Lessons in Statistical Thinking

Table of contents

1 Preliminaries

This Lesson covers some preliminaries: techniques we use throughout the remaining lessons.

1. Using regression to summarize relationships.
2. The presentation of descriptions using **intervals** rather than a number like the mean or proportion.
3. Promotion of the point plot (sometimes jittered) as the standard form for displaying data. Statistical summaries appear as annotations on top of the data layer.
4. A modern format called a "**violin plot**" for displaying the *distribution* of a quantitative variable. Unlike a histogram, the violin plot can be layered as an annotation on top of a point plot.
5. Extending regression modeling to handle *categorical* response variables.

1.1 Regression as a data summary

The first half of this course emphasized data wrangling and visualization. The well-named `summarize()` function is the natural wrangling choice to compute summaries such as means, medians, or standard deviations. For instance, this command calculates four summary statistics on the `net` running time recorded in the `TenMileRace` data frame:

```
TenMileRace %>%
  summarize(ave = mean(net), middle = median(net), sd = sd(net), n = n())
```

ave	middle	sd	n
5599.065	5555	969.6564	8636

`summarize()` works hand-in-hand with `group_by()` to calculate groupwise summaries. The following wrangling statement, for instance, looks at the average `net` running time broken up according to the runner's state of residence and presents the results from the fastest state downwards:

Wrangling is essential for many statistical purposes, not just summarizing but also setting up for making graphical displays, cleaning data, and assembling data from multiple sources.

Regression modeling is used only for summarizing. The summary describes the *relationship* between the response and explanatory variables. Think of it as a kind of substitute for `summarize()` when you want to describe relationships.

As we saw above, `summarize()` and `group_by()` are two different stages of wrangling that, used together, produce a separate summary for each group. Regression modeling, however, offers a rich alternative to grouping. The spirit of what one accomplishes in wrangling with `group_by()` is achieved in regression by *using additional explanatory variables*.

The mean as summary

In its simplest mode, a regression can calculate means. The result will be identical to `group_by()`/`summarize()` but in a different format.
Here is the mean `net` running time calculated using both approaches.

```
TenMileRace %>% lm(net ~ 1, data=.) %>% coef()
```

```
(Intercept)
   5599.065
```

```
TenMileRace %>% summarize(mn = mean(net))
```

mn
5599.065

The groupwise calculations also produce equivalent results, although the results are formatted in different ways.

```
TenMileRace %>% lm(net ~ sex, data=.) %>% coef()
```

```
(Intercept)        sexM
  5916.3979   -635.6958
```

Notice something new in the above command: the `data=.` argument inside `lm()`. The simple `.` is doing something important, carrying the output of the earlier stages of the pipeline into the `data=` argument of `lm()`.

```
TenMileRace %>% group_by(sex) %>% summarize(mn = mean(net))
```

sex	mn
F	5916.398
M	5280.702

Regression of this sort calculates the mean of a reference group and the **difference in means** between the two groups, whereas the wrangling command presents the mean of each group.

In regression, "grouping" is extended to quantitative variables. For instance,

```
TenMileRace %>% lm(net ~ age, data = .) %>% coef()
```

```
(Intercept)         age
5297.219248    8.189886
```

This report indicates a trend of `net` running time increasing with `age` by about 8 seconds per year.

The `group_by()` function can use a quantitative variable, but the result is a different number for each group rather than a trend.

```
TenMileRace %>% group_by(age) %>% summarize(mn = mean(net))
```

age	mn
10	5640
12	5980
13	5410
14	5620
15	5170

With multiple grouping variables, say `age` and `sex`, the output of `summarize()` becomes increasingly complicated. For example:

```
TenMileRace %>%
  group_by(sex, age) %>%
  summarize(mn = mean(net))
```

sex	age	mn
M	16	5288
F	26	5738
M	30	5079
F	38	5979
F	39	5895
F	47	6135
M	55	5648
M	59	5819

Regression keeps things simpler, reporting on trends:

```
TenMileRace %>% lm(net ~ sex + age, data = .) %>% coef()
```

```
(Intercept)        sexM         age
    5339.16     -726.62       16.89
```

The trend reported from this regression model is an increase in `net` of about 16 seconds per year of age. Regression can summarize relationships in more detailed ways as well. The following model looks at the trend with age separately for males and females:

```
TenMileRace %>% lm(net ~ sex * age, data = .) %>% coef()
```

```
(Intercept)        sexM         age    sexM:age
    5371.00     -785.15       15.96        1.61
```

Here, the age trend for women is an increase in `net` running time of 16 seconds per year of age, while for men, that increase is bigger, an extra 1.6 seconds per year of age.

There are good reasons why `lm()` organizes summaries the way it does. The `lm()` paradigm can make much more efficient use of data than `group_by()`. It also offers much more flexibility. `lm()` can handle multiple "grouping" variables together and even lets you "group" by quantitative variables.

1.2 Presentation using intervals

Statistical thinking often involves quantifying uncertainty. Uncertainty appears where a newcomer to statistical thinking might not expect it. For example, consider "**point**" summaries such as the mean or median. So long as the arithmetic is correct, the result is inevitable; everyone doing the calculation will get the same result. The statistical thinker, however, includes the *data collection process* in the calculation. Each person carrying out his or her data collection process will get different results. The study-to-study variation calls for an interval display, where the interval covers the likely range of results.

Prediction is another context benefiting from an interval display. Prediction is imperfect. The predicted result—for instance, the baby's due date—is typically different from the actual outcome. The statistical thinker knows how to estimate the likely range of the difference between the predicted and actual outcomes.

i Example: 1.6 seconds per year?

It is appropriate to be skeptical of a claim that male runners slow down by 1.6 seconds per year compared to females. After all, people differ; some age more gently than others. As we will see in Lesson 2, the results presented from a regression model depend partly on the play of chance in determining the particular people represented in the data. It is helpful to know *how much* chance affects the results. A summary can indicate this by a range of plausible values, in other words, an "**interval**" summary.

Here is an interval summary on the coefficients from the running time versus age model:

```
TenMileRace %>% lm(net ~ sex * age, data = .) %>% confint()
```

	lwr	upr
(Intercept)	5269.77	5472.23
sexM	-927.31	-642.98
age	13.11	18.82
sexM:age	-2.14	5.36

Notice that the interval on `sexM:age` includes zero. Construct interval summaries using the appropriate *extractor* on a regression model. For instance, `confint()` generates an interval summary suggesting there might be no difference in the age trend for males and females.

1.3 Point plot as a graphical foundation

Regression models, which will be the primary means of summarizing data in these Lessons, always have a response variable and typically have one or more explanatory variables.[1]

In these Lessons, we will place graphical depictions of model summaries in the context of actual data. Consequently, the graphical frame will reflect the choice of response and explanatory variables. The vertical axis will *always* represent the response variable. The horizontal axis will represent one of the explanatory variables. A point plot will display the data, or a jitter plot when there are categorical variables to be shown.

Another aspect of our unified data graphic format is that it will *always* be a point plot or, closely related, a jitter plot.

[1] Actually, the previous sentence should say, "**zero** or more explanatory variables." The model with no explanatory variables (and `y` as the response variable) is denoted by `y ~ 1`. This simple model represents the hypothesis that nothing can explain the variability in `y`.

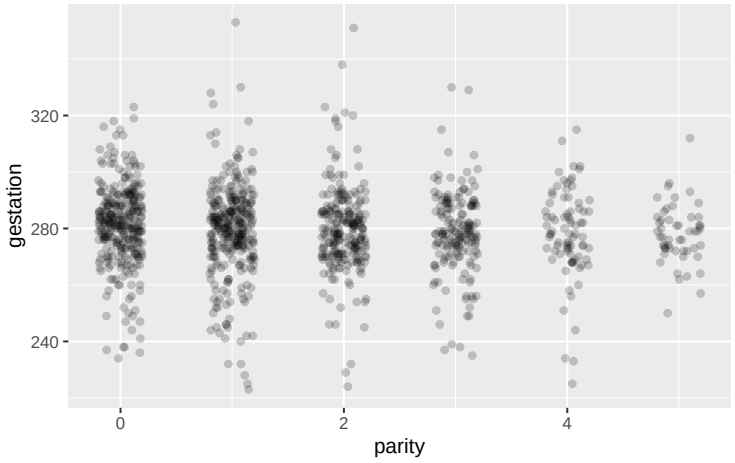
Example: When is the baby due?

As all expecting parents know, a baby's "due date" is hardly exact. Pregnancies vary in length. What accounts for this variation?

In this example, we will entertain the hypothesis that experienced mothers have systematically different gestation periods than first-time mothers. An appropriate response variable is duration of `gestation`. The explanatory variable needs to measure "experience," which is a vague idea. We will make it concrete by representing it by the number of the mother's pregnancies before the one reported in the data.

The `Gestation` data frame records more than 1200 births. `gestation` records the length of the pregnancy and will be our response variable. `parity` gives the number of previous births to the mother, starting at zero for a first-time mother. Although `parity` is encoded as a number, it has only *discrete* values—0, 1, 2, … We will therefore graph it as a *categorical* variable, using jittering to avoid overplotting. There are not many rows with parity greater than five; we will focus on those.

```
Gestation %>%
  filter(parity <= 5) %>%
  #mutate(parity = as.character(parity)) %>%
  ggplot(aes(x=parity, y=gestation)) +
  geom_jitter(alpha=0.2, width=0.2, height=0)
```



This graph shows some things at a glance. For example, a typical gestation period is about 275 days (about nine months), and it is much more common to have a low parity than a very high one.

Figure 1: Gestational period for pregnancies where the mother had five or fewer previous pregnancies. The `width=0.2` controls the amount of horizontal jittering. We chose it to make the columns of data clear. There is no need to jitter in the vertical direction, so we set `height=0`

1.4 Displaying density

It is easy to see a pattern in Figure 1: It looks like mothers with high parity tend to have gestation periods more reliably close to 280 days than mothers with low parity. However, on the other hand, maybe this pattern is an illusion, an artifact of the small number of pregnancies with parity 3, 4, or 5 and, therefore, less opportunity to see extreme values for `gestation`.

One way to explore this idea is to plot the density of the dots as a function of gestation for each of the parity levels individually. A "violin" layer will make it easier to compare the distributions in the different columns, despite the unevenness in the case count. Figure 2 gives an example.

The violin plot is a more flexible display of the distribution of gestation period than a histogram. The histogram has all those bars that clutter up the display. Even worse, one of the axes in the frame of a histogram plot is "count" or maybe "density." Such a frame is inconsistent with the response/explanatory axes used for the data. The violin is drawn in the no-mans-land

```
Gestation %>%
  filter(parity <= 5) %>%
  ggplot(aes(x=parity, y=gestation)) +
  geom_jitter(alpha=0.2, width=0.2, height=0) +
  geom_violin(aes(group=parity), fill="blue", alpha=0.2, color=NA)
```

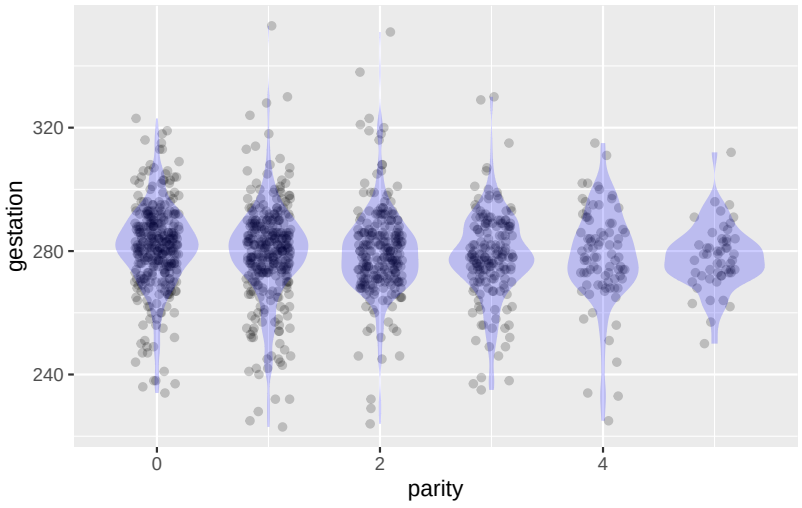


Figure 2: A violin plot. The long axis of the violin-like shape is oriented along the response-variable axis (that is, the vertical axis in our standard format). The width of the violin for each possible value of the response variable is proportional to the density of data near that value.

between the different levels of parity, just as the jittering moves data away from a single vertical line into that same no-mans-land.

This idea of using the graphical no-mans-land between levels of a categorical explanatory variable is not new. You encountered it earlier when you drew box plots. Figure 3 adds a box-plot annotation layer on top of the violin-plot layer.

```
Gestation %>%
  filter(parity <= 5) %>%
  ggplot(aes(x=parity, y=gestation)) +
  geom_jitter(alpha=0.2, width=0.2, height=0) +
  geom_violin(aes(group=parity), fill="blue", alpha=0.2, color=NA) +
  geom_boxplot(aes(group=parity), color="blue", fill=NA, alpha=.5)
```

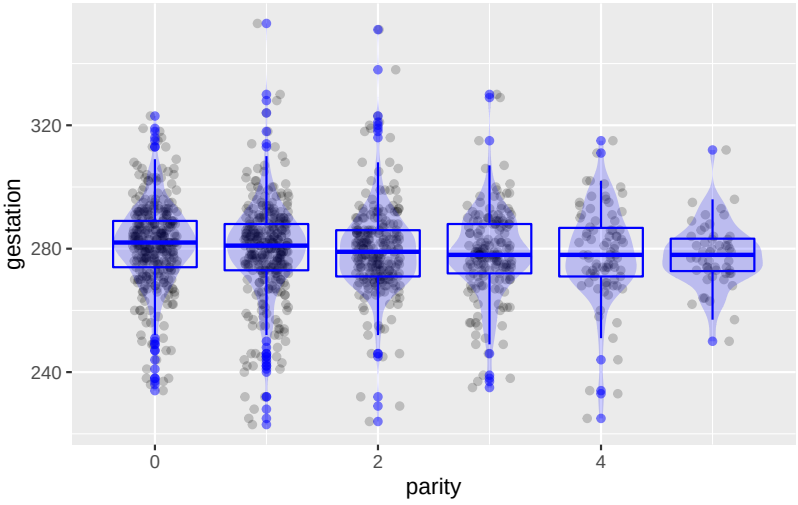


Figure 3: A box and whisker plot uses the no-mans-land between levels of a categorical explanatory variable.

In practice, there is little reason to layer a box plot on top of a violin. The violin does the job nicely on its own.

1.5 Categorical response variables

Regression modeling will be a fundamental tool in these Lessons for summarizing data. Regression models always have a quantitative response variable, although explanatory variables can be either quantitative or categorical.

Often, the modeling situation calls for a response variable that is *categorical*. Expert modelers can use specialized modeling methods to handle such situations. However, categorical response variables often have just two levels, e.g., Alive/Dead, Promoted/Not, or Win/Loss. We will name the general class of such variables as "**yes/no**" or, equivalently, "**zero-one**" variables.

More formally, they are called "**binomial**" variables.

Yes/no response variables can be represented using 0 for one level and 1 for the other. This numerical "**0/1 encoding**" is directly suited for regression modeling and enables us to extend the scope of regression models. The *output* of the regression model is always numerical. Nothing in the regression technique restricts those outputs to exactly zero or one, even when the response variable is of the yes/no type. Usually, the modeler interprets such numerical output as probabilities or, more generally, as measures to be converted to probabilities.

i R technique: Using `zero_one()`.

The `zero_one()` function converts a yes/no variable to the numerical zero-one format. `zero_one()` allows you to specify which of the two levels is represented by 1.
To illustrate, consider the `mosaicData::Whickham` data frame, which records a 1972-1974 survey, part of a study of the relationship between smoking and mortality. Twenty years after the initial survey, a follow-up established whether or not each person was still alive. Here are a few rows from the data frame:

outcome	smoker	age
Alive	Yes	23
Alive	Yes	18
Dead	Yes	71
Alive	No	67
Alive	No	64
Alive	Yes	38

The `outcome` variable in `Whickham` records the result of the follow-up survey. It is a categorical variable with levels "Alive" and "Dead." To examine what the data have to say

about the relationship between smoking and mortality, we construct a model with `outcome` as the response variable and `smoking` as an explanatory variable. Before doing so, we translate `outcome` into a zero-one format. Like this:

```
Whickham %>%
  mutate(alive = zero_one(outcome, one="Alive"))
```

outcome	smoker	age	alive
Alive	Yes	23	1
Alive	Yes	18	1
Dead	Yes	71	0
Alive	No	67	1
Alive	No	64	1
Alive	Yes	38	1

Note the correspondence between the `outcome` and the newly created `alive` variable.

⚠ Demonstration: Predicting calorie content

Starbucks is a famous coffee-shop franchise with more than 30,000 branches (as of 2021). People go to Starbucks for coffee, but they often buy something to eat as well. In this demonstration, we will look at the calorie content of Starbucks' food offerings. As always, when conducting a statistical analysis, it is helpful to have in mind the motivation for the task. So we will imagine, tongue in cheek, that we want to make food recommendations for the calorie-conscious consumer.
First, a **point summary** of the calories in the different types of food products available at Starbucks:

```
df_stats(calories ~ type,
         data = openintro::starbucks, mean)
```

response	type	mean
calories	bakery	369
calories	bistro box	378
calories	hot breakfast	325
calories	parfait	300
calories	petite	178
calories	salad	80
calories	sandwich	396

This summary supports the sensible advice to choose salads or smaller portions (type "petite") to avoid calories. One might go further, for example, concluding that a sandwich is a poor choice (in terms of calorie content), so lean toward parfaits or hot breakfasts. We can even imagine someone concluding from this summary that a bistro box is a better calorie-conscious choice than a sandwich.
Figure 4 shows the point summary, using the raw data to put things in context.

```
openintro::starbucks %>%
  ggplot(aes(x=type, y=calories)) +
  geom_jitter(width=0.2, alpha=0.5) +
  geom_errorbar(data=point_summary, aes(ymin=mean, ymax=mean),
                y=NA, color="blue") +
  geom_point(data=point_summary, aes(y=mean), color="red")
```

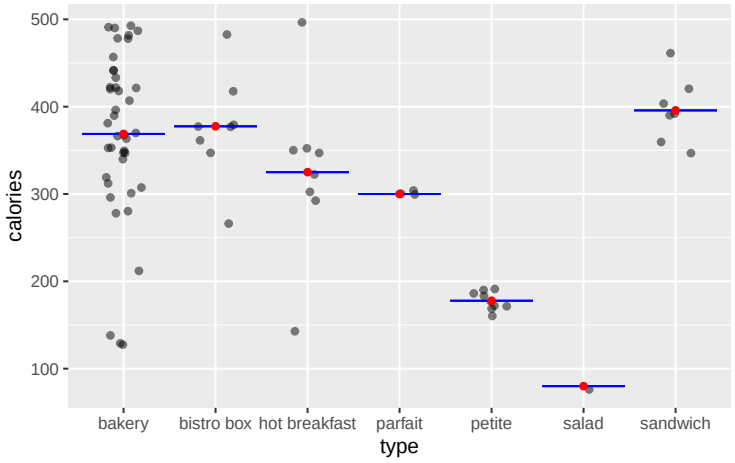


Figure 4: Calories of the various food items sold by Starbucks, annotated with point and interval summaries.

Plotting the point summary in the context of the raw data shows at a glance that the point summary is not of any genuine use. For instance, using the point summary without the data, we might conclude that hot breakfasts are better than sandwiches. However, the data display suggests otherwise; there is just one low-calorie breakfast. The others are much like sandwiches.

A point summary is compact but cannot represent the *variation* within each food type. An interval summary, as in Figure 5, does show this variation.

```
openintro::starbucks %>%
  ggplot(aes(x=type, y=calories)) +
  geom_jitter(width=0.2, alpha=0.5) +
  geom_errorbar(data=point_summary, aes(ymin=mean, ymax=mean),
                y=NA, color="blue")
```

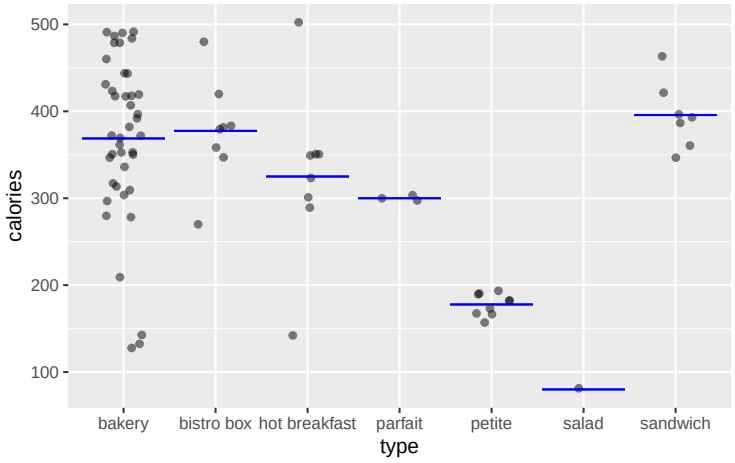


Figure 5: Calories of the various food items sold by Starbucks, annotated with point and interval summaries.

Unlike point summaries, interval summaries can overlap. Such overlap indicates that the groups are not all that different. Here, the interval summary indicates an appropriate conclusion; "Don't make your diet choices based on food type. Look at the calorie content of individual items before choosing."

Admittedly, in this simple setting the data themselves would lead to the conclusion. However, as we move into more complicated settings, it will become infeasible to see patterns quickly straight from the data.

2 Measuring and simulating variation

> *Variation itself is nature's only irreducible essence. Variation is the hard reality, not a set of imperfect measures for a central tendency. Means and medians are the abstractions.* —— Stephen Jay Gould (1941- 2002), paleontologist and historian of science.

This Lesson introduces two techniques: i) how to quantify variation and ii) how to represent our ideas about causal connections.

2.1 Measuring variation

A data frame records observations. In a properly constructed data frame, each row holds the observations from one "unit of observation." Each column stores the values of one "variable," one type of observation. The "unit of observation" might be a winner in a Scottish Hill race (`Hill_racing`), it might be a pregnancy and the resulting birth (`Gestation`), it might be a nurse participating in a survey (`Whickham`), or any of an infinite range of things. All the rows of a data frame record the same kind of unit of analysis.

Similarly, within a variable, all the recorded observations have the same type denominated in the same units: an age in years, a distance in kilometers, a price in dollars, or whatever.

Although the unit of observation is the same *type* in every row, the units themselves can differ from one another. Consequently, the values recorded in a variable can differ; they are not all the same: they vary. The words "variable" and "variation" go hand in hand: a variable displays variation.[2]

Statistical models decompose the response variable into components. Each component is associated with one or more explanatory variables, with one exception. The remaining component, called the "residual," represents the part of the variation that is

[2] A potential source of confusion is the unfortunate use of the word "variable" in algebra to mean a symbol, like x, that represents a quantity symbolically.

still unexplained. In order to make sure that we are not double-counting or omitting variation, it helps to measure the **size** of variation in each component.

There are many possible ways to measure the "size." [3] In these Lessons, we will use a standard measure with an awkward name: the "**standard deviation**." In practice, however, we will work mainly with the *square of the standard deviation*, called the "**variance**."

Both variance and standard deviation are **quantities**, that is, a single number with associated units. The standard deviation of any variable has units that are the same as the variable itself. For instance, height is often denominated in cm. Therefore, the standard deviation of height, as it varies from person to person, will also be in cm.

Since the variance is the square of the standard deviation, variance has units of the square of the units of the variable. So the variance of height, for instance, will be measured in cm^2.

Remember that variance and standard deviation are *summaries* of a variable. A variable in a data frame consists of multiple values, one for each row. The variance (or standard deviation) of that variable is a single number (with units), summarizing all of the values in the variable.

i Why the variance?

