — the points should be a random scatter. When there is a strong relationship, even though it might be complicated, the model may be missing some important term.

3. Plot the residuals versus the values of an important explanatory variable. (If there are multiple explanatory variables, there would be multiple plots to look at.) Again, ideally there should be no evident relationship. If there is, there is something to think about.

Using the world-record swim data, `swim100m.csv` construct the model time ~ year + sex + year:sex. This model captures some of the variability in the record times, but doesn't reflect something that's obvious from a plot of the data: that records improved quickly in the early years (especially for women) but the improvment is much slower in recent years. The point of this exercise is to show how the residuals provide information about this.

- Find the cases in the residuals that are outliers. Explain what it is about these cases that fits in with the failure of the model to reflect the slowing improvement in world records.

- Plot the residuals versus the fitted model values. What pattern do you see that isn't consistent with the idea that the residuals are unrelated to the fitted values?

- Plot the residuals versus year. Describe the pattern you see.

Now use the kids-feet data `kidsfeet.csv` and the model width ~ length + sex + length:sex.

Look at the residuals in the three suggested ways. Are there any outliers? Describe any patterns you see in relationship to the fitted model values and the explanatory variable length.

Chapter Nine Reading Questions

- How does R^2 summarize the extent to which a model has captured variability?

- What does it mean for one model to be nested in another?

- How does the correlation coefficient differ from R^2?

Prob 9.01

The R^2 statistic is the ratio of the variance of the fitted values to the variance of the response variable.

Using the `kidsfeet.csv` data:

1. Find the variance of the response variable in the model width ~ sex + length + sex:length .
   0.053 0.119 0.183 0.260 0.346

2. Find the variance of the fitted values from the model
   0.053 0.119 0.183 0.260 0.346

3. Compute the R^2 as the ratio of these two variances.
   0.20 0.29 0.46 0.53 0.75

4. Is this the same as the "Multiple R^2" given in the `summary(mod)` report? Yes No

Prob 9.02

The variance of a response variable A is 145 and the variance of the residuals from the model A ~ 1+B is 45.

- What is the variance of the fitted model values?
  45 100 145 190 Cannot tell

- What is the R^2 for this model?
  0 45/145 100/145 100/190 145/190 Cannot tell

Prob 9.03

For each of the following pairs of models, mark the statement that is most correct.

Part 1

Model 1 . A ~ B+C

Model 2 . A ~ B*C

A Model 1 is nested in Model 2.
B Model 2 is nested in Model 1.
C The two models are the same.
D None of the above is true.

Part 2

Model 1 . A ~ B

Model 2 . B ~ A

A Model 1 is nested in Model 2.
B Model 2 is nested in Model 1.
C The two models are the same.
D None of the above is true.

Part 3

Model 1 . A ~ B + C

Model 2 . B ~ A * C

A Model 1 is nested in Model 2.
B Model 2 is nested in Model 1.
C The two models are the same.
D None of the above is true.

Part 4

Model 1 . A ~ B + C + B:C

Model 2 . A ~ B * C

A Model 1 is nested in Model 2.
B Model 2 is nested in Model 1.
C The two models are the same.
D None of the above is true.

Prob 9.04

For each of the following pairs of models, mark the statement that is most correct.

Part 1

Model 1 . A ~ B+C

Model 2 . A ~ B*C

A Model 1 can have a higher R^2 than Model 2
B Model 2 can have a higher R^2 than Model 1
C The R^2 values will be the same.
D None of the above are necessarily true.

Part 2

Model 1 . A ∼ B + C

Model 2 . B ∼ A * C

A Model 1 can have a higher R^2 than Model 2

B Model 2 can have a higher R^2 than Model 1

C The R^2 values will be the same.

D None of the above are necessarily true.

Part 3

Model 1 . A ∼ B + C + B:C

Model 2 . A ∼ B * C

A Model 1 can have a higher R^2 than Model 2

B Model 2 can have a higher R^2 than Model 1

C The R^2 values will be the same.

D None of the above are necessarily true.

[From a suggestion by a student]

Going further

In answering this question, you might want to try out a few examples using real data: just pick two quantitative variables to stand in for A and B.

What will be the relationship between R^2 for the following two models?

Model 1 . A ∼ B

Model 2 . B ∼ A

A Model 1 can have a higher R^2 than Model 2

B Model 2 can have a higher R^2 than Model 1

C The R^2 values will be the same.

D None of the above are necessarily true.

Prob 9.06

Which of the following statements is true about R^2?

1. True or False R^2 will never go down when you add an additional explanatory term to a model.
2. For a perfectly fitting model,
   A R^2 is exactly zero.
   B R^2 is exactly one.
   C Neither of the above.
3. In terms of the variances of the fitted model points, the residual, and the response variable, R^2 is the:
   A Variance of the residuals divided by the variance of the fitted.
   B Variance of the response divided by the variance of the residuals.
   C Variance of the fitted divided by the variance of the residuals.
   D Variance of the fitted divided by the variance of the response.
   E Variance of the response divided by the variance of the fitted.

Prob 9.07

Consider models with a form like this

```
> lm( response ~ 1, data=whatever)
```

The R^2 of such a model will always be 0. Explain why.

Prob 9.08

Consider the following models where a response variable A is modeled by explanatory variables B, C, and D.

1 A ∼ B
2 A ∼ B + C + B:C
3 A ∼ B + C
4 A ∼ B * C
5 A ∼ B + D
6 A ∼ B*C*D

Answer the following:

(a) Model 1 is nested in model 2. True or False

(b) Model 5 is netsted in model 3. True or False

(c) Model 1 is nested in model 3. True or False

(d) Model 5 is nested in model 1. True or False

(e) Model 2 is nested in model 3. True or False

(f) Model 3 is nested in model 4.True or False

(g) All the other models are nested in model 6. True or False

Prob 9.09

Consider two models, Model 1 and Model 2, with the same response variable.

1. Model 1 is nested in Model 2 if the variables model terms of Model 1 are a subset of those of Model 2.
2. True or False If Model 1 is nested in Model 2, then model 1 cannot have a higher R^2 than model 2.
3. Which of the following are nested in A ∼ B*C + D?

   True or False A ∼ B
   True or False A ∼ B + D
   True or False B ∼ C
   True or False A ∼ B+C+D
   True or False A ∼ B*D + C
   True or False A ∼ D

Prob 9.10

Here is a set of models:

Model A: wage $\sim$ 1
Model B: wage $\sim$ age + sex
Model C: wage $\sim$ 1 + age*sex
Model D: wage $\sim$ educ
Model E: wage $\sim$ educ + age - 1
Model F: wage $\sim$ educ:age
Model G: wage $\sim$ educ*age*sex

You may want to try fitting each of the models to the Current Population Survey data `cps.csv` to make sure you understand how the * shorthand for interaction and main effects expands to a complete set of terms. That way you can see exactly which coefficients are calculated for any of the models.

Answer the following:

1. B is nested in A. TRUE or FALSE
2. D is nested in E. TRUE or FALSE
3. B is nested in C. TRUE or FALSE
4. All of the models A-F are nested in G. TRUE or FALSE
5. D is nested in F. TRUE or FALSE
6. At least one of the models A-G is nested in educ $\sim$ age. TRUE or FALSE

Prob 9.11

A data set on US Congressional Districts (provided by Prof. Julie Dolan), `congress.csv` contains information on the population of each congressional district in 2004. There are 436 districts listed (corresponding to the 435 voting members of the House of Representatives from the 50 states and an additional district for Washington, D.C., whose citizens have only a non-voting "representative."

The US Supreme Court (Reynolds v. Sims, 377 US 533, 1964) ruled that state legislature districts had to be roughly equal in population: the one-person one-vote principle. Before this ruling, some states had grossly unequally sized districts. For example, one district in Connecticut for the state General Assembly had 191 people, while another district in the same state had 81,000. Los Angeles County had one representative in the California State Senate for a population of six million, while another county with only 14,000 residents also had one representative.

Of course, exact equality of district sizes is impossible in every district, since districts have geographically defined boundaries and the population can fluctuate within each boundary. The Supreme Court has written, "... mathematical nicety is not a constitutional requisite..." and "so long as the divergences from a strict population standard are based on legitimate considerations incident to the effectuation of a rational state policy, some deviations from the equal-population principle are constitutionally permissible ...." (Reynolds v. Simms)

The situation in the US House of Representatives is more complicated, since congressional districts are required to be entirely within a single state.

Let's explore how close the districts for the US House of Representatives comes to meeting the one-person one-vote principle.

One way to evaluate how far districts are from equality of population size is to examine the standard deviation across districts.

- What is the standard deviation of the district populations across the whole US?
  4823 9468 28790 342183 540649

Another way to look at the spread is to try to account for the differences in populations by modeling them and looking at how much of the difference remains unexplained.

Let's start with a very simple model that treats all the districts as the same: population $\sim$ 1.

What is the meaning of the single coefficient from this model?

A The mean district population across all states.
B The mean district population across all districts.
C The median population across all districts.
D The median population across all states.
E None of the above.

Calculate the standard deviation of the residuals. How does this compare to the standard deviation of the district population itself?

A It's much larger.
B It's somewhat larger.
C It's exactly the same.
D It's much smaller.

Now model the district size by the state population $\sim$ 1 + state.

What is the standard deviation of the residuals from this model?

#

A box plot of the residuals shows a peculiar pattern. What is it?

A The residuals are all the same.
B Every residual is an outlier.
C The residuals are almost all very close to zero, except for a few outliers.

The variable state accounts for almost all of the variability from district to district. That is, districts within a state are almost exactly the same size, but that size differs from state to state. Why is there a state-to-state difference? The number of districts within a state must be a whole number: 1, 2, 3, and so on. Ideally, the district populations are the state population divided by the number of districts. The number of districts is set to make the district population as even as possible between states, but exact equality isn't possible since the state populations differ. Notice that the largest and smallest districts (Montana and Wyoming, respectively) are in states with only a single district. Adding a second district to Montana would dramatically reduce the district size below the national mean. And even though Wyoming has a very

low-population district, it's impossible to take a district away since Wyoming only has one.

Prob 9.12

Consider this rule of thumb:

> In comparing two models based on the same data, the model with the larger R^2 is better than the model with the smaller R^2.

Explain what makes sense about this rule of thumb and also what issues it might be neglecting.

Prob 9.14

We're going to use the ten-mile-race data to explore the idea of redundancy: Why redundancy is a problem and what we can do about it.

Read in the data:

```
> run = fetchData("ten-mile-race.csv")
```

The data includes information about the runner's age and sex, as well as the time it took to run the race.

I'm interested in how computer and cell-phone use as a child may have affected the runner's ability. I don't have any information about computer use, but as a rough proxy, I'm going to use the runner's year of birth. The assumption is that runners who were born in the 1950s, 60s, and 70s, didn't have much chance to use computers as children.

Add in a new variable: yob. We'll approximate this as the runner's age subtracted from the year in which the race was run: 2005. That might be off by a year for any given person, but it will be pretty good.

```
> run$yob = 2005 - run$age
```

Each of the following models has two terms.

```
mod1 = lm( net ~ age + yob - 1, data=run)
mod2 = lm( net ~ 1 + age, data = run)
mod3 = lm( net ~ 1 + yob, data=run )
```

- Fit each of the the models and interpret the coefficients in terms of the relationship between age and year of birth and running time. Then look at the R^2 and the sum of square residuals in order to decide which is the better model.

Using special software that you don't have, I have fitted a model — I'll call it mod4 — with all three terms: the intercept, age, and year of birth. The model coefficients are:

My Fantastic Model: mod4

Intercept	age	yob
-20050	20.831891052	12.642004612

My conclusion, based on the mod4 coefficients, is that people slow down by 20.8 seconds for every year they age. Making up for this, however, is the fact that people who were born earlier in the last century tend to run slower by 12.6 seconds for every year later they were born. Presumably this is because those born earlier had less opportunity to use computers and cell phones and therefore went out and did healthful, energetic, physical play rather than typing.

- Using these coefficients, calculate the model values. The statement will look like this:

```
mod4vals = -20050 + 20.831891052*run$age +
    12.642004612*run$yob
```

- Calculate the residuals from mod4 by substracting the model values from the response variable (net running time). Compare the size of the residuals using a sum of squares or a standard deviation or a variance to the size of the residuals from models 1 through 3. Judging from this, which is the better model?

I needed special software to find the coefficients in mod4 because R won't do it. See what happens when you try the models with three terms, like this:

```
lm( net ~ 1 + age + yob, data=run )
lm( net ~ 1 + yob + age, data=run )
```

- Can you get three coefficients from the R software?

I'm very pleased with mod4 and the special methods I used to find the coefficients.

Unfortunately, my statistical arch-enemy, Prof. Nalpak Ynnad, has proposed another model. He claims that computer and cell-phone use is helpful. According to his twisted theory, people actually run faster as they get older. Impossible! But look at his model coefficients.

Ynnad's Evil Model: mod5

Intercept	age	yob
60150	-19.16810895	-27.35799539

Ynnad's ridiculous explanation is that the natural process of aging (that you run faster as you age), is masked by the beneficial effects of exposure to computers and cell phones as a child. That is, today's kids would be even slower (because they are young) except for the fact that they use computers and cell phones so much. Presumably, when they grow up, they will be super fast, benefiting both from their advanced age and from the head start they got as children from their exposure to computers and cell phones.

- Looking at Ynnad's model in terms of the R^2 or size of residuals, how does it compare to my model? Which one should you believe?

- Give an explanation of why both my model and Ynnad's model are bogus. See if you can also explain why we shouldn't take the coefficients in mod1 seriously at face value.

Chapter Ten Reading Questions

- What is a covariate? Why use a special word for it when it is just a variable?

- What is the difference between a *partial* change and a *total* change?

- In the experimental method, how are covariates dealt with?

- What is the modeling approach to dealing with covariates?
- What is Simpson's paradox?

Prob 10.01

Consider the data set on kids' feet in `kidsfeet.csv`.

First, you're going to look at a total change. Build a model of foot width as a function of foot length: width ~ length. Fit this model to the kids' feet data.

According to this model, how much does the typical width change when the foot length is increased from 22 to 27 cm?
0.187 0.362 0.744 0.953 1.060 1.105 1.240 1.487

This is a total change, because it doesn't hold any other variable constant, e.g. sex. That might sound silly, since obviously a kid's sex doesn't change as his or her foot grows. But the model doesn't know that. It happens that most of the kids with foot lengths near 22 cm are girls, and most of the kids with foot lengths near 27 cm are boys. So when you compare feet with lengths of 22 and 27, you are effectively changing the sex at the same time as you change the foot length.

To look at a partial change, holding sex constant, you need to include sex in the model. A simple way to do this is width ~ length + sex. Using this model fitted to the kids' feet data, how much does a typical foot width change if the foot length is increased from 22 to 27 cm?
0.187 0.362 0.744 0.953 1.060 1.105 1.142 1.240 1.487

You can also build more detailed models, for example a model that includes an interaction term: width ~ length * sex. Using this model fitted to the kids' feet data, how much will a typical **girl's** foot width change if the foot length is increased from 22 to 27 cm?
0.187 0.362 0.744 0.953 1.060 1.105 1.142 1.240 1.487

Prob 10.02

In each of the following, a situation is described and a question is asked that is to be answered by modeling. Several variables are listed. Imagine an appropriate model and identify each variable as either the response variable, an explanatory variable, a covariate, or a variable to be ignored.

> EXAMPLE: *Some people have claimed that police foot patrols are more effective at reducing the crime rate than patrols done in automobiles. Data from several different cities is available; each city has its own fraction of patrols done by foot, its own crime rate, etc. The mayor of your town has asked for your advice on whether it would be worthwhile to shift to more foot patrols in order to reduce crime. She asks, "Is there evidence that a larger fraction of foot patrols reduces the crime rate?"*
>
> Variables:
>
> 1. Crime rate (e.g., robberies per 100000 population) variable.
> 2. Fraction of foot patrols
> 3. Number of policemen per 1000 population
> 4. Demographics (e.g., poverty rate)
>
> The question focuses on how the fraction of foot patrols might influence crime rate, so crime rate is the response variable and fraction of foot patrols is an explanatory variable.
>
> But, the crime rate might also depend on the overall level of policing (as indicated by the number of policemen), or on the social conditions that are associated with crime (e.g., demographics). Since the mayor has no power to change the demographics of your town, and probably little power to change the overall level number of policemen, in modeling the data from the different cities, you would want to hold constant number of policemen and the demographics. You can do this by treating number of policement and demographics as covariates and including them in your model.

Alcohol and Road Safety

Fifteen years ago, your state legislature raised the legal drinking age from 18 to 21 years. An important motivation was to reduce the number of car accident deaths due to drunk or impaired drivers. Now, some people are arguing that the 21-year age limit encourages binge drinking among 18 to 20 year olds and that such binge drinking actually increases car accident deaths. But the evidence is that the number of car accident deaths has gone down since the 21-year age restriction was introduced. You are asked to examine the issue: Does the reduction in the number of car-accident deaths per year point to the effectiveness of the 21-year drinking age?

Variables:

1. Drinking age limit. Levels: 18 or 21.
   response explanatory covariate ignore
2. Number of car-accident deaths per year.
   response explanatory covariate ignore
3. Prevalence of seat-belt use.
   response explanatory covariate ignore
4. Fraction of cars with air bags.
   response explanatory covariate ignore
5. Number of car accidents (with or without death).
   response explanatory covariate ignore

Rating Surgeons

Your state government wants to guide citizens in choosing physicians. As part of this effort, they are going to rank all the surgeons in your state. You have been asked to build the rating system and you have a set of variables available for your use. These variables have been measured for each of the 342,861 people who underwent surgery in your state last year: one person being treated by one doctor. How should you construct a rating system that will help citizens to choose the most effective surgeon for their own treatment?

Variables:

- Outcome score. A high score means that the operation did what it was supposed to. A low score reflects failure, e.g. death. Death is a very bad outcome, post-operative infection a somewhat bad outcome.)

  response explanatory covariate ignore

- Surgeon. One level for each of the operating surgeons.

  response explanatory covariate ignore

- Experience of the surgeon.

  response explanatory covariate ignore

- Difficulty of the case.

  response explanatory covariate ignore

School testing

Last year, your school district hired a new superintendent to "shake things up." He did so, introducing several controversial new policies. At the end of the year, test scores were higher than last year. A representative of the teachers' union has asked you to examine the score data and answer this question: Is there reason to think that the higher scores were the result of the superintendent's new policies?

Variables:

1. Superintendent (levels: New or Former superintendent)

   response explanatory covariate ignore

2. Exam difficulty

   response explanatory covariate ignore

3. Test scores

   response explanatory covariate ignore

Gravity

In a bizarre twist of time, you find yourself as Galileo's research assistant in Pisa in 1605. Galileo is studying gravity: Does gravity accelerate all materials in the same way, whether they be made of metal, wood, stone, etc.? Galileo hired you as his assistant because you have brought with you, from the 21st century, a stop-watch with which to measure time intervals, a computer, and your skill in statistical modeling. All of these seem miraculous to him.

He drops objects off the top of the Leaning Tower of Pisa and you measure the following:

Variables

1. The size of the object (measured by its diameter).

   response explanatory covariate ignore

2. Time of fall of the object.

   response explanatory covariate ignore

3. The material from which the object is made (brass, lead, wood, stone).

   response explanatory covariate ignore

[Thanks to James Heyman.]

Prob 10.03

Economists measure the inflation rate as a percent change in price per year. Unemployment is measured as the fraction (percentage) of those who want to work who are seeking jobs.

According to economists, in the short run — say, from one year to another — there is a relationship between inflation and unemployment: all other things being equal, as unemployment goes up, inflation should go down. (The relationship is called the "Phillips curve," but you don't need to know that or anything technical about economics to do this question.)

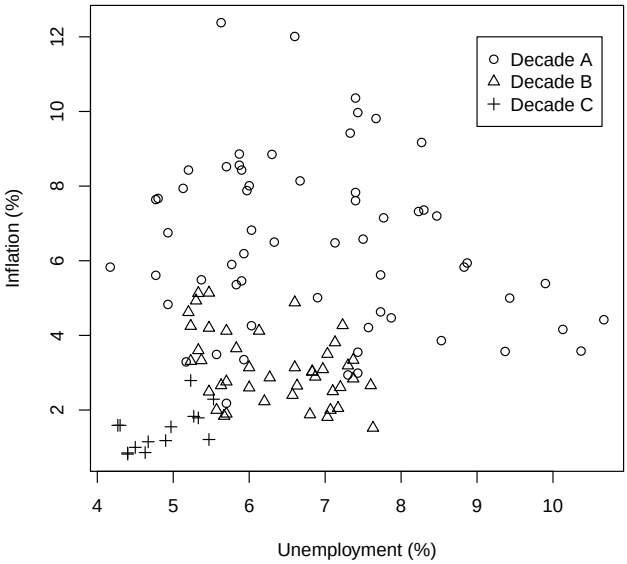
If this is true, in the model Inflation $\sim$ Unemployment, what should be the sign of the coefficient on Unemployment?
positive zero negative

But despite the short term relationship, economists claim that In the long run — over decades — unemployment and inflation should be unrelated.