The name "variance" is a good reminder of what it measures: the variation.
But why use a squared measure?
Consider this familiar equation: $A^2 + B^2 = C^2$. The Pythagorean Theorem states that the equation describes the lengths of the sides of a right triangle: C is the hypotenuse, while A and B are the other two sides of the triangle. Surprisingly, the Pythagorean Theorem is highly

[3] The *OpenIntro* text introduced the standard deviation in Chapter 3, where the authors described it as a measure of "spread." In Chapter 6, *OpenIntro* introduced the variance as the square of the variance. All this is right, so far as it goes, but it dramatically understates the importance of the two measures. These measures are as crucial to statistical thinking as the Pythagorean Theorem is to geometry.

relevant to statistical models.
OpenIntro chapters 5 & 6 point out that the linear modeling technique produces two columns of numbers: the **fitted values** and the **residuals**. These columns have the same number of rows as the data frame used for training. The residuals are the row-by-row numerical *difference* between the response variable and the fitted values.
These three columns of numbers—the response variable, the fitted model values, and the residuals—are exactly analogous to the three sides of a right triangle. (This is not an obvious fact, but it is important to remember.) In particular, the following numerical relationship is as true for linear models as it is for triangles:

$$\text{sd(fitted)}^2 + \text{sd(residuals)}^2 = \text{sd(response)}^2$$

where sd() refers to the standard deviation. Consequently, $\text{sd}()^2$ is the variance.
The particular mathematical definition of variance and the standard deviation makes the Pythagorean relationship always describe models constructed using the `lm()` technique. (One of the names used for this technique, **least squares**, provides a hint that the Pythagorean relationship applies.)

i Calculating variance

Almost always, people use software for calculations. The relevant R functions are `sd()` and `var()` and are used in a `summarize()` statement, for instance

```
mtcars %>% summarize(v = var(hp))
```

v
4701

```
mtcars %>% summarize(s = sd(hp))
```

s
68.6

Regrettably, the software does not indicate the units of the quantity. For that, read the documentation for the data frame.

To understand *what* is being calculated by `var()`, we will describe an algorithm. This algorithm is not numerically efficient, but it highlights the essential feature of variation.

Starting material:

- A single column of numbers created by pulling out from the data frame the variable whose variance is to be calculated.
- A long roll of paper on which to write numbers, one after the other.

Repeat a basic calculation for each and every row in the column of numbers. To illustrate, let us detail the basic calculation for the k^{th} row.

i. Take the data value from the k^{th} row, and call it the "reference value."

ii. Subtract the reference value from each and every other value in the column and *square* the results.

iii. Write those numbers, all of them, on the roll.

Using the same roll of paper for all, carry out the basic calculation starting at each row in the data column. With this done, the paper roll contains many numbers, each of which is the square difference between the values from two rows in the data column. The mathematically inclined might like to know that there will be exactly $n(n-1)$ numbers written on the roll. (The term $n-1$ in that count might perk up the ears of statistics instructors.)

The final result—the variance—will be half the average of the numbers on the roll.

i Deconstructing "standard deviation"

"Standard deviation" is an antique term and is misleading to people who think about "deviation" in the ordinary sense. Nonetheless, "standard deviation" is so widely used in statistics that we can hardly avoid it. To limit the confusion, we will deconstruct the term.
Step 1 in the deconstruction makes clear what "standard" means:

> standard deviation = accepted *measure of* deviation.

Step 2 in the deconstruction replaces the archaic word "deviation" with something more descriptive:

> standard deviation = accepted *measure of* variation in the variable.

2.2 Representing causal connections

Often, but not always, our interest in studying data is to reveal the causal connections between variables. Understanding causality is essential, for instance, if we are planning to intervene in the world and want to anticipate the consequences. Interventions are things like "increase the dose of medicine," "stop smoking!", "lower the budget," "add more cargo to a plane (which will increase fuel consumption and reduce the range)."

Historically, mainstream statisticians were hostile to using data to explore causal relationships. (The one exception was **experiment**, which gathers data from an actual intervention in the world. See Lesson 13.) Statistics teachers encouraged students to use phrases like "associated with" or "correlated with" and reminded them that "correlation is not causation."

Regrettably, this attitude made statistics irrelevant to the many situations where intervention is the core concern and experiment was not feasible. A tragic episode of this sort likely caused millions of unnecessary deaths. Starting in the 1940s, doctors

and epidemiologists saw evidence that smoking causes lung cancer. In stepped the most famous statistician of the age, Ronald Fisher, to insist that the statement should be, "smoking is associated with lung cancer." He speculated that smoking and lung cancer might have a common cause, perhaps genetic. Fisher argued that establishing causation requires running an experiment where people are randomly assigned to smoke or not smoke and then observed for decades to see if they developed lung cancer. Such an experiment is unfeasible and unethical, to say nothing of the need to wait decades to get a result.

Fortunately, around 1960, a researcher at the US National Institutes of Health, Jerome Cornfield, was able to show mathematically that the strength of the association between smoking and cancer ruled out any genetic mechanism. Cornfield's work prompted the development of a new area in statistics: "**causal inference**."

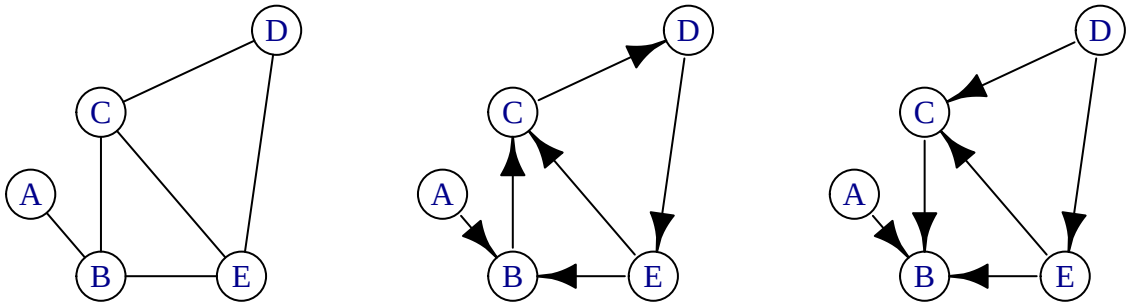
Causal inference is not about proving that one thing causes another but about formal ways to say something about how the world works that can be used, along with data, to make responsible conclusions about causal relationships.

A core tool in thinking about causal connections is a mathematical structure called a "directed acyclic graph" (DAG, for short). DAGs are one of the most popular ways for statistical thinkers to express their ideas about what might be happening in the real world. Despite the long name, DAGs are very accessible to a broad audience.

DAGs, despite the G for "graph," are not about data graphics. The "graph" in DAG is a mathematical term of art; a suitable synonym is "network." Mathematical graphs consist of a set of "nodes" and a set of "edges" connecting the nodes. For instance, Figure 6 shows three different graphs, each with five nodes labeled A, B, C, D, and E.

The nodes are the same in all three graphs of Figure 6, but each graph is different from the others. It is not just the nodes that define a graph; the edges (drawn as lines) are part of the definition as well.

The left-most graph in Figure 6 is an "**undirected**" graph; there is no suggestion that the edges run one way or another.



(a) undirected graph (b) directed but cyclic (c) directed acyclic graph

Figure 6: Graphs of various types

In contrast, the middle graph has the same nodes and edges, but the edges are **directed**. An excellent way to think about a directed graph is that each node is a pool of water; each directed edge shows how the water flows between pools. This analogy is also helpful in thinking about causality: the causal influences flow like water.

Look more carefully at the middle graph. There is a couple of loops; the graph is **cyclic**. In one loop, water flows from E to C to D and back again to E. The other loop runs B, C, D, E, and back to B. Such a flow pattern cannot exist without pumps pushing the water back uphill.

The rightmost graph reverses the direction of some of the edges. This graph has no cycles; it is **acyclic**. Using the flowing and pumped water analogy, an acyclic graph needs no pumps; the pools can be arranged at different heights to create a flow exclusively powered by gravity. The node-D pool will be the highest, E lower. C has to be lower than E for gravity to pull water along the edge from E to C. The node-B pool is the lowest, so water can flow in from E, C, and A.

Directed acyclic graphs represent causal influences; think of "A causes B," meaning that causal "water" flows naturally from A to B. In a DAG, a node can have multiple outputs, like D and E, and it might have multiple inputs, like B and C. In terms of causality, a node—like B—having multiple inputs means that more than one factor is responsible for the value of that node. A real-world example: the rising sun causes a rooster to crow,

but so can another intruder to the coop.

Often, nodes do not have any inputs. These are called "**exogenous factors**"at least by economists. The "genous" means "originates from." "Exo" means "outside." The value of an exogenous node is determined by something, just not something that we are interested in (or perhaps capable of) modeling. No edges are directed into an exogenous node since none of the other nodes influence its value.

The point of a DAG is to make a clear statement of a hypothesis about causation. Drawing a DAG does not mean that the hypothesis is correct, just that we believe the hypothesis is, in some sense, a possibility. Different people might have different beliefs about what causes what in real-world systems. Comparing their different DAGs can help, sometimes, to discuss and resolve the disagreement.

We are going to use DAGs for two distinct purposes. One purpose is to inform responsible conclusions from data about what causes what. The data on its own is insufficient to demonstrate the causal connections. However, data *combined with* a DAG can tell us something. Sometimes a DAG includes a causal connection that should create an association between variables. The DAG is incomplete if the association does not appear in the data.

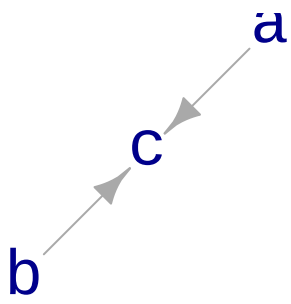
DAGs are also valuable aids for building models. For example, analysis of the paths in a DAG, as in Lesson 11, can tell us which explanatory variables to include and which to exclude from a model if our modeling goal is to represent the hypothetical causal connections.

In these Lessons, we have a second, entirely different, use for DAGs: learning modeling technique. Our approach will be to outfit DAGs with specific formulas representing the mechanism imbued in each node. DAGs equipped with formulas can be used to generate simulated data.[4] Training a model on those data leads to a model function that we can compare to the DAG's formulas. Then check whether the formulas and the model function match. This practice helps us learn what can

[4]The value of exogenous nodes is usually set randomly, without input from the other nodes in the DAG.

go right or wrong in building a model, just as practice in an aircraft simulator trains pilots to handle real-world situations in real aircraft.

We start with a simple example. The DAG called `dag09` has two exogenous nodes (`a` and `b`) and another node, `c`, that gets input from both `a` and `b`.




```
dag_draw(dag09)
```

```
print(dag09)
```

```
a ~ exo()
b ~ exo()
c ~ binom(2 * a + 3 * b)
```

The `dag_draw()` command draws a picture of the graph. Printing the dag displays the formulas that set the values of the nodes.

The formulas for `dag09` show that the nodes `a` and `b` are exogenous, their values set randomly and independently of one another by the `exo()` function. In contrast, the formula for node `c` says that the value of `c` will be a linear combination of the values of `a` and `b`, translated into a zero-one format.

Generate simulated data using the `sample()` function. For instance,

```
sample(dag09, size=5)
```

a	b	c
-0.326	1.174	1
0.552	0.619	0
-0.675	-0.113	0
0.214	0.917	1
0.311	-0.223	0

Each row in the sample is one trial; in each trial, the node's formula sets the value for that node. For example, the formula might use the values of other nodes as input. Alternatively, the formula might specify that the node is exogenous, without input from any other nodes.

Models can be trained on the simulated data using the same techniques as for any other data. To illustrate, here we generate a sample of size $n = 10,000$, then fit the model `c ~ a + b` and summarize by taking the coefficients.

```
sample(dag09, size=10000) %>%
  lm(c ~ a + b, data = .) %>%
  confint()
```

	lwr	upr
(Intercept)	0.500	0.513
a	0.193	0.206
b	0.295	0.308

The coefficients on `a` and `b` are inconsistent with the `dag09` formulas. This discrepancy suggests the existence of a flaw in the modeling technique. In the following box, we demonstrate another modeling technique that can do the job.

Demonstration: Modeling binomial variables

Keep in mind that this is just a demonstration. There is no need to master (or even understand) the calculations presented in this box.

The printed version of `dag09` shows that the value of node `c` is a linear combination of `a` and `b` converted into a zero-one, binomial value. Unfortunately, the linear modeling trainer, `lm()`, is not well-tuned to work with binomial data. Another modeling technique, "logistic regression," does a better job. The `glm()` function trains logistic regression models on data.

```
sample(dag09, size=10000) %>%
  glm(c ~ a + b, data = ., family="binomial") %>%
  confint()
```

	lwr	upr
(Intercept)	-0.03	0.096
a	1.87	2.060
b	2.97	3.231

When we use the appropriate modeling technique, we can, in this case, recover the coefficients in the DAG formula: 2 for `a` and 3 for `b`.

⚠ Reality check: DAGs and data

DAGs represent hypotheses about the connections between variables in the real world. They are a kind of scratchpad for constructing alternative scenarios and, as seen in Lesson 9, thinking about how models might go wrong in the face of a plausible alternative causal mechanism.

In this book, we extend the use of DAGs beyond their scope in professional statistics; we use them as simulations from which we can generate data. Such simulations provide one way to learn about statistical methodology.

DAGs are aides to reasoning, scratchpads that help us play out the consequences of our hypotheses about possible real-world mechanisms. However, take caution to distinguish data from DAG simulations from data from reality.

Finding out about the real world requires collecting data from the real world. The proper role of DAGs in real work is to guide model building **from real data**.

In this course, we sample from DAGs to learn statistical techniques. But never to make claims about real-world phenomena.

2.3 Samples, summaries, and samples of summaries

Beginners sometimes think that each row in a data frame is a sample. Better to say that each row is a "**specimen**." A "**sample**" is a collection of specimens, the set of rows in a data frame.

The "**sample size**" is the number of rows. "**Sampling**" is the process of collecting the specimens to be put into the data frame.

The following command illustrates computing a summary of a sample from `dag09`.

```
sample(dag09, size=10000) %>%
  glm(c ~ a + b, data = ., family="binomial") %>%
  confint()
```

	lwr	upr
(Intercept)	-0.074	0.049
a	1.902	2.091
b	2.812	3.055

An essential question in statistics is how the summary depends on the incidental specifics of a particular sample. DAGs provide a convenient way to address this question since we can generate multiple samples from the same DAG, summarize each, and compare those summaries.

To generate a sample of summaries, re-run many trials of the summary. The `do()` function automates this process, accumulating the results from the trials in a single data frame: a "**sample of summaries**." We will use `do()` mostly in demonstrations.

⚠ Demonstration: Conducting many trials with `do()`

In this demonstration, we will revisit a model used earlier in this Lesson to see how much the coefficients vary from one sample to another. Each trial consists of drawing a sample from `dag09`, training a model, and summarizing

with the model coefficients. Curly braces (`{` and `}`) surround the commands needed for an individual trial. Preceding the curly braces, we have placed `do(5) *`. This instruction causes the trial to be repeated five times.

```
do(5) * {
  sample(dag09, size=10000) %>%
    glm(c ~ a + b, data = ., family="binomial") %>%
    coef()
}
```

Intercept	a	b
-0.013	2.04	3.05
-0.010	1.99	3.08
0.028	1.94	3.04
-0.035	1.97	2.97
-0.099	1.98	2.97

The five trials are collected together by `do()` into the five rows of a single data frame. Such a data frame can be considered a "**sample of summaries**."

One of the things we will do with a "sample of summaries" is to … wait for it … summarize it. For instance, in the following code chunk, a sample of 40 summaries is stored under the name `Trials`. Then we will summarize `Trials`, in this case, to see how much the values of the `a` and `b` coefficients vary from trial to trial.

```
Trials <- do(40) * {
  sample(dag09, size=10000) %>%
    glm(c ~ a + b, data = ., family="binomial") %>%
    coef()
}
Trials %>%
  summarize(mean_a = mean(a), spread_a = sd(a),
            mean_b = mean(b), spread_b = sd(b))
```

mean_a	spread_a	mean_b	spread_b
2	0.044	3.01	0.064

The result of summarizing the trials is a "summary of a sample of summaries." This phrase is admittedly awkward, but we will use this technique often: summarizing trials, where each trial is a "summary of a sample" Often, the clue will be the use of `do()`, which repeats trials as many times as you ask.

⚠ For the statistically experienced reader

Warning! This box contains mathematical formulas that are **not needed for the course**. The formulas might interest mathematically inclined statistics instructors, but others can skip this material.
The algorithm for the variance described previously is not used by any statistical software package; there are much faster ways to arrive at the result. One way to see this is to compare the traditional formula for the variance to the formula version of the above algorithm:

$$\underbrace{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}_{\text{traditional}} \quad \text{versus} \quad \underbrace{\frac{1}{2n}\sum_{i=1}^{n}\left[\frac{1}{n-1}\sum_{j\neq i}(x_i-x_j)^2\right]}_{\text{our algorithm}}$$

The inefficiency of the algorithm stems from the double sum. The advantage of the algorithm is conceptual and two-fold:

i. The $n-1$ in the formula for the variance comes from the $j \neq i$ in the inner sum. Why not $\sum_j = 1^n$? Because that would put n zeros on the roll and bias the result downward. We want to average the square distance between each value and *every other value*.

ii. There is no need to introduce the mean $\bar{x}$ of the values. Of course, $\bar{x}$ is easy and fast to calculate, so

there is no numerical reason to avoid using it in the calculation. There is, however, a philosophical reason based on Stephen J. Gould's observation, quoted at the start of this Lesson: "Variation is the hard reality. Means are the abstractions."

The usual definition given for the variance uses the mean as the reference point for calculating deviations. This use imbues the mean with an unjustified veneer of reality. The pairwise square difference algorithm demonstrates that the mean is not needed to calculate the variance; it is just, as Gould says, a mathematical abstraction.
There is a factor $\frac{1}{2}$ in the formula for the variance based on the pairwise square-differences. The $\frac{1}{2}$ may seem inelegant to some readers. Stripping out the $\frac{1}{2}$, the quantity has a name dating from at least 1885: the **modulus**. Writing the modulus as $c = 2s^2$ makes the formula for the gaussian function cleaner than the formula usually given:

$$\underbrace{\frac{1}{\sqrt{2\pi s^2}}\exp\left[-\frac{(x-m)^2}{2s^2}\right]}_{\text{using standard deviation}} \quad \text{vs} \quad \underbrace{\frac{1}{\sqrt{\pi c^2}}\exp\left[-\frac{(x-m)^2}{c^2}\right]}_{\text{using modulus}}$$

In mathematical and scientific nomenclature, "modulus" is roughly synonymous with "size." For example, in algebra, $|x|$ is the modulus of x. "Modulus" is less off-putting than "standard deviation" and less redolent of the unfortunate early ties between statistics and eugenics. Looking back, perhaps "modulus" would have been the better choice for parameterizing the gaussian.

3 Signal and noise

Imagine being transported back to June 1940. The family is in the living room, sitting around the radio console, waiting for it to warm up. The news today is from Europe, the surrender of the French in the face of the German invasion. Press the play button and listen to recording #103.

You may have to scroll down to see the play button and the recordings.

The spoken words from the recording are discernable despite the hiss and clicks of the background noise. The situation is similar to a conversation in a sports stadium. The crowd is loud, so the speaker has to shout. The listener ignores the noise (unless it is too loud) and recovers the shouted words.

Engineers and others make a distinction between **signal** and noise. The engineer aims to separate the signal from the noise. That aim applies to statistics as well.

There are many sources of noise in data; every variable has its own story, part of which is noise from measurement errors and recording blunders. For instance, economists use national statistics, like GDP, even though the definition is arbitrary (a Hurricane can raise GDP!), and early reports are invariably corrected a few months later. Historians go back to original documents, but inevitably many of the documents have been lost or destroyed: a source of noise. Even in elections where, in principle, counting is straightforward, the voters' intentions are measured imperfectly due to "hanging chads," "butterfly ballots," broken voting machines, spoiled ballots, and so on.

The statistical thinker is well advised to know about the sources of noise in the system she is studying. Analysis of data will be better the more the modeler knows about how measurements are made and data collected.

> **i** Noise in hiring
>
> The author has, on several occasions, testified in legal hearings as a statistical expert. In one case, the US Department of Labor audited the records of a contractor with several hundred employees and high employee turnover. The records led the Department to bring suit

against the contractor for discriminating against Hispanics. The hiring records showed that many Hispanics applied for jobs; the company hired none. An open-and-shut case.
The lawyers for the defense asked me, the statistical expert, to review the findings from the Department of Labor. The lawyers thought they were asking me to check the arithmetic in the hiring spreadsheets. As a statistical thinker, I know that arithmetic is only part of the story; the origin of the data is critically important. So I asked for the complete files on all applicants and hires the previous year.
The spreadsheet files and the paper job applications were in accord; there were many Hispanic applicants. But the data on the paper job application form was not always consistent with the data on hiring spreadsheets. It turned out that whenever an applicant was hired, the contractor (per regulation) got a report on that person from the state police. The report returned by the state police had only two available race/ethnicities: white and Black. The contractor's personnel office filled in the hired-worker spreadsheet based on the state police report. So all the Hispanic applicants who were hired had been transformed into white or Black by the state police. Noise.

3.1 Signal and noise

To illustrate the statistical problem of signal and noise, let us turn to a DAG simulation: `dag01`. Here's a sample from `dag01`:

```
Tiny <- sample(dag01, size=2)
```

x	y
-0.326	2.84
0.552	5.04

The DAG simulation implements a relationship between `x` and `y`. In statistics, this *relationship* is the signal.

> Look at the 2-row sample (**?@tbl-tiny-dag01**) from the DAG and guess what the relationship might be.

Any of an infinite number of possible relationships could account for the `x` and `y` data. The noise reduction problem of statistics is to make a guess that is as good as possible. Unfortunately, for a sample with $n = 2$, as "good as possible" is not very good!

More data—a bigger sample—gives us a better shot at revealing the relationship hidden by the noise. **?@tbl-small-dag01** shows a sample of size $n = 10$:

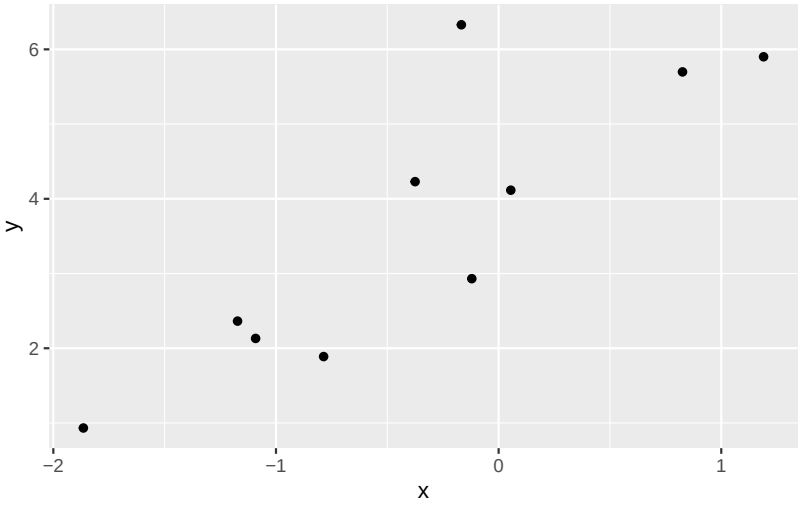
```
Small <- sample(dag01, size=10)
```

x	y
-0.786	1.889
0.055	4.115
-1.173	2.363
-0.167	6.329
-1.865	0.933
-0.120	2.931
0.826	5.698
1.190	5.901
-1.091	2.131
-0.375	4.230

A careful perusal of the `Small` sample suggests some patterns. `x` is never larger than about 2 in magnitude and can be positive or negative. `y` is always positive. Furthermore, when `x` is negative, the corresponding `y` value is relatively small compared to the `y` values for positive `x`.

A sample of size $n = 10$ provides more information than a sample of $n = 2$, so we can make a more informed guess about the relationship between variables `x` and `y`.