If this is true, in the model Inflation $\sim$ Unemployment, what should be the sign of the coefficient on Unemployment?
positive zero negative

The point of this exercise is to figure out how to arrange a model so that you can study the short-term behavior of the relationship, or so that you can study the long term relationship.

For your reference, here is a graph showing a scatter plot of inflation and unemployment rates over about 30 years in the US. Each point shows the inflation and unemployment rates during one quarter of a year. The plotting symbol indicates which of three decade-long periods the point falls into.



The relationship between inflation and unemployment seems to be different from one decade to another — that's the short term.

Which decade seems to violate the economists' Phillips Curve short-term relationship?
A B C none all

Using the modeling language, express these different possible relationships between the variables Inflation,

Unemployment, and Decade, where the variable Decade is a categorical variable with the three different levels shown in the legend for the graph.

1. Inflation depends on Unemployment in a way that doesn't change over time.
   - [A] Inflation ~ Decade
   - [B] Inflation ~ Unemployment
   - [C] Inflation ~ Unemployment+Decade
   - [D] Inflation ~ Unemployment*Decade

2. Inflation changes with the decade, but doesn't depend on Unemployment.
   - [A] Inflation ~ Decade
   - [B] Inflation ~ Unemployment
   - [C] Inflation ~ Unemployment+Decade
   - [D] Inflation ~ Unemployment*Decade

3. Inflation depends on Unemployment in the same way every decade, but each decade introduces a new background inflation rate independent of Unemployment.
   - [A] Inflation ~ Decade
   - [B] Inflation ~ Unemployment
   - [C] Inflation ~ Unemployment+Decade
   - [D] Inflation ~ Unemployment*Decade

4. Inflation depends on Unemployment in a way that differs from decade to decade.
   - [A] Inflation ~ Decade
   - [B] Inflation ~ Unemployment
   - [C] Inflation ~ Unemployment+Decade
   - [D] Inflation ~ Unemployment*Decade

Whether a model examines the short-term or the long-term behavior is analogous to whether a partial change or a total change is being considered.

Suppose you wanted to study the long-term relationship between inflation and unemployment. Which of these is appropriate?

- [A] Hold Decade constant. (Partial change)
- [B] Let Decade vary as it will. (Total change)

Now suppose you want to study the short-term relationship. Which of these is appropriate?

- [A] Hold Decade constant. (Partial change)
- [B] Let Decade vary as it will. (Total change)

Prob 10.04

Consider two models that you are to fit to a single data set involving three variables: A, B, and C.

Model 1 A ~ B

Model 2 A ~ B + C

(a) When should you say that Simpson's Paradox is occuring?

- [A] When Model 2 has a lower R^2 than Model 1.
- [B] When Model 1 has a lower R^2 than Model 2.
- [C] When the coef. on B in Model 2 has the opposite sign to the coef. on B in Model 1.
- [D] When the coef. on C in Model 2 has the opposite sign to the coef. on B in Model 1.

(b) True or False: If B is uncorrelated with A, then the coefficient on B in the model A ~ B must be zero.

TRUE or FALSE

(c) True or False: If B is uncorrelated with A, then the coefficient on B in a model A ~ B+C must be zero.

TRUE or FALSE

(d) True or False: Simpson's Paradox can occur if B is uncorrelated with C.

TRUE or FALSE

Based on a suggestion by student Atang Gilika.

Prob 10.05

Standard & Poor's is a RATING AGENCY that provides information about various financial instruments such as stocks and bonds. The S&P 500 Stock Index, for instance, provides a summary of the value of stocks.

Bonds issued by governments, corporations, and other entities are rated using letters. As described on the Standard & Poor's website, the ratings levels are AAA, AA+, AA, AA-, A+, A, A-, BBB+, BBB, BBB-, BB+, BB, BB-, B+, B, B-, CCC+, CCC, CCC-, CC, C, and D. The AAA rating is the best. ("The obligor's capacity to meet its financial commitment on the obligation is extremely strong.") D is the worst. ("The 'D' rating category is used when payments on an obligation are not made on the date due ....)

- The bond ratings are a categorical variable. TRUE or FALSE
- The bond ratings are an ordinal variable. TRUE or FALSE

Bonds are a kind of debt; they pay interest and the principal is paid back at the end of a maturity period. The people and institutions who invest in bonds are willing to accept somewhat lower interest payments in exchange for greater security. Thus, AAA-rated bonds tend to pay the lowest interest rates and worse-rated bonds pay more. A report on interest rates on bonds (www.fmsbonds.com, for 8/21/2008) listed interest rates on municipal bonds:

Issue	Maturity	Rate
	AAA Rated	
National	10 Year	3.75
National	20 Year	4.60
National	30 Year	4.75
Florida	30 Year	4.70
	AA Rated	
National	10 Year	3.90
National	20 Year	4.70
National	30 Year	4.85
Florida	30 Year	4.80
	A Rated	
National	10 Year	4.20
National	20 Year	5.05
National	30 Year	5.20
Florida	30 Year	5.15

How many explanatory variables are given in this table to account for the interest rate:

A Two: Issue and Maturity
B Three: Issue, Maturity, and S & P Rating
C Four: Issue, Maturity, S & P Rating, and Interest Rate

Judging from the table, and holding all other explanatory variables constant, what is the change in interest rate associated with a change from AAA to AA rating?

0.05 0.15 0.25 0.30 0.40

Again, holding all other explanatory variables constant, what is the change in interest rate for a 10-year compared to a 20-year maturity bond? (Pick the closest answer.)

0.15 0.50 0.85 1.20 1.45

Sometimes it is unclear when a variable should be considered quantitative and when it should be taken as categorical. For example, the maturity variable looks on the surface to be quantitative (10-year, 20-year, 30-year, etc.). What is it about these data that suggests that it would be unrealistic to treat maturity as a quantitative variable in a model of interest rate?

Prob 10.06

A study on drug D indicates that patients who were given the drug were **less** likely to recover from their condition C. Here is a table showing the overall results:

Drug	# recovered	# died	Recovery Rate
Given	1600	2400	40%
Not given	2000	2000	50%

Strangely, when investigators looked at the situation separately for males and females, they found that the drug **improves** recovery for each group:

Females

Drug	num recovered	# died	Recovery Rate
Given	900	2100	30%
Not given	200	800	20%

Males

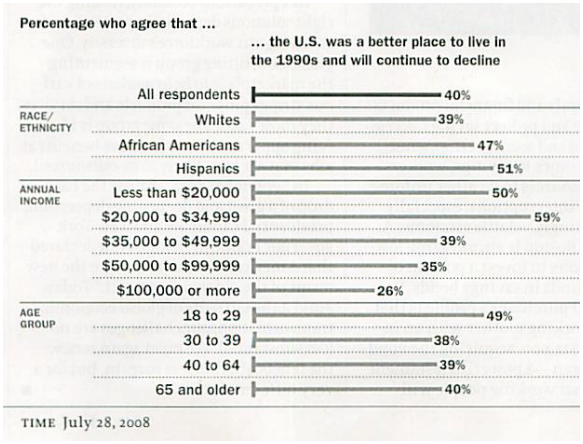
Drug	# recovered	# died	Recovery Rate
Given	700	300	70%
Not given	1800	1200	60%

Which is right? Does the drug improve recovery or hinder recovery? What advice would you give to a physician about whether or not to prescribe the drug to her patients? Give enough of an explanation that the physician can judge whether your advice is reasonable.

Based on an example from Judea Pearl (2000) *Causality: Models, Reasoning, and Inference*, Cambridge Univ. Press, p. 175

Prob 10.08

Time Magazine reported the results of a poll of people's opinions about the U.S. economy in July 2008. The results are summarized in the graph.



[Source: *Time*, July 28, 2008, p. 41]

In a typical news media report of a poll, the results are summarized using one explanatory variable at a time. The point of this exercise is to show that such *univariate* explanations can be misleading.

The poll involves three explanatory variables: ethnicity, income, and age. Regretably, the reported results treat each of these explanatory variables separately, even though there are likely to be correlations among them. For instance, relatively few people in the 18 to 29 age group have high incomes.

The original data set from which the *Time* graphic was made contains the information needed to study the multiple explanatory variables simultaneously, for example looking at the connection between pessimism and age while adjusting for income. This data set is not available, so you will need to resort to a simulation which attempts to mimic the poll results. Of course, the simulation doesn't necessarily describe people's attitudes directly, but it does let you see how the conclusions drawn from the poll might have been different if the results for each explanatory variable had been presented in a way that adjusts for the other explanatory variables.

The following statement will run a simulation of a poll in which 10,000 people are asked to rate their level of pessimism (on a scale from 0 to 10) and to indicate their age group and income level:

```
> poll = run.sim(economic.outlook.poll, 10000)
```

The output of the simulation will be a data frame that looks something like this:

```
> head(poll)
              age                 income pessimism
```

```
1     [18 to 29]   [less than $20000]        10
2     [40 to 64] [$50,000 to $99,999]         5
3     [40 to 64]   [less than $20000]         9
4     [40 to 64] [$50,000 to $99,999]         7
5 [65 and older] [$50,000 to $99,999]         7
6     [18 to 29]   [less than $20000]        10
```

Your output will differ because the simulation reflects random sampling.

- Construct the model pessimism $\sim$ age-1. Look at the coefficients and choose the statement that best reflects the results:

  A Middle aged people have lower pessimism than young or old people.
  B Young people have the least pessimism.
  C There is no relationship between age and pessimism.

- Now construct the model pessimism $\sim$ income-1. Look at the coefficients and choose the statement that best reflects the results:

  A Higher income people are more pessimistic than low-income people.
  B Higher income people are less pessimistic than low-income people.
  C There is no relationship between income and pessimism.

- Construct a model in which you can look at the relationship between pessimism and age while *adjusting for* income. That is, include income as a covariate in your model. Enter your model formula here: .

  Look at the coefficients from your model and choose the statement that best reflects the results:

  A Holding income constant, older people tend to have higher levels of pessimism than young people.
  B Holding income constant, young people tend to have higher levels of pessimism than old people.
  C Holding income constant, there is no relationship between age and pessimism.

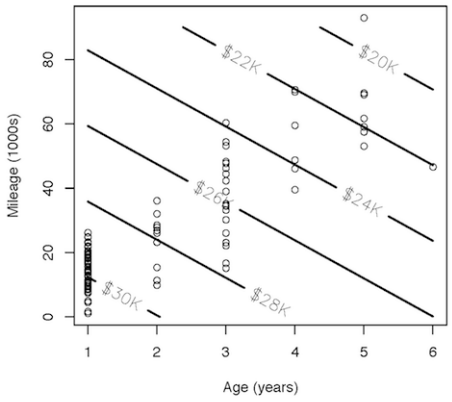
- You can also interpret that same model to see the relationship between pessimism and income while *adjusting for* age. Which of the following statements best reflects the results? (Hint: make sure to pay attention to the sign of the coefficients.)

  A Holding age constant, higher income people are more pessimistic than low-income people.
  B Holding age constant, higher income people are less pessimistic than low-income people.
  C Holding age constant, there is no relationship between income and pessimism.

Prob 10.10

Whenever you seek to study a partial relationship, there must be at least three variables involves: a response variable, an explanatory variable that is of direct interest, and one or more other explanatory variables that will be held constant: the co-variates. Unfortunately, it's hard to graph out models involving three variables on paper: the usual graph of a model just shows one variable as a function of a second.

One way to display the relationship between a response variable and two quantitative explanatory variables is to use a contour plot. The two explanatory variables are plotted on the axes and the fitted model values are shown by the contours. The figure shows such a display of the fitted model of used car prices as a function of mileage and age.



The dots are the mileage and age of the individual cars — the model Price is indicated by the contours.

The total relationship between Price and mileage involves how the price changes for typical cars of different mileage. Pick a dot that is a typical car with about 10,000 miles. Using the contours, find the model price of this car.

Which of the following is closest to the model price (in dollars)?

18000 21000 25000 30000

Now pick another dot that is a typical car with about 70,000 miles. Using the contours, find the model price of this car.

18000 21000 25000 30000

The total relationship between Price and mileage is reflected by this ratio: change in model price divided by change in mileage. What is that ratio (roughly)?

A $\frac{30000-21000}{70000-10000} = 0.15$ dollars/mile
B $\frac{70000-10000}{25000-21000} = 15.0$ dollars/mile
C $\frac{25000-18000}{70000-10000} = 0.12$ dollars/mile

In contrast, the partial relationship between Price and mileage holding age constant is found in a different way, by comparing two points with different mileage but exactly the same age.

Mark a point on the graph where age is 3 years and mileage is 10000. Keep in mind that this point doesn't need to be an actual car, that is, a data point in the graph typical car. There might be no actual car with an age of 3 years and mileage

10000. But using the contour model, find the model price at this point:

22000 24000 26000 28000 30000

Now find another point, one where the age is exactly the same (3 years) but the mileage is different. Again there might not be an actual car there. Let's pick mileage as 80000. Using the contours, find the model price at this point:

22000 24000 26000 28000 30000

The partial relationship between price and mileage (holding age constant) is reflected again reflected by the ratio of the change in model price divided by the change in mileage. What is that ratio (roughly)?

A $\frac{80000-10000}{25000-21000} = 17.50$ dollars/mile
B $\frac{28000-22000}{80000-10000} = 0.09$ dollars/mile
C $\frac{26000-24000}{80000-10000} = 0.03$ dollars/mile

Both the total relationship and the partial relationship are indicated by the slope of the model price function given by the contours. The total relationship involves the slope between two points that are typical cars, as indicated by the dots. The partial relationship involves a slope along a different direction. When holding age constant, that direction is the one where mileage changes but age does not (vertical in the graph).

There's also a partial relationship between price and age holding mileage constant. That partial relationship involves the slope along the direction where age changes but mileage is held constant. Estimate that slope by finding the model price at a point where age is 2 years and another point where age is 5 years. You can pick whatever mileage you like, but it's key that your two points be at exactly the same mileage.

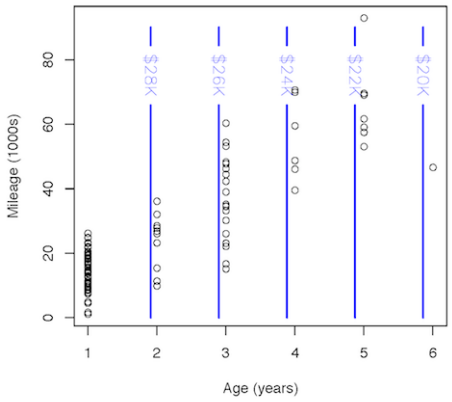
Estimate the slope of the price function along a direction where age changes but mileage is held constant (horizontally on the graph).

A 100 dollars per year
B 500 dollars per year
C 1000 dollars per year
D 2000 dollars per year

The contour plot above shows a model in which both mileage and age are explanatory variables. By choosing the direction in which to measure the slope, one determines whether the slope reflects a total relationship (a direction between typical cars), or a partial relationship holding age constant (a direction where age does not change, which might not be typical for cars), or a partial relationship holding mileage constant (a direction where mileage does not change, which also might not be typical for cars).

In calculus, the partial derivative of price with respect to mileage refers to an infinitesimal change in a direction where age is held constant. Similarly, the partial derivative of price with respect to age refers to an infinitesimal change in a direction where mileage is held constant.

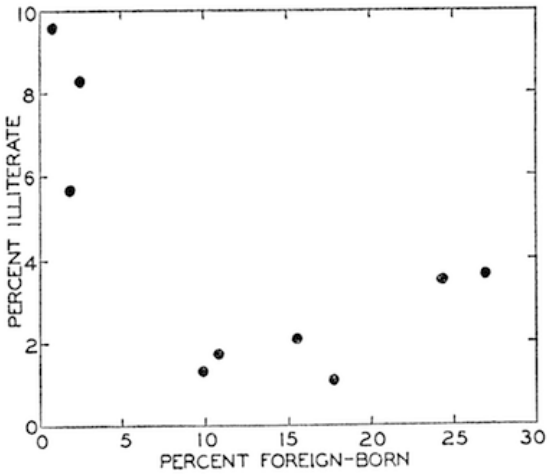
Of course, in order for the directional derivatives to make sense, the price function needs to have **both** age and mileage as explanatory variables. The following contour plot shows a model in which only age has been used as an explanatory variable: there is no dependence of the function on mileage.



Such a model is incapable of distinguishing between a partial relationship and a total relationship. Both the partial and the total relationship involve a ratio of the change in price and change in age between two points. For the total relationship, those two points would be typical cars of different ages. For the partial relationship, those two points would be different ages at exactly the same mileage. But, because the model depend on mileage, the two ratios will be exactly the same.

Prob 10.11

Sometimes people use data in aggregated form to draw conclusions about individuals. For example, in 1950 W.S. Robinson described the correlation between immigration and illiteracy done in two different ways.[**?**] In the first, the unit of analysis is individual US states as shown in the figure — the plot shows the fraction of people in each state who are illiterate versus the fraction of people who are foreign born. The correlation is negative, meaning that states with higher foreign-born populations have *less* illiteracy.



Robinson's second analysis involves the same data, but takes the unit of analysis as an individual person. The table gives the number of people who are illiterate and who are foreign born in the states included in the scatter plot.

TABLE 3. THE INDIVIDUAL CORRELATION BETWEEN NATIVITY AND ILLITERACY FOR THE UNITED STATES, 1930
(for the population 10 years old and over)

	Foreign Born	Native Born	Total
Illiterate	1,304	2,614	3,918
Literate	11,913	81,441	93,354
Total	13,217	84,055	97,272

The data in the table leads to a different conclusion than the analysis of states: the foreign born people are *more* likely to be illiterate.

This conflict between the results of the analyses, analogous to Simpson's paradox, is called the **ecological fallacy**. (The word "ecological" is rooted in the Greek word *oikos* for house — think of the choice between studying individuals or the groups of individuals in their houses.)

The ecological fallacy is not a paradox; it isn't a question of what is the correct unit of analysis. If you want to study the characteristic of individuals, your unit of analysis should be individuals. If you want to study groups, your unit of analysis should be those groups. It's a fallacy to study groups when your interest is in individuals.

One way to think about the difference between Robinson's conclusions with groups (the states) and the very different conclusions with individuals, is the factors that create the groups. Give an explanation, in everyday terms, why the immigrants that Robinson studied might tend to be clustered in states with low illiteracy rates, even if the immigrants themselves had high rates of illiteracy.

Chapter Eleven Reading Questions

- What is a probability model?
- What are some of the different probability models and what different situations do they describe?
- What is a "parameter" in a probability model? Give some examples.

Prob 11.01

Two basic operations that you need to perform on probability models are these:

percentile Given a value, what is the probability (according to the model) of the outcome from one trial being that value or less?

quantile Given a probability, what is the value whose percentile is that probability?

To illustrate by example, suppose that you are dealing with a probability model for an IQ test score that is a normal distribution with these parameters: mean = 100 and standard deviation = 15.

Percentile question: What is the percentile that corresponds to a test score of 120? Answer: 0.91 or, in other words, the 91st percentile.

```
> pnorm(120, mean=100,sd=15)
[1] 0.9087888
```

Quantile question: What score will 95% of scores be less than or equal to? Answer: a score of 125.

```
> qnorm(0.95, mean=100, sd=15)
[1] 124.6728
```

Here are two very basic questions about percentile and quantile calculations:

(a) TRUE or FALSE The output of a percentile question will always be a probability, that is, a number between 0 and 1.

(b) TRUE or FALSE The output of a quantile question will always be a value, that is, something in the same units as the random variable.

Sometimes to answer more complicated questions, you need first to answer one or more percentile or quantile questions.

Answer the following questions, using the normal probability model with the parameters given above:

(a) What's the test score that almost everybody, say, 99% of people, will do better than?