Human cognition is not well suited to looking at long columns of numbers. Often, we can make better use of our natural human talents by translating the sample into a graphic:



Collecting more data can make the relationship clearer. Figure 7 displays an $n = 10,000$ sample.

```
Large <- sample(dag01, size=10000)
```

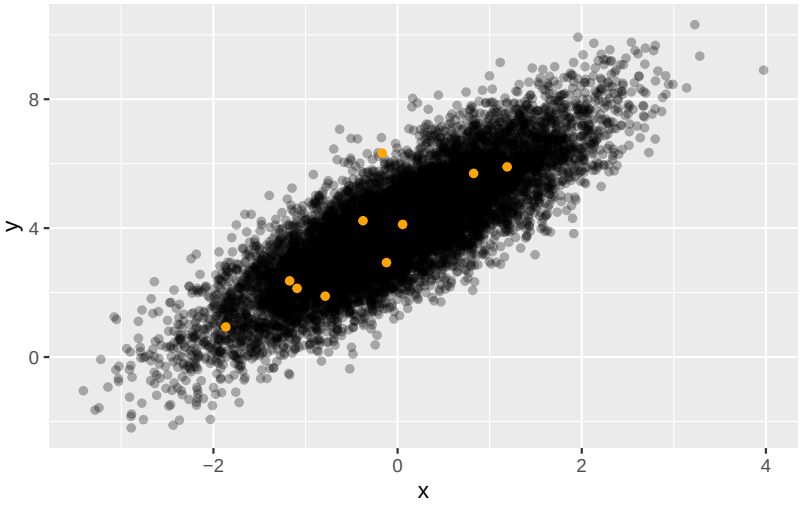


Figure 7: With $n = 10,000$ rows, the relationship between `x` and `y` is evident graphically. (The original `Small` sample is shown in orange.)

There are many possible ways to describe the x-y relationship in Figure 7. For instance, we can see that when `x` is positive, `y` is almost always greater than 4, but for negative `x`, the value of

y tends to be less than 4. Such a description might be apt for some purposes, but in these Lessons, we describe relationships by fitting models to data. For example, the following command uses the small sample (n=10) as the training data:

```
lm(y ~ x, data = Small) %>% coef()  # n = 10 sample
```

```
(Intercept)           x
       4.26        1.74
```

The coefficients provide the information needed to construct the formula for the *model function*:

$$y = 4.26 + 1.74x \ .$$

This formula is a guess of the *signal*—the relationship between the two variables in `dag01`. Unfortunately, the formula tells us nothing about the noise obscuring the signal nor how good the guess is.

The model coefficients produced by training the model on a much larger sample will presumably be a better guess:

```
lm(y ~ x, data = Large) %>% coef() #  n = 10,000 sample
```

```
(Intercept)           x
       4.01        1.50
```

Unfortunately, we cannot tell from the coefficients how good the guess is.

Luckily for us, since the data are a simulation from a DAG, we can see what the coefficients *should be* as well as the origin of the noise mixing in with the signal.

```
print(dag01)
```

```
x ~ exo()
y ~ 1.5 * x + 4 + exo()
```

The `Large` sample produced coefficients much closer than the `Small` sample to the mechanism in the DAG. The idea that larger samples lead to better accuracy has been appreciated since the 16th century and now has the prestige of being a "Law": the Law of Large Numbers.

However, "better accuracy" does not tell us whether the accuracy suffices for any given purpose. The model filters out some of the noise. However, the model coefficients still display a noisy legacy.

The challenge of real-world data is that we cannot open the black box that generated the data; all we have is the data! So how can we tell whether the data at hand are sufficient for giving a usefully accurate description of the actual relationships?

The key to the puzzle is the variation *within* the sample.

3.2 Measuring variation

Lesson 2 introduced the standard way to measure variation in a single variable: the **variance** or its square root, the **standard deviation**. For instance, we can measure the variation in the variables from the `Large` sample using `sd()` and `var()`:

```
Large %>%
  summarize(sx = sd(x), sy = sd(y), vx = var(x), vy = var(y))
```

sx	sy	vx	vy
0.983	1.78	0.966	3.17

According to the standard deviation, the size of the `x` variation is about 1. The size of the `y` variation is about 1.7.

Look again at the formulas that compose `dag01`:

```
print(dag01)
```

```
x ~ exo()
y ~ 1.5 * x + 4 + exo()
```

The formula for `x` shows that `x` is endogenous, its values coming from a random number generator, `exo()`, which, unless otherwise specified, generates noise of size 1.

As for `y`, the formula includes two sources of variation:

1. The part of `y` determined by `x`, that is $y = \mathbf{1.5x} + 4 + \text{exo}()$
2. The noise added directly into `y`, that is $y = 1.5\mathbf{x} + 4 + \mathbf{exo}()$

The 4 in the formula does not add any *variation* to `y`; it is just a number.

We already know that `exo()` generates random noise of size 1. So the amount of variation contributed by the `+ exo()` term in the DAG formula is 1. The remaining variation is contributed by `1.5 * x`. The variation in `x` is 1 (coming from the `exo()` in the formula for `x`). A reasonable guess is that `1.5 * x` will have 1.5 times the variation in `x`. So, the variation contributed by the `1.5 * x` component is 1.5. The overall variation in `y` is the sum of the variations contributed by the individual components. This suggests that the variation in `y` should be

$$\underbrace{1}_{\text{from exo()}} + \underbrace{1.5}_{\text{from 1.5 x}} = \underbrace{2.5}_{\text{overall variation in y}} .$$

Simple addition! Unfortunately, the result is wrong. In the previous summary of the `Large`, we measured the overall variation in `y` as about 1.72.

The *variance* will give a better accounting than the standard deviation. Recall that `exo()` generates variation whose standard deviation is 1, so the variance from `exo()` is $1^2 = 1$. Since `x` comes entirely from `exo()`, the variance of `x` is 1. So is the variance of the `exo()` component of `y`.

Turn to the `1.5 * x` component of `y`. Since variances involve squares, the variance of `1.5 * x` works out to be $1.5^2 \text{ var}(x) = 2.25$. Adding up the variances from the two components of `y` gives

$$\text{var}(y) = \underbrace{2.25}_{\text{from 1.5 exo()}} + \underbrace{1}_{\text{from exo()}} = 3.25$$

This result that the variance of `y` is 3.25 closely matches what we found in summarizing the `y` data generated by the DAG.

The lesson here: When adding two sources of variation, the variances of the individual sources add to form the overall variance of the sum. Just like $A^2 + B^2 = C^2$.

3.3 DAGs from data

In modeling data from `dag01` we could recover a good approximation to the formula for `y`.

```
Large %>%
  lm(y ~ x, data = .) %>%
  coef()
```

```
(Intercept)           x
       4.01        1.50
```

A DAG describes the causal links between variables. Data modeling reveals the formula implementing the causal link in `dag01`. Nevertheless, it is wrong to think we can determine the DAG that generated the data from the data alone. Only if we already know the structure of the data-generation DAG can we recover the mechanism inside that DAG. For instance, another statistical thinker might believe that the causal mechanism behind the data is `y` causing `x`. Based on this assumption, she also can find the mechanism inside her hypothesized DAG:

```
sample(dag01, size=10000) %>%
  lm(x ~ y, data = .) %>%
  coef()
```

```
(Intercept)           y
     -1.826       0.456
```

A DAG is a **hypothesis**, a statement that might or might not be true. DAGs are part of the statistical apparatus for thinking responsibly about **causality**. Use a DAG—or, potentially, multiple DAGs—when the issue of what causes what is relevant to the purpose behind the work.

When there are only two variables involved in the system under consideration—we will call them X and Y for simplicity—there are only two possible DAGs:

$$X \to Y \qquad \text{and} \qquad X \leftarrow Y$$

Our understanding of the world sometimes allows us to focus on one of these and not the other. Example: Does the rooster crowing cause the sun to rise, or does the rising sun cause the rooster to crow?

Beyond the two DAGs $X \to Y$ and $X \leftarrow Y$, additional DAG possibilities can account for the relationship between X and Y. For instance, if we introduce another variable, C, located between X and Y, four other DAGs need to be considered:

$$X \to C \to Y \quad \text{and} \quad X \leftarrow C \leftarrow Y \quad \text{and} \quad X \leftarrow C \to Y \quad \text{and} \quad X \to C \leftarrow Y$$

There are many other DAG configurations involving three variables. To keep things simple, we will restrict things to DAGs where X might or might not cause Y, but Y never causes X.[5] Figure 8 shows the ten configurations of 3-variable DAGs where Y does not cause X.

With the conceptual tool of DAGs, the statistical thinker can consider multiple possibilities for what might cause what. Sometimes she can discard some of the possibilities based on common sense. (Think: roosters and the sun.) However, in other settings, there may be possibilities that she does not favor but might be plausible to other people. In Lesson 9, we will explore how each configuration of DAG has implications for which model formulas can or cannot reveal the hypothesized causal mechanism.

[5]We do not lose generality by this restriction. The modeler gets to choose which real-world variable corresponds to X and which one to Y.

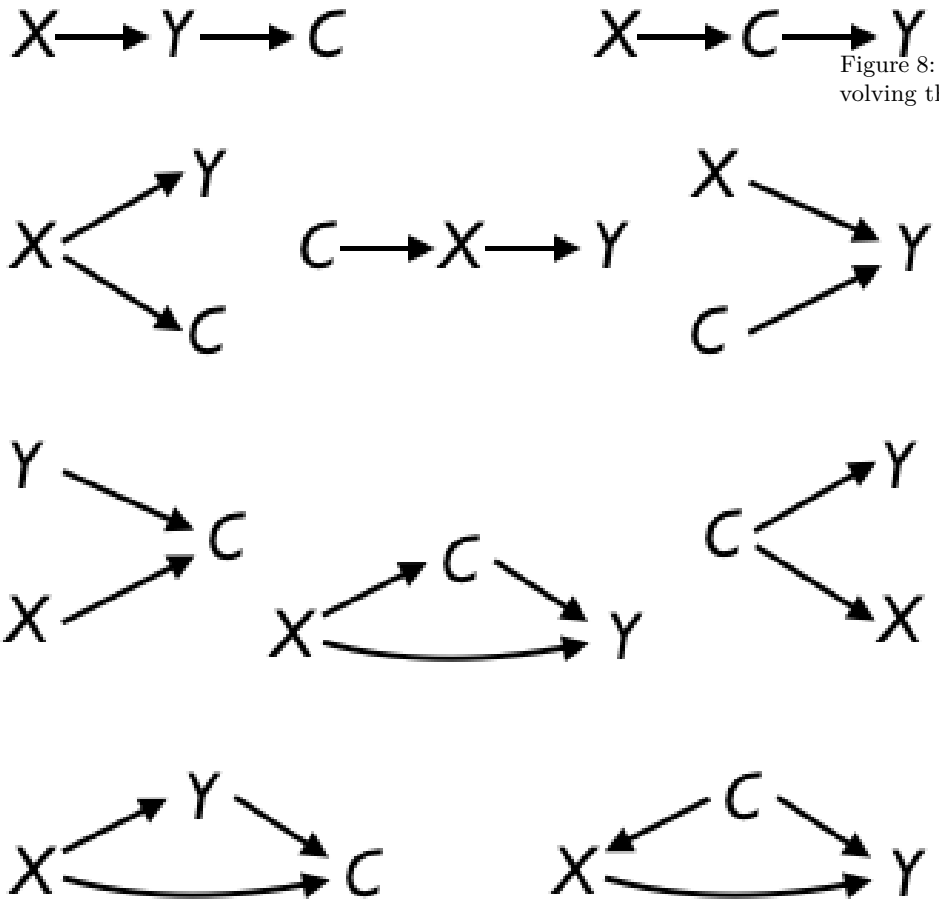


Figure 8: Ten DAG configurations involving three variables X, Y, and C.

4 Sampling and sampling variation

We use "sample" as a near synonym for "data frame." However, a data frame may not be a sample but contains a row for every possible unit of observation. Such a complete enumeration—the inventory records of a merchant, the records kept of student grades by the school registrar—has a technical name: a "**census**." Famously, many countries conduct a census of the population in which they try to record every resident of the country. For example, the US, UK, and China carry out a census every ten years.

In a typical setting, it is unfeasible to record every possible unit of observation.[6] Such incomplete records constitute a "**sample**." One of the great successes of statistics is the means to draw useful information from a sample, at least when the sample is collected correctly.

Sampling is called for when we want to find out about a large group but lack time, energy, money, or the other resources needed to contact every group member. For instance, France collects samples at short intervals to collect up-to-date data while staying within a budget. The name used for the process—*recensement en continu* or "rolling census"—signals the intent. Over several years, the French rolling census contacts about 70% of the population.

Sometimes, as in quality control in manufacturing, the measurement process is destructive: the measurement process consumes the item. In a destructive measurement situation, it would be pointless to measure every single item. Instead, a sample will have to do.

4.1 Sampling bias

Collecting a reliable sample is usually considerable work. An ideal is the "simple random sample" (SRS), where all of the items are available, but only some are selected—completely at random—for recording as data. Undertaking an SRS requires assembling a "sampling frame," essentially a census. Then,

[6] Even a population "census" inevitably leaves out some individuals.

with the sampling frame in hand, a computer or throws of the dice can accomplish the random selection for the sample.

Understandably, if a census is unfeasible, constructing a perfect sampling frame is hardly less so. In practice, the sample is assembled by randomly dialing phone numbers or taking every 10th visitor to a clinic or similar means. Unlike genuinely random samples, the samples created by these practical methods are not necessarily representative of the larger group. For instance, many people will not answer a phone call from a stranger; such people are underrepresented in the sample. Similarly, the people who can get to the clinic may be healthier than those who cannot. Such unrepresentativeness is called "**sampling bias**."

Professional work, such as collecting unemployment data, often requires government-level resources. Assembling representative samples uses specialized statistical techniques such as stratification and weighting of the results. We will not cover the specialized techniques in this introductory course, even though they are essential in creating representative samples. The table of contents of a classic text, William Cochran's *Sampling techniques* shows what is involved.

All statistical thinkers, whether expert in sampling techniques or not, should be aware of factors that can bias a sample away from being representative. In political polls, many (most?) people will not respond to the questions. If this non-response stems from, for example, an expectation that the response will be unpopular, then the poll sample will not adequately reflect unpopular opinions. Such **non-response bias** can be significant, even overwhelming, in surveys.

Survival bias plays a role in many settings. The `mosaicData::TenMileRace` data frame provides an example, recording the running times of 8636 participants in a 10-mile road race and including information about each runner's age. Can such data carry information about changes in running performance as people age? The data frame includes runners aged 10 to 87. Nevertheless, a model of running time as a function of age from this data frame is seriously biased. The reason? As people age, casual runners tend to drop out of such races. So the older runners are skewed toward higher

performance. (We can see this by taking a different approach to the sample: collecting data over multiple years and tracking individual runners as they age.

i Examples: Returned to base

An inspiring story about dealing with survival bias comes from a World War II study of the damage sustained by bombers due to enemy guns. The sample, by necessity, included only those bombers that survived the mission and returned to base. The holes in those surviving bombers tell a story of survival bias. Shell holes on the surviving planes were clustered in certain areas, as depicted in Figure 9. The clustering stems from survivor bias. The unfortunate planes hit in the middle of the wings, cockpit, engines, and the back of the fuselage did not return to base. Shell hits in those areas never made it into the record.

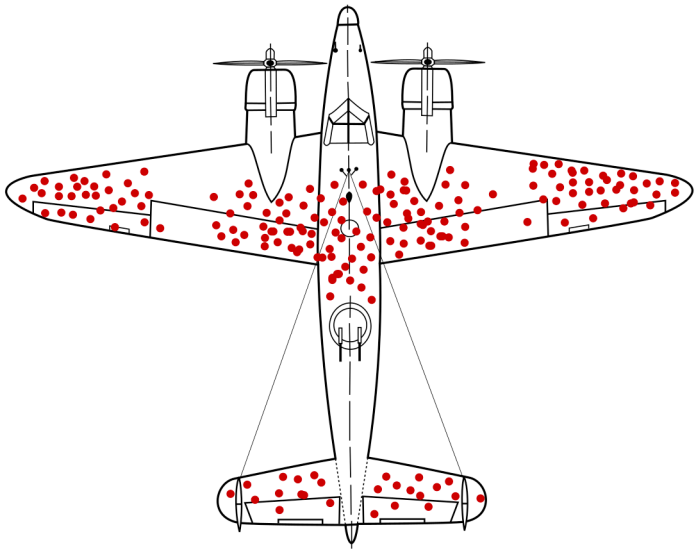


Figure 9: An illustration of shell-hole locations in planes that returned to base. Source: Wikipedia

4.2 Measuring sampling variation

Sampling variation is a form of noise. Unlike some other forms of noise, modeling cannot filter out sampling variation or reduce

its magnitude. Sampling variation is easiest to see by collecting multiple samples from the same source and summarizing each one. The summaries likely will vary from sample to sample: sampling variation.

Typically, the data frame at hand is our only sample. With no other samples to compare it to, it may seem impossible to measure sampling variation. In this Lesson, we will use simulations from DAGs to study sampling variation. DAG simulations are suited to this because we can effortlessly collect as many samples as we wish from a DAG. In Lesson 5, we will use the knowledge gained from the simulations to see how to measure sampling variation even when there is only one sample.

In the spirit of starting simply, we return to `dag01`. This DAG is `x` ⟶ `y`. The causal formula setting the value of `y` is `y ~ 4 + 1.5 * x + exo()`.

It is crucial to remember that sampling variation is not about the row-to-row variation in a single sample. Rather, it is about the variation in the summary from one sample to another. So our initial process for exploring sampling variation will be to carry out many trials, each trial being a summary of a sample.

4.3 Demonstration: Samples and specimens

To illustrate, here is one trial using a sample of $n = 25$ and a simple model, `y ~ 1`.

```
Sample <- sample(dag01, size=25)
Sample %>%
  lm(y ~ 1, data = .) %>%
  coef()
```

```
(Intercept)
       3.65
```

We cannot see sampling variation directly in the above result because there is only one trial. The sampling variation becomes

evident when we run *many* trials. In each trial, a new sample (of size $n = 25$ is taken and summarized.)

```
Trials <- do(100) * {
  Sample <- sample(dag01, size=25)
  Sample %>%
    lm(y ~ 1, data = .) %>%
    coef()
}
```

`Trials` is a *sample of summaries*. (See Lesson 2). The row-to-row variation in`Trials` comes from sampling variation. We can *summarize* the variation in the *sample of summaries*. As always, our standard measure of variation is the standard deviation (or, equivalently, variance):

```
Trials %>%
  summarize(se = sd(Intercept))
```

se
0.357

This summary quantity, which we have named `se`, has a technical name in statistics: the **standard error**. The standard error is an ordinary standard deviation in a particular context: the standard deviation of a sample of summaries. The words **standard error** should be followed by a description of the summary and the size of the individual samples involved. Here it would be, "The standard error of the Intercept coefficient from a sample of size $n = 25$ is around 0.36."

It is easy to confuse "standard error" with "standard deviation." Adding to the potential confusion is another related term, the "margin of error." To avoid this confusion, we will tend to use an *interval* description of the sampling variation called the "**confidence interval**." However, for the present, we will continue with the standard error, sometimes written SE for short.

Table 1: Trials for seeing sampling variation.

Intercept
3.75
4.55
3.95
4.01
3.58
4.01
4.26
3.87
3.92
3.93
3.73
4.09
3.65
4.02
3.51
4.28
3.86
3.84
4.34
4.09
4.67
3.59
3.73
3.88
4.28

Table 2: Results of repeating the sampling variability trials for samples of varying sizes.

n	se
25	0.360
100	0.190
400	0.091
1600	0.043
6400	0.023
25000	0.011
100000	0.006

4.4 SE depends on the sample size

We found an SE of 0.36 on the Intercept in a sample of size $n = 25$. We can see how the SE depends on sample size by repeating the trials for several different sizes, say, $n = 25$, 100, 400, 1600, 6400, 25,000, and 100,000.

The following command estimates the SE a sample of size 400: ::: {.cell}

```
Trials <- do(1000) * {
  Sample <- sample(dag01, size=25)
  Sample %>%
    lm(y ~ 1, data = .) %>%
    coef()
}
Trials %>% summarize(se400 = sd(Intercept))
```

se400
0.354

:::

We repeated this process for each of the other sample sizes. Table 2 reports the results.

There is a pattern in Table 2. Every time we double n, the standard error goes down by a factor of 2, that is, $\sqrt{4}$. (The pattern is not exact because there is also sampling variation in the trials.)

Conclusion: The larger the sample size, the smaller the SE. For a sample size of n, the SE will be proportional to $1/\sqrt{n}$.

4.5 The confidence interval

The "confidence interval" is a more user-friendly format than SE for describing the amount of sampling variation. Being an interval, write it either as [lower, upper] or center±half-width. These styles are equivalent; both styles are correct. (The preferred style can depend on the field or the journal publishing the report.)

In practice, confidence intervals are calculated using special-purpose software such as the `confint()` function, for instance:

```
Hill_racing %>%
  lm(time ~ distance + climb, data=.) %>%
  confint()
```

	lwr	upr
(Intercept)	-533.43	-406.52
distance	246.39	261.23
climb	2.49	2.73

Notice that there is a separate confidence interval for each model coefficient. The sampling variation is essentially the same, but that variation appears different when translated to the various coefficients' units.

> ⚠ Demonstration: How many digits?
>
> The confidence intervals on the model `time ~ distance + climb`, report the results to many digits. Such a report is appropriate for further calculations that might need do-

ing, but it is usually not appropriate for a human reader. To know how many digits are worth reporting to humans, look toward the standard error. The standard error is a part of a different kind of summary of a model: the "regression report." We will only need to look at regression reports in the last few Lessons of the course. Here we want to point out how many digits are worth reporting to humans. That requires looking at the standard error itself.

Previously, we looked at the confidence intervals on coefficients from the `Hill_racing` model. Now we look at the regression summary, which contains the information on sampling variation in a different format.

```
Hill_racing %>%
  lm(time ~ distance + climb, data=.) %>%
  regression_summary()
```

term	estimate	std.error	statistic	p.value
(Intercept)	-469.98	32.358	-14.5	0
distance	253.81	3.784	67.1	0
climb	2.61	0.059	43.9	0

Each coefficient's standard error appears in the `std.error` column of the regression summary.

For the human reader, only the first two significant digits of the standard error are worth reporting. (This is true regardless of the data and model design.) Here, the SE is 32 for the Intercept, 3.8 for the distance coefficient, and 0.059 for the climb coefficient. The confidence interval will be the coefficient (column labeled `estimate`) plus or minus "twice" the `std.error`. It is appropriate to round the confidence interval (for a human reader) to the first two significant digits of the standard error.

For example, the confidence interval on the distance coefficient will be $253.808295 \pm 2 \times 3.78433220$. Keep only the digits before the first two significant digits of the SE, so the reported interval can be 253.8 ± 3.8.

5 Estimating sampling variation from a single sample

Lesson 3 introduced separating data into components: signal and noise. The *signal* is a summary of the data that tells us something we want to know. Often, the signal will be one or more coefficients from a regression report, but it might be something as simple as the mean, median, or standard deviation of a variable in a data frame.

The *noise* comes into the data from various sources: e.g., error in measurement or a data-entry blunder. Another source of noise is omnipresent (except in a perfect census): sampling variation as discussed in Lesson 4. Sampling variation arises because the particular sample we happen to be working with is, so far as sampling is concerned, the play of luck. If we had happened to select another sample, the results would be different.

The Greek philosopher Heraclitus (c. 500 BC) said, "You can't step into the same river twice." A step into a river might be at the same place on the bank, but the water flowing by will be different. A data sample is like collecting water from a river or lake using a dipper. Imagine ten people standing side by side on the shore of a lake, each person dipping into the water to acquire a specimen and making one or more measurements from the specimen, for instance, the temperature, pH, and bacteria count. Each person collects a sample—that is, a series of specimens. These might be taken right after one another or by some protocol, say a weekly tracking of lake conditions over time.

The ten people are each doing the same thing in approximately the same place and time, but each person's sample will be different, even if only by a little bit. That sample-to-sample variation will be noise.

If the ten people were fishing, each specimen would be the catch from one cast of the rod. Typically this is just an empty hook, lake weeds, or a stick, but sometimes it will be a fish. Each fisherman will have a sample at the end of the fishing day. These samples will not be identical; the fishermen's varying skills (or

luck) will produce different results. That is sampling variation. To fishermen, the question of interest, the signal they want to measure, might be, "How good is the fishing today?" The fishermen's varying catches are the sampling variation.

In Lesson 3, we repeated trials over and over again to gain some feeling for sampling variation. Each trial consisted of collecting a sample and summarizing it. The individual trial is a summary of a sample. Then, to quantify the sampling variation, we summarized the set of individual trial results using the standard measure of variation: the standard deviation.

It is time to take off the DAG simulation training wheels and measure sampling variation from a *single* data frame. Our first approach will be to turn the single sample into several smaller samples: subsampling. Later, we will turn to another technique, resampling, which draws a sample of full size from the data frame.

5.1 Subsampling

To "subsample" means to draw a smaller sample from a large one. "Small" and "large" are relative. For our example, we turn to the `TenMileRace` data frame containing the record of thousands of runners' times in a race, along with basic information about each runner. There are many ways we could summarize `TenMileRace`. Any summary would do for the example. We will summarize the relationship between the runners' ages and their start-to-finish times (variable `net`), that is, `net ~ age`. To avoid the complexity of a runner's improvement with age followed by a decline, we will limit the study to people over 40.

```
TenMileRace %>% filter(age > 40) %>%
  lm(net ~ age, data = .) %>% coef()
```

```
(Intercept)         age
     4278.2        28.1
```

The units of `net` are seconds, and the units of `age` are years. The model coefficient on `age` tells us how the `net` time changes for each additional year of `age`: seconds per year. Using the entire data frame, we see that the time to run the race gets longer by about 28 seconds per year. So a 45-year-old runner who completed this year's 10-mile race in 3900 seconds (about 9.2 mph, a pretty good pace!) might expect that, in ten years, when she is 55 years old, her time will be longer by 280 seconds.

It would be asinine to report the ten-year change as 281.3517 seconds. The runner's time ten years from now will be influenced by the weather, crowding, the course conditions, whether she finds a good pace runner, the training regime, improvements in shoe technology, injuries, and illnesses, among other factors. There is little or nothing we can say from the `TenMileRace` data about such factors.

There's also sampling variation. There are 2898 people older than 40 in the `TenMileRace` data frame. The way the data was collected (radio-frequency interrogation of a dongle on the runner's shoe) suggests that the data is a census of finishers. However, it is also fair to treat it as a sample of the kind of people who run such races. People might have been interested in running but had a schedule conflict, lived too far away, or missed their train to the start line in the city.

We see sampling variation by comparing multiple samples. To create those multiple samples from `TenMileRace`, we will draw, at random, subsamples of, say, one-tenth the size of the whole, that is, $n = 290$

```
Over40 <- TenMileRace %>% filter(age > 40)
lm(time ~ age, data = Over40 %>% sample(size=290)) %>% coef()
```

```
(Intercept)         age
     4896.0        19.7
```

```
lm(time ~ age, data = Over40 %>% sample(size=290)) %>% coef()
```

```
(Intercept)         age
     4073.0        36.7
```

The age coefficients from these two subsampling trials differ one from the other by about 0.5 seconds. To get a more systematic view, run more trials:

```
# a sample of summaries
Trials <- do(1000) * {
  lm(time ~ age, data = sample(Over40, size=290)) %>% coef()
}
# a summary of the sample of summaries
Trials %>%
  dplyr::summarize(se = sd(age))
```

se
8.97

We used the name `se` for the summary of samples of summaries because what we have calculated is the standard error of the age coefficient from samples of size $n = 290$.

In Lesson 4 we saw that the standard error is proportional to $1/\sqrt{n}$, where n is the sample size. From the subsamples, know that the SE for $n = 290$ is about 9.0 seconds. This tells us that the SE for the full $n = 2898$ samples would be about $9.0\frac{\sqrt{290}}{\sqrt{2898}} = 2.85$.

So the interval summary of the `age` coefficient—the *confidence interval*— is

$$\underbrace{28.1}_{\text{age coef.}} \pm 2\times \underbrace{2.85}_{\text{standard error}} = 28.1 \pm \underbrace{5.6}_{\text{margin of error}} \quad \text{or, equivalently, 22.6 to 33.6}$$

5.2 Bootstrapping

There is a trick, called "**resampling**," to generate a random subsample of a data frame with the same n as the data frame: draw the new sample randomly from the original sample **with**

replacement. An example will suffice to show what the "with replacement" does:

```
example <- c(1,2,3,4,5)
# without replacement
sample(example)
```

```
[1] 1 4 3 5 2
```

```
# now, with replacement
sample(example, replace=TRUE)
```

```
[1] 2 4 3 3 5
```

```
sample(example, replace=TRUE)
```

```
[1] 3 5 4 4 4
```

```
sample(example, replace=TRUE)
```

```
[1] 1 1 2 2 3
```

```
sample(example, replace=TRUE)
```

```
[1] 4 3 1 4 5
```

The "with replacement" leads to the possibility that some values will be repeated two or more times and other values will be left out entirely.

The calculation of the SE using resampling is called "**bootstrapping**."

⚠ Demonstration: Bootstrapping the standard error

We will apply bootstrapping to find the standard error of the `age` coefficient from the model `time ~ age` fit to the `Over40` data frame.
There are two steps:

1. Run many trials, each of which fits the model `time ~ age` using `lm()`. From trial to trial, the data used for fitting is a resampling of the `Over40` data frame. The result of each trial is the coefficients from the model.

2. Summarize the trials with the standard deviation of the `age` coefficients.

```
# run many trials
Trials <- do(1000) * {
  lm(time ~ age, data = sample(Over40, replace=TRUE)) %>%
       coef()
}
# summarize the trials to find the SE
Trials %>% summarize(se = sd(age))
```

se
2.86

6 Effect size

Regression modeling and confidence intervals provide a substantial toolbox to support statistical thinking. This Lesson starts to develop methods using modeling to inform decision-making. Decision-making takes many guises: whether to administer medicine, change a budget, raise or lower a price, respond to an evolving situation, and so on.

A useful simplification splits support for decision-making into two broad categories.

1. **Making a prediction** for an individual choice. The need for predictions arises in both mundane and critical settings. For instance, an airline needs to set prices. They want to maximize revenue. Higher prices will bring in more money per seat but may reduce the number of people flying. To make the pricing decision, the airline needs a prediction about what the demand will be for those seats, which may vary based on price, day of the week, time of day, time of year, origin and destination of the flight, and so on. Another example: Merchants and social media sites must choose what products or posts to display to a viewer. Merchants have many products, and social media has many news feeds, tweets, and competing blog entries. The people who manage these websites want to promote the products or postings most likely to cause a viewer to respond. To identify viable products or postings, the site managers construct predictive models based on earlier viewers' choices. We will study prediction models in Lessons 25 and 26,
2. **Intervening** in a system. Such interventions occur on both grand scales and small: changes in government policies such as funding for preschool education or subsidies for renewable energy, closing a road to redirect traffic or opening a new highway or bus line, changing the minimum wage, etc. Before making such interventions, it is wise to know what the consequences are likely to be. Figuring this out is often a matter of understanding how the system works: what causes what. As interventions often affect multiple individuals, influencing the overall trend

of the effect across individuals might be the goal instead of predicting how each individual will be affected.

This Lesson focuses on "**effect size**," a measure of how changing an explanatory variable will play out in the response variable. Built into the previous sentence is an assumption that the explanatory variable *causes* the response variable. In Lessons 28 through 31, we will look into ways to make responsible claims about whether a connection between variables is causal. Here, we will focus on the calculation and interpretation of effect size.

6.1 Effect size: Input to output

An intervention changes something in the world. Some examples are the budget for a program, the dose of a medicine, or the fuel flow into an engine. The thing being changed is the *input*. In response, something else in the world changes, for instance, the reading ability of students, the patient's serotonin levels (a neurotransmitter), or the power output from the engine. The thing that changes in response to the change in input is called the "output."

"**Effect size**" describes the change in the output with respect to the change in the input. The simplest case is when the output is a quantitative variable. In this case, the change in the output is a difference between two numbers. The form of the effect size depends on the input type. For example, for a quantitative input, the effect size will be a *ratio*, that is, a rate. (For calculus students: the effect size is a derivative of the output with respect to the input.)

To measure an effect size from data, construct a model with the output as the response variable and the input as an explanatory variable.

> **i** Example: Fuel economy
>
> A person buying a car typically has multiple objectives in mind. Perhaps the buyer is deciding whether to order a more powerful engine. This decision has conse-

quences, including a reduction in fuel economy. The decision variable—the engine size—is the input; the fuel economy is the output.

Since both input and output are quantitative, the effect size will be a rate: change in fuel economy per change in engine size. To inform a decision, use data such as the `math300::MPG` data frame, which compares various car models. `MPG` records the engine size in terms of `displacement`, in liters. Fuel economy is listed in miles per gallon, differently for city versus highway driving.

The buyer is debating between a 2-liter and a 3-liter engine. Most driving will be in the city. To calculate the effect size, first build a model with the output (`mpg_city`) as the response variable and the input (`displacement`) as an explanatory variable.

```
Mod <- lm(mpg_city ~ displacement, data=MPG)
```

Second, evaluate that model for the range of inputs under consideration.

```
model_eval(Mod, displacement=c(2, 3))
```

displacement	.output	.lwr	.upr
2	24.0	15.9	32.1
3	20.9	12.8	29.0

The change in the input from 3 liters displacement to 2 liters leads to a change in fuel economy of $24.0 - 20.9 = -3.1$ miles per gallon. The change in displacement is $3 - 2 = 1$ liters. The effect size is the ratio between the output change and the input change. Here, that is -3.1 miles per gallon per liter.

The decision-maker may be more concerned about the cost of driving than with the miles per gallon. Then the appropriate response variable might be `EPA_fuel_cost`, denominated in dollars per year.

```
Mod2 <- lm(EPA_fuel_cost ~ displacement, data=MPG)
model_eval(Mod2, displacement=c(2, 3))
```

displacement	.output	.lwr	.upr
2	1586	1001	2171
3	1883	1297	2468

The change in output is about $300 per year. However, the change in input is still 1 liter. The effect size is, therefore, $300 per year per liter.

Some decision variables are categorical. For instance, the buyer might like the idea of an engine that automatically turns off when the car is stopped at a light or in traffic. The `start_stop` variable, which has categorical levels "Yes" and "No," records whether the car has this feature. Effect size estimation is slightly different when the input is categorical rather than quantitative. Still, build a model and compare the change in output to the change in input:

```
Mod3 <- lm(EPA_fuel_cost ~ start_stop, data=MPG)
model_eval(Mod3, start_stop=c("No", "Yes"))
```

start_stop	.output	.lwr	.upr
No	1872	916	2828
Yes	1945	989	2901

In this case, the change in output is $73 per year; the change in input is "Yes" - "No." But, of course, it is meaningless to subtract one categorical level from another. Consequently, the effect size of `start_stop` on fuel cost cannot be quantified as a ratio. So, instead, the effect size is simply the difference in the output: a $73 per year increase with the Start/Stop feature.

The statistical thinker knows to pay attention to whether a calculated result makes sense. It seems unlikely that the Start/Stop feature causes more fuel to be consumed. Was

there an error? Perhaps we did the subtraction backward? Check the report from `model_eval()` to make sure.

Here, the problem is not arithmetic. However, there is another possibility. It might be that manufacturers include the Start/Stop feature with big cars but not little ones. Then, even if Start/Stop might save gas when everything else is held constant, because the big cars use more fuel than little cars, it only *appears* that Start/Stop hurts fuel economy. This theory is, at this point, speculation: a hypothesis. Such a mixture of effects—big versus small car mixed with availability of Start/Stop—is called "**confounding**." In Lessons 28 through 30, we discuss identifying and dealing with possible confounding.

> ⚠ Confounding?
>
> The surprising positive effect size of the Start/Stop feature caused a double take and led us to think of ways to make sense of the result. Right now, we simply have a hypothesis that Start/Stop is associated with bigger cars. (We will check that out in a little bit.)
> The effect size of annual fuel cost with respect to engine displacement, $300 per year per liter, did not surprise us. Perhaps it should have. After all, larger vehicles tend to have larger engines. This relationship might lead to confounding between vehicle size and engine displacement. We think we are looking at engine displacement, but instead, the effect might be due to vehicle size. Again, just a hypothesis at this point. The statistical thinker knows to consider possible confounding from the start.

6.2 Categorical outputs

Sometimes the relevant effect size involves a categorical output variable. A case in point is the possible confounding of the Start/Stop feature with vehicle size. To investigate this, we should build a model with Start/Stop as the output and vehicle size as the input.

In this case, the issue of whether vehicle size causes Start/Stop is not essential. We are not concerned with the decisions made by automobile designers so much as with the possible confounding.

When the output variable is categorical, it is not reasonable to calculate the change in output as the difference in categories. As before, "Yes" - "No" is not a number. Still, there is a meaningful and helpful way to quantify a change in a categorical output.

The essential insight is quantifying the change in output in terms of probabilities. For instance, a small effect size would reflect a slight chance of the output changing from one level to another.

The appropriate model type for a categorical output is to transform the output to a zero-one variable, as introduced in Lesson 1. We will present this in a demonstration here and return to the topic more fully in Lesson 34.

⚠ Demonstration: Start/Stop and vehicle size

As described earlier, we are interested in the possibility that Start/Stop is available mainly on large, higher-fuel-consumption cars. If so, that would explain why the effect size we calculated of fuel cost with respect to Start/Stop was positive.
The model we build will have a zero-one encoding of Start/Stop as the response and the vehicle's fuel cost as the explanatory variable.

```
MPG <- MPG %>%
    mutate(has_start_stop = zero_one(start_stop, one="Yes"))
Mod4 <- lm(has_start_stop ~ EPA_fuel_cost, data = MPG)
model_eval(Mod4, EPA_fuel_cost=c(1600, 2000))
```

EPA_fuel_cost	.output	.lwr	.upr
1600	0.490	-0.489	1.47
2000	0.521	-0.458	1.50

The `.output` here is interpreted as a *probability* of `start_stop` having the value "Yes." (That is because we set `one="Yes"` in the `zero_one()` conversion.) The `model_eval()` report indicates $400 per year increase in fuel cost is associated with a three percentage point increase in the probability of a vehicle having a Start/Stop feature. That is a small effect, so we see little support for our hypothesis that Start/Stop tends to be installed on larger, more fuel-efficient vehicles.

6.3 Multiple explanatory variables

When a model has more than one explanatory variable, each has a different effect size.

As an example, consider the price of books. We have some data that might be informative, `moderndive::amazon_books`. What is the effect size of page count on price. The appropriate model here is `list_price ~ num_pages`. The effect size is easy to compute:

```
Mod1 <- lm(list_price ~ num_pages, data = moderndive::amazon_books)
model_eval(Mod1, num_pages = c(200, 400))
```

num_pages	.output	.lwr	.upr
200	15.8	-11.64	43.3
400	19.8	-7.64	47.2

We elected to compare 200-page books with 400-page books, simply because those seem like reasonable book lengths. However, the longer book costs about 4 dollars more. So the effect size, to judge from this model, is $4 divided by 200 more pages, which comes to 2 cents per page.

Another effect size is needed to address the question: Are hardcovers more expensive than paperbacks? The output is still price. But now, the input is categorical. In the `moderndive::amazon_books` data frame, the variable

`hard_paper` has levels "P" and "H." A possible model: `list_price ~ hard_paper`.

```
Mod2 <- lm(list_price ~ hard_paper, data = amazon_books)
model_eval(Mod2, hard_paper = c("P", "H"))
```

hard_paper	.output	.lwr	.upr
P	17.1	-10.62	44.9
H	22.4	-5.46	50.2

A hardcover book costs about $5.25 more than a paperback book. Since the input is categorical, there is no change of input to divide by, so the effect size is $5.25 when going from a paperback to a hardcover.

We can look at the effects of page length and cover-type separately. Instead, we can include both as explanatory variables.

```
Mod3 <- lm(list_price ~ hard_paper + num_pages, data = amazon_books)
model_eval(Mod3, hard_paper = c("P", "H"),  num_pages=c(200, 400))
```

hard_paper	num_pages	.output	.lwr	.upr
P	200	14.5	-12.64	41.7
H	200	19.5	-7.79	46.8
P	400	18.4	-8.71	45.6
H	400	23.4	-3.85	50.6

This output requires some interpretation. We have got short and long paperback books and short and long hardcover books. What should we compare to what?

The convention is to consider each of the two inputs separately and hold the other input constant when we compare.

Effect size of ***num_pages*** *on* ***list_price***. To hold `hard_paper` constant, we will compare the two rows of the `model_eval()` report that have a "P" value for `hard_paper`. The difference in output for these two rows is $3.90. The effect size divides by the change in input—200 pages—so the effect size is just under 2 cents per page. *Effect size of* ***hard_paper*** *on* ***list_price***.

This time we will hold `num_pages` constant, say at 200 pages. Comparing the corresponding rows in the `model_eval()` output shows a change in list price of $4.96 when going from paper back to hard cover. There is no special reason we decided to hold `hard_paper` constant at "P" rather than "H" or hold `num_pages` constant at 200 rather than 400. In general, the effect size will depend on the value being held constant. Choose a value that's relevant to the purpose at hand.

In these Lessons we are building models with additive effects.That is what the `+` means in, say, `list_price ~ hard_paper + num_pages`. We do this to keep the effect-size story as simple as possible. (Occasionally, you will see examples with *multiplicative* effects, called "**interactions**." The tilde expressions for such models involve `*` rather than `+`, as in `list_price ~ hard_paper + num_pages.`

6.4 Interval estimates

Statistical thinkers know that any estimate they make, including estimates of effect sizes, involves sampling variation. Consequently, give an *interval* estimate. The interval communicates to the decision-maker the uncertainty in the estimated quantity. Sophisticated decision-makers keep this uncertainty in mind, considering the range of outcomes likely from whatever use they make of effect size. On the other hand, statistically naive decision makers—even highly educated decision-makers can be statistically naive—look at the interval and sometimes ask the modeler, "Just give me a number. I don't know what to do with two numbers." Such a request might elicit a frank response: "If you don't know what to do with two numbers, you also won't know what to do with one number." Unfortunately, that kind of frankness is not often well received; a reasonable alternative is: "The interval indicates the amount of uncertainty in the result. We'll need to collect more data if you want to reduce the uncertainty." (Lesson 10 introduces a not-always-available alternative to collecting more data: building a better model!)

For the additive models that we are mainly using in these Lessons, the effect size is identical to a model coefficient. For

these models, the confidence interval on the coefficient is the confidence interval on the effect size. For instance,

```
Mod3 %>% confint()
```

	lwr	upr
(Intercept)	7.04	14.189
hard_paperH	1.56	8.357
num_pages	0.01	0.029

7 Mechanics of prediction {?@sec-lesson-25}

An effect size describes the relationship between two variables in an input/output format. Lesson 6 introduced effect size in the context of causal connections as if turning a knob to change the input will produce a change in the output. Such mechanistic connections make for a nice mental image for those considering intervening in the world but can be misleading.

First, the mere calculation of an effect size does not establish a causal connection. The statistical thinker has more work to do to justify a causal claim, as we will see in Lesson 11.

Second, owing to noise, the input/output relationship quantified by an effect size may not be evident in a single intervention, say, increasing a drug dose for any given individual patient. Instead, effect sizes are descriptions of *average* effects—trends—across a large group of individuals.

This Lesson is about *prediction*: what a model can properly say about the outcome of an individual case. Often, the setting is that we know values for some aspects of the individual but have yet to learn some other aspect of interest.

The word "prediction" suggests the future but also applies to saying what we can about an unknown current or past state. Synonyms for "prediction" include "classification" (Lessons 34 and 35), "conjecture," "guess," and "bet." The phrase "informed guess" is a good description of prediction: using available information to support decision-making about the unknown.

> **i** Example: Differential diagnosis
>
> A patient comes to an urgent-care clinic with symptoms. The healthcare professional tries to diagnose what disease or illness the patient has. A diagnosis is a prediction. The inputs to the prediction are the symptoms—neck stiffness, a tremor, and so on—as well as facts about the person, such as age, sex, occupation, and family history. The prediction output is a set of probabilities, one for each medical condition that could cause the symptoms.

> Doctors learn to perform a *differential diagnosis*, where the current set of probabilities informs the choices of additional tests and treatments. The probabilities are updated based on the information gained from the tests and treatments. This update may suggest new tests or treatments, the results of which may drive a new update. The television drama *House* provides an example of the evolving predictions of differential diagnosis in every episode.

Differential diagnosis is a cycle of prediction and action. This Lesson, however, is about the mechanics of prediction: taking what we know about an individual and producing an informed guess about what we do not yet know.

7.1 The prediction machine

A statistical prediction is the output of a kind of special-purpose machine. The inputs given to the machine are values for what we already know; the output is a value (or interval) for the as-yet-unknown aspects of the system.

There are always two phases involved in making a prediction. The first is building the prediction machine. The second phase is providing the machine with inputs for the individual case, turning the machine crank, and receiving the prediction as output.

These two phases require different sorts of data. Building the machine requires a "historical" data set that includes records from many instances where we already know two things: the values of the inputs and the observed output. The word "historical" emphasizes that the machine-building data must already have known values for each of the inputs and outputs of interest.

The evaluation phase—turning the crank of the machine—is simple. Take the input values for the individual to be predicted, put those inputs into the machine, and receive a predicted value as output. Those input values may come from pure speculation or the measured values from a specific case of interest.

7.2 Building and using the machine

To illustrate building a prediction machine, we turn to a problem first considered quantitatively in the 1880s: the relationship between parents' heights and their children's heights at adulthood. The `Galton` data frame records the heights of about 900 children, along with their parents' heights. Suppose we want to predict a child's adult height (variable name: `height`) from his or her parents' heights (`mother` and `father`). An appropriate model formula is `height ~ mother + father`. We use the model-training function`lm()` to transform the model formula and the data into a model.

```
Mod1 <- lm(height ~ mother + father, data = Galton)
```

As the output of an R function, `Mod1` is a computer object. It incorporates a variety of information organized in a somewhat complex way. There are several often-used ways to extract this information in ways that serve specific purposes.

One of the most common ways to see what is in a computer object like `Mod1` is by printing:

```
print(Mod1)
```

```
Call:
lm(formula = height ~ mother + father, data = Galton)

Coefficients:
(Intercept)       mother       father
     22.310        0.283        0.380
```

Newcomers to technical computing tend to confuse the printed form of an object with the object itself. For example, the `Mod1` object contains many components, but the printed form displays only two: the model coefficients and the command used to construct the object.

We have already used some other functions to extract information from a model object. For instance,

```
Mod1 %>% coef()
```

```
(Intercept)      mother      father
     22.310       0.283       0.380
```

```
Mod1 %>% confint()
```

	lwr	upr
(Intercept)	13.857	30.76
mother	0.187	0.38
father	0.290	0.47

```
Mod1 %>% regression_summary()
```

term	estimate	std.error	statistic	p.value
(Intercept)	22.310	4.307	5.18	0
mother	0.283	0.049	5.76	0
father	0.380	0.046	8.28	0

Another extractor, `model_eval()`, is particularly convenient for prediction. Perhaps the most common use is to provide new input values to the model function, with `model_eval()` producing a data frame showing the output of the model function. To illustrate, here is how to calculate the predicted height of the child of a 63-inch-tall mother and a 68-inch father.

```
Mod1 %>% model_eval(mother = 63, father=68)
```

mother	father	.output	.lwr	.upr
63	68	66	59.3	72.6

The data frame includes the input values along with a point value for the prediction (`.output`) and a prediction interval (`.lwr` to `.upr`).

Naturally, the predictions depend on the explanatory variables used in the model. For example, here is a model that uses only `sex` to predict the child's height:

```
Mod2 <- lm(height ~ sex, data = Galton)
Mod2 %>% model_eval(sex=c("F", "M"))
```

sex	.output	.lwr	.upr
F	64.1	59.2	69.0
M	69.2	64.3	74.2

This model includes three explanatory variables:

```
Mod3 <- lm(height ~ mother + father + sex, data = Galton)
Mod3 %>% model_eval(mother=63, father=68, sex=c("F", "M"))
```

mother	father	sex	.output	.lwr	.upr
63	68	F	63.2	59.0	67.4
63	68	M	68.4	64.2	72.7

In Lesson 8, we will look at the components that make up the prediction interval and some ways to use it.