- Which kind of calculation is this? percentile quantile
- What is the answer? 55 65 75 95 115 125 135

(b) To calculate a coverage interval on a probability model, you need to calculate two quantities: one for the left end of the interval and one for the right. Which type of calculation are these probabilities from: percentile quantile

(c) Calculate a 50% coverage interval on the test scores, that is the range from the 0.25 quantile to the 0.75 quantile:

- Left end of interval: 80 85 90 95 100 105 110 115 120
- Right end of interval: 80 85 90 95 100 105 110 115 120

(d) Calculate an 80% coverage interval, that is the range from the 0.10 to the 0.90 quantile:

- Left end of interval: 52 69 73 81 85 89
- Right end of interval: 117 119 123 128 136

(e) To calculate the probability of an outcome falling in a given range, you need to do two percentile calculations, one for each end of the range. Then you substract the two different probabilities. What is the probability of a test score falling between 100 and 120? 0.25 0.37 0.41 0.48 0.52 0.61 0.73

Prob 11.02

A coverage interval gives a range of values. The "level" of the interval is the probability that a random trial will fall inside that range. For example, in a 95% coverage interval, 95% of the trials will fall within the range.

To construct a coverage interval, you need to translate the level into two quantiles, one for the left side of the range and one for the right side. For example, a 50% coverage interval runs from the 0.25 quantile on the left to the 0.75 quantile on the right; a 60% coverage interval runs from 0.20 on the left to 0.80 on the right. The probabilities used in calculating the quantiles are set so that

- the difference between them is the *level* of the interval. For instance, 0.75 and 0.25 give a 50% interval.
- they are symmetric. That is, the left probability should be exactly as far from 0 as the right probability is from 1

A classroom of students was asked to calculate the left and right probabilities for various coverage intervals. Some of their answers were wrong. Explain what is wrong, if anything, for each of these answers.

(a) For a 70% interval, the 0.20 and 0.90 quantiles

- [A] The difference between them isn't 0.70
- [B] They are not symmetrical.
- [C] Nothing is wrong.

(b) For a 95% interval, the 0.05 and 0.95 quantiles.

- [A] The difference between them isn't 0.95
- [B] They are not symmetrical.
- [C] Nothing is wrong.

(c) For a 95% interval, the 0.025 and 0.975 quantiles.

- [A] The difference between them isn't 0.95
- [B] They are not symmetrical.
- [C] Nothing is wrong.

Prob 11.03

For each of the following probability models, calculate a 95% coverage interval. This means that you should specify a left value and a right value. The left value corresponds to a probability of 0.025 and the right value to a probability of 0.975.

(a) The number of cars driving along a highway in one hour, when the mean number of cars is 2000 per hour. Hint: Poisson model

Left side of interval: 1812 1904 1913 1928 1935

Right side of interval: 2064 2072 2077 2088 2151

(b) The number of heads out of 100 flips of a fair coin. Hint: Binomial model.

Left side of interval: 36 38 40 42 44 46

Right side of interval: 54 56 58 60 62 64

(c) The angle of a random spinner, ranging from 0 to 360 degrees. Hint: Uniform model.

Left side of interval: 9 15 25 36 42 60

Right side of interval: 300 318 324 335 345 351

Prob 11.04

For each of these families of probability distributions, what are the parameters used to describe a specific distribution?

(a) Uniform distribution

- [A] Mean and Standard Deviation
- [B] Max and Min
- [C] Probability and Size
- [D] Average Number per Interval

(b) Normal distribution

- [A] Mean and Standard Deviation
- [B] Max and Min
- [C] Probability and Size
- [D] Average Number per Interval

(c) Exponential distribution

- [A] Mean and Standard Deviation
- [B] Max and Min
- [C] Probability and Size
- [D] Average Number per Interval

(d) Poisson distribution

- [A] Mean and Standard Deviation
- [B] Max and Min
- [C] Probability and Size
- [D] Average Number per Interval

(e) Binomial distribution

- [A] Mean and Standard Deviation
- [B] Max and Min
- [C] Probability and Size
- [D] Average Number per Interval

Prob 11.05

College admissions offices collect information about each year's applicants, admitted students, and matriculated students. At one college, the admissions office knows from past years that 30% of admitted students will matriculate.

The admissions office explains to the administration each year that the results of the admissions process vary from year to year due to random sampling fluctuations. Each year's results can be interpreted as a draw from a random process with a particular distribution.

Which family of probability distribution can best be used to model each of the following situations?

(a) 1500 students are offered admission. The number of students who will actually matriculate is:

[A] Normal
[B] Uniform
[C] Binomial
[D] Poisson
[E] Exponential
[F] Lognormal

(b) The average SAT score of the admitted applicants:

[A] Normal
[B] Uniform
[C] Binomial
[D] Poisson
[E] Exponential
[F] Lognormal

(c) The number of women in the matriculated class:

[A] Normal
[B] Uniform
[C] Binomial
[D] Poisson
[E] Exponential
[F] Lognormal

Prob 11.06

In 1898, Ladislaus von Bortkiewicz published *The Law of Small Numbers*, in which he demonstrated the applicability of the Poisson probability law. One example dealt with the number of Prussian cavalry soldiers who were kicked to death by their horses. The Prussian army monitored 10 cavalry corps for 20 years and recorded the number X of fatalities each year in each corps. There were a total of $10 \times 20 = 200$ one-year observations, as shown in the table:

Number of Deaths X	Number of Times X Deaths Were Observed
0	109
1	65
2	22
3	3
4	1

(a) From the data in the table, what was the mean number of deaths per year per cavalry corps?

[A] (109+65+22+3+1)/5
[B] (109+65+22+3+1)/200
[C] (0*109 + 1*65 + 2*22 + 3*3 + 4*1)/5
[D] (0*109 + 1*65 + 2*22 + 3*3 + 4*1)/200
[E] Can't tell from the information given.

(b) Use this mean number of deaths per year and the Poisson probability law to determine the theoretical proportion of years that 0 deaths should occur in any given calvary corps.

0.3128 0.4286 0.4662 0.5210 0.5434

(c) Repeat the probability calculation for 1, 2, 3, and 4 deaths per year per calvary corps. Multiply the probabilities by 200 to find the expected number of calvary corps with each number of deaths. Which of these tables is closest to the theoretical values:

[A]	112.67	64.29	16.22	5.11	1.63
[B]	108.67	66.29	20.22	4.11	0.63
[C]	102.67	70.29	22.22	6.11	0.63
[D]	106.67	68.29	17.22	6.11	1.63

Prob 11.07

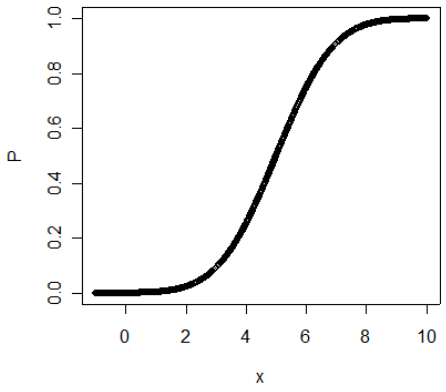
Experience shows that the number of cars entering a park on a summer's day is approximately normally distributed with mean 200 and variance 900. Find the probability that the number of cars entering the park is less than 195.

(a) Which type of calculation is this?
percentile quantile

(b) Do the calculation with the given parameters. (Watch out! Look carefully at the parameters and make sure they are in a standard form.)
0.3125 0.4338 0.4885 0.5237 0.6814 0.7163

Prob 11.08

The graph shows a cumulative probability.



(a) Use the graph to estimate by eye the 20th percentile of the probability distribution. (Select the closest answer.)

0 2 4 6 8

(b) Using the graph, estimate by eye the probability of a randomly selected x falling between 5 and 8? (Select the closest answer.)

0.05 0.25 0.50 0.75 0.95

Prob 11.09

Ralph's bowling scores in a single game are normally distributed with a mean of 120 and a standard deviation of 10.

(1) He plays 5 games. What is the mean and standard deviation of his total score?

Mean: 10 120 240 360 600 710
Standard deviation: 20.14 21.69 22.36 24.31 24.71

(2) What is the mean and standard deviation of his average score for the 5 games?

Mean: 10 60 90 120 150 180

Standard deviation: 4.47 4.93 5.10 5.62 6.18

Lucky Lolly bowls games that with scores randomly distributed with a mean of 100 and standard deviation of 15.

(3) What is the z-score of 150 for Lolly?
1.00 2.00 2.33 3.00 3.33 7.66 120 150

(4) What is the z-score of 150 for Ralph?
1.00 2.00 2.33 3.00 3.33 7.66 120 150

(5) Is Lolly or Ralph more likely to score over 150?

[A] Lolly
[B] Ralph
[C] Equally likely
[D] Can't tell from the information given.

(6) What is the z-score of 130 for Lolly?
1.00 2.00 2.33 3.00 3.33 7.66 120 150

(7) What is the z-score of 130 for Ralph?
1.00 2.00 2.33 3.00 3.33 7.66 120 150

(8) Is Lolly or Ralph more likely to score over 130?

[A] Lolly
[B] Ralph
[C] Equally likely
[D] Can't tell from the information given.

Prob 11.10

Jim scores 700 on the mathematics part of the SAT. Scores on the SAT follow the normal distribution with mean 500 and standard deviation 100. Julie takes the ACT test of mathematical ability, which has mean 18 and standard deviation 6. She scores 24. If both tests measure the same kind of ability, who has the higher score?

[A] Jim
[B] Julie
[C] They are the same.
[D] No way to tell.

Prob 11.11

For each of the following, decide whether the random variable is binomial or not. Then choose the best answer from the set offered.

(a) Number of aces in a draw of 10 cards from a shuffled deck with replacement.

[A] It is binomial.
[B] It's not because the sample size is not fixed.
[C] It's not because the probability is not fixed for every individual component.
[D] It's not for both of the above reasons.

(b) Number of aces in a draw of 10 cards from a shuffled deck without replacement.

[A] It is binomial.
[B] It's not because the sample size is not fixed.
[C] It's not because the probability is not fixed for every individual component.
[D] It's not for both of the above reasons.

(c) A broken typing machine has probability of 0.05 to make a mistake on each character. The number of erroneous characters in each sentence of a report.

[A] It is binomial.
[B] It's not because the sample size is not fixed.
[C] It's not because the probability is not fixed for every individual component.
[D] It's not for both of the above reasons.

(d) Suppose screws produced by a certain company will be defective with probability 0.01 independent of each other. The company sells the screws in a package of 10. The number of defective screws in a randomly selected pack.

[A] It is binomial.
[B] It's not because the sample size is not fixed.
[C] It's not because the probability is not fixed for every individual component.
[D] It's not for both of the above reasons.

(e) Observe the sex of the next 50 children born at a local hospital. Let x= # of girls among them.

[A] It is binomial.
[B] It's not because the sample size is not fixed.
[C] It's not because the probability is not fixed for every individual component.
[D] It's not for both of the above reasons.

(f) A couple decides to continue to have children until their first daughter. Let x = # of children the couple has.

[A] It is binomial.
[B] It's not because the sample size is not fixed.
[C] It's not because the probability is not fixed for every individual component.
[D] It's not for both of the above reasons.

(g) Jason buys the state lottery ticket every month using his favorite combination based on his birthday and his wife's. x= # of times he wins a prize in one year.

[A] It is binomial.
[B] It's not because the sample size is not fixed.
[C] It's not because the probability is not fixed for every individual component.
[D] It's not for both of the above reasons.

Prob 11.12

Just before a referendum on a school budget, a local newspaper plans to poll 400 random voters out of 50,000 registered voters in an attempt to predict whether the budget will pass. Suppose that the budget actually has the support of 52% of voters.

(a) What is the probability that the newspaper's sample will wrongly lead them to predict defeat, that is, less than 50% of the poll respondents will indicate support?

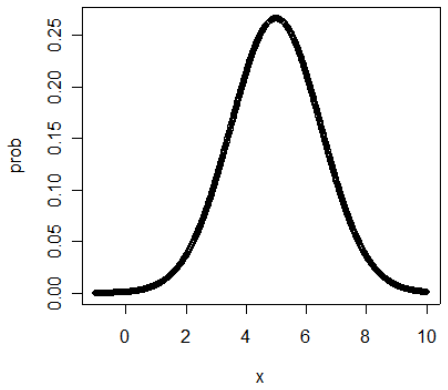
A `qbinom(.5,size=400,prob=.52)`
B `pbinom(.52,size=400,prob=.50)`
C `rnorm(400,mean=0.52,sd=.50)`
D `pbinom(199,size=400,prob=0.52)`
E `qnorm(.52,mean=400,sd=.50)`

(b) What is the probability that more than 250 of those 400 voters will support the budget?

A `pbinom(250,size=400,prob=.50)`
B `pbinom(249,size=400,prob=0.52)`
C `1-pbinom(250,size=400,prob=0.52)`
D `qbinom(.5,size=400,prob=.52)`
E `1-qnorm(.52,mean=400,sd=.50)`

Prob 11.13

Here is a graph of a probability density.



(a) Using the graph, estimate by eye the probability of a randomly selected x falling between 2 and 4. (Give the closest answer.)

0.05 0.25 0.50 0.75 0.95

(b) Using the graph of probability density above, estimate by eye the probability of a randomly selected x being less than 2. (Give the closest answer.)

0.05 0.25 0.50 0.75 0.95

Prob 11.14

A student is asked to calculate the probability that $x = 3.5$ when x is chosen from a normal distribution with the following parameters: mean=3, sd=5. To calculate the answer, he uses this command:

```
> dnorm(3.5, mean=3, sd=5)
[1] 0.0794
```

This is not right. Why not?

A He should have used pnorm.
B The parameters are wrong.
C The answer is zero since the variable x is continuous.
D He should have used qnorm.

Prob 11.15

A paint manufacture has a daily production, x, that is normally distributed with a mean 100,000 gallons and a standard deviation of 10,000 gallons. Management wants to create an incentive for the production crew when the daily production exceeds the 90th percentile of the distribution. You are asked to calculate at what level of production should management pay the incentive bonus?

A `qnorm(0.90,mean=100000,sd=10000)`
B `pnorm(0.90,mean=100000,sd=10000)`
C `qbinom(10000,size=100000,prob=0.9)`
D `dnorm(0.90,mean=100000,sd=10000)`

Prob 11.16

Suppose that the height, in inches, of a randomly selected 25-year-old man is a normal random variable with standard deviation 2.5 inches. In the strange universe in which statistics problems are written, we don't know the mean of this distribution but we do know that approximately 12.5% of 25-year-old man are taller than 6 feet 2 inches. Using this information, calculate the following.

(a) What's the average height of 25-year-old men? That is, find the mean of a normal distribution with standard deviation of 2.5 inches such that 12.5% of the distribution lies at or above 74 inches.

68.54 70.13 71.12 73.82 74.11 75.23 75.88 76.14

(b) Using this distribution, how tall should a man be in order to be in the tallest 5% of 25-year-old men?

68.54 70.13 71.12 73.82 74.11 75.23 75.88 76.14

Prob 11.17

In commenting on the "achievement gap" between different groups in the public school, the Saint Paul Public School Board released the following information:

> Saint Paul Public Schools (SPPS) serve more than 42,000 students. Thirty percent are African American, 30% Asian, and 13% Hispanic. The stark reality is that reading scores for two-thirds of our district's African American students fall below the national average, while reading scores for 90% of their white counterparts surpass it.

The point of this exercise is to translate this information into the point-score increase needed to bring African American students' scores into alignment with the white students.

Imagine that the test scores for white students form a normal distribution with a mean of 100 and a standard deviation of 25. Suppose also that African American students have test scores that form a normal distribution with a standard deviation of 25. What would have to be the mean of the African American students' test scores in order to match the information given by the School Board?

(a) What is the score threshold for passing the test if 90% of white students pass? One of the following R commands will calculate it. Which one?

A `pnorm(0.1,mean=100,sd=25)`
B `pnorm(0.9,mean=100,sd=25)`
C `qnorm(25,mean=100,sd=0.9)`
D `qnorm(0.1,mean=100,sd=25)`
E `qnorm(0.9,mean=100,sd=25)`
F `qnorm(25,mean=100,sd=0.1)`
G `rnorm(25,mean=100,sd=0.9)`

(b) Using that threshold, what would be the mean score for African Americans such that two-thirds (66.7%) are below the threshold? Hint: If you knew the answer, then it would produce this result.

```
> pnorm(67.96,mean=YourAnswer,sd=25)
[1] 0.667
```

Start by proposing an answer; a guess will do. Look at the resulting response and use that to guide refining your proposal until you hit the target response: 0.667. When you are at the target, your proposal will be close to one of these:
47 57 63 71 81

(c) Suppose scores for the African American students were to increase by 15 points on average. What would be the failure rate (in percent)?
21 35 44 53 67 74

(d) A common way to report the difference between two groups is the number of standard deviations that separate the means. How big is this for African American students in the Saint Paul Public Schools compared to whites (under the assumptions made for this problem)?
0.32 0.64 1.18 1.72 2.66 5.02

It would be more informative if school districts gave the actual distribution of scores rather than the passing rate.

Prob 11.18

A manufacturer of electrical fiber optic cables produces spools that are 50,000 feet long. In the production process, flaws are introduced randomly. A study of the flaws indicates that, on average, there is 1 flaw per 10000 feet.

(a) Which probability distribution describes the situation of how many flaws there will be in a spool of cable.

A normal
B uniform
C binomial
D exponential
E poisson

(b) What's the probability that a 50,000 foot-long cable has 3 or fewer flaws? (Enter your answer to 3 decimal places, e.g., 0.456.)

to within ±0.001

Prob 11.19

To help reduce speeding, the local governments sometimes put up speed signs at locations where speeding is a problem. These signs measure the speed of each passing car and display that speed to the driver. In some countries, such as the UK, the devices are equiped with a camera which records an image of each speeding car and a speeding ticket is sent to the registered owner.

At one location, the data recorded from such a device indicates that between 7 and 10 PM, 32% of cars are speeding and that 4.3 cars per minute pass the intersection, on average.

Which probability distributions can be best used to model each of the following situations for 7 to 10PM?

(a) The number of speeding cars in any 1 hour period.

A Normal
B Uniform
C Binomial
D Poisson
E Exponential
F Lognormal

(b) The time that elapses between cars:

A Normal
B Uniform
C Binomial
D Poisson
E Exponential
F Lognormal

(c) Out of 100 successive cars passing the device, the number that are speeding

A Normal
B Uniform
C Binomial
D Poisson
E Exponential
F Lognormal

(d) The mean speed of 100 successive cars passing the device.

A Normal
B Uniform
C Binomial
D Poisson
E Exponential
F Lognormal

Prob 11.20

As part of a test of the security inspection system in an airport, a government supervisor adds 5 suitcases with illegal materials to an otherwise shipping load, bringing the total to 150 suitcases.

In order to determine whether the shipment should be accepted, security officers randomly select 15 of the suitcases and X-rays them. If one or more of the suitcases is found to contain the materials, the entire shipment will be searched.

1. What probability model best applies here in describing the probability that at one or more of the five added suitcases will be X-rayed?

   A Normal
   B Uniform
   C Binomial
   D Poisson
   E Exponential

2. What is the probability that one or more of the five added suitcases will be X-rayed?

   A `1-pnorm(5,mean=150,sd=15)`
   B `1-pnorm(15,mean=150,sd=5)`
   C `1-punif(5,min=0,max=150)`
   D `1-punif(15,min=0,max=150)`
   E `1-pbinom(0,size=15,prob=5/150)`
   F `1-ppois(0,45/150)`
   G `1-ppois(0,5/150)`
   H Not enough information to tell.

Prob 11.21

Do the best job you can answering this question. The information provided is not complete, but that's the way things often are.

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in antinuclear rallies.

(a) Rank the following possibilities from most likely to least likely, for instance "A B C D E":

   (a) Linda is a teacher.
   (b) Linda works in a bookstore and takes yoga classes.
   (c) Linda is a bank teller.
   (d) Linda sells insurance.
   (e) Linda is a bank teller and is active in the feminist movement.

(b) Is there any relationship between the probabilities of the above items that you can be absolutely sure is true?

From 'Judgments of and by representativeness,' in "Judgment under uncertainty : heuristics and biases" / edited by Daniel Kahneman, Paul Slovic, Amos Tversky. Pub info Cambridge ; New York : Cambridge University Press, c1982.

Prob 11.22

According to the website `http://www.wikihealth.com/Pregnancy`, approximately 3.6% of pregnant women give birth on the predicted date (using the method that calculates gestational duration starting at the time of the last menstrual period). Assume that the probability of giving birth is a normal distribution whose mean is at the predicted date. The standard deviation quantifies the spread of gestational durations.

Using just the 3.6% "fact," make an estimate of the standard deviation of all pregnancies assuming that pregnancies are distributed as a normal distribution centered on the predicted date. Hint: Think of the area under the distribution over the range that covers one 24-hour period.

Your answer should be in days. Select the closest value:
8 9 10 11 12

Prob 11.23

You have decided to become a shoe-maker. Contrary to popular belief, this is a highly competitive field and you had to take the Shoe-Maker Apprenticeship Trial (SAT) as part of your apprenticeship application.