8 Constructing a prediction interval

?@sec-lesson-25 introduced predictions in two forms:

1. a **point quantity**, the direct output of the model function.
2. the **prediction interval**, which indicates a range of likely values for the quantity being predicted.

To clarify this distinction, consider this three-step procedure that trains a model, extracts the model function, and applies the model function to inputs to generate a prediction in point-estimate form.

```
Time_mod <- lm(time ~ distance + climb, data = Hill_racing)
Time_mod_fun <- makeFun(Time_mod)
Time_mod_fun(distance=10, climb=500)
```

In the first line, `lm()` is used to train a model.

The second line, `Time_mod_fun <- makeFun(Time_mod)`, creates and names a *function* that implements the input/output relationship defined by the model.

The third line uses the ordinary parentheses notation to apply the newly created `Time_mod_fun()` to specific values of the argument, generating the corresponding output value.

```
# applying the function to arguments
Time_mod_fun(distance=10, climb=500)
```

```
   1
3373
```

In these Lessons, whenever we refer to the "model function," we mean a model translated into the form of a function. The point is to emphasize the input-to-output relationship implied by a model.

In topics like calculus, functions are the primary objects of interest. Calculus operations such as differentiation,

anti-differentiation, and zero-finding always act on functions. However, calculus software hardly ever lets one interrogate a function to find properties such as the range, domain, continuity, and asymptotes. Instead, students are expected to look at the formula of a function to deduce these properties.

In statistical modeling, model functions are *not* an object of primary interest. Why? Because there are several other properties of models are essential to interpreting the results of using a model. These properties include the residuals from model training and more abstract and advanced ones, such as the model's "degrees of freedom." People design software to construct model objects—for us, objects of class "lm"—from which these properties can be accessed and translated by software into valuable forms.

For this reason, there is no need to construct the model function explicitly. Consequently, one generally does not use the function application syntax directly as we did with `Time_mod_fun(distance=10, climb=500)`. Instead, one invokes the model function with other software that can use all the information in a model object. For us, that software will be the `model_eval()` extractor.

Use `model_eval()` as you do the other familiar extractors such as `coef()` or `confint()`. To generate a prediction, give `model_eval()` arguments specifying the desired inputs to use with the model function.

```
Time_mod %>% model_eval(distance=10, climb=500)
```

distance	climb	.output	.lwr	.upr
10	500	3373	1664	5082

Notice that the result from the above command includes a column `.output` which will always be an exact match to the output the model function will have generated. However, there is more to the output of `model_eval()`. The interval form of the prediction is of particular importance, contained in the columns `.lwr` and `.upr`.

Many people prefer a point prediction, possibly because the single number suggests a single, correct answer, which is somehow emotionally comforting. *But the comfort is unjustified.*

The proper form for a prediction is a **prediction interval**: two numbers bounding the lower and upper limits for the likely outcome. For the hill-racing model, the point prediction is 3372.985 seconds, which is a running time of just under one hour. Nothing about this single number even tells us how many digits are appropriate. The prediction interval tells a different story. The interval, 1700 to 5100 seconds, conveys the appropriate uncertainty in the prediction.

8.1 Where does the prediction interval come from

The prediction interval has two distinct components:

1. The uncertainty in the model function and hence in the output of the model function.
2. The size of the residuals found when training the model.

Consider first the model function. For the running-time model, we can construct the model function from the coefficients of the linear model. These are:

```
Time_mod %>% coef()
```

```
(Intercept)    distance       climb 
    -469.98      253.81        2.61 
```

The algebraic expression for the model function is straightforward:

$$t(d, c) \equiv -470 + 254d + 2.61c \ .$$

The statistical thinker knows that such coefficients have uncertainty due to sampling variation. That uncertainty is, of course, quantified by the confidence interval.

```
Time_mod %>% confint()
```

	lwr	upr
(Intercept)	-533.43	-406.52
distance	246.39	261.23
climb	2.49	2.73

Since we cannot legitimately claim to know the values of the coefficients any better than indicated by these confidence intervals, we ought to temper our claims about the model function so that it reflects the uncertainty in the coefficients. For instance, we might provide an interval for the model output, using in an "upper" function the high ends of the confidence intervals on the coefficients and another "lower" function that uses the low ends of the confidence interval. Like this:

$$t_{upr}(d, c) \equiv -407 + 261d + 2.72c \quad t_{lwr}(d, c) \equiv -533 + 246d + 2.49c$$

Evaluate both the lower and upper functions to get an *interval* on the model output. That would give us $t_{lwr}(10, 500) = 3172$ and $t_{upr}(10, 500) = 3569$.

This idea for generating the "lower" and "upper" functions has the right spirit but is not on target mathematically. The reason is that using the low end of the confidence interval for all coefficients is overly pessimistic; usually, the uncertainty in the different coefficients cancels out to some extent.

The mathematics for the correct "lower" and "upper" functions are well understood but too advanced for the general reader. For our purposes, it suffices to know that `model_eval()` knows how to do the calculations correctly.

The prediction interval produced by `model_eval()` includes both components (1) and (2) listed above. Insofar as we are interested in component (1) in isolation, the correct sort of interval—a *confidence interval*—can be requested from `model_eval()`.

```
Time_mod %>%
    model_eval(distance=10, climb=500, interval="confidence")
```

distance	climb	.output	.lwr	.upr
10	500	3373	3335	3411

This report shows that the confidence interval on the model output—that is, just component (1) of the prediction interval—is pretty narrow: 3335 seconds to 3411 seconds, or, in plus-or-minus format, 3373 ± 38 seconds.

The prediction interval—that is, the sum of components (1) and (2)—is comparatively huge: 1700 to 5100 seconds or, in plus-or-minus format, 3400 ± 1700 seconds. That is almost 50 times wider than the confidence interval.

Why is the prediction interval so much more comprehensive than the confidence interval? The confidence interval reports on the sampling variation of a model constructed as an average over all the data, the $n = 2236$ participants recorded in the `Hill_racing` data frame. However, each runner in `Hill_racing` has their own individual time: not an average but just for the individual. The individual value might be larger or smaller than the average. How much larger or smaller? The residuals for the model provide this information. As always, we can measure the individual-to-individual variation with the standard deviation.

```
Time_mod %>% model_eval() %>% summarize(se_residuals = sd(.resid))
```

se_residuals
871

Keeping in mind that the overall spread of the residuals is plus-or-minus "twice" the standard deviation of the residuals, we can say that the residuals indicate an additional uncertainty in the prediction for a runner of about $\pm$1700 seconds. This $\pm$1700 second is our estimate of the **noise** in the measurements. In contrast, the **confidence interval** is about the sampling variation in the signal.

In this case, the **prediction interval** is wholly dominated by noise; the sampling variability contributes only a tiny amount of additional uncertainty.

i Example: Graphics for the prediction interval

We shift the running scene from Scotland to Washington, DC. The race now is a single 10-miler with almost 9000 registered participants. We wish to predict the running time of an individual based on his or her `age`.

```
Age_mod <- lm(net ~ age, data = TenMileRace)
```

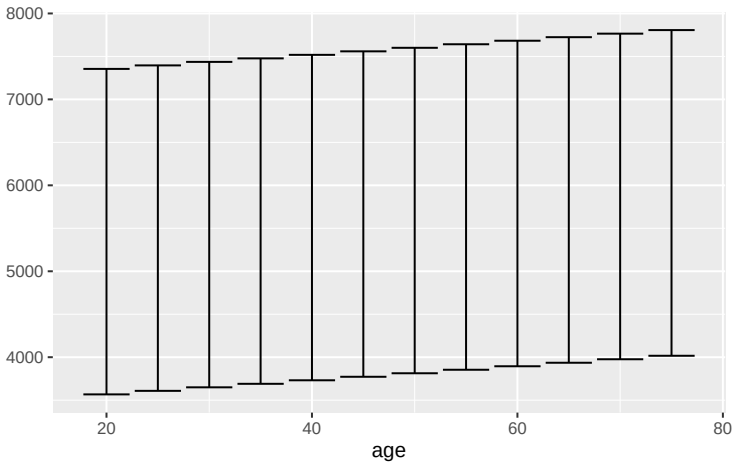
We can see the prediction interval for an individual runner using `mod_eval()`. For example, here it is for a 23-year-old.

```
Age_mod %>% model_eval(age=23)
```

age	.output	.lwr	.upr
23	5486	3592	7379

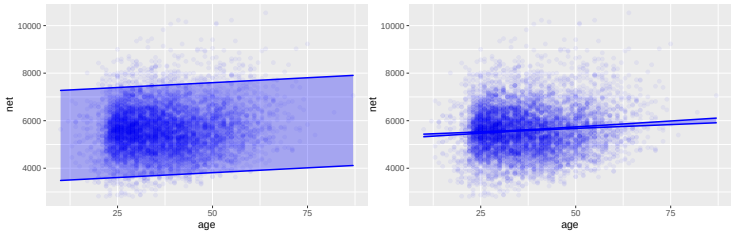
We can also calculate the prediction interval for several different ages and graph out the results with the "errorbar" glyph:

```
Age_mod %>%
  model_eval(age=c(20,25,30,35,40,45,50,55,60,65,70,75)) %>%
  ggplot(aes(x=age)) +
  geom_errorbar(aes(ymin=.lwr, ymax=.upr))
```



For convenience, the `model_plot()` function will do this work for us, plotting the prediction interval along with the training data. We can also direct `model_plot()` to show the confidence interval.

```
### #| column: page-right
model_plot(Age_mod, x=age, interval="prediction", data_alpha=0.05)
model_plot(Age_mod, x=age, interval="confidence", data_alpha=0.05)
```



(a) prediction interval (b) confidence interval

Figure 10: The prediction and model-function confidence intervals for the model `net ~ age`.

Since we are looking at the intervals as a function of an input variable, what we formerly showed using the "error-bar" glyph is now shown using a ribbon or **band**.

Notice that the prediction interval covers almost all of the data points. There are hundreds of data points outside the interval, but with almost 9000 rows in the `TenMileRace` data frame, an interval that covers 95% of the data will have about 450 rows *outside* the interval.

Such a prediction interval is of little use; it cannot give a precise prediction about the running time of an individual. The honest prediction of an individual's outcome needs to reflect the spread of all the individuals with a similar age.

In contrast, the *confidence* band on the model function is pleasingly narrow and precise. It covers only a tiny fraction of the raw data. For this very reason, the confidence interval is *inappropriate* for presenting a prediction. As always, confidence intervals only show general trends in the data, not the range of results for an individual prediction. For instance, Figure 10 shows a clear upward trend in running time with age. There is no flat or negatively sloping line compatible with the confidence interval.

To summarize:

1. When making a prediction, report a prediction interval.
2. The prediction interval is always larger than the confidence interval and is usually *much* larger.

The confidence interval is not for predictions. Use a confidence interval when looking at an effect size. Graphically, the confidence interval is to indicate whether there is an overall trend in the model.

9 Covariates

Dr. Mary Meyer is a statistics professor at Colorado State University. In 2006, she published an article recounting an episode from family life:

> *When my daughter was in fourth grade, I took her shopping for dress shoes. I was disappointed in the quality of girls' shoes at every store in the mall. The shoes for boys were sturdy and had plenty of room in the toes. On the other hand, shoes for girls were flimsy, narrow, and had pointed toes. In spite of the better construction for boys, the costs of the shoes were similar! For children the same age, boys had shoes they could run around in, while girls' shoes were clearly for style and not comfort.*
>
> *Upon complaining about this state of affairs, I was told by sales representatives in two stores that boys actually had wider feet than girls, so needed wider shoes. Being very skeptical, I thought I would test this claim.*

We will return to Dr. Meyer's project in a little bit. However, for now, imagine how this situation might be addressed by someone who still needs to develop good statistical thinking skills. We will call this imagined protagonist "Mr. Shoebuyer." Since the salesmen claimed that girls' feet are narrower than boys, Mr. Shoebuyer heads out to measure the widths of girls' and boys' shoes.

A shoe store provides a convenient place to measure the widths of many different shoe styles. Mr. Shoebuyer gets to the shoe store, heads to the children's section, and starts measuring. For each shoe on display, he records the shoe width and whether the shoe is for girls or boys. Here are his data:

sex	width
G	9.0
G	8.5
G	9.0
G	9.5
B	8.6
B	8.4
B	8.8
B	9.4

Once back home, Mr. Shoebuyer uses his calculator to find the mean width of the shoes in each group. His results surprise him:

sex	mean width
Girls	9.0 cm
Boys	8.8 cm

Mr. Shoebuyer happens to be your uncle. He knows you are taking a statistics course and asks you to check his arithmetic. Putting on a statistical thinking hat to the effect size of sex on shoe width, you note the absence of a confidence interval. This omission is easy to fix.

```
Shoebuyer_data %>% lm(width ~ sex, data=.) %>% confint()
```

	lwr	upr
(Intercept)	8.286	9.314
sexG	-0.527	0.927

Your uncle is at the table at Thanksgiving break. "Sorry, Uncle, but you don't have nearly enough data to conclude that girls' feet are wider than boys'." Translating the confidence interval into plus-or-minus format, you point out that the difference between the sexes is 0.2 ± 0.8 cm. "You'll need enough data to get that 0.8 margin of error down to something like 0.2." You also point out that there might be a better place to collect data than a shoe store. "It's the feet, not the shoes, that you want to look at."

Aware of these pitfalls, Dr. Meyer worked with the third- and fourth-grade teachers at her daughter's school to collect data. Being a statistical thinker, she thought about what data would illuminate the matter before carrying out the data collection. Her data, a sample of size $n = 39$, are recorded in the `KidsFeet` data frame.

```
lm(width ~ sex, data = KidsFeet) %>% confint()
```

	lwr	upr
(Intercept)	8.976	9.404
sexG	-0.713	-0.099

In plus-or-minus format, this confidence interval is -0.4 ± 0.3. Whatever the format, Dr. Meyer's data provides some evidence that girls' feet are narrower than boys'.

As a statistical thinker, Dr. Meyer knows that even though the foot width is the original quantity of interest, other factors might play a role in the system. For example, boys' feet might trend longer or shorter than girls' feet. This possibility should be taken into account by looking at the effect size of `sex` on width, holding length constant. After all, a shoe buyer first tells the salesperson their foot length (or "size"); the salesperson then brings shoes of that size to try on.

```
lm(width ~ sex + length, data=KidsFeet) %>% confint()
```

	lwr	upr
(Intercept)	1.105	6.178
sexG	-0.495	0.030
length	0.120	0.322

Although `sex` is the explanatory variable of primary interest to Dr. Meyer's question, she knows to include other explanatory variables that might play a role. Such explanatory variables, not of direct interest, are called "**covariates**." Dr. Meyer's statistical expertise led her to consider possible covariates *before*

collecting her data and took the trouble of measuring both foot length and width.

The confidence interval on the `sexG` coefficient includes zero when `length` is taken into account. Dr. Meyer's little study provides evidence that even if girls' shoes tend to be narrower than boys', the feet inside them have about the same shape for both sexes.

9.1 All other things being equal

The common phrase "all other things being equal" is a critical qualifier in describing relationships. To illustrate: A simple claim in economics is that a high price for a commodity reduces the demand. For example, increasing the price of heating fuel will reduce demand as people turn down thermostats to save money. Nevertheless, the claim can be considered obvious only with the qualifier *all other things being equal*. For instance, the fuel price might have increased because winter weather has increased the demand for heating compared to summer. Thus, higher prices may be associated with higher demand. Therefore, increased price may not be associated with lower demand unless holding other variables, such as weather conditions, constant.

In economics, the Latin equivalent of "all other things being equal" is sometimes used: "**ceteris paribus**". So, the economics claim would be, "higher prices are associated with lower demand, *ceteris paribus*."

Although the phrase "all other things being equal" has a logical simplicity, it is impractical to implement "all." So instead of the blanket "all other things," it is helpful to consider just "some other things" to be held constant, being explicit about what those things are. Other phrases along the same lines are "taking into account …" and "controlling for …." Those additional variables that are to be considered are called "**covariates**.

> **i** Example: Covariates and Death
>
> This news report appeared in 2007:
>
> **Heart Surgery Drug Carries High Risk,**

Study Says. A drug widely used to prevent excessive bleeding during heart surgery appears to raise the risk of dying in the five years afterward by nearly 50 percent, an international study found. The researchers said replacing the drug—aprotinin, sold by Bayer under the brand name Trasylol—with other, cheaper drugs for a year would prevent 10,000 deaths worldwide over the next five years.
Bayer said in a statement that the findings are unreliable because Trasylol tends to be used in more complex operations, and the researchers' statistical analysis did not fully account for the complexity of the surgery cases. The study followed 3,876 patients who had heart bypass surgery at 62 medical centers in 16 nations. Researchers compared patients who received aprotinin to patients who got other drugs or no antibleeding drugs. Over five years, 20.8 percent of the aprotinin patients died, versus 12.7 percent of the patients who received no antibleeding drug. [This is a 64% increase in the death rate.] When researchers adjusted for other factors, they found that patients who got Trasylol ran a 48 percent higher risk of dying in the five years afterward. The other drugs, both cheaper generics, did not raise the risk of death significantly. The study was not a randomized trial, meaning that it did not randomly assign patients to get aprotinin or not. In their analysis, the researchers took into account how sick patients were before surgery, but they acknowledged that some factors they did not account for may have contributed to the extra deaths. - Carla K. Johnson, Associated Press, 7 Feb. 2007

"Significant" has a specialized meaning in statistical language. It is *not* a synonym for "important." See Lessons 36 through 38

The report involves several variables. Of primary interest is the relationship between (1) the risk of dying after surgery and (2) the drug used to prevent excessive

bleeding during surgery. Also potentially important are (3) the complexity of the surgical operation and (4) how sick the patients were before surgery. Bayer disputes the published results of the relationship between (1) and (2) holding (4) constant, saying that it is also essential to hold variable (3) constant.

The total relationship involves a death rate of 20.8 percent of patients who got aprotinin versus 12.7 percent for the patients taking the generic drugs: an increase in the death rate by a factor of 1.64. However, when the researchers looked at a partial relationship (holding constant patient sickness before the operation), the effect size of aprotinin on mortality was less: a factor of 1.48. In other words, the model `death ~ aprotinin` shows a 64% increase in the death rate, but the model `death ~ aprotinin + sickness` shows a slightly smaller increase in death rate: 48%. The difference between the two estimates reflects doctors being more likely to give aprotinin to sicker patients.

The story's last paragraph states that the choice of patients receiving aprotinin versus the generic drugs was not made at random. Some readers may find this reassuring. Why in the world would anyone prescribe a drug at random? The point, however, is to select randomly who gets which drug *among the patients for whom the drugs would be appropriate*. The phrase "randomized trial" used in the paragraph means specifically an *experiment* in which one treatment or the other—aprotinin versus the generic drugs—is assigned at random. The virtues of experiment and the vital role of random assignment are detailed in Lesson 13.

9.2 Letting things change as they will

Using covariates in models enables the relationship between a response and an explanatory variable to be described *ceteris paribus*, that is, "all other things being equal." Another phrase used in news stories is "after adjusting for ...." This is appropriate since the *all* in "all other things" is, in reality, refers only

to those particular factors used as the covariates in the model. So, Dr. Meyer's foot width results might be stated in everyday language as, "After adjusting for foot width, she found no difference in the widths of girls' and boys' feet."

Not including covariates in a model amounts to "letting other things change as they will." In Latin, this is "*mutatis mutandis*." In the foot-width example, the model `width ~ sex` looks at the differences in foot width for the two sexes. However, sex is not the only thing associated with foot width. The model `width ~ sex` ignores all other factors than sex; it compares boys and girls *mutatis mutandis*, that is, letting other things change as they will. In this case, comparing boys and girls involves not just the possible differences in foot width but also the differences in other factors such as foot length and body weight.

i Example: One change can bring another

I was once involved in a budget committee that recommended employee health benefits for the college where I worked. At the time, college employees who belonged to the college's insurance plan received a generous subsidy for their health insurance costs. Employees who did not belong to the plan received no subsidy but were given a modest monthly cash payment. After the stock market crashed in 2000, the college needed to cut budgets. One proposal called for eliminating the cash payment to employees who did not belong to the insurance plan. Proponents of the plan claimed that this would save money without reducing health benefits. I argued that this claim was an "all other things being equal" analysis: how expenditures would change assuming the number of people belonging to the insurance plan remained constant. In reality, however, the policy change would play out *mutatis matandis*; the loss of the cash payment would cause some employees, who currently received health benefits through their spouse's health plan, to switch to the college's health plan. That is what happened, contributing to an overall increase in healthcare expenses.

Example: Spending and student performance

To illustrate how covariates set context, consider an issue of interest to public policy-makers in many societies: How much money to spend on children's education? State lawmakers in the US are understandably concerned with the quality of public education provided. However, they also have other concerns and constraints and constituencies who give budget priority to other matters.

In evaluating their various trade-offs, lawmakers could benefit by knowing how increased educational spending will shape educational outcomes. What can available data tell us? Unfortunately, there are various political constraints that work against states adopting and publishing data on a standard, genuine measure of educational outcome. Instead, we have high-school graduation rates, student grades, and other non-standardized data. These data might have some meaning but can also reflect system gaming by administrators and teachers, for which there is little systematic data.

Although imperfect, college admissions tests such as the ACT and SAT provide consistent data between states. For example, Figure 11 shows the average SAT score in 2010 in each state versus expenditures per pupil in public elementary and secondary schools. Layered on top of the data is a flexible linear model (and its confidence band) of SAT score versus expenditure.

The overall impression given by the model is that the relationship is negative, with lower expenditures corresponding to higher SAT scores. However, the confidence band is broad; it is possible to find a smooth path with almost zero slope through the confidence band. Either way, this graph does not support the conventional wisdom that higher spending produces better school outcomes.

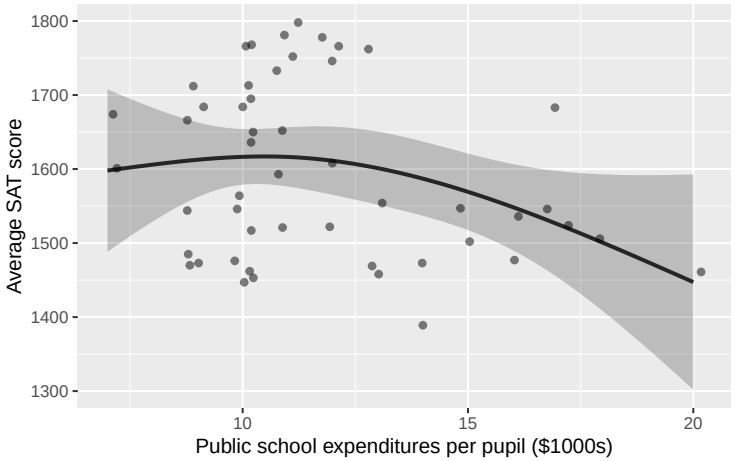


Figure 11: State by state data (from 2010) on average SAT college admissions test scores and expenditures for public education.

Of course, other factors play a role in shaping education outcomes: for instance, poverty levels, parental education, and how the educational money is spent (higher pay for teachers or smaller class sizes? administrative bloat?).
At first glance, it is tempting to ignore these additional factors. For instance, we may not have data on them. Moreover, as our interest is in understanding the relationship between expenditures and education outcomes, we are not directly concerned with the additional factors. However, the lack of direct concern does not imply that we should ignore the factors but that we should do what we can to "hold them constant".
To illustrate, consider the fraction of eligible students (those in their last year of high school) who take the college admission test. This fraction varies widely from state to state. In a poor state where few students go to college, the fraction can be tiny (Alabama 8%, Arkansas 5%, Mississippi 4%, Louisiana 8%). In some other states, the large majority of students take the SAT (Maine 93%, Massachusetts 89%, New York 89%). In states with low SAT participation rates, the students who take the test tend to be those applying to schools with competitive admissions.