(a) The Shoe-maker's union has told you that among all the people taking the SAT, the mean score is 700 and the standard deviation is 35.

   According to this information, what is the percentile corresponding to your SAT score of 750?

   A `1-pnorm(750,mean=700,sd=35)`
   B `pnorm(750,mean=700,sd=35)`
   C `qnorm(750,mean=700,sd=35)`
   D `1-qnorm(750,mean=700,sd=35)`
   E Not enough information to answer.

(b) Your friend just told you that she scored at the 95th percentile, but she can't remember her numerical score. Using the information from the Shoe-maker's union, what was her score?

   A `pnorm(.95,mean=700,sd=35)`
   B `pnorm(.95,mean=750,sd=35)`
   C `qnorm(.95,mean=700,sd=35)`
   D `qnorm(.95,mean=750,sd=35)`
   E Not enough information to answer.

Prob 11.24

Both the poisson and binomial probability distributions describe a count of events.

The binomial distribution describes a series of identical discrete events with two possible outcomes to each event: yes or no, true or false, success or failure, and so on. The number of "success" or "true" or "yes" events in the series is given by the binomial distribution, so long as the individual events are independent of one another. An example is the number of heads that occur when flipping a coin ten times in a row. Each flip has a heads or tails outcome. The individual flips are independent.

There are two parameters to the binomial distribution: the number of events ("size") and the probability of the outcome that will be counted as a success. For the distribution to be binomial, the number of events must be fixed ahead of time and the probability of success must be the same for each event. The outcome whose probability is represented by the binomial distribution is the number of successful events.

Example: You flip 10 fair coins and count the number of heads. In this case the size is 10 and the probability of success is 1/2.

Counter-example: You flip coins and count the number of flips until the 10th head. This is not a binomial distribution because the size is not fixed.

The poisson probability model is different. It describes a situation where the rate at which events happen is fixed but there is no fixed number of events.

Example: Cars come down the street in a random way but at an average rate of 3 per minute. The poisson distribution describes the probability of seeing any given number of cars in one minute. Unlike the binomial distribution, there is no fixed number of events; potentially 50 cars could pass by in one minute (although this is very, very unlikely).

Both the poisson and binomial distributions are discrete. You can't have 5.5 heads in 10 flips of a coin. You can't have 4.2 cars pass by in one minute. Because of this, the basic way to use the tabulated probabilities is as a probability assigned to each possible outcome.

Here is the table of outcomes for the number of heads in 6 flips of a fair coin:

```
> dbinom(0:5, size=6, prob=.5)
[1] 0.016 0.094 0.234 0.313 0.234 0.094 0.016
```

So, the probability of exactly zero heads is 0.016, the prob. of 1 head is 0.094, and so on.

You may also be interested in the cumulative probability:

```
> pbinom(0:5, size=6, prob=.5)
[1] 0.016 0.109 0.344 0.656 0.891 0.984 1.000
```

So, the probability of 1 head or fewer is 0.109, just the sum of the probability of exactly 0 heads and exactly one head.

Note that in both cases, there was no point in asking for the probability of more than 6 heads; six is the most that could possibly happen. If you do ask, the answer will be "zero": it can't happen.

```
> dbinom(7, size=6, prob=.5)
[1] 0
```

Similarly, there is no point in asking for the probability of 3.5 heads, that can't happen either.

```
> dbinom(3.5, size=6, prob=.5)
[1] 0
Warning message:
non-integer x = 3.500000
```

The software sensibly returns a probability of zero, but warns that you are asking something silly.

The poisson distribution is similar, but different in important ways. If cars pass by a point randomly at an average rate of 3 per minute, here is the probability of seeing 0, 1, 2, ... cars in any randomly selected minute.

```
> dpois(0:6, 3)
[1] 0.05 0.15 0.22 0.22 0.17 0.10 0.05
```

So, there is a 5% chance of seeing no cars in one minute.

But unlike the binomial situation, where the maximum number of successful outcomes is fixed by the number of events, it's possible for a very large number of cars to pass by.

```
> dpois(0:9, 3)
 [1] 0.0498 0.1494 0.2240 0.2240 0.1680
 [6] 0.1008 0.0504 0.0216 0.0081 0.0027
```

For instance, there is a 0.2% chance that 9 cars will pass by in one minute. That's small, but it's definitely non-zero.

The poisson model, like the binomial, describes a situation where the outcome is a whole number of events. It makes little sense to ask for a fractional outcome. The probability of a fractional outcome is always zero.

```
> dpois(3.5, 3)
[1] 0
Warning message:
non-integer x = 3.500000
```

Often one wants to consider a poisson event over a longer or shorter interval than the one implicit in the specified rate. For example, when you say that the average rate of cars passing a spot is 3 per minute, the interval of one-minute is implicit. Suppose, however, that you want to know the number of cars that might pass by a spot in one hour. To calculate this, you need to find the rate in terms of the new interval. Since one hour is 60 minutes, the rate of 3 per minute is equivalent to 180 per hour. You can use this to find the probability.

For example, the probability that 150 or fewer cars will pass by in one hour (when the average rate is 3 per minute) is given by a cumulative probability:

```
> ppois(150, 180)
[1] 0.0122061
```

It can be hard to remember whether the above means "150 or fewer" or "fewer than 150." When in doubt, you can always make the situation explicit by using a non-integer argument

```
> ppois(150.1, 180)  # includes 150
[1] 0.0122061
> ppois(149.9, 180)  # excludes 150
[1] 0.00991012
```

This works only when asking for cumulative probabilities, since 150.1 or less includes the integers 150, 149, and so on. Were you to ask for the probability of getting exactly 150.1 cars in one hour, using the `dpois` operator, the answer would be zero:

```
> dpois(150.1, 180)
[1] 0
Warning message:
non-integer x = 150.100000
```

For each of the following, figure out the computer statement with which you can compute the probability.

1. If cars pass a point randomly at an average rate of 10 per minute, what is the probability of exactly 15 cars passing in one minute?

   0.000 0.0026 0.035 0.053 0.087 0.263 0.337 0.334 0.437 0.559

2. If cars pass a point randomly at an average rate of 10 per minute, what is the probability of 15 or fewer cars passing in one minute?

   0.000 0.0026 0.035 0.053 0.087 0.263 0.334 0.337 0.437 0.559

3. If cars pass a point randomly at an average rate of 10 per minute, what is the probability of 20 or fewer cars passing in two minutes?

   0.000 0.0026 0.035 0.053 0.087 0.263 0.334 0.337 0.437 0.559 0.915 0.951

4. If cars pass a point randomly at an average rate of 10 per minute, what is the probability of 1200 or fewer cars passing in two hours and ten minutes (that is, 130 minutes)?

   0.000 0.0026 0.035 0.053 0.087 0.263 0.334 0.337 0.437 0.559 0.915 0.951

5. A department at a small college has 5 faculty members. If those faculty are effectively random samples from a population of potential faculty that is 40% female, what is the probability that 1 or fewer of the five deparment members will be female?

   0.000 0.0026 0.035 0.053 0.087 0.263 0.334 0.337 0.437 0.559 0.915 0.951

6. What is the probability that 4 or more will be female?

   0.000 0.0026 0.035 0.053 0.087 0.263 0.334 0.337 0.437 0.559 0.915 0.951

Prob 11.25

Government data indicates that the average hourly wage for manufacturing workers in the United States is $14. (Statistics Abstract of the United States, 2002) Suppose the distribution of the manufacturing wage rate nationwide can be approximated by a normal distribution. If a worker did a nationwide job search and found that 15% of the jobs paid more than $15.30 per hour. In order to find the standard deviation of the hourly wage for manufacturing workers, what process should we try?

- A `qnorm(0.15, mean=14, sd=15.3)`
- B Look for x such that `pnorm(15.3, mean=14,sd=x)` gives 0.85
- C Calculate a z-score using 1.3 as the standard deviation
- D Not enough information is being given.

Prob 11.26

In many social issues, policy recommendations are based on cases from the extremes of a distribution. Consider, for example, a news story (Minnesota Public Radio, March 19, 2006) comparing the high-school graduation rates of Native Americans (85%) and whites (97%). The disparity becomes more glaring when one compares high-school drop-out rates, 3% for whites and 15% for Native Americans.

One way to compare these two drop-out rates is simply to take the ratio: five times as many children in one group drop out as in the other. Or, one could claim that the graduation rate for one group is "only" a factor of 1.14 higher than that for the other. While both of these descriptions are accurate, neither of them has a unique claim to truth.

Another way to interpret the data is to imagine a student's high-school experience as a point on a quantitative continuum. If the experience is below a threshold, the student does not graduate. We can imagine the outcome as being the sum of many contributions: quality of the school and teachers, support from family, peer influences, personality of the student, and so on.

Suppose that we model the high-school experience as a normal distribution with the same standard deviation for whites and Native Americans but with different means. For the sake of specificity, let the mean for whites be 100 with a standard deviation of 20.

1) What is the threshold for graduation? Find a number such that 97% of whites are below this.

2) Using the threshold you found above, find the mean for Native Americans such that 15% of children are below the threshold.

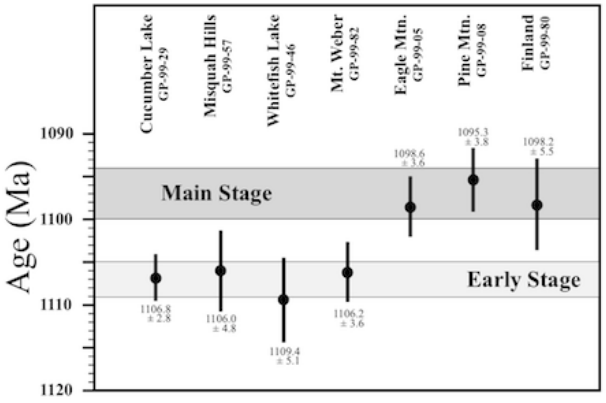
This model is, of course, arbitrary. We don't know that there is anything that corresponds to a quantitative high-school experience," and we certainly don't know that even if there were it would be distributed according to a normal distribution. Nevertheless, this can be a helpful way to interpret data about the "extremes" when making comparisons of the means.

3) Suppose, contrary to fact, that the drop-out rate for group A is 15% and that for group B were five times as high: 75%. If group A has a high-school experience with mean 100 and standard deviation 20, and group B has a standard deviation of 20, what should be the mean of group B to produce the higher drop-out rate.

Enter the work you used to answer these questions in the box. You can cut and paste from the computer output, but make sure to indicate clearly and concisely what your answers are.

Prob 11.27

Geology Professor Karl Wirth studies the age of rocks as determined by ratios of isotropes. The figure shows the results of an age assay of rocks collected at seven sites. Because of the intrinsically random nature of radioactive decay, the measured age is a random variable and has been reported as a mean (in millions of years before the present) and a standard deviation (in the same units).



From the geology of the sites, four of them have been classified as "early stage" and three as "main stage." The graph

clearly indicates that the early stage rocks tend to be younger than the main stage rocks. But perhaps this is just the luck of the draw.

Professor Wirth wants to calculate a new random variable: the difference in mean ages between the early and main stage rocks. Since the age difference is a random variable, Prof. Wirth needs to know both the variable's mean and it's standard deviation.

To get you started on the calculation, here's the formula for the difference in mean ages, Δ_{age} of the rocks from the two different stages.

$$\Delta_{\text{age}} = \frac{1}{3}(M_1 + M_2 + M_3) + \frac{1}{4}(E_1 + E_2 + E_3 + E_4)$$

where M_i is a rock from the main stage and E_i is a rock from the early stage.

To remind you, here are the arithmetic rules for the means and variances of random variables V and W when summed and multiplied by fixed constants a and b:

- mean(aV) = a mean(V)
- var(aV) = a^2 var(V)
- mean($aV + bW$) = a mean(V) + b mean(W)
- var($aV + bW$) = a^2 var(V) + b^2 var(W)

Do calculations based on the above formulas to answer these questions:

1. What is the mean of Δ_{age}? What are the units?
2. What is the variance of Δ_{age}? What are the units?
3. What is the variance of Δ_{age}? What are the units?
4. A skeptic claims that the two stages do not differ in age. He points out, correctly, that since Δ_{age} is a random variable, there is some possibility that its value is zero? What is the z-score of the value 0 in the distribution of Δ_{age}?

Prob 11.28

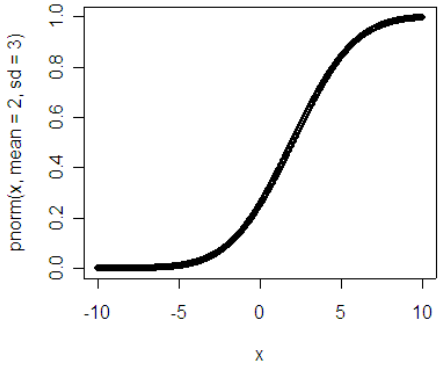
Below are two different graphs of cumulative probability distributions. Using the appropriate graph, not the computer, estimate the items listed below. Your estimates are not expected to be perfect, but do mark on the graph to show the reasoning behind your answer:

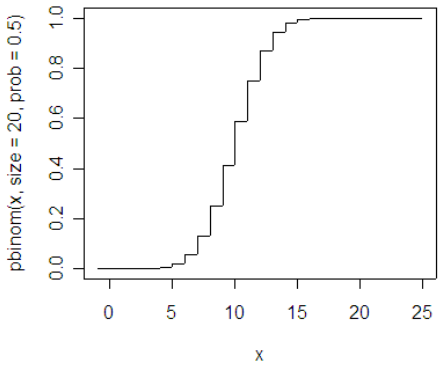
(a) The 75th percentile of a normal distribution with mean 2 and standard deviation 3.

(b) When flipping 20 fair coins, the probability of getting 7 or fewer heads.

(c) The probability of x being 1 standard deviation or more below the mean of a normal distribution.

(d) The range that covers 90% of the most likely number of heads when flipping 20 fair coins.





Chapter Twelve Reading Questions

- What is the difference between a "standard error" and a "confidence interval?"
- Why are confidence intervals generally constructed at a 95% level? What does the level mean?
- What is a sampling distribution?
- What is resampling? What is it used for?
- How does a confidence interval describe what might be called the reliability of a measurement?
- Does collinearity between explanatory vectors tend to make confidence intervals smaller or larger?

Prob 12.01

Here's a confidence interval: 12.3 ± 9.8 with 95% confidence.

(a) What is 12.3?

- margin.of.error
- point.estimate
- standard.error
- confidence.level
- confidence.interval

(b) What is 9.8?

- margin.of.error
- point.estimate
- standard.error
- confidence.level
- confidence.interval

(c) What is 95%?

- margin.of.error
- point.estimate
- standard.error
- confidence.level
- confidence.interval

Prob 12.02

Look at this report from a model of the kids' feet data,

```
summary(lm(width~length+sex,data=kids))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.6412     1.2506   2.912  0.00614
length        0.2210     0.0497   4.447 8.02e-05
sexG         -0.2325     0.1293  -1.798  0.08055
```

(a) Based on the output of the report, which of these statements is a correct confidence interval on the `sexG` coefficient?

- A $-.23 \pm 0.13$ with 95 percent confidence
- B $-.23 \pm 0.13$ with 50 percent confidence
- C $-.23 \pm 0.13$ with 68 percent confidence
- D $-.23 \pm 0.0805$ with 95 percent confidence
- E $-.23 \pm 0.23$ with 68 percent confidence
- F None of the above

(b) Based on the output of the report, which of these statements is a correct confidence interval on the `length` coefficient?

- A 0.22 ± 0.050 with 95 percent confidence
- B 0.22 ± 0.050 with 68 percent confidence
- C 0.22 ± 0.100 with 50 percent confidence
- D 0.22 ± 0.070 with 50 percent confidence
- E None of the above

Prob 12.03

A confidence interval is often written as a point estimate plus or minus a margin of error: $P \pm M$ with C percent confidence. How does the size of the margin of error M depend on the confidence level C?

- A It doesn't.
- B It increases as C increases.
- C It decreases as C increases.

Prob 12.05

A statistics professor sends her students out to collect data by interviewing other students about various quantities, e.g., their SAT scores, GPAs and other data for which the college registrar retains official records. Each student is assigned one quantity and one target population, for example, the verbal SAT scores of all female students, or the cumulative grade point average for sophomore males.

Each student interviews some students — how many depends on the student's initiative and energy. The student then reports the results in the form of a confidence interval, $P \pm M$ with 95% confidence.

After the students hand in their reports, the professor contacts the registrar to find the population parameters for each group that the students surveyed. **Assuming** that the interviewed students provided accurate information, what fraction of the students' confidence intervals will contain the corresponding population parameter?

- A 95%
- B 50%
- C 5%
- D Can't tell, it depends on how much data each student collected.
- E Can't tell, the students each looked at different quantities, not all of them at the same quantity.

Prob 12.07

Here are three different model statements for the kids' feet data.

- width $\sim$ 1
- width $\sim$ sex
- width $\sim$ sex - 1

Each of the above models for kids' feet is relevant to one of the problems below. Fit the model to the data in `kidsfeet.csv` and interpret your results to give a **95% confidence interval** on these quantities written in the standard form: point estimate $\pm$ margin of error.

1. The mean width of boys' feet.

   Point estimate: 8.76 9.19 9.37 9.98 10.13

   Margin of error: 0.041 0.176 0.211 0.352 1.430 6.540

2. The mean width of all children's feet.

   Point estimate: 8.15 8.99 9.13 9.86 12.62

   Margin of error: 0.16 0.18 0.22 0.35 1.74

3. The difference between the means of boys' and girls' foot widths. (The differences can be either positive or negative, depending on whether it is boys minus girls or girls minus boys. State your difference as a positive number.)

   Point estimate: 0.406 0.458 0.514 0.582 0.672

   Margin of error: 0.16 0.18 0.22 0.30 1.74

Prob 12.11

What's wrong with the following statement written by a student on an exam?

> The the larger the number of cases examined and taken into account, the more likely your estimation will be accurate. Having more cases decreases your risk of having a bias and increases the probability that your sample accurately represents the real world.

Prob 12.22

In 1882, Charles Darwin wrote about earthworms and their importance in producing soil.

> *Hensen, who has published so full and interesting an account of the habits of worms, calculates, from the number which he found in a measured space, that there must exist 133,000 living worms in a hectare of land, or 53,767 in an acre.* — p. 161, "The Formation of Vegetable Mould, through the Action of Worms with Observations on their Habits"

While 133,000 seems sensibly rounded, 53,767 is not. This problem investigates some of the things you can find out about the precision of such numbers and how to report them using modern notation, which wasn't available to Darwin or his contemporaries.

Background: A hectare is a metric unit of area, 10,000 square meters. An acre is a traditional unit of measure, with one acre equivalent to 0.4046863 hectares. That is, an acre is a little less than half a hectare.

The implicit precision in Hensen's figure is $133,000 \pm 500$, since it is rounded to the thousands place. Correctly translate the Hensen figure to be in worms per acre.

1. Literally translating 133,000 worms per hectare to worms per acre gives what value?
   53760 53767 53770 53823 53830
2. Literally translating ± 500 worms per hectare to worms per acre gives what value?
   197 200 202 205 207
3. Which one of these reports gives a proper account for the number of worms per acre?
   - [A] 53767 ± 202
   - [B] 53823 ± 200
   - [C] 53820 ± 200
   - [D] 53830 ± 200

Of course, it's just an assumption that Hensen's precision is ± 500. Imagine that the way Hensen made his estimate was to dig up 10 different patches of ground, each of size one square meter. In each patch, Hensen counted the worms found then added these together to get the total number of worms in 10 square meters. Since Hensen reported 133,000 worms per hectare, he would have found a total of 133 worms in the ten square meters he actually dug up.

Of course, if Hensen had picked a different set of 10 patches of soil in the same field, he would likely not have found exactly 133 worms. There is some sampling variability to the number of worms found.

Using an appropriate probability model for the number of worms to be found in 10 square meters of soil, estimate the standard deviation of the number found, assuming that on average the number is 133 per 10 square meters.

1. What is an appropriate probability model?
   gaussian uniform exponential poisson binomial
2. Using the appropriate probability model, what standard deviation corresponds to a mean of 133 per 10 square meters? (Hint: You can use a random number generator to make a large sample of draws and then find the standard deviation of this sample.)
   2.1 7.9 11.5 15.9 58.2 102
3. Using your standard deviation, and recalling that the number of worms in one hectare will be 1000 times that found in 10 square meters, give an appropriate 95% confidence interval to use today in reporting Hensen's result.
   - [A] $133,000 \pm 23000$
   - [B] $133,000 \pm 2100$
   - [C] $133,000 \pm 16000$
   - [D] $133,000 \pm 20000$
   - [E] $130,000 \pm 120000$
4. Now imagine, almost certainly contrary to fact, that Hensen had actually dug up an entire hectare and found 133,201 worms, and rounded this to 133,000 just for the sake of not seeming silly. Of course, this would have been a heroic effort just to gain precision on the count. It would also be futile, since the number in a "hectare of soil" presumably differs widely depending on the soil conditions. But if Hensen had calculated a 95% confidence interval using an appropriate probability model on the count of 133,201 worms, rather than just rounding to what seems reasonable, what would have been his margin of error?
   730 2100 16000 58000 190000

Chapter Thirteen Reading Questions

- What is a "null hypothesis?" Why does it play a special role in hypothesis testing?
- Why is it good to have a hypothesis test with a low "significance level?"
- Why is it good to have a hypothesis test with a high "power?"
- What is a p-value?
- Why are there two kinds of potential errors in hypothesis testing, Type I and Type II?

Prob 13.01

Which of the following are legitimate possible outcomes of a hypothesis test? Mark as "true" any legitimate ones.

(a) TRUE or FALSE accept the alternative hypothesis

(b) TRUE or FALSE accept the null hypothesis

(c) TRUE or FALSE reject the alternative hypothesis

(d) TRUE or FALSE fail to reject the alternative hypothesis

(e) TRUE or FALSE reject the null hypothesis

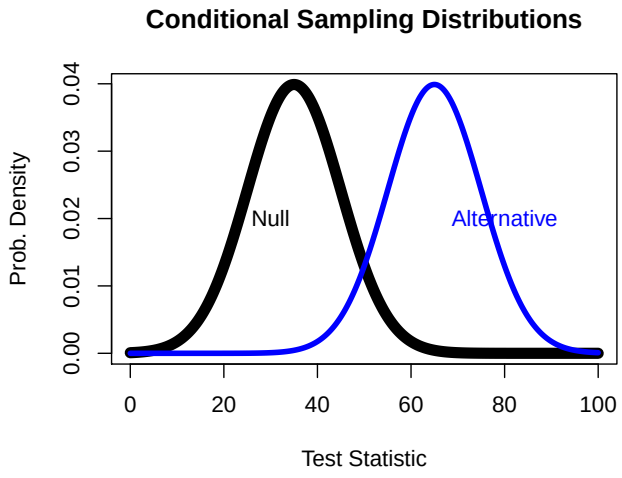
(f) TRUE or FALSE fail to reject the null hypothesis

(g) TRUE or FALSE indeterminate result

Pick one of the illegitimate outcomes and explain why it is illegitimate.

Prob 13.03

Consider the two conditional sampling distributions shown in the figure.



Imagine that you set rejection criteria so that the power was 99%. What would be the significance of the test? (Choose the best answer.)

A About 0.05.
B About 0.30
C About 0.95
D Can't tell (even approximately) from the information given.

We sometimes speak of the probability of Type I or Type II errors. We want to know what sort of probability these are. To simplify notation, we'll define the following outcomes:

N The Null Hypothesis is correct.

A The Alternative Hypothesis is correct.

Fail Fail to reject the Null Hypothesis

Reject Reject the Null Hypothesis.

What is the probability of a Type I error?

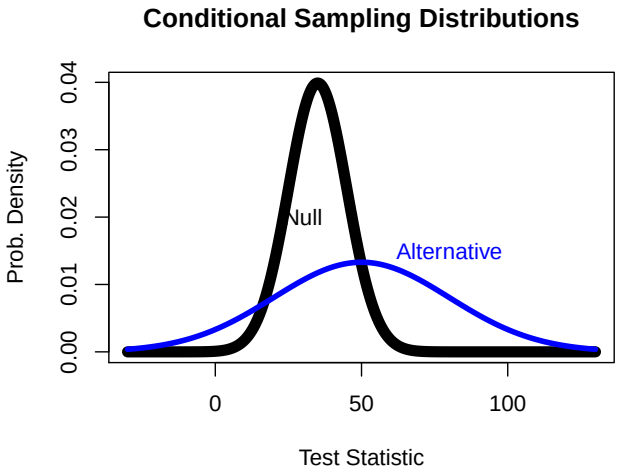
A A joint probability: p(Reject and N)
B A joint probability: p(Fail and A)
C A conditional probability: p(Reject given N)
D A conditional probability: p(Reject given A)
E A marginal probability: p(Reject)
F A marginal probability: p(Fail)

What is the probability of a Type II error?

A A joint probability: p(Reject and A)
B A joint probability: p(Fail and N)
C A conditional probability: p(Fail given N)
D A conditional probability: p(Fail given A)
E A marginal probability: p(N)
F A marginal probability: p(A)

Prob 13.23

Consider these two sampling distributions:



Suppose that you insisted that every hypothesis test have a significance of 0.05. For the conditional sampling distributions shown above, what would be the power of the most powerful possible one-tailed test?

A About 5%
B About 50%
C About 95%
D Can't tell from the information given.

What would be the power of the most powerful possible two-tailed test?

A About 10% bigger than the one-tailed test.
B About 10% smaller than the one-tailed test.
C About 50% bigger than the one-tailed test.
D About 50% smaller than the one-tailed test.
E Can't tell from the information given.

Prob 13.34

The Ivory-billed woodpecker has been thought to be extinct; the last known observation of one was in 1944. A new sighting in the Pearl River forest in Arkansas in 2004 became national news and excited efforts to confirm the sighting. These have not been successful and there is skepticism whether the reported 2004 sighting was correct. This is not an easy matter since the 2004 sighting was fleeting and the Ivory-billed woodpecker is very similar to a relatively common bird, the Pileated woodpecker.

Pileated (left) and Ivory-billed Woodpeckers
Drawings by Ernest S Thompson & Rex Brasher

This problem is motivated by the controversy over the Ivory-billed sighting, but the problem is a gross simplification.

Ivory-billed woodpeckers are 19 to 21 inches long, while Pileated woodpeckers are 16 to 19 inches. Ivory-bills are generally glossy blue-black, whereas Pileated woodpeckers are generally dull black, slaty or brownish-black. Ivory-bills have white wing tops while Pileated woodpeckers have white underwings. These differences would make it easy to distinguish between the two types of birds, but sightings are often at a distance in difficult lighting conditions. It can be difficult, particularly in the woods, to know whether a distant bird is flying toward the observer or away.

Imagine a study where researchers display bird models in realistic observing conditions and record what the observers saw. The result of this study can usefully be stated as conditional probabilities:

Observed	Code	Alternative Ivory-billed	Null Pileated
Short & Dull	A	0.01	0.60
Long & Dull	B	0.10	0.13
Short & Glossy	C	0.04	0.20
Long & Glossy	D	0.60	0.05
Short & White Back	E	0.05	0.01
Long & White Back	F	0.20	0.01

For simplicity, each of the six possible observations has been given a code: A, B, C, and so on.

The Bayesian Approach The table above gives conditional probabilities of this form: *given that the bird is an Ivory-billed woodpecker*, the probability of observation D is 0.60.

In the Bayesian approach, you use the information in the table to compute a different form of conditional probability, in this form: *given that you observed D*, what is the probability of the bird being an Ivory-billed.

In order to switch the form of the conditional probability from "given that the bird is ..." to "given that the observation is ...", you need some additional information, called the **prior probability**. The prior probability reflects your view of how likely a random bird is to be an Ivory-billed or Pileated **before** you make any observation. Then, working through the Bayesian calculations, you will find the **posterior probability**, that is, the probability **after** you make your observation.

Suppose that, based on your prior knowledge of the history and biology of the Ivory-bill, that your prior probability that the sighting was really an Ivory-bill is 0.01, and the probability that the sighting was a Pileated is 0.99. With this information, you can calculate the **joint** probability of any of the 12 outcomes in the table.

Then, by considering each row of the table, you can calculate the **marginal** probability of Ivory-bill vs Pileated for each of the possible observations.

(a) What is the joint probability of a sighting being both D and Ivory-billed? (Pick the closest one.)

impossible 0.006 0.01 0.05 0.60

(b) What is the joint probability of a sighting being both D and Pileated? (Pick the closest one.)

impossible 0.006 0.01 0.05 0.60

(c) What is the conditional probability of the sighting being an Ivory-billed GIVEN that it was D? (Pick the closest one.)

0.01 0.05 0.10 0.60

(d) Which of the possible observations would provide the largest posterior probability that the observation was of an Ivory-bill?

A B C D E F

(e) What is that largest posterior probability? (Pick the closest one.)

0.02 0.08 0.17 0.73

The Hypothesis-Testing Approach The hypothesis-testing approach is fundamentally different from the Bayesian approach. In hypothesis testing, you pick a null hypothesis that you are interested in disproving. Since the Pileated woodpecker is relatively common, it seems sensible to say that the null hypothesis is that the observation is a Pileated woodpecker, and treat the Ivory-billed as the alternative hypothesis.

Once you have the null hypothesis, you choose rejection criteria based on observations that are unlikely given the null hypothesis. You can read these off directly from the conditional probability table given above.

1. The two least likely observations are E and F. If observing E or F is the criterion for rejecting the null hypothesis, what is the significance of the test?

   0.01 0.02 0.05 0.25 0.60 0.65

   What would be the power of this test?

   0.01 0.05 0.25 0.60 0.65

2. Now suppose the rejection criteria are broadened to include D, E, and F.

   What would be the significance of such a test?

   0.01 0.02 0.05 0.07 0.10 0.20 0.25 0.60 0.85 other

   What would be the power of such a test?

   0.01 0.02 0.05 0.07 0.10 0.20 0.25 0.60 0.85 other

Comparing the Bayesian and hypothesis-testing approaches, explain why you might reject the null hypothesis even if the observation was very likely to be a Pileated woodpecker.

Chapter Fourteen Reading Questions

- A permutation test involves randomizing in order to simulate eliminating the relationship, if any, between the explanatory variables and the response variable. What's the point of this?
- The F statistic compares two quantities. What are they?
- What is a "degree of freedom?"

Prob 14.01

Here is the report of a simple model of the foot-length data:

```
> kids = fetchData("kidsfeet.csv");
> summary( lm( length ~ 1, data=kids ) )
Coefficients:
            Estimate   Std. Err.  t value Pr(>|t|)
(Intercept) 24.723       0.211    117.2   <2e-16
```

The summary report includes a p-value (written as `Pr(>|t|)`. What is the Null Hypothesis corresponding to this p-value:

A The mean cannot be calculated.
B The sample mean is zero.
C The population mean is zero.
D The sample mean is greater than zero.
E The sample mean is less than zero.
F The population mean is greater than zero.
G The population mean is less than zero.

Prob 14.02

Your friend is interested in using natural teas to replace commercial drugs. He has been testing out his own tea blend, which he calls "Special Cooling Tea" to reduce temperatures of people with the flu. Fortunately for him, he had a bad flu with fever that lasted for 3 days. During this time, he alternated taking Special Cooling Tea and taking a conventional fever reducing drug, Paracetamol. On each occasion, he measured the change in his body temperature over one hour. His data were

Treatment	Change in Temp. (F)
Tea	-1.2
Drug	-2.0
Tea	-0.7
Drug	-1.5
Tea	+0.0
Drug	-0.7

Your friend is excited by these results. When he fit the model temperature change $\sim$ Treatment he got the following reports:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
treat	1	0.13	0.13	0.11	0.7555
Residuals	4	4.85	1.21		

Based on these report, your friend claims, "This shows that my Special Cooling Tea is just as effective as Paracetamol."

- Which of the following features of the ANOVA report supports your friend's conclusion that the two treatments are effective:
  A There are 4 degrees of freedom in the residual.
  B The p-value is very small.
  C The p-value is not small.
  D There is no p-value on the residuals.
- What's an appropriate Null Hypothesis is this setting:
  A The Tea and Paracetamol have the same effect.
  B The Tea is more effective than Paracetamol.
  C The Tea is less effective than Paracetamol.

Of course, your friend's conclusion that his experiment shows that Tea and Paracetamol are the same is actually unjustified. Failing to reject the null hypothesis is not the same as accepting the null hypothesis. You explain this to your friend.

He's disappointed, of course. But then he gets a bit annoyed. "How can you ever show that two things are the same if your never get to accept the null hypothesis?"

"Well," you respond, "the first thing is to decide what 'the same' means. In this case, you're interested in the fever reduction. How big a difference in the temperature reduction would be medically insignificant?"

Your friend thinks for a while. "I suppose no one would care about 0.2 degrees."

"Then you have to make your measurement precise to about 0.2 degrees. If such a precise measurement shows no difference, then it's reasonable to claim that your Tea and the drug are equivalent." Using your friend's data, you fit the model and look at the regression report.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.9333	0.6360	-1.47	0.2161
treatTea	0.3000	0.8994	0.33	0.7555

The margin of error on the difference in temperature reduction is ±1.8 degrees, based on the 6 cases your friend recorded.

- To reduce the margin of error to ±0.2 degrees, about how many cases altogether should your friend plan to collect? (Choose the closest one. For instance, 12 would be a doubling from the 6 actually collected.)
  12 75 150 500 2000

Prob 14.03

You can test the null hypothesis that the population mean of a variable x is zero by constructing a model `x~1` and interpreting the coefficient on the intercept term `1`.

Often, the null hypothesis that the population mean is zero is irrelevant. For example, in the kid's feet data, the mean foot width is 8.99 cm. There is no physical way for the population mean to be zero.

But, suppose you want to test the null hypothesis that the population mean (for the population of school-aged children from whom the sample was drawn) is 8.8 cm. To do this, create a new variable that would be zero if the population mean were 8.8 cm. This is simple: make the new variable by subtracting 8.8 cm from the original variable. The new variable can then be tested to see if its mean is different from zero.

Using the kids-feet data:

1. Find the p-value on the null-hypothesis that the population mean of the foot width is 8.8 cm.
   0.000 0.024 0.026 0.058 0.094 0.162 0.257
2. Find the p-value on the null-hypothesis that the population mean of the foot width is 8.9 cm.
   0.00 0.23 0.27 0.31 0.48 0.95

Prob 14.04

You are performing an agricultural experiment in which you will study whether placing small bits of iron in the soil will increase the yield of barley. Randomly scattered across an enclosure owned by the state's agricultural field office, you have marked off ten plots, each of size 10 meters by 10 meters. You randomly assign five of the plots to be controls, and in the other five you place the iron.

A colleague hears of your experiment and asks to join it. She would like to test zinc. Still another colleague wants to test copper and another has two different liquid fertilizers to test. The head of the field office asks you to share the data from your control plots, so that the other researchers won't have to grow additional control plots. This will reduce the overall cost of the set of experiments. He points out that the additional experiments impose absolutely no burden on you since they are being grown in plots that you are not using. All you need to do is provide the yield measurements for the control plots to your colleagues so that they can use them in the evaluation of their own data.

You wonder whether there is a hidden cost. Will you need to adjust the p-value thresholdhold? For instance, a Bonferroni correction would reduce the threshold for significance from 0.05 to 0.01, since there are five experiments altogether (iron, zinc, copper, and two fertilizers, each compared to a control). What's the cost of making such an adjustment?

- [A] It reduces the significance level of your experiment.
- [B] It reduces the power of your experiment.
- [C] It invalidates the comparison to the control.
- [D] None of the above.

What's the argument for making an adjustment to the p-value threshold?

- [A] Doing multiple experiments increases the probability of a Type I error.
- [B] Doing multiple experiments increase the probability of a Type II error.
- [C] Doing multiple experiments introduces interaction terms.
- [D] The other experiments increase the power of your experiment.

What's the argument for not making an adjustment to the p-value threshold?

- [A] The additional experiments (zinc, copper, fertilizers) have nothing to do with my iron experiment, so I can reasonably ignore them.
- [B] You don't think the other experiments will be successful.
- [C] A p-value of 0.05 is always appropriate.

Suppose that the other experimenters decided that they had sufficient budget to make their own control plots, so that their experiments could proceed independently of yours (although all the plots would be in the same field). Would this change the arguments for and against adjusting the p-value?

- [A] No. There are still multiple tests being done, even if the data sets don't overlap.
- [B] Yes. Now the tests aren't sharing any data.

The idea that a researcher's evaluation of a p-value should depend on whether or not other experiments are being conducted has been termed "inconsistent" by Saville. [**?**] He writes, "An experiment is no more a natural unit than a project consisting of several experiments or a research program consisting of several projects." Why should you do an adjustment merely because a set of experiments happened to be performed by the same researcher or in the same laboratory? If you're going to adjust, why shouldn't you adjust for all the tests conducted in the same institution or in the same country?

This quandry illustrates that data do not speak for themselves. The evaluation of evidence needs to be made in the context in which the data were collected. In my view, the iron researcher would be justified in not adjusting the p-value threshold because the other experiments are irrelevant to his. However, the field office head, who is supervising multiple experiments, should be making adjustments. This is paradoxical. If the p-value from the iron experimental were 0.04, the individual researcher would be justified in declaring the effect of iron significant, but the office head would not be. Same data, different results.

This paradox has an impact on how you should interpret the reports that you read. If you are reading hundreds of reports, you need to keep a skeptical attitude about individual reports with a p-value of 0.01. A small p-value is only one piece of evidence, not a proof.

Prob 14.05

One of the difficulties in interpreting p-values comes from what is sometimes called **publication bias**: only those studies with sufficiently low p-values are published; we never see studies on the same subject that didn't have low p-values.

We should interpret a p-value of, say, 0.03 differently in these two situations:

1. The study that generated the p-value was the only one that has been done on that subject. In this case, the p-value can reasonably be taken at face value.

2. The study that generated the p-value was one of 20 different studies but was the only one of the 20 studies that generated a p-value below 0.05. In this case, the p-value has little genuine meaning since, even if the null hypothesis were true we wouldn't be surprised to see a p-value like 0.03 from the "best" of 20 studies.

It's sometimes the case that "para-normal" phenomena — mind-reading, for instance — are subjected to studies and that the studies sometimes give low p-values. It's hard to interpret the p-value directly because of the publication bias effect. We can, of course, ask that the study be repeated, but if there are many repeats that we never hear about (because they happened to generate large p-values), it can still be difficult to interpret the p-value.

There is one way out, however. In addition to asking that the study be repeated, we can ask that the sample size be increased. The reason is that the larger sample size should generate much smaller p-values if the studied effect is genuine. However, if the effect is spurious, we'll still expect to see published p-values around 0.03.

To explore this, consider the following result of a ficticious study of the possible link between "positive thinking" (posThinking) and the t-cell count (tcells) in the immune system of 50 different people. The study has produced a "suggestive" p-value of about 0.08:

```
Model:    tcells ~ posThinking
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.6412     1.2506   2.912  0.00614
posThinking         0.2325     0.1293   1.798  0.07847
```

Here is a statement that can generate the p-value:

```
> 2*(1-pt(1.798,df=48))
[1] 0.07846795
```

Why is the the output of `pt` subtracted from 1?

A This is always the way `pt` is used.
B Because the t-value is positive.
C Because we want a one-tailed test.
D Because the t-value is less than 2.

The `df` in the statement stands for the degrees of freedom of the residual, just as in the F test. There were 50 cases and two model vectors (the intercept and the posthinking indicator vector), so 48 degrees of freedom in the residual.

Now imagine that the same study had been repeated with four times as much data: 200 cases. Assuming that everything else remained the same, how big would the standard error on posThinking be given the usual relationship between the size of the standard error and the number of cases n:

- Standard error with $n = 200$ cases:

  $\underline{0.52}$ $\underline{0.26}$ $\underline{0.13}$ $\underline{0.065}$ $\underline{0.032}$

- With $n = 200$ cases, how many degrees of freedom will there be in the residual?

  $\underline{48}$ $\underline{50}$ $\underline{98}$ $\underline{100}$ $\underline{198}$ $\underline{200}$

- Using the standard error with $n = 200$ cases, and assuming that the coefficient remains exactly as reported in the table, recalculate the p-value on `posThinking`. Select the one of these that is closest to the p-value:

  $\underline{0.4}$ $\underline{0.04}$ $\underline{0.004}$ $\underline{0.0004}$ $\underline{0.00004}$ $\underline{0.000004}$

Prob 14.06

The t- and F-distributions are closely related. That is, you can calculate the p-value on a t-statistic by using the F distribution, if you do it right.

Generate a large random sample from a t-distribution with 10 degrees of freedom. Then generate another random sample from the F-distribution with 1 and 10 degrees of freedom.

Graphically compare the distributions of the F and t values. What do you see?

- TRUE or FALSE They are the same.
- TRUE or FALSE The F distribution is much more skew to the right.
- TRUE or FALSE The t distribution is symmetrical around zero.

Now compare the distributions of the F values and the **square** of the t values. What do you see?

- TRUE or FALSE They are the same.