Such strong students will get high scores. In contrast, the scores in states with high participation rates reflect both strong and weak students. Consequently, the scores will be lower on average than in the low-participation states.
Putting the relationship between expenditure and SAT scores in the context of the fraction taking the SAT is accomplished with the model `SAT ~ expenditure + fraction` rather than just `SAT ~ expenditure`. Figure 12 shows a model with `fraction` as a covariate.

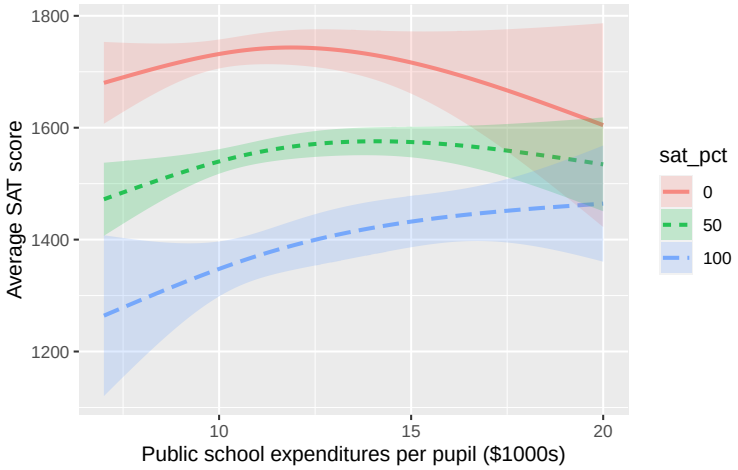


Figure 12: The model of SAT score versus expenditures, including as a covariate the fraction of eligible students in the state who take the SAT.

Note that the effect size of spending on SAT scores is positive when the expenditure level is less than $10,000 per pupil. Notice as well that when the fraction taking the SAT is tiny, the average scores do not depend on expenditure. This flat relationship suggests that, among elite students, state expenditure does not make a discernable difference. Perhaps the college-bound students in such states have other educational resources to draw on.
The relationship shown in Figure 11 is genuine. However, so is the very different relationship seen in Figure 12. How can the same data be consistent with two utterly different displays? The answer, perhaps unexpectedly, has to

do with the connections among the explanatory variables. Whatever the relationship between an individual explanatory variable and the response variable, the *appearance* of that relationship will depend on which covariates the modeler chooses to include.

10 Covariates eat variance

In Lesson 9, we introduced covariates to set the relationship between an explanatory variable and the response variable in the correct context. In Lesson 30, we will return to this context-setting role to show that the appropriate choice of covariates to include in a model depends on the modeler's opinion about the relevant structure of a DAG. Here, we will treat covariates as commodity items to show a surprising property of models. This property is a boon to the modeler, helping to enable sound decisions about whether to include any given covariate. However, it is also a pitfall lieing in wait for the wishful thinker.

10.1 How much variation is explained

We start by returning to the definition of statistical thinking introduced at the start of these Lessons:

> *Statistic thinking is the explanation or description of variation* in the context of *what remains unexplained or undescribed.*

In this Lesson, we will work with a straightforward measure of "what remains unexplained or undescribed." The fitted model values represent the explained part of the variation. The residuals are what is left over, the difference between the actual values of the response variable and the fitted model values.

As a reminder, we will construct a simple model of the list price of books as a function of the number of pages and whether the book is a paperback or hardcover.[7]

```
Price_model <- lm(list_price ~ num_pages + hard_paper,
                  data = amazon_books)
```

The `model_eval()` function can extract the fitted model values and the residuals from the model. We show just a few rows

[7] If you seek to duplicate the results presented in this chapter, please note that we have deleted six rows from 'amazon_books because the rows are either duplicates or have one of the variables missing. The deleted rows are 62, 103, 205, 211, 242, and 303.

here, but we will use the entire report from `model_eval()`. Remember that when `model_eval()` is not given input values, it uses the model *training* data as input.

```
Results <- model_eval(Price_model)
```

list_price	num_pages	hard_paper	.output	.resid	.lwr	.upr
12.9	304	P	16.6	-3.65	-10.73	43.9
15.0	273	P	16.0	-0.98	-11.36	43.3
1.5	96	P	12.4	-10.95	-14.97	39.9
16.0	672	P	23.9	-7.96	-3.57	51.5
30.5	720	P	24.9	5.59	-2.67	52.5
28.9	460	H	24.6	4.38	-2.89	52.0

The first book in the training data is a 304-page paperback with a list price of $12.95. The fitted model value for that book is $16.60. (Ordinarily, we refer to the output of the model function simply as the "output" or the "model output." However, the output of the model function, when applied to rows from the *training* data also called the *fitted model value*.)

At $16.60, the fitted model value is $3.65 *higher* than the list price. This difference is the residual for that book, the sign reflecting the definition

$$\text{residual} \equiv \text{response value} - \text{fitted model value} \ .$$

When the residual is small in magnitude, the fitted model value is close to the response value. Conversely, a large residual means the model was way off target for that book.

The standard measure of the typical size of a residual is the standard deviation or, equivalently, the variance.

```
Results %>% summarize(se_resids = sd(.resid), v_resids=var(.resid))
```

se_resids	v_resids
13.8	191

As always, the standard deviation is easier to read because it has sensible units, in this case, dollars. On the other hand, the variance has strange units (square dollars) because it is the square of the standard deviation. We will use the variance for measuring the typical size of a residual for the reasons described in Lesson 2; variances add nicely in a manner analogous to the Pythagorean Theorem.

Similarly, the total amount of variation in the list price is simply the variance of the list price:

```
Results %>% summarize(v_response = var(list_price))
```

v_response
207

A simple measure of how much of the variation in list price remains unexplained is the ratio of these variances 190.96/206.51 = 92.5%. More than 90% of the variation remains unexplained by the `Price_model`. This high fraction of unexplained variance suggests the model has little to tell us. In the spirit of putting a positive spin on things, statisticians typically work with the complement of the unexplained fraction. Since the unexplained fraction is 92.5%, the complement is 7.5%. This number is written R^2 and pronounced "R-squared." (It also has a formal name: the "coefficient of determination." In Lesson 11, we will meet the inventor of the coefficient of determination, Sewall Wright, who is an early hero of causal reasoning.)

R^2 is such a widely used summary of how the explanatory variables account for the response variable that a software extractor calculates it and some related values.

```
Price_model %>% R2()
```

n	k	Rsquared	F	adjR2
317	2	0.075	12.8	0.069

Many modelers act as if their goal is to build a model that makes R^2 as big as possible. Their thinking is that large R^2 means that the explanatory variables account for much of the response variable's variance. Unfortunately, it is a naive goal. Instead, always focus on the model's suitability for the purpose at hand. Often, shooting for a large R^2 imposes costs that can undermine the purpose for the model. Furthermore, even models with the largest possible R^2 sometimes have nothing to say about the response variable.

10.2 Getting to 1

R^2 can range from zero to one. Zero means that the model accounts for *none* of the variation in the response variable. We can construct such a model quickly enough: `list_price ~ 1` has no explanatory variables and, therefore, no ability to distinguish one book from another.

```
Null_model <- lm(list_price ~ 1, data = amazon_books)
Null_model %>% R2()
```

n	k	Rsquared	F	adjR2
319	0	0	NaN	0

We are using the word "null" to name this model. "Null" is part of the statistics tradition. The dictionary definition of "null" is "having or associated with the value zero" or "lacking distinctive qualities; having no positive substance or content."[8]

In the null model, the fitted model values are all the same; all the variation is in the residuals.

```
Null_model %>% model_eval()
```

[8]Source: Oxford Languages

list_price	.output	.resid	.lwr	.upr
12.9	18.6	-5.65	-9.69	46.9
15.0	18.6	-3.60	-9.69	46.9
1.5	18.6	-17.10	-9.69	46.9
16.0	18.6	-2.61	-9.69	46.9
30.5	18.6	11.90	-9.69	46.9
28.9	18.6	10.35	-9.69	46.9

At the other extreme, where $R^2 = 1$, the explanatory variables account for every bit of variation in the response variable. We can try various combinations of explanatory variables to see if we can accomplish this. For example, `publisher` explains 67% of the variation in list price.

```
lm(list_price ~ publisher, data = amazon_books) %>% R2()
```

n	k	Rsquared	F	adjR2
319	158	0.675	2.1	0.354

We can also check whether `author` has anything to say about the list price.

```
lm(list_price ~ author, data = amazon_books) %>% R2()
```

n	k	Rsquared	F	adjR2
319	250	0.943	4.53	0.735

Incredible! How about if we use *both* `publisher` and `author` as explanatory variables? We get very close to $R^2 = 1$.

```
lm(list_price ~ publisher + author, data = amazon_books) %>%
  R2()
```

n	k	Rsquared	F	adjR2
319	281	0.982	7.25	0.847

The modeler discovering this tremendous explanatory power of `publisher` and `author` can be forgiven for thinking he or she

has found a meaningful explanation. But, unfortunately, the high R^2 is an illusion in this case.

To see why, consider another possible explanatory variable, the International Standard Book Number (ISBN). The ISBN is a ten- or thirteen-digit number that marks each book with a unique number.

Figure 13: The ISBN number from one of the Project MOSAIC textbooks.

There is a system behind ISBNs, but despite the "N" standing for "number," an ISBN is a character string or word (written using only digits). Consequently, the `isbn_10` variable in `amazon_books`is categorical.

```
ISBN_model <- lm(list_price ~ isbn_10, data = amazon_books)
ISBN_model %>% R2()
```

n	k	Rsquared	F	adjR2
319	318	1	NaN	NaN

The `isbn_10` explains all variation in the list price!

Given that the ISBN is, as we have said, an arbitrary sequence of characters, why does it do such a good job of accounting for the list price? The answer lies not in the content of the ISBN but in another fact: each book has a unique ISBN. As well, each book has a single price. So the ISBN identifies the price of each book. Cleverness is not involved; the list price could be anything, and the ISBN would still identify it precisely. The model coefficients store the whole set of ISBNs and the corresponding set of list prices.

We can substantiate the claim just made—that the list price could be anything at all—by synthesizing a data frame with random list prices:

```
amazon_books %>%
  mutate(random_list_price = rnorm(nrow(.))) %>%
  lm(random_list_price ~ isbn_10, data = .) %>%
  R2()
```

n	k	Rsquared	F	adjR2
319	318	1	NaN	NaN

Similar randomization can be accomplished by *shuffling* the `isbn_10` column of the data frame so that each ISBN points to a random book. Of course, such shuffling destroys the link between the ISBN and the list price. Even so, the R^2 remains high.

```
lm(list_price ~ shuffle(isbn_10), data=amazon_books) %>% R2()
```

n	k	Rsquared	F	adjR2
319	318	1	NaN	NaN

```
lm(shuffle(list_price) ~ isbn_10, data=amazon_books) %>% R2()
```

n	k	Rsquared	F	adjR2
319	318	1	NaN	NaN

Statistical nomenclature is obscure here. So we will make up a name for such incidental alignment with no true explanatory power: the "**ISBN-effect**."

Statistical thinkers know to be aware of situations where categorical variables have many levels and check whether the ISBN effect is in play.

10.3 The ISBN effect as a benchmark

Shuffling an explanatory variable (while keeping the response variable in the original order) voids any possible explanatory connection between the two. An R^2=0, as we get from any model of the form `y ~ 1`, signals that the `1` cannot account for any variation. However, this does not mean shuffling will lead to $R^2 = 0$. Instead, there is a systematic relationship between

the number of model coefficients associated with the shuffled variable, the sample size n, and R^2.

We can demonstrate this relationship by conducting many trials of modeling the `list_price` with a shuffled explanatory variable: either `publisher`, `author`, or `isbn_10`.

⚠ Demonstration: Counting coefficients

The `amazon_books` data frame has $n = 319$ rows.[a] In the next computing chunk, we fit the model `list_price ~ publisher` and collect the coefficients for counting:

```
Publisher_model <- lm(list_price ~ shuffle(publisher),
                      data=amazon_books)
Coefficients <- Publisher_model %>% coef() %>% data.frame()
nrow(Coefficients)
```

```
[1] 159
```

There are 161 coefficients in the model, the first one being the "Intercept." We will show only the first few.

```
Coefficients %>% head()
```

	value
(Intercept)	14.95
shuffle(publisher)Adams Media	0.05
shuffle(publisher)Akashic Books	13.00
shuffle(publisher)Aladdin	15.05
shuffle(publisher)Albert Whitman & Company	-0.95
shuffle(publisher)Alfred A. Knopf	0.05

Altogether, there are $k = 160$ coefficients relating to `shuffle(publisher)`.

[a] The data frame in the `moderndive` package has six additional rows, which we have deleted as duplicates or because of missing data.

The theory relating R^2 to the number of coefficients associated

is straightforward for shuffled explanatory variables: R^2 will be random with mean value $\frac{k}{n-1}$.

Demonstration: The mean R^2 across many trials

For the `shuffle(publisher)` model, the theoretical mean across many trials will be $R^2 = 158/324 = 0.49$. The demonstration below confirms this using 100 trials:

```
Pub_trials <- do(100) * {
  lm(list_price ~ shuffle(publisher), data=amazon_books) %>%
    R2()
}
Pub_trials %>% summarize(meanR2 = mean(Rsquared))
```

meanR2
0.496

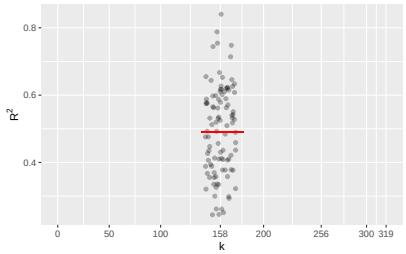


Figure 14: 100 trials of R^2 from `list_price ~ shuffled(publisher)`. The theoretical value $k/n = 160/324 = 0.49$ is marked in red.

We can carry out similar trials for the models `list_price ~ shuffle(author)` and `list_price ~ shuffle(isbn_10)`, which have $k = 251$ and $k = 319$ respectively.

The blue diagonal line in Figure 15 shows the theoretical average R^2 as a function of the number of model coefficients when the explanatory variable is randomized. R^2 will always be 1.0 when $k = n$, that is, when the number of coefficients is the same as the sample size.

Figure 15 suggests a way to distinguish between R^2 resulting from the ISBN-effect and R^2 that shows some true explanatory power: Check if R^2 is substantially above the blue diagonal line, that is, check if $R^2 \gg \frac{k}{n-1}$ where k is the number of model coefficients.

10.4 The F statistic

k and n provide the necessary context for proper interpretation of R^2; all three numbers are needed to establish whether $R^2 \gg \frac{k}{n-1}$ to rule out the ISBN effect. The calculation is not

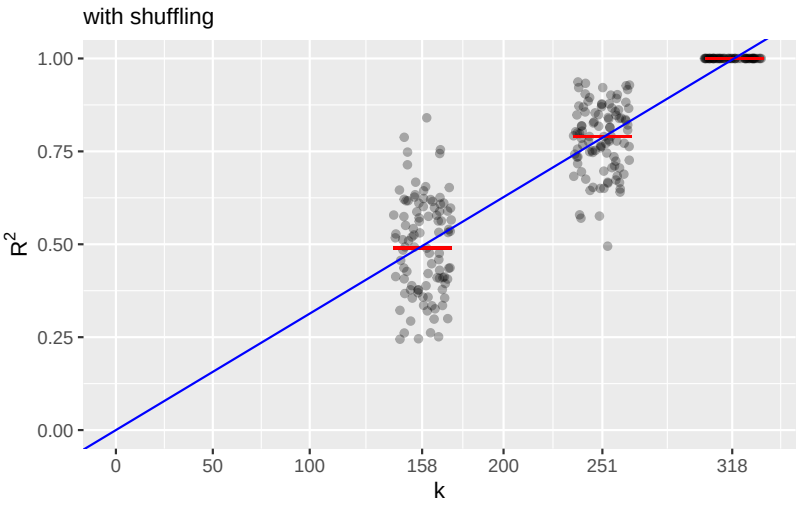


Figure 15: R^2 from many trials of three models, `list_price ~ shuffle(publisher)` and `~ shuffle(author)` and `~shuffle(isbn_10)`.

difficult; the modeler always knows the size n of the training data and can find k as the number of coefficients in the model (not counting the Intercept term).

Perhaps a little easier than interpreting R^2 is the interpretation of another statistic, named F, which folds in the k, n, and R^2 into a single number:

$$F \equiv \frac{n-k-1}{k} \frac{R^2}{1-R^2}$$

Figure 16 is a remake of Figure 15 but using F instead of R^2. The blue line, which had the formula $R^2 = k/(n-1)$ in Figure 15, gets translated to the constant value 1.0 in Figure 16, regardless of k. To decide when a model points to a connection stronger than the ISBN effect, the threshold $F > 3$ is a good rule of thumb. (Lesson 18 introduces a more precise calculation for the F threshold, which is built into statistical software and presented as a "**p-value**.")

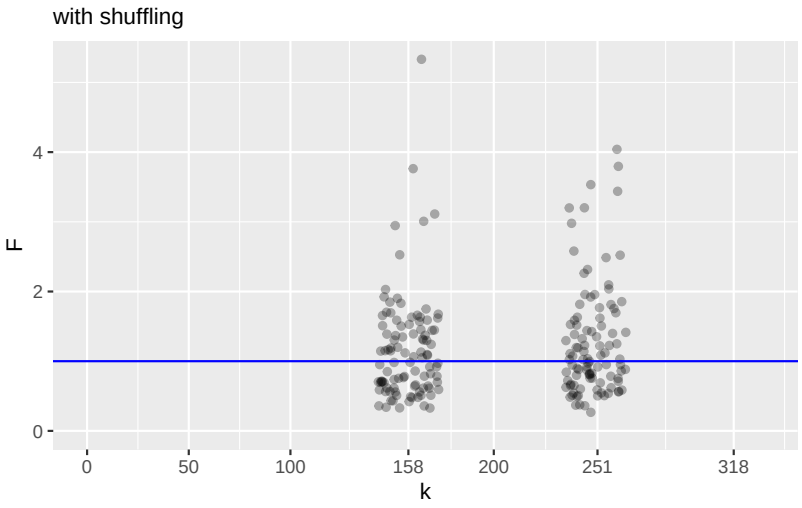


Figure 16: Like Figure 15, but using the F statistic to summarize each trial.

i Adjusted R^2

Some fields, notably economics, prefer an alternative to F called "**adjusted R^2**" (or R^2_{adj}). The adjustment comes from moving the raw R^2 downward and leftward, more-or-less in the direction of the blue line in Figure 15. This movement adjusts a raw R^2 that lies on the blue line to $R^2_{\text{adj}} = 0$.

We leave the debate on the relative merits of using F or R^2_{adj} their respective boosters. However, before getting wrapped up in such debates, it is worth pointing out that R^2_{adj} is just a rescaling of F.

$$R^2_{\text{adj}} = 1 - \frac{n-1}{k}\frac{R^2}{F} \ .$$

10.5 Comparing models

Modelers are often in the position of having a model that they like but are contemplating adding one or more additional explanatory variables. To illustrate, consider the following models:

All the explanatory variables in the smaller models also apply to the bigger models. Such sets are said to be "**nested**" in

Figure 17: Nesting Russian dolls

- Model 1: `list_price ~ 1`
- Model 2: `list_price ~ 1 + hard_paper`
- Model 3: `list_price ~ 1 + hard_paper + num_pages`
- Model 4: `list_price ~ 1 + hard_paper + num_pages + weight_oz`

much the same way as for Russian dolls.

For a nested set of models, R^2 can never decrease when moving from a smaller model to a larger one—almost always, there is an increase in R^2. To demonstrate:

```
amazon_books <- amazon_books %>%
  select(list_price, weight_oz, num_pages, hard_paper) %>%
  filter(complete.cases(.))
model1 <- lm(list_price ~ 1, data=amazon_books)
model2 <- lm(list_price ~ 1 + weight_oz, data = amazon_books)
model3 <- lm(list_price ~ 1 + weight_oz + num_pages, data=amazon_books)
model4 <- lm(list_price ~ 1 + weight_oz + num_pages + hard_paper, data=amazon_books)
```

```
R2(model1)
```

n	k	Rsquared	F	adjR2
309	0	0	NaN	0

```
R2(model2)
```

n	k	Rsquared	F	adjR2
309	1	0.16	57	0.15

```
R2(model3)
```

n	k	Rsquared	F	adjR2
309	2	0.17	30	0.16

```
R2(model4)
```

n	k	Rsquared	F	adjR2
309	3	0.17	21	0.16

When adding explanatory variables to a model, a good question is whether the new variable(s) add to the ability to account for the variability in the response variable. R^2 never goes down when moving from a smaller to a larger model, so we cannot rely on the increase in R^2. A valuable technique called "**Analysis of Variance**" (ANOVA for short) looks at the incremental change in variance explained from a smaller model to a larger one. The increase can be presented as an F statistic. To illustrate:

```
anova_summary(model1, model2, model3, model4)
```

term	df.residual	rss	df	sumsq	statistic
list_price ~ 1	308	54531	NA	NA	NA
list_price ~ 1 + weight_oz	307	46032	1	8499	57.2
list_price ~ 1 + weight_oz + num_pages	306	45466	1	566	3.8
list_price ~ 1 + weight_oz + num_pages + hard_paper	305	45277	1	189	1.3

Focus on the column named `statistic`. This records the F statistic. The move from Model 1 to Model 2 produces F=57, well above the threshold described above and clearly indicating that the `weight_oz` variable accounts for some of the list price. Moving from Model 2 to Model 3 creates a much less impressive F of 3.8. It is as if the added explanatory variable, `num_pages`, is just barely pulling its own "weight." Finally, moving from Model 3 to Model 4 produces a below-threshold F of 1.3. In other words, in the context of `weight_oz` and `num_pages`, the `hard_paper` variable does not carry additional information about the list price.

The last column of the report, labeled `Pr(>F)`, translates F into a universal 0 to 1 scale called a p-value. A large F produces a small p-value. The rule of thumb for reading p-values is that a value $p < 0.05$ indicates that the added variable brings new

information about the response variable. We will return to p-values and the controversy they have entailed in Lessons 36 through 38.

11 Confounding

Many people are concerned that the chemicals used by lawn-greening companies are a source of cancer or other illness. Imagine designing a study that could confirm or refute this concern. The study would sample households, some with a history of using lawn-greening chemicals and others that have never used them. The question for the study designers: What variables to record?

An obvious answer: record both chemical use and a measure of health outcome, say whether anyone in that household has developed cancer in the last five years. We will suppose that the two possible levels of grass treatment are "organic" or "chemicals." As for illness, the levels will be "cancer" or "not."

Here are two very simple DAGs describing possible theories:

illness $\leftarrow$ grass treatment or illness $\rightarrow$ grass treatment

The DAG on the left expresses the belief among people who think chemical grass treatment might cause cancer. But belief is not necessarily reality, so we should consider the right-hand DAG. For example, one way to avoid the possibility of illness $\rightarrow$ grass treatment is to include only households where cancer (if any) started *after* the grass treatment. Note that we are not ignoring the right-hand DAG; we are using the study design to disqualify it.

The statistical thinker knows that covariates are important. But which covariates? Answering that requires knowing a lot about the "domain," that is, how things connect in the world. Such knowledge helps in thinking about the bigger picture and, in particular, possible covariates that connect plausibly to the response variable and the primary explanatory variable, grass treatment.