Prob 14.07

You know that you can compute the sample mean m and variance s^2 of a variable with simple calculations

$$m = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{or} \quad s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - m)^2$$

Such calculations are implemented with simple computer commands, e.g.

```
feet = fetchData("kidsfeet.csv")
mean( feet$length )

[1] 24.723

var( feet$length )
```

```
[1] 1.736
```

This is a fine way to calculate these quantities. But, just to show the link of these quantities to modeling, I ask you to do the following with the kidsfeet data: `kidsfeet.csv` :

1. Fit a model where one of the coefficients will be the mean of the length. Enter the model report below.
2. Perform ANOVA on the same model. One of the numbers in the ANOVA report will be the variance. Which one is it?
3. Based on the above, explain why the calculation of the sample variance involves dividing by $N-1$ rather than N.

Prob 14.08

Zebra mussels are a small, fast reproducing species of freshwater mussel native to the lakes of southeast Russia. They have accidentally been introduced in other areas, competing with native species and creating problems for people as they cover the undersides of docks and boats, clog water intakes and other underwater structures. Zebra mussels even attach themselves to other mussels, sometimes starving those mussels.

Zebra mussels growing on a larger, native mussel.

Ecologists Shirley Baker and Daniel Hornbach examined whether zebra mussels gain an advantage by attaching to other mussels rather than to rocks.[**?**] The ecologists collected samples of small rocks and *Amblema plicata* mussels, each of which had a collection of zebra mussels attached. The samples were transported to a laboratory where the group of mussels from each individual rock or *Amblema* were removed and placed in an aquarium equipped to measure oxygen uptake and ammonia excretion. After these physiological measurements were made, the biochemical composition of the mussel tissue was determined: the percentage of protein, lipid, carbohydrate, and ash.

Baker and Hornbach found that zebra mussels attached to *Amblema* had greater physiological activity than those attached to rocks as measured by oxygen uptake and ammonia excretion. But this appears to be a sign of extra effort for the *Amblema*-attached zebra mussels, since they had lower carbohydrate and lipid levels. In other words, attaching to *Amblema* appears to be disadvantageous to the zebra mussels compared to attaching to a rock.

In this problem, you will use the data collected by Baker and Hornbach to reprise their statistical analysis. The data are in the file `zebra-mussels.csv`.

```
  GroupID dry.mass count attachment  ammonia   O2
1       1     0.55    20       Rock    0.075 0.82
2       2     0.45    19       Rock    0.066 0.70
3       3     0.37    20       Rock    0.074 0.62
... and so on ...
28     28     0.89    28    Amblema    0.248 2.68
29     29     0.70    31    Amblema    0.258 2.26
30     30     0.68    64    Amblema    0.235 2.05
```

There are 30 cases altogether. The unit of analysis was not an individual zebra mussel but the group of zebra mussels that were attached to individual rocks or *Amblema*.

The variables are:

- dry.mass The mass of the dried tissue of the group, in grams.
- count How many mussels in the group.
- attachment The substrate to which the mussels in the group were attached: a rock or an *Amblema* mussel.
- O2 Oxygen uptake in mg per hour for the group.
- ammonia – Nitrogen excretion measured as ammonia in mg per hour for the group.
- lipid, protein, carbo, ash Biochemical makeup as percent of the dry weight for each of these.
- Kcal Total calorific value of the tissue in kilo-calories per gram.

A first question is whether the amount of mussel tissue is greater for the rock-attached or the *Amblema*-attached zebra mussels. Here is the report of a model:

```
> z = fetchData("zebra-mussels.csv")
> summary(lm(dry.mass ~ attachment, data=z))
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.6846     0.0554   12.36  1.2e-12
attachmentRock   -0.1983     0.0770   -2.58    0.016
```

Based on this report, which of these claims do data support?

- [A] Smaller mass for rock-attached mussels, as a group.
- [B] Larger mass for rock-attached mussels, as a group.
- [C] No statistically significant difference in mass between the rock- and *Amblema*-attached mussels.

The dry.mass variable gives the mass of the entire group of mussels. It might be that the mussels were systematically different in size on the two different substrates. One way to look at this is to compute the dry mass per individual mussel. This can be calculated by dividing dry.mass by count:

```
> z$ind.mass = z$dry.mass / z$count
```

Fit a model of ind.mass versus attachment. Which of the claims does this model support?

- [A] Smaller mass for rock-attached mussels, as a group.
- [B] Larger mass for rock-attached mussels, as a group.
- [C] No statistically significant difference in mass between the rock- and *Amblema*-attached mussels.

In summarizing their results about oxygen uptake and nitrogen excretion, Baker and Hornbach wrote:

> *Attachment substrate had significant effects on ammonia excretion and oxygen uptake rates. Dreissenids [that is, Zebra mussels] attached to Amblema had higher ammonia excretion rates ($p < 0.0001$) and higher oxygen uptake rates ($p < 0.0001$) compared to dreissenids attached to rocks.*

To check their statistics, fit these models for oxygen uptake per individual and ammonia excretion per individual:

```
> mod1 = lm( O2/count ~ attachment, data=z )
> mod2 = lm( ammonia/count ~ attachment, data=z )
```

The coefficients on the attachment variable indicate that:

- ammonia excretion is lower higher for rock-attached zebra mussels than for *Amblema*-attached mussels.
- oxygen uptake is lower higher for rock-attached than for *Amblema-attached.*
- TRUE or FALSE These results are consistent with Baker and Hornbach's claims.

Look at the p-values on the attachment variable. What are they?

- Ammonia: 0.01 0.04 0.10 0.25 0.40
- Oxygen: 0.01 0.04 0.10 0.25 0.40

Perhaps you are surprised to see that the p-values from your models are quite different from those reported by Baker and Hornbach. That's because Baker and Hornbach included ind.mass as a covariate in their models, in order to take into account how the metabolic rate might depend on the mass of the mussel.

Following Baker and Hornbach's procedure, fit these models:

```
> mod3 = lm( O2/count ~ ind.mass + attachment, data=z )
> mod4 = lm( ammonia/count ~ ind.mass + attachment, data=z )
```

You'll get different p-values for the attachment variable in these models depending on whether the analysis is done using ANOVA (which is often called "Analysis of Covariance" (ANCOVA) when there is a covariate in the model, or is done using a regression report. In ANCOVA, the p-value depends on whether ind.mass is put before or after attachment.

What is the regression report p-value on attachment?

- Oxygen 0.0000016 0.000016 0.0016 0.016 0.16
- Ammonia 0.0000005 0.0005 0.0001 0.001 0.01 0.1

The reason that including ind.mass as a covariate has improved the p-value is that ind.mass has soaked up the variance in the response variable. Show an ANOVA report for one of the two models mod3 or mod4 and explain what in the report demonstrates that ind.mass is soaking up variance.

In contrast to the dramatic effect of ind.mass on the p-values, including it as a covariate does not seem to affect much the coefficient on attachment. What is it about ind.mass that explains this?

- A ind.mass is almost entirely uncorrelated with the response variable.
- B ind.mass is almost entirely uncorrelated with attachment.
- C ind.mass is strongly correlated with the response variable.
- D ind.mass is strongly correlated with attachment.

Prob 14.09

The table shows a brief data set that we will use to trace through a simple one-way ANOVA calculation by hand.

Age	Group
18	Student
22	Student
45	Parent
51	Parent
35	Professor
57	Professor

Our model will be Age $\sim 1 +$ Group

The ANOVA report breaks down the model into a series of nested models. In this case, the series is simple:

1. Age ~ 1
2. Age $\sim 1 +$ Group

For each of these nested models, you need to calculate the "degrees of freedom" and the sum of squares of the fitted model values. The degrees of freedom is the number of model vectors involved. The sum of squares is found by fitting the model and adding up the squares of the fitted values.

We will do this by hand. First, fit Age ~ 1 and enter the fitted values into the table. "Fitting" is easy here, since the coefficient on Age ~ 1 is just the grand mean of the Age variable.

For Age ~ 1

Age	Group	Fitted
18	Student	
22	Student	
45	Parent	
51	Parent	
35	Professor	
57	Professor	

Once you have written down the fitted values for each case, compute the sum of squares. Since there is one model vector, there is just one degree of freedom for this model.

Next, fit the model Age $\sim 1 +$ Group and enter the fitted values into the table. This model is also easy to fit, since the fitted values are just the groupwise means.

For Age $\sim$ 1 + Group

Age	Group	Fitted
18	Student	
22	Student	
45	Parent	
51	Parent	
35	Professor	
57	Professor	

Compute the sum of squares for this model. Since there are three groups, there are three model vectors and hence three degrees of freedom.

Finally, compute the sum of squares of the Age variable itself. You might like to think of this as the sum of squares of the fitted values to a "perfect" model, a model that captures all of the variability in Age. We know that such a model can always be constructed so long as we allow N model vectors, even if they are junk, so the degrees of freedom ascribed to this "perfect" model is N. (We put "perfect" in quotation marks to emphasize that this model is perfect only in the sense that the residuals are zero. That can happen anytime we have as many model vectors as cases, that is, when $m = N$.)

Enter the degrees of freedom and the sums of squares of the fitted values into the table below:

Model	D.F.	Sum of Squares of Fitted
Age $\sim$ 1		
Age $\sim$ 1 + Group		
"Perfect" model		

We are almost at the point where we can construct the ANOVA table. At the heart of the ANOVA table is the degree-of-freedom and sum-of-squares information from the above table. **But** rather than entering the sum of squares directly, we subtract from each of the sum of squares the total of all the sums of squares of the models that appear above it. That is, we give each successive nested model credit only for the amount that it *increased* the sum of squares from what it had been in the previous model. A similar accounting is done for the degrees of freedom. In the following table, we'll mark the header as Δ to emphasize that you should the *change* in sum of squares and the *change* in degrees in freedom from the previous model.

Fill in the first two columns of the table. Then go back and fill in the "mean square" column, which is just the sum of squares divided by the degrees of freedom. Similarly, fill in the F column, which is the mean square for each model divided by the mean square for the perfect model.

Model	Δ D.F.	Δ S.S.	M.S.	F ratio	p-value
Age $\sim$ 1					
Age $\sim$ 1 + Group					
"Perfect" model					

You can use software to compute the p-value. The relevant parameters of the F distribution are the degrees of freedom of the model and the degrees of freedom of the "perfect" model.

Of course, in a real ANOVA table, rows are labeled differently. Each row gives a name to the terms that were added in making the model. So, in the above table, the labels will be "Intercept," "Group," and "Residuals."

Compare your table to one constructed with software. Enter the data into a spreadsheet, save it to disk in an appropriate format (e.g., CSV), then read it into the statistical software and create the ANOVA table.

Prob 14.10

The special program `rand()` generates random vectors for use in modeling. It's special because it works only within the `lm` command. For example, suppose we want to create three random model vectors along with an intercept term to model the kids' foot width data:

```
> lm( width ~ rand(3), data=kids)
Coefficients:
(Intercept)     rand(3)1     rand(3)2     rand(3)3
    9.00838      0.01648     -0.05185      0.01627
> lm( width ~ rand(3), data=kids)
(Intercept)     rand(3)1     rand(3)2     rand(3)3
    8.99367     -0.09795     -0.06916      0.05676
```

The coefficients are different in the two models because the "explanatory" vectors are random.

We're going to study the R^2 due to such collections of random vectors. This can be calculated with the `r.squared` program:

```
> r.squared(lm( width ~ rand(3), data=kids))
[1] 0.1770687
> r.squared(lm( width ~ rand(3), data=kids))
[1] 0.03972449
```

Note that the R^2 values vary from trial to trial, because the vectors are random.

According to the principle of equipartition, **on average**, each random vector should contribute $1/(n-1)$ to the R^2 and the effect of multiple vectors is additive. So, for example, with $n = 39$ cases, the three random vectors in the above example should result in an R^2 that is near $3/(39-1) = 0.08$.

Repeat the above calculation of R^2 many times for $p = 3$ random vectors and find the mean of the resulting R^2. You can do the repetitions 100 times with a statement like this:

```
> samps=do(100)*r.squared(lm(width~rand(3),data=kids))
```

Now do the same thing for $p = 1, 3, 10, 20, 37$ and 38. Are the results consistent with the theory that, on average, R^2 should be $p/(n-1)$?

Enter all your results in the table:

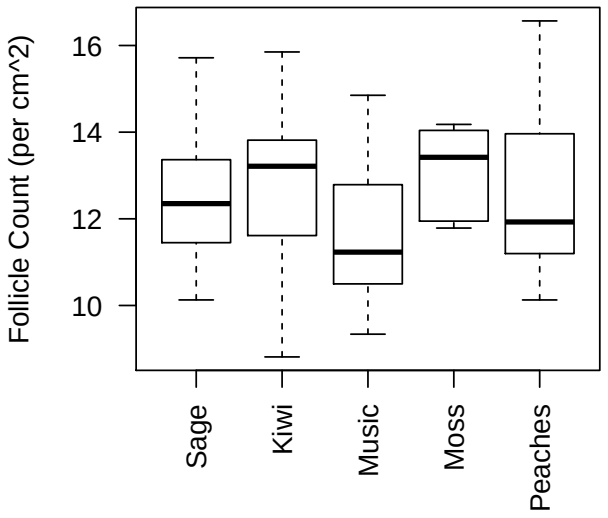
p	$p/(n-1)$	Mean R^2
1		
3		
10		
20		
37		
38		

Note that for $p = 38$ all of the trials produced the same R^2 value. Explain why.

Prob 14.11

Suppose that you have worked hard to carry out an experiment about baldness that involves several different treatments: rubbing various herbs and spices on the scalp, playing hairy music, drinking kiwi juice, laying fresh moss on the head, eating peaches. Then, you measure the density of hair follicles after 2 months.

You plot out your data:



It looks like the music treatment produced the lowest number of follicles, while moss produced the highest. (Kiwi is a close second, but the range is so large that you aren't sure about it.)

You decide to extract the Music and Moss groups and build a model of follicle density versus treatment. (This sort of test, comparing the means of two groups, is called a t-test.)

```
dd = subset(d, group=='Music' | group=="Moss")
summary(lm(val ~ group, data=dd))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	93.6316	3.5882	26.09	0.0000
groupMoss	11.1991	5.0745	2.21	0.0405

Aha! $p = 0.04$. You have evidence that such treatments can have an effect.

Or do you? Admittedly, a p-value of 0.04 will not be terribly compelling to someone who doubts that music and moss affect follicle density in different ways. But it is below the conventional threshold of 0.05. Consider some of the things that could be wrong:

- What's the null hypothesis in the hypothesis test presented above?
  - A That Moss and Music are no different than a Control group.
  - B That Moss and Music are the same as each other.
  - C That Moss and Music have no effect.
- Assuming that the null hypothesis is correct, what's the probability of seeing a p-value as small or smaller than the one actually observed?
  0 0.01 0.04 0.05 0.10 0.50 0.94 0.99
- You chose the two groups to compare based on the graph. There were many other possible pairs of groups: Sage vs Peaches, Kiwi vs Moss, etc. Altogether, there are $5 * 4/2 = 10$ possible pairs of groups. (In general, when there are k different groups, there $k(k-1)/2$ possible pairs of groups. So if $k = 10$ there are 45 different possible pairs.)

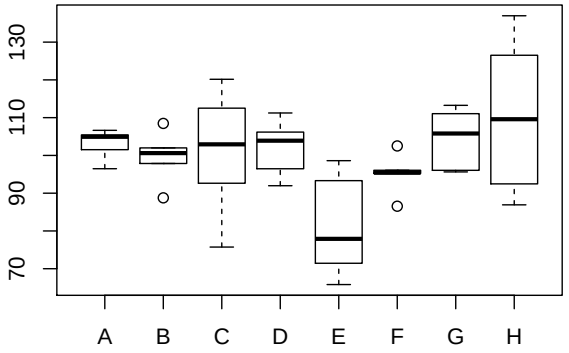
  The Bonferroni correction calls for the p-value to be adjusted based on the number of comparisons done. There are two, equivalent ways to do it. One way is to divide the conventional threshold by the number of comparisons and use the new threshold. So for this test, with 10 pairs of groups, the threshold for rejecting the null would be $0.05/10 = 0.005$. The other way is just to multiply the p-value by the number of comparisons and then compare this to the standard threshold.

  Using the "multiply by" approach to the Bonferroni correction, what is the p-value value on the Music-vs-Moss comparison?
  0.005 0.01 0.04 0.05 0.20 0.40 0.80

Prob 14.12

This exercise concerns assessing the appropriateness of the Bonferroni correction in a setting where you are choosing one pair out of several possible groups to compare to each other.

Recall that the Bonferroni correction relates to a situation where you are conducting k hypothesis tests and choosing the p-value that is most favorable. For example, you might have tested several different treatments: B, C, D, ... against a control, A. An example is shown in the figure.



Group E seems to stand out as different from the control, so you decide to focus your attention on comparing the control A to group E.

To generate the data in the graph, use this command:

```
d = run.sim(bogus.groups, ngroups=8, seed=230)
```

When you do this, your results should match these exactly (subject only to round-off):

```
head(d)
```

```
  group     val
1     A  96.477
2     A 101.519
3     A 104.930
4     A 105.426
5     A 106.640
6     B 100.612
```

```
tail(d)
```

```
   group     val
35     G  96.053
36     H  86.914
37     H  92.466
38     H 109.589
39     H 126.523
40     H 136.980
```

As you might guess from the name of the simulation, `bogus.groups`, the data are being sampled from a population in which the assigned `group` is unrelated to the quantitative variable `val`. That is, the null hypothesis is true.

The point of the `seed=230` statement is to make sure that exactly the same "random" numbers are generated as in the example.

Using the data, extract the subset of data for groups A and E, and carry out a hypothesis test on `group` using the model `val` $\sim$ `group`. (Hint: in the `subset` command, use the selector variable `group=="A''|group=="E''` to select cases that are from either group A *or* group E. Then apply `lm` to this subset of data.)

- What is the p-value that compares groups A to E.
  0.011 0.032 0.057 0.128 0.193 0.253

As a convenience, you are provided with a special-purpose function, `compare.many.groups` that does each of the hypothesis tests comparing the control group to another group. You can run it this way:

```
compare.many.groups( d, control="A")
```

```
  group1 group2     diff   pvalue
1      A      B  -3.4563 0.377303
2      A      C  -2.2038 0.789874
3      A      D  -1.0472 0.795413
4      A      E -21.5810 0.011020
5      A      F  -7.8415 0.036872
6      A      G   1.3614 0.749352
7      A      H   7.4958 0.464852
```

For the simulated data, your results should match exactly (subject to round-off). Confirm that the p-value from this report matches the one you got by hand above.

- As it happens, the p-value for group F is also below the 0.05 threshold. What is that p-value?
  0.021 0.031 0.037 0.042 0.049

The Bonferroni correction takes the "raw" p-value and corrects it by multiplying by the number of comparisons. Since there are 7 comparisons altogether (groups B, C, ..., H against group A), multiply the p-value by 7.

- What is the Bonferroni-corrected p-value for the A vs E comparison?
  0.011 0.037 0.077 0.11 0.37 0.77

Now you are going to do a simulation to test whether the Bonferroni correction is appropriate. The idea is to generate data from a simulation that implements the null hypothesis, calculate the p-value for the "best looking" comparison between the control group A and one of the other groups, and see how often the p-value is less than 0.05. It will turn out that the p-value on the best looking comparison will be below 0.05 much too often. (If the p-value is to be taken at face value, when the null hypothesis is true then $p < 0.05$ should happen only 5% of the time.) Then you'll apply the Bonferroni correction and see if this fixes things.

Because this is a somewhat elaborate computation, the following leads you through the development of the R statement that will generate the results needed. Each step is a slight elaboration on the previous one.

1. Generate one realization from the `bogus.groups` simulation and find the A vs other group p-values. In this example, you will use `seed=111` in the first few steps so that you can compare your output directly to what's printed here. This can be done in two statements, like this:

   ```
   d = run.sim( bogus.groups, ngroups=8,seed=111 )
   compare.many.groups( d, control="A")
   ```

   ```
     group1 group2     diff   pvalue
   1      A      B  -2.78824 0.73112
   2      A      C  16.01192 0.11136
   3      A      D   0.11785 0.98931
   4      A      E  11.83709 0.12047
   5      A      F   9.73525 0.49863
   6      A      G -10.98126 0.43226
   7      A      H   6.99226 0.46696
   ```

2. It's helpful to combine the two statements above into a single statement. That will make it easier to automate later on. To do this, just put the statement that created `d` in place of `d`, as in the following. (The statement is so long that it's spread over 3 lines. You can cut it out of this page and into an R session, if you wish.)

   ```
   compare.many.groups(
              run.sim( bogus.groups, ngroups=8,seed=11
                       control="A")
   ```

   ```
     group1 group2     diff   pvalue
   1      A      B  -2.78824 0.73112
   2      A      C  16.01192 0.11136
   3      A      D   0.11785 0.98931
   4      A      E  11.83709 0.12047
   5      A      F   9.73525 0.49863
   6      A      G -10.98126 0.43226
   7      A      H   6.99226 0.46696
   ```

It happens that none of the p-values from this simulation were < 0.05.

3. Extend the statement above to extract the smallest p-value. This involves prefacing the statement with `min(` and following it with `$pvalue)`

```
min(compare.many.groups(
       run.sim( bogus.groups, ngroups=8,seed=111 ),
              control="A")$pvalue)
```

```
[1] 0.11136
```

4. Now, run this statement a dozen times, using `do(12)*`. Assign the results to an object `s`.