For now, suppose that the study designers have not yet become statistical thinkers and have rushed out to gather data on illness and grass treatment. Here are a few rows from the data (which we have simulated for this example):

grass	illness
organic	not
chemicals	not
chemicals	not
chemicals	not
organic	not
organic	not
organic	not
organic	not
chemicals	cancer
organic	not

Analyzing such data is straightforward. First, check the overall cancer rate:

```
# overall cancer rate
lm(zero_one(illness, one="cancer") ~ 1, data = Cancer_data) %>% coef()
```

```
(Intercept)
      0.026
```

In these data, 2.6% of the sampled households had cancer in the last five years. How does the grass treatment affect that rate?

```
mod <- lm(zero_one(illness, one="cancer") ~ grass, data = Cancer_data)
coefficients(mod)
```

	coefficient
(Intercept)	0.0125
grassorganic	0.0226

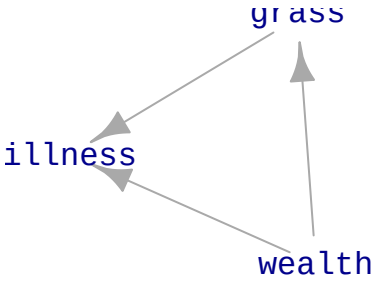
For households whose lawn treatment is "organic," the risk of cancer is higher by 2.3 percentage points compared to households that treat their grass with chemicals. We were expecting the reverse, but it is what the data show. On the other hand, there is sampling variability to take into account. Look at the confidence intervals:

```
confint(mod)
```

	lwr	upr
(Intercept)	-0.0031	0.0280
grassorganic	0.0025	0.0427

The confidence interval on `grassorganic` does not include zero, but it comes close. So might the chemical treatment of grass be protective against cancer? Only at this point do the study designers do what they should have from the start: think about covariates.

One theory—just a theory—is this: Green grass is not a necessity, so the households who treat their lawn with chemicals tend to have money to spare. Wealthier people also tend to have better health, partly because of better access to health care. Another factor is that wealthier people can live in less polluted neighborhoods and are less likely to work in dangerous conditions, such as exposure to toxic chemicals. Such a link between wealth and illness points to a DAG hypothesis where "`wealth`" influences how the household's `grass` is treated and `wealth` similarly influences the risk of developing `cancer`. Like this:



A description of this structure of causality is, "The effect of grass treatment on illness is **confounded** by wealth." The Oxford Languages dictionary offers two definitions of "confound."

1. *Cause surprise or confusion in someone, especially by acting against their expectations.*
2. *Mix up something with something else so that the individual elements become difficult to distinguish.*

This second definition carries the statistical meaning of "confound."

The first definition seems relevant to our story since the protagonist expected that chemical use would be associated with higher cancer rates and was surprised to find otherwise. But, the statistical thinker does not throw up her hands when dealing with mixed-up causal factors. Instead, she uses modeling techniques to untangle the influences of various factors.

Using covariates in models is one such technique. Our wised-up study designers go back to collect a covariate representing household wealth. Here is a glimpse at the updated data.

wealth	grass	illness
1.4284	organic	not
0.0629	chemicals	not
0.4383	chemicals	not
0.6084	chemicals	not
0.8034	organic	not
-0.9367	organic	not
0.6664	organic	not
-1.2446	organic	not
-1.3195	chemicals	cancer
-1.6162	organic	not

Having measured `wealth`, we can use it as a covariate in the model of `illness`:

```
lm(zero_one(illness, one="cancer") ~ grass + wealth, data = Cancer_data) %>%
  confint()
```

	lwr	upr
(Intercept)	0.0247	0.0575
grassorganic	-0.0451	-0.0010
wealth	-0.0568	-0.0356

With `wealth` as a covariate, the model shows that (all other things being equal) "organic" lawn treatment reduces cancer risk. However, we do not see this directly from the `grass`

and `illness` variables because all other things are not equal: wealthier people are more likely to use chemical lawn treatment. (Keep in mind that this is **simulated data**. Do not conclude from this example anything about the safety of the chemicals used for lawn greening.)

i Example: The flu vaccine

As you know, people are encouraged to get vaccinated before flu season. This recommendation is particularly emphasized for older adults, say, 60 and over.
In 2012, the *Lancet*, a leading medical journal, published a systematic examination and comparison of many previous studies. The *Lancet* article describes a hypothesis that existing flu vaccines may not be as effective as was originally found.

Such a study of earlier studies is called a *meta-analysis*.

> *A series of observational studies undertaken between 1980 and 2001 attempted to estimate the effect of seasonal influenza vaccine on rates of hospital admission and mortality in [adults 65 and older]. Reduction in all-cause mortality after vaccination in these studies ranged from 27% to 75%. In 2005, these results were questioned after reports that increasing vaccination in people aged 65 years or older did not result in a significant decline in mortality. Five different research groups in three countries have shown that these early observational studies had substantially overestimated the mortality benefits in this age group because of unrecognized confounding. This error has been attributed to a healthy vaccine recipient effect: reasonably healthy older adults are more likely to be vaccinated, and a small group of frail, undervaccinated elderly people contribute disproportionately to deaths, including during periods when influenza activity is low or absent.*

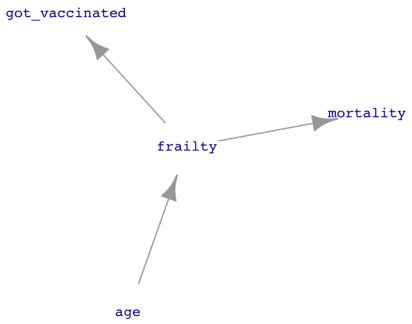


Figure 18: A DAG diagramming the "healthy vaccine recipient" effect

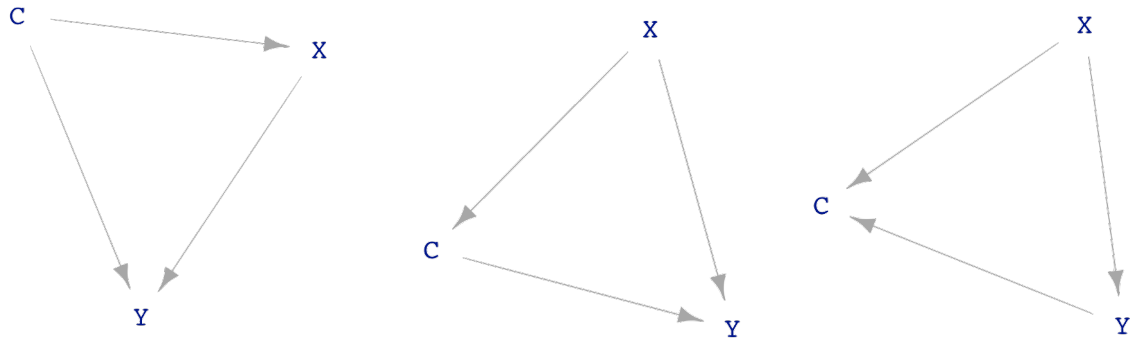
Figure 18 presents a network of causal influences that could shape the "healthy vaccine recipient." People are more likely to become frail as they get older. Frail people are *less* likely to get vaccinated, but more likely to die in the next few months. The result is that vaccination is associated with reduced mortality, even if there is no direct link between vaccination and mortality.

11.1 Block that path!

Let us look more generally at the possible causal connections among three variables, which we will call X, Y, and C. We will stipulate that X points causally toward Y and that C is a possible covariate. Like all DAGs, there cannot be a cycle of causation. These conditions leave three distinct DAGs that do not have a cycle, shown in Figure 19.

C plays a different role in each of the three dags. In sub-figure (a), C causes both X and Y. In (b), part of the way that X influences Y is *through* C. We say, in this case, "C is a mechanism by which X causes Y. In sub-figure (c), C does not cause either X or Y. Instead, C is a consequence of both X and Y.[9]

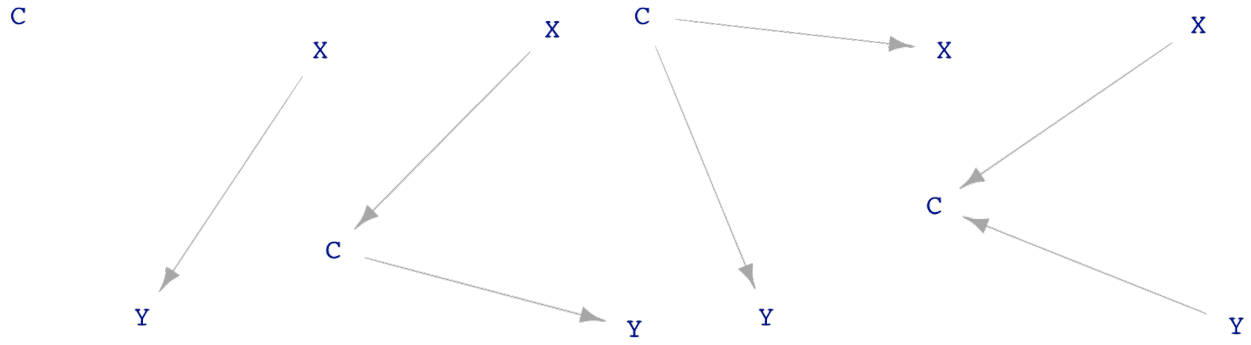
[9]In any given real-world context, good practice calls for considering each possible DAG structure and concocting a story behind it. Such stories will sometimes be implausible, but there can also be surprises that give the modeler new insight.



(a) C is a confounder. (b) C is a mechanism. (c) C is a consequence.

Figure 19: Three different DAGs connecting X, Y, and C.

To understand how a DAG informs whether or not to include a covariate, It will help to give general names to some of the sub-structures seen in the Figure 19 DAGs. **?@fig-dag-paths** shows some of these sub-structures, removing other links that are not part of the structure.



(a) Direct causal link from X to Y (b) Causal path from X through C to Y (c) Correlating path connecting X and Y via C (d) C as a consequence of X and Y

Figure 20: Sub-structures seen in Figure 19.

- A "**direct causal link**" between X and Y. There are no intermediate nodes.
- A "**causal path**" from C to X and on to Y. A causal path is one where, starting at the originating node, flow along the arrows can get to the terminal node, passing through

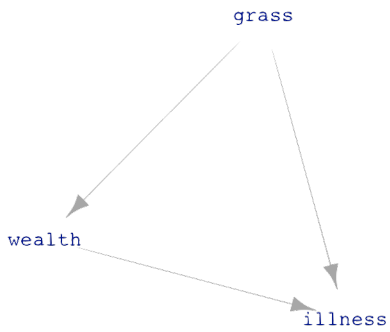
all intermediate nodes.

- A "**correlating path**" from Y through X to C. Correlating paths are distinct from causal paths because, in a correlating path, there is no way to get from one end to the other by following the flows.
- A "**collider**" `wealth`. In other words, both X and Y are causes of C.

Look back to Figure 19(a), where `wealth` is a confounder. A confounder is always an intermediate node in a *correlating path*.

Including a covariate either blocks or opens the pathway on which that covariate lies. Which it will be depends on the kind of pathway. A causal path, as in Figure 20(b), is blocked by including the covariate. Otherwise, it is open. A correlating path (**?@fig-dags-path**(c)) is similar: the path is open unless the covariate is included in the model. A colliding path, as in Figure 20(d), is blocked *unless* the covariate is included—the opposite of a causal path.

Often, covariates are selected to block all paths except the direct link between the explanatory and response variable. This means *do* include the covariate if it is on a correlating path and *do not* include it if the covariate is at the collision point.



As for a causal path, the choice depends on what is to be studied. Consider the DAG drawn in Figure 19(b), reproduced here for convenience:

`grass` influences `illness` through two distinct paths:

i. the direct link from `grass` to `illness`.
ii. the causal pathway from `grass` through `wealth` to `illness`.

Admittedly, it is far-fetched that choosing to green the grass makes a household wealthier, but focus on the topology of the DAG and not the unlikeliness of this specific causal scenario.

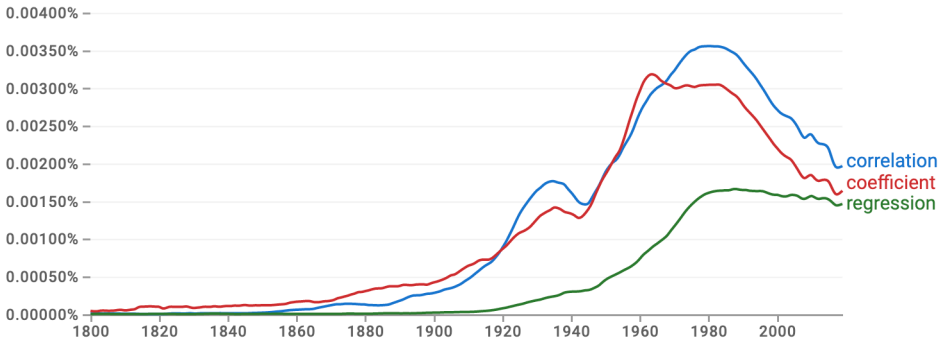
There is no way to block a direct link from an explanatory variable to a response. If there were a reason to do this, the modeler probably selected the wrong explanatory variable.

But there is a genuine choice to be made about whether to block pathway (ii). If the interest is the purely biochemical link between grass-greening chemicals and illness, then block pathway (ii). However, if the interest is in the *total* effect of `grass` and `illness`, including both biochemistry and the sociological reasons why `wealth` influences `illness`, then leave the pathway open.

> ⚠ In draft: Some resources
>
> https://towardsdatascience.com/causal-effects-via-dags-801df31da794
> https://towardsdatascience.com/causal-effects-via-the-do-operator-5415aefc834a

12 Spurious correlation

Google NGram provides a quick way to track word usage in books over the decades. Figure 21 shows the NGram for three statistical words: coefficient, correlation, and regression.



: Google NGram for "co- "correlation," and "regres-

The use of "correlation" started in the mid to late 1800s, reached an early peak in the 1930s, then peaked again around 1980. "Correlation" is tracked closely by "coefficient." This parallel track might seem evident to historians of statistics; the quantitative measure called the "**correlation coefficient**" was introduced by Francis Galton in 1888 and quickly became a staple of statistics textbooks.

In contrast to mainstream statistics textbooks, "correlation" barely appears in these lessons (until this chapter). There is a good reason for this. Although the correlation coefficient measures the "strength" of the relationship between two variables, it is a special case of a more general and powerful method that appears throughout these Lessons: regression modeling.

Figure 21 shows that "regression" got a later start than correlation. That is likely because it took 30-40 years before it was appreciated that correlation could be generalized. Furthermore, regression is more mathematically complicated than correlation, so practical use of regression relied on computing, and computers started to become available only around 1950.

12.1 Correlation

A dictionary is a starting point for understanding the use of a word. Here are four definitions of "correlation" from general-purpose dictionaries.

> "*A relation existing between phenomena or things or between mathematical or statistical variables which tend to vary, be associated, or occur together in a way not expected on the basis of chance alone*" Source: Merriam-Webster Dictionary

> "*A connection between two things in which one thing changes as the other does*" Source: Oxford Learner's Dictionary

> "*A connection or relationship between two or more things that is not caused by chance. A positive correlation means that two things are likely to exist together; a negative correlation means that they are not.*" Source: Macmillan dictionary

> "A mutual relationship or connection between two or more things," "interdependence of variable quantities." Source: [Oxford Languages]

All four definitions use "connection" or "relation/relationship." That is at the core of "correlation." Indeed, "relation" is part of the word "correlation." One of the definitions uses "causes" explicitly, and the everyday meaning of "connection" and "relation" tend to point in this direction. The phrase "one thing changes as the other does" is close to the idea of causality, as is "interdependence.:

Three of the definitions use the words "vary," "variable," or "changes." The emphasis on variation also appears directly in a close statistical synonym for correlation: "covariance."

Two of the definitions refer to "chance," that correlation "is not caused by chance," or "not expected on the basis of chance alone." These phrases suggest to a general reader that correlation, since not based on chance, must be a matter of fate: pre-determination and the action of causal mechanisms.

We can put the above definitions in the context of four major themes of these Lessons:

- Quantitative description of relationships
- Variation
- Sampling variation
- Causality

Correlation is about relationships; the "correlation coefficient" is a way to describe a straight-line relationship quantitatively. The correlation coefficient addresses the tandem variation of quantities, or, more simply stated, how "one thing changes as the other does."

To a statistical thinker, the concern about "chance" in the definitions is not about fate but reliability. Sampling variation can lead to the appearance of a pattern in some samples of a process that is not seen in other samples of that same process. Reliability means that the pattern will appear in a large majority of samples.

> **Note**
>
> One of the better explanations of "correlation" appears in an 1890 article by Francis Galton, who invented the correlation coefficient. Since the explanation is more than a century old, some words will be unfamiliar to the modern reader. For example, a "clerk" is an office worker. An "omnibus" is merely a means of public transportation today.
>
> > *Two clerks leave their office together and travel homewards in the same and somewhat unpunctual omnibus every day. They both get out of the omnibus at the same halting-place, and thence both walk by their several ways to their respective homes. … The upshot is that when either clerk arrives at his home later than his average time, there is some reason to expect that the other clerk will be late also, because the retardation of the first clerk may have been wholly or partly due to slowness of the omnibus*

> *on that day, which would equally have retarded the second clerk. Hence their unpunctualities are related. If the omnibus took them both very near to their homes, the relation would be very close. If they lodged in the same house and the omnibus dropped them at its door, the relation would become identity.*
>
> *The problems of … correlation deal wholly with departures or variations ; they pay no direct regard to the central form from which the departures or variations are measured. If we were measuring statures, and had made a mark on our rule at a height equal to the average height of the race of persons whom we were considering, then it would be the distance of the top of each man's head from that mark, upward or downward as the case might be, that is wanted for our use, and not its distance upward from the ground.*[a]

[a] Francis Galton (1890) "Kinship and Correlation" *The North American Review* 150(401) URL

12.2 Spurious causation

The "Spurious correlations" website http://www.tylervigen.com/spurious-correlations provides entertaining examples of correlations gone wrong. The running gag is that the two correlated variables have no reasonable association, yet the correlation coefficient is very close to its theoretical maximum of 1.0. Typically, one of the variables is morbid, as in Figure 22.

According to Aldrich (1995)^[John Aldrich (1994) "Correlations Genuine and Spurious in Pearson and Yule" *Statistical Science* 10(4) URL the idea of **spurious correlations** appears first in an 1897 paper by statistical pioneer and philosopher of science Karl Pearson. The correlation coefficient method was published only in 1888, and, understandably, early users encountered pitfalls. One very early user, W.F.R. Weldon, pub-

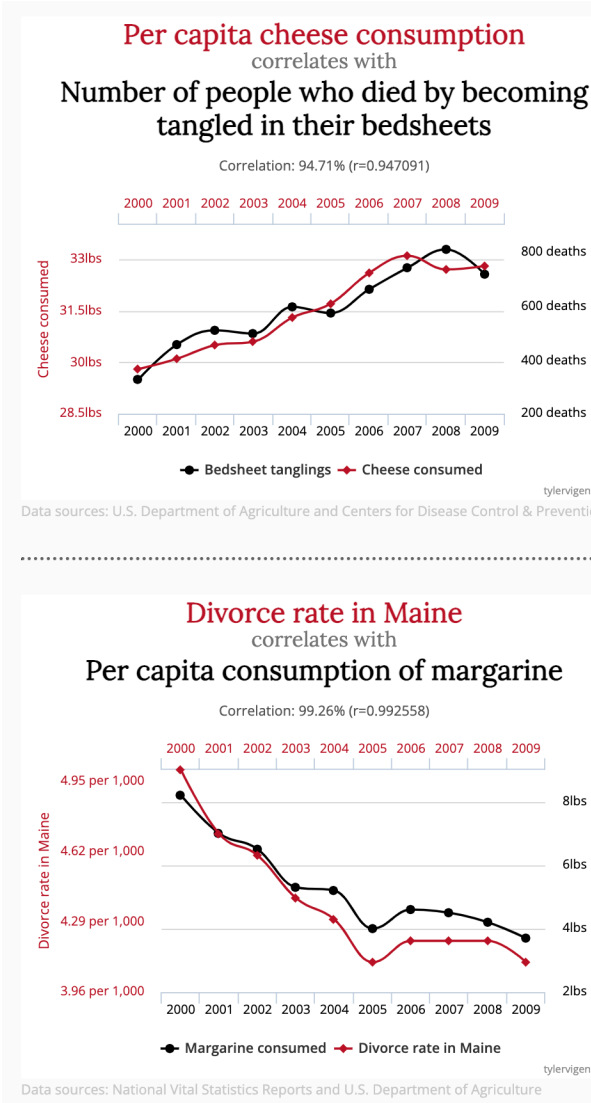
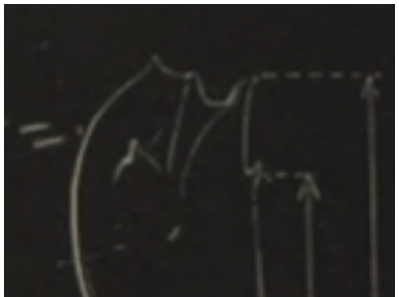
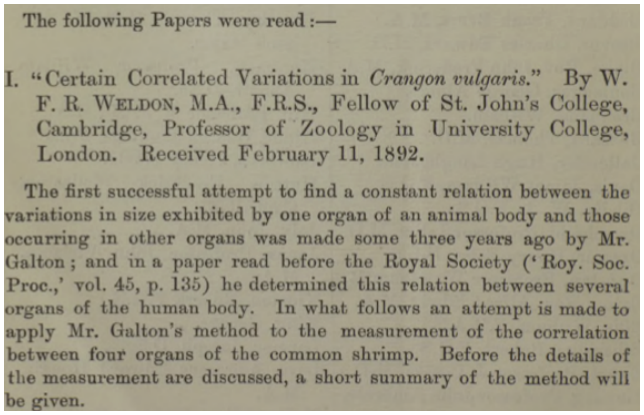


Figure 22: Two examples from the Spurious correlations website



lished a study in 1892 on the correlations between the sizes of organs, such as the tergum and telson in shrimp. (See Figure 23.)



The following Papers were read :—

I. "Certain Correlated Variations in *Crangon vulgaris*." By W. F. R. Weldon, M.A., F.R.S., Fellow of St. John's College, Cambridge, Professor of Zoology in University College, London. Received February 11, 1892.

The first successful attempt to find a constant relation between the variations in size exhibited by one organ of an animal body and those occurring in other organs was made some three years ago by Mr. Galton; and in a paper read before the Royal Society ('Roy. Soc. Proc.,' vol. 45, p. 135) he determined this relation between several organs of the human body. In what follows an attempt is made to apply Mr. Galton's method to the measurement of the correlation between four organs of the common shrimp. Before the details of the measurement are discussed, a short summary of the method will be given.

Pearson noticed a distinctive feature of Weldon's method. Weldon measured the tergum and telson as a fraction of the overall body length.

Figure 24 shows one possible DAG interpretation where `telson` and `tergum` are *not* connected by any causal path. Similarly, `length` is exogenous with no causal path between it and either `telson` or `tergum`. ::: {.cell .column-margin}

```
shrimp_dag <- dag_make(
  tergum ~ unif(min=2, max=3),
  telson ~ unif(min=4, max=5),
  length ~ unif(min=40, max=80),
  x ~ tergum/length + exo(.01),
  y ~ telson/length + exo(.01)
)
# dag_draw(shrimp_dag, seed=101, vertex.label.cex=1)
knitr::include_graphics("www/telson-tergum.png")
```

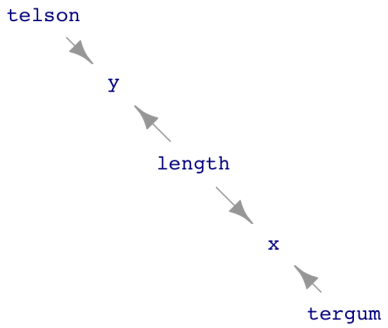


Figure 24: DAG for the shrimp measurements.

:::

The Figure 24 shows a hypothesis where there is no causal relationship between telson and tergum. Pearson wondered whether dividing those quantities by `length` to produce variables `x` and `y`, might induce a correlation. Weldon had found a correlation coefficient between `x` and `y` of about 0.6. Pearson estimated that dividing by `length` would induce a correlation between `x` and `y` of about 0.4-0.5, even if telson and tergum are not causally connected.

We can confirm Pearson's estimate by sampling from the DAG and modeling `y` by `x`. The confidence interval on `x` shows a relationship between `x` and `y`. In 1892, before the invention of regression, the correlation coefficient would have been used. In retrospect, we know the correlation coefficient is a simple scaling of the `x` coefficient.

```
Sample <- sample(shrimp_dag, size=1000)
lm(y ~ x, data=Sample) %>% confint()
```

	lwr	upr
(Intercept)	0.0458	0.0523
x	0.6148	0.7566

```
cor(y ~ x, data=Sample)
```

```
[1] 0.5148
```

Pearson's 1897 work precedes the earliest conception of DAGs by three decades. An entire century would pass before DAGs came into widespread use. However, from the DAG of Figure 24] in front of us, we can see that `length` is a common cause of `x` and `y`.