```
s = do(12)*min(compare.many.groups(
          run.sim( bogus.groups, ngroups=8,seed=111 ),
                 control="A")$pvalue)
s
```

```
 [1] 0.11136 0.11136 0.11136 0.11136 0.11136
 [6] 0.11136 0.11136 0.11136 0.11136 0.11136
[11] 0.11136 0.11136
```

Don't be too surprised by the result. You got exactly the same answer on each trial because the same random "seed" was used every time.

5. Delete the random seed, so that new random numbers will be generated on each trial.

```
s = do(12)*min(compare.many.groups(
           run.sim( bogus.groups, ngroups=8 ),
                 control="A")$pvalue)
s
```

```
 [1] 0.204602 0.052472 0.013789 0.037551
 [5] 0.035872 0.261042 0.133712 0.080999
 [9] 0.033976 0.167735 0.431789 0.166764
```

6. The point of running a dozen trials was just to produce a manageable amount of output for you to read. Now generate a thousand trials and, storing the results in `s`. It will take about a minute for the computer to complete the calculation, more on older computers.

```
s = do(1000)*min(compare.many.groups(
            run.sim( bogus.groups, ngroups=8 ),
                 control="A")$pvalue)
```

Then answer the following questions:

(a) About what fraction of the "raw" (that is, uncorrected) p-values are below 0.05? (Pick the closest answer.)
0.01 0.02 0.05 0.10 0.20 0.30 0.40 0.50

(b) Now calculate the Bonferroni correction. Remember that there were 7 comparisons done in each trial, of which the one resulting in the smallest p-value was selected. What fraction of trials led to $p < .05$? (Pick the closest answer.)
0.01 0.02 0.05 0.10 0.20 0.30 0.40 0.50

Prob 14.13

You and your statistics-class friends are having a disagreement. Does the result of a hypothesis test depend on the units in which quantities are measured? For example, in the kids' feet data, `kidsfeet.csv` the length and width are measured in cm. Would it matter if they had been measured in inches or meters?

Do a numerical experiment to figure out whether the units make a difference. Explain your strategy concisely and give the corresponding commands and the results here.

Chapter Fifteen Reading Questions

- What is a "covariate" and how does it differ from any other kind of variable?
- Why is there a separate F statistic for each explanatory term in a model?
- How can covariates make an explanatory term look better (e.g., more significant) in an F test?
- How can covariates make an explanatory term look worse (e.g., less significant) in an F test?
- Why does the the sum of squares of the various model terms change when there is collinearity among the model terms?

Prob 15.01

Often we are interested in whether two groups are different. For example, we might ask if girls have a different mean footlength than do boys. We can answer this question by constructing a suitable model.

```
> kids = fetchData("kidsfeet.csv")
> summary( lm( length ~ sex, data=kids ) )
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  25.1050     0.2847  88.180   <2e-16
sexG         -0.7839     0.4079  -1.922   0.0623
```

Interpret this report, keeping in mind that the foot length is reported in centimeters. (The reported value `<2e-16` means $p < 2 \times 10^{-16}$.)

1. What is the point estimate of the difference between the lengths of boys and girls feet.
   - [A] Girls' feet are, on average, 25 centimeters long.
   - [B] Girls' feet are 0.4079 cm shorter than boys'.
   - [C] Girls' feet are 0.7839 cm shorter than boys'.
   - [D] Girls' feet are 1.922 cm shorter than boys'.

2. The confidence interval can be written as a point estimate plus-or-minus a margin of error: $P \pm M$. What is the 95% margin of error, M, on the difference between boy's and girl's foot lengths. -0.78 0.28 0.41 0.60 0.80

3. What is the Null Hypothesis being tested by the reported p-value 0.0623?

   A Boys' feet are, on average, longer than girls' feet.
   B Girls' feet are, on average, shorter than boys' feet.
   C All boys' feet are longer than all girls' feet.
   D No girl's foot is shorter than all boys' feet.
   E There is no difference, on average, between boys' footlengths and girls' footlengths.

4. What is the Null Hypothesis being tested by the p-value on the intercept?

   A Boys' and girls' feet are, on average, the same length
   B The length of kids' feet is, on average, zero.
   C The length of boys' feet is, on average, zero.
   D The length of girls' feet is, on average, zero.
   E Girls' and boys' feet don't intercept.

Here is the report from a related, but slightly different model:

```
> summary( lm( length~sex-1, data=kids ))
Coefficients:
     Estimate Std. Error t value Pr(>|t|)
sexB  25.1050     0.2847   88.18   <2e-16
sexG  24.3211     0.2921   83.26   <2e-16
```

Note that the p-values for both coefficients are practically zero, $p < 2 \times 10^{-16}$.

What is the Null Hypothesis tested by the p-value on `sexG`?

A Girls' feet have a different length, on average, than boys'.
B Girls' feet are no different in length, on average, than boys'.
C Girls' footlengths are, on average, zero.
D Girls' footlengths are, on average, greater than zero.

Prob 15.02

Here is an ANOVA table (with the "intercept" term included) from a fictional study of scores assigned to various flavors, textures, densities, and chunkiness of ice cream. Some of the values in the table have been left out. Figure out from the rest of the table what they should be.

	Df	Sum-Sq	Mean-Sq	F-value	p-value
(intercept)	1	_A_	200	_B_	_C_
flavor	8	640	80	_D_	0.134
density	_E_	100	100	2	0.160
fat-content	1	300	_F_	6	0.015
chunky	1	200	200	4	0.048
Residuals	100	5000	50		

(a) The value of A:
1 2 100 200 400 600

(b) The value of B:
1 2 3 4 5 6 7 8 10 20 200

(c) The value of C: (Hint: There's enough information in the table to find this.)
0.00 0.015 0.030 0.048 0.096 0.134 0.160 0.320 0.480

(d) The value of D:
0.0 0.8 1.6 3.2 4.8

(e) The value of E:
0 1 2 3 4 5 6

(f) The value of F:
100 200 300 400 500

(g) How many cases are involved altogether?
50 100 111 112 200 5000

(h) How many different flavors were tested?
1 3 5 8 9 10 12 100

Prob 15.03

Consider the following analysis of the kids' feet data looking for a relationship between foot width and whether the child is left or right handed. The variable `domhand` gives the handedness, either `L` or `R`. We'll construct the model in two different ways. There are 39 cases altogether.

```
> anova( lm(width ~ domhand, data=kids))
Response: width
            Df  Sum Sq Mean Sq    F value Pr(>F)
(Intercept)  1 3153.60 3153.60 12228.0064 <2e-16 ***
domhand      1    0.33    0.33     1.2617 0.2686
Residuals   37    9.54    0.26

> anova( lm(width ~ domhand - 1, data=kids))
Response: width
          Df  Sum Sq Mean Sq F value    Pr(>F)
domhand    2 3153.93 1576.96  6114.6 < 2.2e-16 ***
Residuals 37    9.54    0.26
```

- Explain why, in the first case, the p-value is not significant, but in the second case it is.

- Why does domhand have 1 degree of freedom in the first ANOVA report, but 2 degrees of freedom in the second?

Prob 15.05

A statistics student wrote:

I'm interested in the publishing business, particularly magazines, and thought I would try a statistical analysis of some of the features of magazines. I looked at several different magazines, and recorded several variables, some of which I could measure from a single copy and some of which I deduced from my knowledge of the publishing business.

magazine a number to identify the magazine

pages the number of pages in a typical issue

color the number of pages with a color picture

age the age group of the target audience

sex the sex of the intended audience

sentenceLength the average number of words in a sentence in the articles in the magazine.

Most people find it hard to believe, but most mass-market magazines are very deliberately written and composed graphically to be attractive to the target audience. The distinctive "styles" of magazines is no accident.

I was interested to see if there is a relation between the average sentence length and any of the other variables. I made one linear model and had a look at the ANOVA table, as shown below.

```
Analysis of Variance Table

Response: sentenceLength
           Df Sum Sq Mean Sq F value  Pr(>F)
sex         1  0.222   0.222  0.0717 0.80626
age         3 71.067  23.689  7.6407   ????
color       1  0.299   0.299  0.0964 0.77647
Residuals   3  9.301   3.100
```

Answer each question based on the information given above.

1. What model structure did I use to generate this table?
   - A `sentenceLength ~ age + sex + color`
   - B `sentenceLength ~ sex * age + color`
   - C `sentenceLength ~ sex + age + color`
   - D `color ~ sentenceLength + sex + age`

2. How many cases are there altogether?
   - A 8
   - B 9
   - C 10
   - D No way to tell from the information given.

3. Is the variable `age` categorical or quantitative?
   - A categorical
   - B quantitative
   - C Could be either.
   - D Can't know for sure from the data given.

4. The p-value for `age` is missing. What should this value be?
   - A 0.93553 from `pf(7.6407, 3, 3)`
   - B 0.06446 from `1-pf(7.6407, 3, 3)`
   - C 0.98633 from `pf(23.689, 3, 3)`
   - D 0.01367 from `1-pf(23.689, 3, 3)`
   - E 0.99902 from `pnorm(23.689, 0, 7.6507)`
   - F 0.00098 from `1-pnorm(23.689, 0, 7.6507)`

5. Based on the ANOVA table, what null hypothesis can be rejected in this study at the significance level $p < 0.10$?
   - A An average sentence has zero words.
   - B There is no relationship between the number of color pages and the sex of the intended audience.
   - C The number of color pages is not related to the sentence length.
   - D There is no relation between the average number of words per sentence in an article and the age group that the magazine is intended for, after taking sex into account.
   - E None of the above, because there is a different null hypothesis corresponding to each model term in the ANOVA report.

6. I really wanted to look to see if there is an interaction between `sex` and `age`. What would be the point of including this term in the model?
   - A To see if the different sexes have a different distribution of age groups.
   - B To see if there is a difference in average sentence length between magazines for females and males.
   - C To see if magazines for different age groups are targeted to different sexes.
   - D To see if the difference in average sentence length between magazines for females and males changes from one age group to another.

7. I tried to include an interaction between `sex` and `age` into the model. This didn't work out. Just using the information in the printed ANOVA report, judge what might have gone wrong.
   - A The term was included as the last term in the ANOVA report and didn't have a significant sum of squares.
   - B I discovered that `sex` and `age` were redundant.
   - C The p-values disappeared from the report.
   - D None of the above.

Prob 15.07

P-values concern the "statistical significance" of evidence for a relationship. This can be a different thing from the real-world importance of the observed relationship. It's possible for a weak connection to be strongly statistically significant (if there is a lot of data for it) and for a strong relationship to lack statistical significance (if there is not much data).

Consider the data on the times it took runners to complete the Cherry Blossom ten-mile race in March 2005:

```
> run = fetchData("ten-mile-race.csv")
> names(run)
[1] "state" "time"  "net"   "age"   "sex"
```

Consider the `net` variable, which gives the time it took the runners to get from the start line to the finish line.

Answer each of the following questions, giving both a quantitative argument and also an everyday English explanation. Assessing statistical significance is a technical matter,

but to interpret the substance of a relationship, you will have to put it in a real-world context.

1. What is the relationship between net running time and the runner's age? Is the relationship significant? Is it substantial?

2. What is the relationship between net running time and the runner's sex? Is the relationship significant? Is it substantial?

3. Is there an interaction between sex and age? Is the relationship significant? Is it substantial?

Prob 15.09

You are conducting an experiment of a treatment for balding. You measure the hair follicle density before treatment and again after treatment. The data table has the following variables (with a few examples shown):

Subject.ID	follicle.density	when	sex
A59	7.3	before	M
A59	7.9	after	M
A60	61.2	before	F
A60	61.4	after	F
and so on, 100 entries altogether			

1. Here is an ANOVA table for a model of these data:

```
> anova(lm(follicle.density~when,data=hair))
           Df  Sum Sq Mean Sq F value  p
when        1    33.7   33.7  0.157 0.693
Residuals  98 21077.5  215.1
```

   Does this table suggest that the treatment makes a difference? Why or why not?

2. Here's another ANOVA table

```
> anova(lm(follicle.density~when+Subject.ID,
+  data=hair))
           Df  Sum Sq Mean Sq  F value p
when        1    33.7   33.7  14.9  0.0002
Subject.ID 49 20858.6  425.7 185.0  zero
Residuals  97   218.9    2.3
```

   Why is the F-value on when different in this model than in the previous one?

3. What overall conclusion do you draw about the effectiveness of the treatment? Is the effect of the treatment statistically significant? Is it significant in practice?

Prob 15.14

During a conversation about college admissions, a group of high-school students starts to wonder how reliable the SAT score is, that is, how much an **individual student's** score could be expected vary just do to random factors such as the day on which the test was taken, the student's mood, the specific questions on the test, etc. This variation within an individual is quite different from the variation from person to person.

The high-school students decide to study the issue. A simple way to do this is to have one student take the SAT test several times and examine the variability in the student's scores. But, it would be better to do this for many different students. To this end, the students propose to pool together their scores from the times they took the SAT, producing a data frame that looks like this:

Student	Score	Sex	Order
PersonA	2110	F	1
PersonB	1950	M	1
PersonC	2080	F	1
PersonA	2090	F	2
PersonA	2150	F	3
... and so on			

The order variable indicates how many times the student has taken the test. 1 means that it is the student's first time, 2 the second time, and so on.

One student suggests that they simply take the standard deviation of the score variable to measure the variability in the SAT score. What's wrong with this for the purpose the students have in mind?

A There's nothing wrong with it.
B Standard deviations don't measure random variability.
C It would confound variability between students with variability.

Another student suggests looking at the sum of square residuals from the model score $\sim$ student. What's wrong with this:

A There's nothing wrong with it.
B It's the coefficients on student that are important.
C Better to look at the mean square residual.

The students' statistics teacher points out that the model score $\sim$ student will exactly capture the score of any student who takes the SAT only once; the residuals for those students will be exactly zero. Explain why this isn't a problem, given the purpose for which the model is being constructed.

Still another student suggests the model score $\sim$ student + order in order to adjust for the possibility that scores change with experience, and not just at random. The group likes this idea and starts to elaborate on it. They make two main suggestions:

- Elaboration 1: score $\sim$ student + order + sex
- Elaboration 2: score $\sim$ student + order + student:order

Why not include sex as an additional covariate, as in Elaboration 1, to take into account the possibility that males and females might have systematically different scores.

A It's a good idea.
B Bad idea since probably a person's sex has nothing to do with his or her score.
C Useless, since sex is redundant with student.

Regarding Elaboration 2, which of the following statements is correct?

1. TRUE or FALSE It allows the model to capture how the change of score with experience itself might be different from one person to another.
2. TRUE or FALSE It assumes that all the students in the data frame are taking the test multiple times.
3. TRUE or FALSE With the interaction term in place, the model would capture the exact scores of all students who took the SAT just once or twice, so the mean square residual would reflect only those students who took the SAT three times or more.

Prob 15.15

In conducting a hypothesis test, we need to specify two things:

- A Null Hypothesis
- A Test Statistic

The numerical output of a hypothesis test is a p-value.

In modeling, a sensible Null Hypothesis is that one or more explanatory variables are unrelated to the response variable. We can simulate a situation in which this Null applies by shuffling the variables. For example, here are two trials of a simulation of the Null in a model of the kidsfeet data:

```
> kids = fetchData("kidsfeet.csv")
> lm( width ~ length + shuffle(sex), data=kids)
Coefficients:
(Intercept)      length  shuffle(sex)G
    3.14406     0.23828       -0.07585
> lm( width ~ length + shuffle(sex), data=kids)
Coefficients:
(Intercept)      length  shuffle(sex)G
    2.74975     0.25106        0.08668
```

The test statistic summarizes the situation. There are several possibilities, but here we will use R^2 from the model since this gives an indication of the quality of the model.

```
> r.squared(lm( width ~ length + shuffle(sex), data=kids))
[1] 0.4572837
> r.squared(lm( width ~ length + shuffle(sex), data=kids))
[1] 0.4175377
> r.squared(lm( width ~ length + shuffle(sex), data=kids))
[1] 0.4148968
```

By computing many such trials, we construct the sampling distribution under the Null — that is, the sampling distribution of the test statistic in the world in which the Null holds true. We can automate this process using do:

```
> samps = do(1000) *
+   r.squared(lm( width ~ length + shuffle(sex), data=kids
```

Finally, to compute the p-value, we need to compute the test statistic on the model fitted to the actual data, not on the simulation.

```
> r.squared( lm( width ~ length + sex, data=kids))
[1] 0.4595428
```

The p-value is the probability of seeing a value of the test statistic from the Null Hypothesis simulation that is more extreme than our actual value. The meaning of "more extreme" depends on what the test statistic is. In this example, since a better fitting model will always have a larger R^2 we check the probability of getting a larger R^2 squares from our simulation than from the actual data.

```
> table( samps >= 0.4595428)

FALSE  TRUE
  912    88
```

Our p-value is about 9%.

Here are various computer modeling statements that implement possible Null Hypotheses. Connect each computer statement to the corresponding Null.

1. `lm(width ~ length + shuffle(sex),data=kids)`
2. `lm(width ~ shuffle(length) + shuffle(sex), data=kids)`
3. `lm(width ~ shuffle(length), data=kids)`
4. `lm(width ~ shuffle(sex), data=kids)`
5. `lm(width ~ length + sex, data=shuffle(kids))`

- Foot width is unrelated to foot length or to sex.
  1 2 3 4 5
- Foot width is unrelated to sex, but it is related to foot length.
  1 2 3 4 5
- Foot width is unrelated to sex, and we won't consider any possible relationship to foot length.
  1 2 3 4 5
- Foot width is unrelated to foot length, and we won't consider any possible relationship to sex.
  1 2 3 4 5
- This isn't a hypothesis test; the randomization won't change anything from the original data.
  1 2 3 4 5

Prob 15.16

I'm interested in studying the length of gestation as a function of the ages of the mother and the father. In the gestation data set, (`gestation.csv`) the variable `age` records the mother's age in years, and `dage` gives the father's age in years. The variable `gestation` is the length of the gestation in days. I hypothesize that the older the mother and father, the shorter the gestational period. So, I fit a model to those 599 cases where all the relevant data were recorded:

```
> summary(lm( gestation ~ age+dage, data=b))
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 282.62201    3.25821  86.741   <2e-16
age          -0.21313    0.19947  -1.068    0.286
dage          0.06085    0.17372   0.350    0.726
```

1. Describe in everyday language the relationship between age and gestation indicated by this model.

2. I note that the two p-values are nothing great. But I wonder whether if I treated mother's and father's age together — lumping them together into a single term with two degrees of freedom — I might not get something significant. Using the ANOVA reports given below, explain how you might come up with a single p-value summarizing the joint contribution of mother's and father's age. Insofar as you can, try to calculate the p-value itself.

```
> anova( lm(gestation ~ age+dage, data=b))
             Df Sum Sq Mean Sq F value Pr(>F)
age           1    486     486  1.9091 0.1676
dage          1     31      31  0.1227 0.7262
Residuals   596 151758     255
```

```
> anova( lm( gestation ~ dage+age, data=b))
             Df Sum Sq Mean Sq F value Pr(>F)
dage          1    227     227  0.8903 0.3458
age           1    291     291  1.1416 0.2858
Residuals   596 151758     255
```

Chapter Sixteen Reading Questions

- What's the difference between a "link value" and a "probability value" in a logistic regression model? How are they related to one another?

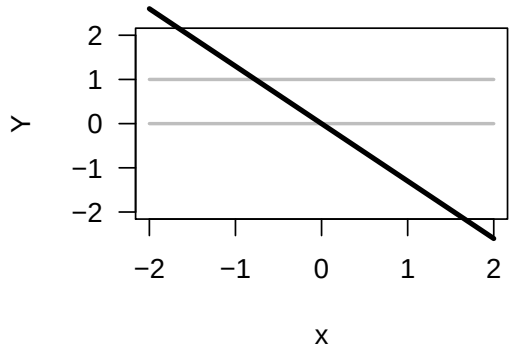
- How does the logistic function serve to keep the fitted probability values always within the range 0 to 1?

- What is maximum likelihood and how is it used as a criterion to fit a model?

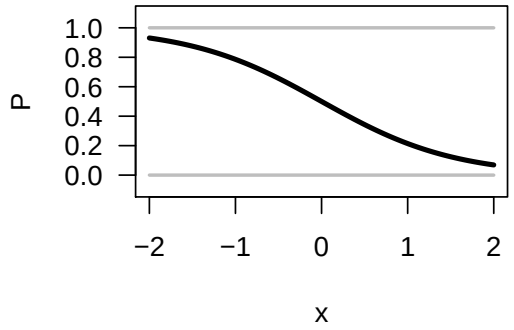
Prob 16.01

The graphs show the link values and the corresponding probability values for a logistic model where x is the explanatory variable.

- Link Values



- Probability Values



Use the graphs to look up answers to the following. Choose the closest possibility to what you see in the graphs.

- At what value of x is the link value 0?
  -2 -1 0 1 2
- What probability corresponds to a link of 0?
  0.0 0.1 0.5 0.9 1.0
- At what value of x is the link value 1?
  -1.50 -0.75 0.00 1.25 1.75
- What probability corresponds to a link of 1?
  0.0 0.25 0.50 0.75 1.00
- What probability corresponds to a link of -1?
  0.0 0.25 0.50 0.75 1.00
- What probability corresponds to a link of ∞? (This isn't on the graph.)
  0.0 0.25 0.50 0.75 1.00
- What probability corresponds to a link of $-\infty$? (This isn't on the graph.)
  0.0 0.25 0.50 0.75 1.00

Prob 16.02

The NASA space shuttle Challenger had a catastrophic accident during launch on January 28, 1986. Photographic evidence from the launch showed that the accident resulted from a plume of hot flame from the side of one of the booster rockets which cut into the main fuel tank. US President Reagan appointed a commission to investigate the accident. The commission concluded that the jet was due to the failure of an O-ring gasket between segments of the booster rocket.

A NASA photograph showing the plume of flame from the side of the booster rocket during the Challenger launch.

An important issue for the commission was whether the accident was avoidable. Attention focused on the fact that the ground temperature at the time of launch was 31°F, much lower than for any previous launch. Commission member and Nobel laureate physicist Richard Feynman famously demonstrated, using a glass of ice water and a C-clamp, that the O-rings were very inflexible when cold. But did the data available to NASA **before the launch** indicate a high risk of an O-ring failure?

Here is the information available at the time of Challenger's launch from the previous shuttle launches:

Flight	Temp	Damage	Flight	Temp	Damage
STS-1	66	no	STS-2	70	yes
STS-3	69	no	STS-4	80	NA
STS-5	68	no	STS-6	67	no
STS-7	72	no	STS-8	73	no
STS-9	70	no	STS 41-B	57	yes
STS 41-C	63	yes	STS 41-D	70	yes
STS 41-G	78	no	STS 51-A	67	no
STS 51-B	75	no	STS 51-C	53	yes
STS 51-D	67	no	STS 51-F	81	no
STS 51-G	70	no	STS 51-I	76	no
STS 51-J	79	no	STS 61-A	75	yes
STS 61-B	76	no	STS 61-C	58	yes

Using these data, you can fit a logistic model to estimate the probability of failure at any temperature.

```
> mod = glm(Damage ~ Temp, family='binomial')
> summary(mod)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  15.0429     7.3786   2.039   0.0415
Temp         -0.2322     0.1082  -2.145   0.0320
```

Use the coefficients to find the *link* value for these launch temperatures:

- 70°F (a typical launch temperature)
  -2.4 -1.2 1.6 2.7 4.3 7.8 9.4
- 53°F (the previous low temperature)
  -2.4 -1.2 1.6 2.7 4.3 7.8 9.4
- 31°F (the Challenger temperature)
  -2.4 -1.2 1.6 2.7 4.3 7.8 9.4

Convert the link value to a probability value for the launch temperatures:

- 70°F
  0.08 0.23 0.83 0.94 0.985 0.9996 0.9999
- 53°F
  0.08 0.23 0.83 0.94 0.985 0.9996 0.9999
- 31°F
  0.08 0.23 0.83 0.94 0.985 0.9996 0.9999

A more complete analysis of the situation would take into account the fact that there are multiple O-rings in each booster, while the Damage variable describes whether *any* O-ring failed. In addition, there were two O-rings on each booster segment, both of which would have to fail to create a leakage problem. Thus, the probabilities estimated from this model and these data do not accurately reflect the probability of a catastrophic accident.

Prob 16.03

George believes in astrology and wants to check whether a person's sign influences whether they are left- or right-handed. With great effort, he collects data on 100 people, recording their dominant hand and their astrological sign. He builds a logistic model hand ~ sign. The deviance from the model hand ~ 1 is 102.8 on 99 degrees of freedom. Including the sign term in the model reduces the deviance to 63.8 on 88 degrees of freedom.

The sign term only reduced the degrees of freedom by 11 (that is, from 99 to 88) even though there are 12 astrological signs. Why?

A There must have been one sign not represented among the 100 people in George's sample.
B sign is redundant with the intercept and so one level is lost.
C hand uses up one degree of freedom.

According to theory, if sign were unrelated to hand, the 11 degrees of freedom ought to reduce the deviance by how much, on average?

A $11/99 \times 102.8$
B $1/11 \times 102.8$
C to zero
D None of the above.

Prob 16.04

This model traces through some of the steps in fitting a model of a yes/no process. For specificity, pretend that the data are from observations of a random sample of teenaged drivers. The response variable is whether or not the driver was in an accident during one year (birthday to birthday). The explanatory variables are sex and age of the driver. The model being fit is accident $\sim 1 +$ age $+$ sex.

Here is a very small, fictitious set of data.

Case	Age	Sex	Accident?
1	17	F	Yes
2	17	M	No
3	18	M	Yes
4	19	F	No

Even if it weren't fictitious, it would be too small for any practical purpose. But it will serve to illustrate the principles of fitting.

In fitting the model, the computer compares the likelihoods of various candidate values for the coefficients, choosing those coefficients that maximize the likelihood of the model.

Consider these two different candidate coefficients:

Candidate A Coefficients		
Intercept	age	sexF
35	-2	-1

Candidate B Coefficients		
Intercept	age	sexF
35	-2	0

The link value is found by multiplying the coefficients by the values of the explanatory variables in the usual way.

- Using the candidate A coefficients, what is the link value for case 1?
  - [A] $35 - 2 \times 17 - 0 = 1$
  - [B] $35 - 2 \times 17 - 1 = 0$
  - [C] $35 - 2 \times 18 - 1 = -2$
  - [D] $35 - 2 \times 19 - 1 = -4$

- Using the candidate B coefficients, what is the link value for case 3?
  - [A] $35 - 2 \times 18 - 0 = -1$
  - [B] $35 - 2 \times 18 - 1 = -2$
  - [C] $35 - 2 \times 18 + 1 = 0$
  - [D] $35 - 2 \times 18 - 2 = -3$

The link value is converted to a probability value by using the logistic transform.

- The link value under the candidate A coefficients for case 4 is $35 - 2 \times 19 - 1 = -4$. What is the corresponding probability value? (Hint: Plug in the link value to the logistic transform!) 0.004 0.018 0.027 0.047 0.172 0.261

- The link value under the candidate B coefficients for case 4 is $35 - 2 \times 19 - 0 = -3$. What is the corresponding probability value? 0.004 0.018 0.027 0.047 0.172 0.261

The probability value is converted to a likelihood by calculating the probability of the observed outcome according to the probability value. When the outcome is "Yes," the likelihood is just the same as the probability value. But when the outcome is "No," the likelihood is 1 minus the probability value.

- The link value for case 3 using the candidate A coefficients is -1 and the corresponding probability value is 0.269. What is the likelihood of the observed value of case 3 under the candidate A coefficients? 0.000 0.269 0.500 0.731 1.000

- The link value for case 2 using the candidate A coefficients is 1 and the corresponding probability value is 0.731. What is the likelihood of the observed value of case 2 under the candidate A coefficients? 0.000 0.269 0.500 0.731 1.000

To compute the likelihood of the entire set of observations under the candidate coefficients, multiply together the likelihoods for all the cases. Do this calculation separately for the candidate A coefficients and the candidate B coefficients. Show your work. Say which of the two candidates gives the *bigger* likelihood?

In an actual fitting calculation, the computer goes through large numbers of candidate coefficients in a systematic way to find the candidate with the largest possible likelihood: the maximum likelihood candidate. Explain why it makes sense to choose the candidate with the *maximize* rather than the minimum likelihood.

Prob 16.05

The National Osteoporosis Risk Assessment (NORA)[**?**] studied about 200,000 postmenopausal women aged 50 years or old in the United States. When entering the study, 14,412 of these women had osteoporosis as defined by a bone-mineral density "T score." In studying the risk factors for the development of osteoporosis, the researchers fit a logistic regression model.

The coefficients in a logistic regression model can be directly interpreted as the logarithm of an odds ratio — the "log odds ratio." In presenting results from logistic regression, it's common to exponentiate the coefficients, that is, to compute e^{coef} to produce a simple odds ratio.

The table below shows the coefficients in odds ratio form from the NORA model. There were many explanatory variables in the model: Age group, years since menopause, health status, etc. All of these were arranged to be categorical variables, so there is one coefficient for each level of each variable. As always, one level of each variable serves as a reference level. For instance, in the table below, the age group 50-54 is the reference level. In the table below, the odds ratio for the reference level is always given as 1.00. The other odds ratios are always with respect to this reference. So, women in the 55-59 age group have odds of having osteoporosis that are 1.79 time bigger than women in the 50-54 age group. In contrast, women who are 6-10 years since menopause have odds of having osteoporosis that are 0.79 as big as women who are ≤ 5 years since menopause.

An odds ratio of 1 means that the group has the same probability value as the reference group. Odds ratios bigger than 1 mean the group is more likely to have osteoporosis

than the reference group; odds ratios smaller than 1 mean the group is less likely to have the condition.

The 95% confidence interval on the odds ratio indicates the precision of the estimate from the available data. When the confidence interval for a coefficient includes 1.00, the null hypothesis that the population odds ratio is 1 cannot be rejected at a 0.05 significance level. For example, the odds ratio for a self-rated health status level of "very good" is 1.04 compared to those in "excellent" health. But the confidence interval, 0.97 to 1.13, includes 1.00, indicating that the evidence is weak that women in very good health have a different risk of developing osteoporosis compared to women in excellent health.

For some variables, e.g., "college education or higher," no reference level is given. This is simply because the variable has just two levels. The other level serves as the reference.

Age group (years)	Odds Ratio (95% CI)
50-54	1.00 (Referent)
55-59	1.79 (1.56-2.06)
60-64	3.84 (3.37-4.37)
65-69	5.94 (5.24-6.74)
70-74	9.54 (8.42-10.81)
75-79	14.34 (12.64-16.26)
≥80	22.56 (19.82-25.67)

Years since menopause	Odds Ratio (95% CI)
≤ 5	1.00 (Referent)
6-10	0.79 (0.70-0.89)
11-15	0.83 (0.76-0.91)
16-20	0.96 (0.89-1.03)
21-25	1.01 (0.95-1.08)
26-30	1.02 (0.95-1.09)
31-35	1.10 (1.03-1.19)
36-40	1.14 (1.05-1.24)
≥41	1.24 (1.14-1.35)

College educ or higher	Odds Ratio (95% CI)
	0.91 (0.87-0.94)

Self-rated health status	Odds Ratio (95% CI)
Excellent	1.00 (Referent)
Very good	1.04 (0.97-1.13)
Good	1.23 (1.14-1.33)
Fair/poor	1.62 (1.50-1.76)

Fracture history	Odds Ratio (95% CI)
Hip	1.96 (1.75-2.20)
Wrist	1.90 (1.77-2.03)
Spine	1.34 (1.17-1.54)
Rib	1.43 (1.32-1.56)

Maternal history of osteoporosis	Odds Ratio (95% CI)
	1.08 (1.01-1.17)

Maternal history of fracture	Odds Ratio (95% CI)
	1.16 (1.11-1.22)

Race/ethnicity	Odds Ratio (95% CI)
White	1.00 (Referent)
African American	0.55 (0.48-0.62)
Native American	0.97 (0.82-1.14)
Hispanic	1.31 (1.19-1.44)
Asian	1.56 (1.32-1.85)

Body mass index, kg/m2	Odds Ratio (95% CI)
≤ 23	1.00 (Referent)
23.01-25.99	0.46 (0.44-0.48)
26.00-29.99	0.27 (0.26-0.28)
≥ 30	0.16 (0.15-0.17)

Current medication use	Odds Ratio (95% CI)
Cortisone	1.63 (1.47-1.81)
Diuretics	0.81 (0.76-0.85)

Estrogen use Odds Ratio (95% CI)	
Former	0.77 (0.73-0.80)
Current	0.27 (0.25-0.28)

Cigarette smoking Odds Ratio (95% CI)	
Former	1.14 (1.10-1.19)
Current	1.58 (1.48-1.68)

Regular Exercise	Odds Ratio (95% CI)
Regular	0.86 (0.82-0.89)

Alcohol use, drinks/wk	Odds Ratio (95% CI)
None	1.00 (Referent)
1-6	0.85 (0.80-0.90)
7-13	0.76 (0.69-0.83)
≥ 14	0.62 (0.54-0.71)

Technology	Odds Ratio (95% CI)
Heel x-ray	1.00 (Referent)
Forearm x-ray	2.86 (2.75-2.99)
Finger x-ray	4.86 (4.56-5.18)
Heel ultrasound	0.79 (0.70-0.90)

Since all the variables were included simultaneously in the model, the various coefficients can be interpreted as indicating partial change: the odds ratio comparing the given level to the reference level for each variable, adjusting for all the other variables as if they had been held constant.

- For which ethnicity are women least likely to have osteoporosis?
  White African.American Native.American Hispanic Asian

- Is regular exercise (compared to no regular exercise) associated with a greater or lesser risk of having osteoporosis? greater lesser same

- Is current cigarette smoking (compared to never having smoked) associated with a greater or lesser risk of having osteoporosis? greater lesser same

- The body mass index (BMI) is a measure of overweight. For adults, a BMI greater than 25 is considered overweight (although this is controversial) and a BMI greater than 30 is considered "obese." Are women with BMI ≥ 30 (compared to those with BMI < 23) at greater, lesser, or the same risk of having osteoporosis?
  greater lesser same

- There are different technologies for detecting osteoporosis. Since the model adjusts for all the other risk factors, it seems fair to interpret the risk ratios for the different technologies as indicating how sensitive each technology is in detecting osteoporosis.

  Which technology is the most sensitive?
  heel.x-ray forearm.x-ray finger.x-ray heel.ultrasound
  Which technology is the least sensitive?
  heel.x-ray forearm.x-ray finger.x-ray heel.ultrasound

- In combining the odds ratios of multiple variables, you can *multiply* the individual odds ratios. For instance, the odds of a woman in very good health with a body mass index of 24 is 1.04×0.46 as large as a woman in excellent health with a BMI of < 23 (the reference levels for the variables involved). (If log odds ratios were used, rather than the odds ratios themselves, the values would be added, not multiplied.)

  What is the odds ratio of a women having osteoporosis who is in fair/poor health, drinks 7-13 drinks per week, and is Asian?

  A $0.76 \times 0.27 \times 1.00$
  B $1.62 \times 0.27 \times 1.56$
  C $1.62 \times 0.76 \times 1.56$
  D $0.76 \times 1.62 \times 1.00$

- The two variables "age" and "years since menopause" are likely to be somewhat collinear. Explain why. What effect might this collinearity have on the width of the confidence intervals for the various variables associated with those variables? If you were recommending to remove one of the variables in the list of potential risk factors, which one would it be?

  Notice that the table gives no intercept coefficient. The intercept corresponds to the probability of having osteoporosis when belonging to the reference level of each of the explanatory variables. Without knowing this, you cannot use the coefficients calculate the absolute risk of osteoporosis in the different conditions. Instead, the odds ratios in the table tell about relative risk. Gigerenzer [**?**, **?**] points out that physicians and patients often have difficulty interpreting relative risks and encourages information to be presented in absolute terms.

  To illustrate, in the group from whom the NORA subjects was drawn, the absolute risk of osteoporosis was 72 in 1000 patients. This corresponds to an odds of osteoporosis of $72/(1000 - 72) = 0.776$. Now consider a woman taking cortisone. According to the table, this increases her odds of osteoporosis by a factor of 1.63, to $0.776 \times 1.63 = 0.126$. Translating this back into an absolute risk means converting from odds into probability. The probability will be $0.126/(1 + 0.126) = 0.112$, or, in other words, an absolute risk of 112 in 1000 patients.

  Now suppose the woman was taking cortisone to treat arthritis. Knowing the absolute risk (an increase of 40 women per 1000) puts the woman and her physician in a better position to compare the positive effects of cortisone for arthritis to the negative effects in terms of osteoporosis.

  [This problem is based on an item used in a test of the statistical expertise of medical residents reported in [**?**].]

Prob 16.07

The concept of residuals does not cleanly apply to yes/no models because the model value is a probability (of a yes outcome), whereas the actual observation is the outcome itself. It would be silly to try to compute a difference between "yes" and a probability like 0.8. After all, what could it mean to calculate $(\text{yes} - 0.8)^2$?

In fitting ordinary linear models, the criterion used to select the best coefficients for any given model design is "least squares," minimizing the sum of square residuals. The corresponding criterion in fitting yes/no models (and many other types of models) is "maximum likelihood."

The word "likelihood" has a very specific and technical meaning in statistics, it's not just a synonym for "chance" or "probability." A likelihood is the probability of the outcome *according to a specific model*.

To illustrate, here is an example of some yes-no observations and the model values of two different models.

	Model A		Model B		Observed
Case	p(Yes)	p(No)	p(Yes)	p(No)	Outcome
1	0.7	0.3	0.4	0.6	Yes
2	0.6	0.4	0.8	0.2	No
3	0.1	0.9	0.3	0.7	No
4	0.5	0.5	0.9	0.1	Yes

Likelihood always refers to a given model, so there are two likelihoods here: one for Model A and another for Model B. The likelihood for each case under Model A is the probability of the observed outcome according to the model. For example, the likelihood under Model A for case 1 is 0.7, because that is the model value of the observed outcome "Yes" for that case. The likelihood of case 2 under Model A is 0.4 — that is the probability of "No" for case 2 under model A.

- What is the likelihood under Model A for case 3?
  0.1 0.3 0.5 0.7 0.9

- What is the likelihood under Model B for case 3? 0.1 0.3 0.5 0.7 0.9
- What is the likelihood under Model A for case 4? 0.1 0.3 0.5 0.7 0.9
- What is the likelihood under Model B for case 4? 0.1 0.3 0.5 0.7 0.9

The likelihood for the whole set of observations combines the likelihoods of the individual cases: multiply them all together. This is justified if the cases are independent of one another, as is usually assumed and sensible if the cases are the result of random sampling or random assignment to an experimental treatment.

- What is the likelihood under Model A for the whole set of cases?
  - [A] $0.3 \times 0.4 \times 0.9 \times 0.5$
  - [B] $0.7 \times 0.6 \times 0.9 \times 0.5$
  - [C] $0.3 \times 0.4 \times 0.1 \times 0.5$
  - [D] $0.7 \times 0.4 \times 0.9 \times 0.5$
  - [E] $0.7 \times 0.4 \times 0.1 \times 0.5$
- What is the likelihood under Model B for the whole set of cases?
  - [A] $0.4 \times 0.8 \times 0.3 \times 0.9$
  - [B] $0.4 \times 0.2 \times 0.3 \times 0.9$
  - [C] $0.6 \times 0.2 \times 0.3 \times 0.9$
  - [D] $0.4 \times 0.2 \times 0.7 \times 0.9$
  - [E] $0.4 \times 0.2 \times 0.3 \times 0.1$

Chapter Seventeen Reading Questions

1. Why does an observed correlation between two variables not provide compelling evidence that one causes the other?
2. What is a hypothetical causal network? What does the word "hypothetical" signify about these networks?
3. What is the difference between a correlating pathway and a non-correlating pathway?
4. How is the appropriate choice of covariates in models to study causation influenced by the structure of the modeler's hypothetical causal network?
5. When might two modelers legitimately disagree about which covariates to include in a model?

Chapter Eighteen Reading Questions

1. Is the point of a controlled experiment to create variability or to avoid variability?
2. What is an experimental variable? In what ways is it the same and in what ways different from an explanatory variable?
3. What is the purpose of randomization in conducting an experiment? What gets randomized?
4. Why might blocking be preferred to randomization?
5. What is the point of matched sampling and instrumental variables?