Within 20 years of Pearson's publication, a mathematical technique called "**partial correlation**" was in use that could deal with this particular problem of spurious correlation. The key is that the model should include `length` as a covariate. The covariate correctly blocks the path from `x` to `y` via `length`.

```
lm(y ~ x + length, data=Sample) %>% confint()
```

	lwr	upr
(Intercept)	0.1508	0.1635
x	-0.0363	0.0834
length	-0.0014	-0.0013

The confidence interval on the `x` coefficient includes zero once `length` is included in the model. So the data, properly analyzed, show no correlation between telson and tergum.

In this case, "spurious correlation" stems from using an inappropriate method. This situation, identified 130 years ago and addressed a century ago, is still a problem for those who use the correlation coefficient. Although regression allows the incorporation of covariates, the correlation coefficient does not.

> **i** Time series analysis
>
> Some spurious correlations, such as those presented on the eponymous website, can also be attributed to methodological error.
> One source of error was identified in 1904 by F.E. Cave-Browne-Cave in her paper "On the influence of the time factor on the correlation between the barometric heights at stations more than 1000 miles apart," published in the

Proceedings of the Royal Society. "Miss Cave," as she was referred to in 1917 and 1921, respectively by eminent statisticians William Sealy Gosset (who published under the name "Student") and George Udny Yule, also offered a solution to the problem. Her solution is very much in the tradition of "**time-series analysis**," a contemporary specialized area of statistics.

The unlikeliness of the correlations on the website is another clue to their origin as methodological. Nobody woke up one morning with the hypothesis that cheese consumption and bedsheet mortality are related. Instead, the correlation is the product of a search among many miscellaneous records. Imagine that data were available on 10,000 annually tabulated variables for the last decade. These 10,000 variables create the opportunity for 50 million pairs of variables. Even if none of these 50 million pairs have a genuine relationship, sampling variation will lead to some of them having a strong correlation coefficient.

In statistics, such a blind search is called the "multiple comparisons problem." Ways to address the problem have been available since the 1950s. (We will return to this topic under the label "false discovery" in Lesson 19.) Multiple comparisons can be used as a trick, as with the website. However, multiple comparisons also arise naturally in some fields. For example, in molecular genetics, "microarrays" make a hundred thousand simultaneous measurements of gene expression. Correlations in the expression of two genes give a clue to cellular function and disease. With so many pairs available, multiple comparisons will be an issue.

12.3 "Correlation implies causation."

Francis Galton's 1890 example of the clerks on the bus introduces "correlation" as a causality story. The bus trip causes variation in commute times. Two clerks riding the same bus will have correlated commute times. In the dictionary definitions of "correlation" at the start of the Lesson, the words "connection,"

"relationship," and "interdependence" suggests causal connections.

Insofar as the dictionary definitions of correlation suggest a causal relationship, they are at odds with the statistical mainstream, which famously holds that "correlation does not imply causation." This view is so entrenched that it appears on tee shirts, one style of which is available for sale by the American Statistical Association.

The statement "A is not B" can be valid only if we know what A and B are. We have a handle on the meaning of "correlation." So what is the meaning of "causation?"

Dictionaries define "causation" using the word "cause." So we look there for guidance.

> A person or thing that gives rise to an action, phenomenon, or condition. Source: Oxford Languages
>
> An event, thing, or person that makes something happen. Source: Macmillan Dictionary
>
> A person or thing that acts, happens, or exists in such a way that some specific thing happens as a result; the producer of an effect. Source: Dictionary.com

Interpreting these definitions requires making sense of "give rise to," "makes happen," or "happens as a result." All of them are synonyms for "cause."

This circularity produces a muddle. Centuries of philosophical debate have yet to clarify things much.

Still, we can do something. The point of view of these Lessons is to support decision-making. Causation is a valuable concept for decision-making, particularly in cases where the decision-maker is considering an *intervention*. With this as an anchor, a pragmatic definition of "causation" is available:

> Causation describes a class of hypotheses that DAGs can represent. In that representation, a causal relationship between two nodes X and Y is marked by a causal path connecting X to Y. In

> Lesson 11, we defined "causal path" in terms of the directions of arrows in a DAG.[10] A definitive demonstration of a causal relationship between X and Y is that intervening to change X results reliably in a change in Y, *all other nodes not on the causal path being held constant.* (Lesson 13 treats the methodology behind this definitive sign.)

Whether or not a definitive demonstration is feasible is not directly relevant to the decision-maker. A decision-maker acts under the guidance of one or more hypotheses. A good rule of thumb for decision-makers is to be guided only by plausible hypotheses. Whether a hypothesis is plausible is a matter of informed belief. A definitive demonstration should sharpen that belief. If no such definitive demonstration is available, the decision-maker must rely on alternative sources for belief. Austin Bradford Hill (1898-1991), an epidemiologist and eminent statistician, famously published a list of nine criteria that support belief in a causal hypothesis.

Using my definition of causation, and in marked disagreement with many statisticians, I submit that

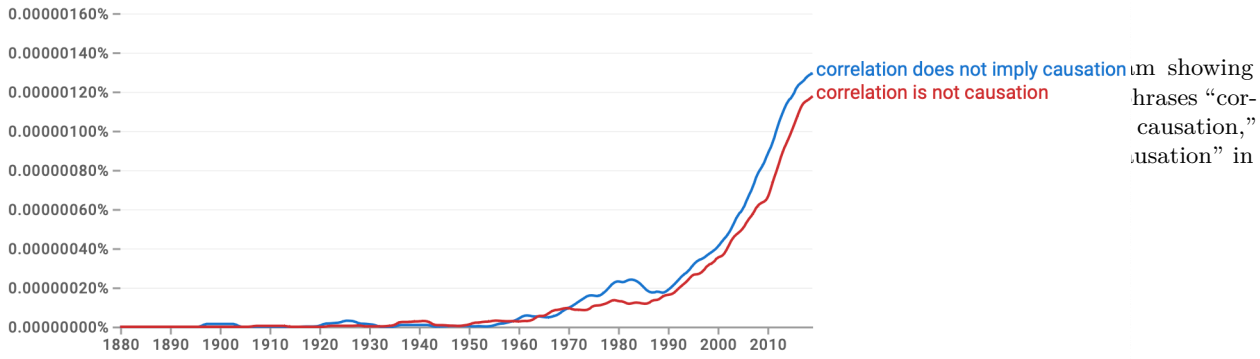
> *Correlation implies causation.*

"Correlation implies causation" is not the same as saying, "A correlation between A and B implies that A causes B." That statement is false. For instance, it might be instead that B causes A. Alternatively, there might be a common cause C for both A and B. Or, C might be a collider between A and B.

There is no mechanism to produce correlation that I am aware of, other than the sources of spurious correlation described previously, that does not involve causation in some way.

> **i** So why do many statisticians say different?
>
> Historically, the rise of the expression "correlation does not imply causation"—Figure 25 shows the ngram since the 1888 invention of the correlation coefficient—comes *after* the peak in the use of the word "correlation."

[10] We will consider a "direct causal link" to be a form of causal path.



m showing
hrases "cor-
causation,"
usation" in

The first documented use of the phrase is from 1900. It comes in a review of the second edition of a book, *The Grammar of Science*, by Karl Pearson (whom we have met before in this Lesson).

The Grammar of Science is a metaphysically oriented prescription for a new type of science. It posited that sciences such as physics or chemistry unnecessarily drew on metaphors for causation, such as "force." Instead, the book advocated another framework as more appropriate, eschewing causation in favor of descriptions of "perceptions" with probability.

Pearson illustrates his antipathy toward causation with an example of an ash tree in his garden:

> [T]he causes of its growth might be widened out into a description of the various past stages of the universe. One of the causes of its growth is the existence of my garden, which is conditioned by the existence of the metropolis [London]; another cause is the nature of the soil, gravel approaching the edge of the clay, which again is conditioned by the geological structure and past history of the earth. The causes of any *individual* thing thus widen out into the unmanageable history of the universe. *The Grammar of Science*, 2/e, p. 131

It should not be surprising that the field of statistics, which uses probability very extensively as a description,

and that developed correlation as a measure of probability, would advocate for more general use of its approach. In this spirit, I read "correlation does not imply causation" as "our new science framework of probability and correlation replaces the antiquated framework of causation." Outside of statistics, however, probability is merely a tool; causation does indeed have practical use. All the more so for decision-makers.

13 Experiment and random assignment

In its everyday meaning, the word "experiment" is similar in meaning to the word "experience." As a verb, to experiment means to "try out new concepts or ways of doing things." As a noun, an experiment is a "course of action tentatively adopted without being sure of the outcome: the farm is an ongoing experiment in sustainable living." Both quotes are from the Oxford Languages, which provides examples of each: "the designers experimented with new ideas in lighting" or "the farm is an ongoing experiment in sustainable living."

From movies and other experiences, people associate experiments with science. Indeed, one of the dictionary definitions of "experiment" is: "a scientific procedure undertaken to make a discovery, test a hypothesis, or demonstrate a known fact."

Almost all the knowledge needed to perform a scientific experiment relates to the science itself: what reagents to use, how to measure the concentration of a neurotransmitter, how to administer a drug safely, and so on. This is why people who carry out scientific procedures are trained primarily in their area of science.

Example: Malaria and bed nets

In many parts of the world, malaria is a major cause of disability and death. Economists who study ways to relieve poverty have a simple, plausible theory: reducing the effect of illnesses such as malaria will have an impact on poverty rates, since healthier people are more productive and reduced uncertainty can help them amass capital to invest to increase production further.

There are many possible ways to reduce the burden of malaria. Vaccination (although effective vaccines have been hard to develop), insect control using pesticides (which can cause environmental problems), etc. One simple intervention is the use of bed nets; screen nets deployed at night by draping over the bed and its occupant. Still, there are reasons why distributing bed nets may not be effective; people might use them incorrectly or for other purposes such as fishing. People might not be able to afford them, but giving them away might signal that they have no value.

To find out, try it: do an experiment. For instance, run a trial program where nets are given away to everyone in an area and observed whether and to what extent rates of malarial illness go down.

Such a trial is certainly an experiment. But it may not be the best way to get meaningful information.

13.1 Replication

To understand some of the contribution that statistical thinking can make to experiment, recall our earlier definition:

> *Statistic thinking is the explanation/description of variation* in the context of *what remains unexplained/undescribed.*

A key concept that statistical thinking brings to experi-

ment is the idea of **variation**. Simply put, a good experiment should involve some variation. The simplest way to create variation is to repeat each experimental trial multiple times. This is called "**replication**."

13.2 Example: Replicated bed net trials

One way to improve the simple experiment bed net described above is to carry out many trials. One reason is that the results from any single trial might be shaped by accidental or particular circumstances: the weather in the trial area was less favorable to mosquito reproduction; another government agency decided to help out by spraying pesticides broadly, and so on. Setting up trials in different areas can help to balance out these influences.
Replicated trials also allow us to estimate the size of the variability caused by the accidental or particular factors. To illustrate, suppose a single trial is done and the rate of malarial illness goes down by 5 percentage points. What can we conclude? The result is promising but we can't rule out that it is due to accidental factors other than bed nets. Why not? Because we have no idea how much unexplained variation is in play.
:::: {.column-margin} ::: {#tbl-bed-net .cell .column-margin tbl-cap='`Bed_net_data`'}

?@tbl-bed-net shows data from four imagined trials of the effect of bed nets. (Reduction by a negative number, like -1, is an *increase*.) The mean reduction is 3 percentage points, but this number is not much use unless we can put it in the context of sampling variation. Conducting multiple trials gives us a handle on the amount of sampling variation. By We can easilyNow we know something about the amount of variation due to site-to-site factors. The replication introduces *observed* variation in results, the observed variation can be quantified and used to place the overall trend in context.

Using the regression framework makes it easy to estimate the amount of sampling variation. The mean reduction corresponds to the coefficient from the model `reduction ~ 1`.

```
lm(reduction ~ 1,
     data=Bed_net_data) %>%
  coef()
```

```
(Intercept)
          3
```

```
lm(reduction ~ 1,
     data=Bed_net_data) %>%
  confint()
```

	lwr	upr
(Intercept)	1	5

The observed 3 percentage point mean reduction in the incidence of malaria does stand out from the noise: the confidence interval does not include zero. In these (imagined) data, we have confidence that we have seen a signal.

13.3 Control

However, there is still a problem with the design of the imagined bed-net experiment. What if the year the experiment was done was unusually dry, reducing the mosquito population and, with it, the rate of malaria infection? Then we don't know whether the observed 3 point reduction is due to the weather or the bed nets, or even something else, e.g. better nutrition due to a drop in international prices for rice.

We need to measure what the change in malarial infection would have been without the bed-net intervention. Care needs to be taken here. If the trial sites were rural, it would not be appropriate to look at malarial rates in urban areas where there was no bed-net program. We want to compare the trial sites with non-trial sites where the intervention was not carried out, so-called "**control**" sites. The `With_controls` data frame imagines what data might look like if in half the sites no bed-net program was involved.

Table 4: With_controls

site	reduction	nets
A	2	control
B	8	treatment
C	4	treatment
D	1	treatment
E	-1	control
F	-2	control
G	0	control
H	2	treatment
I	3	treatment
J	2	control

The proper regression model for the `With_controls` data is `reduction ~ treatment`:

```
lm(reduction ~ nets,
       data=With_controls) %>%
  coef()
```

```
  (Intercept) netstreatment
          0.2           3.4
```

```
lm(reduction ~ nets,
       data=With_controls) %>%
  confint()
```

	lwr	upr
(Intercept)	-2.200	2.6
netstreatment	0.058	6.7

The effect of the bed nets is summarized by the `netstreatment` coefficient, which compares the `reduction` between the `treatment` and `control` groups. In this new (imagined) data frame, the confidence interval on `netstreatment` touches very

close to zero; the signal is just barely discernible from the noise.

The reader might wonder why, in moving to the controlled design, the ten sites were not all treated with nets and another ten or so sites found to use as the control. Perhaps, even, the control sites could be selected as villages nearby to the bed net villages.

One reason is pragmatic: the larger study would require more effort and money. The larger study might be worthwhile; larger n would presumably narrow the confidence interval. Another reason, to be expanded on in the next section, is that the treatment and control sites should be as similar as possible. This can be surprising hard to achieve. Other factors such as the enthusiasm or skepticism of the town leaders toward public-health interventions might be behind the choice of the original sites for the bed-net program. The control sites might be towns that turned down the original offer of the bed-net program and, accordingly, have different attitudes toward public health.

13.4 Example: Testing the Salk polio vaccine

Today, most children are vaccinated against polio, though a smaller fraction than in previous years. This might be because symptomatic polio is very rare, lessening the perceived urgency of protecting against it. Partly, the reduction reflects the growth in the "anti-vax" movement, which became especially notable with the advent of COVID-19.

The first US polio epidemic occurred in 1916, just two years before the COVID-like "Spanish flu" pandemic.[11] Up through the early 1950s, polio injured or killed hundreds of thousands of people, particularly children. Anxiety about the disease was similar to that seen in the first year of the COVID-19 pandemic.

There were many attempts to develop a vaccine against polio. Jonas Salk created the first really promising vaccine, the promise being based on laboratory tests. To establish the safety

[11] "Spanish" is in quotes because Spain was not the source of the pandemic.

and effectiveness of the Salk vaccine, it needed to be tried in the field, with people. Two organizations, the US Public Health Service and the National Foundation for Infantile Paralysis got together to organize a clinical field trial which, all told, involved two-million students in grades 1 through 3.

The two studies involved both a treatment and a control group. In some school districts, students in grades 1 and 3 were held as controls. The treatment group was students in grade 2 whose parents gave consent. We will call this "Study 1." In other school districts, the study design was different: the parents of all students in all three grades were asked for consent. The students with parental consent were then randomly split into two groups: a treatment and a control. Call this "Study 2."

The Study 2 design might seem inefficient; it reduced the number of children receiving the vaccine because half of the children with parental consent were left unvaccinated. On the other hand, it might be that children from families who consent to be given a vaccine are different in a systematic way from children whose families refuse, just as today's anti-vax families might be different from "pro-vax" families.

As reported in Freedman (1998)[12], the different risk of symptomatic polio between children from consent versus refuse families became evident in the study. Table 5 shows the study results from the school districts which used half the consent group as controls.

The difference between treatment and control groups is very evident: a reduction from 71 cases per 100,000 children to 28 cases per 100,000. The no-consent children had a rate between the two, 46 per 100,000. Since both the "control" and "no consent" groups did not get the vaccine, one might expect those rates to be similar. That they are not shows that the "no-consent" children are systematically different from those children whose parents gave consent.

In the other branch of the study, Study 1, where no-consent 2nd-graders were used as control and vaccine was given to all whose parents did consent, the results (Table 6) were different because of confounding between treatment and consent.

[12]D. Freedman, R Pisani, R Purves, *Statistics* 3/e, p.6

Table 5: Results from Study 2.

vaccine	size	rate
Treatment	200000	28
Control	200000	71
No consent	350000	46

Table 6: Results from Study 1

vaccine	size	rate
Treatment	225000	25
No consent	125000	44

The effect of the vaccine from Study 1 under-estimated the biological link between vaccination and reduction of polio risk.

13.5 Random assignment

The example of the Salk vaccine trial is a chastening reminder that care must be taken when assigning `treatment` or `control` to the units in an experiment. Without such care, confounding enters into the picture. Merely the possibility of confounding is damaging to the experiment's result; it invites skepticism and doubt.

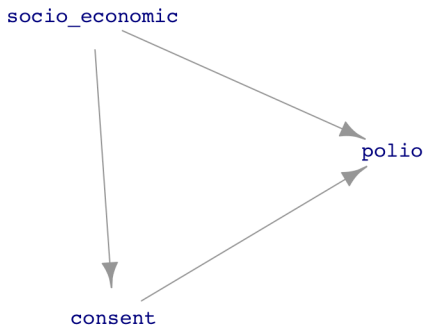


Figure 26: A DAG for the polio vaccine experiment.

It is illuminating to look at the vaccine trial as a DAG. The essential situation is diagrammed in Figure 26. The `socio_economic` node represents the idea that socio-economic status has an influence on susceptibility to symptomatic polio[13] and also is a factor in shaping a family's decision about

[13]In contrast to the usual expectation that lower socio-economic status is associated higher risk of disease, with polio the opposite holds true. The explanation usually given is that children who are exposed to the polio virus as infants do not become sick but do gain immunity to later infection. People later in childhood and in adulthood are at risk of a severe, symptomatic response to exposure. Polio is transmitted mainly via a fecal-oral route. Conditions favoring this route are more common among those of low socio-economic status. Consequently, infants of well-to-do families are less exposed to the virus and do not develop immunity. When they are eventually exposed to polio as children or

giving consent.

The DAG in Figure 26 has two pathways between `treatment` and `polio` that can produce confounding:

- `treatment` ← `consent` → `polio`
- `treatment` ← `consent` ← `socio.economic` → `polio`

The approach emphasized in Lesson 11 to avoid such confounding is to block the relevant pathways. Both can be blocked by including `consent` as a covariate. However, in Study 1, assignment to vaccine was purely a matter of consent; `consent` and `treatment` are essentially the same variable. Figure 27 shows the corresponding DAG, where `consent` and `treatment` are merged into a single variable. Holding `consent` constant deprives the system of the explanatory variable and still introduces confounding through `socio_economic`.

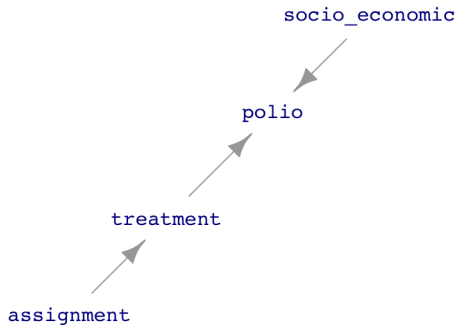


Figure 27: The DAG when `consent` ≡ `vaccine`.

In Study 2, all the children participating had parents give consent. This means that `consent` is not actually a variable; it doesn't vary! The corresponding DAG, without `consent` as a factor, is drawn in Figure 28. This Study 2 DAG is unfolded; there are no confounding pathways! Thus the model `polio ~ treatment` is appropriate.

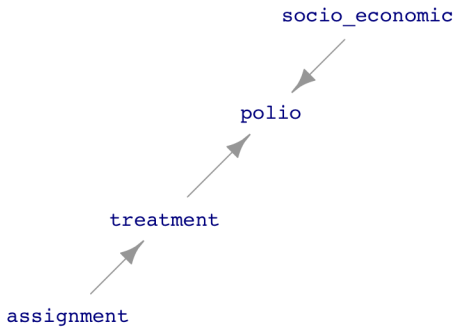


Figure 28: The Study 2 DAG.

The assignment to treatment or control in Figure 28 is made by the people running the study. Although the DAG doesn't show any inputs to `assignment`, the involvement of people in making the assignment opens up a possibility that their assignment of treatment or control might have been influenced by other factors, such as socio-economic status. To guard against this, or even skepticism raised by the possibility, experimentalists have developed a simple safeguard: "**random assignment**." In random assignment, assignment is made by a computer generating random numbers. Nobody believes that the computer algorithm is influenced by socio-economic status or any other factor that might be connected to polio in any way.

> ⚠ Under construction

adults, the well-to-do are at greater risk of developing disease.

13.6 Blocking

14 Measuring and accumulating risk

The *linear models* (`lm()`) we have mostly been using up until now accumulate the model output as a linear combination of model inputs. Consider, for instance, a simple model of fuel economy based on the horsepower and weight of a car:

```
mpg_mod <- lm(mpg ~ hp + wt, data = mtcars)
mpg_mod %>% coef()
```

```
(Intercept)          hp          wt
   37.22727    -0.03177    -3.87783
```

These coefficients mean that the model output is a **sum**. For instance, a 100 horsepower car weighting 2500 pounds has a predicted fuel economy of `37.2 - 0.032*100 - 3.88*2.5`=24.3 miles per gallon.[14] If we're interested in making a prediction, we often hide the arithmetic behind a computer function, but it is the same arithmetic:

```
mod_eval(mpg_mod, hp = 100, wt = 2.5)
```

hp	wt	model_output
100	2.5	24.36

The arithmetic, in principle, lets us evaluate the model for any inputs, even ridiculous ones like a 10,000 hp car weighing 50,000 lbs. There is no such car, but there is a model output.[15]

```
mod_eval(mpg_mod, hp=10000, wt = 50)
```

hp	wt	model_output
10000	50	-474.4

[14] The `wt` variable is measured in units of 1000 lbs, so a 2500 pound vehicle has a `wt` value of 2.5.

[15] A 10,000 hp, 50,000 lbs ground vehicle does have a name: a "tank." Common sense dictates that one not put too much stake in a calculation of a tank's fuel economy based on data from cars!

The prediction reported here means that such a car goes *negative* 474 miles on a gallon of gas. That's silly. Fuel economy needs to be non-negative; the output −474 mpg is *out of bounds*.

A good way to avoid out-of-bounds behavior is to model a *transformation* of the response variable instead of the variable itself. For example, to avoid negative outputs from a model of `mpg`, change the model so that the output is in terms of the logarithm of `mpg`, like this:

```
logmpg_mod <- lm(log(mpg) ~ hp + wt, data = mtcars)
mod_eval(logmpg_mod, hp = 100, wt = 2.5)
```

hp	wt	model_output
100	2.5	3.173

The reported output, 3.17, should **not** be interpreted as `mpg`. Instead, interpret it as `log(mpg)`. If we want output in terms of `mpg`, then we have to undo the logarithm. That's the role of the exponential function, which is the *inverse* of the logarithm.

```
mod_eval(logmpg_mod, hp = 100, wt = 2.5) %>%
  mutate(mpg = exp(model_output))
```

hp	wt	model_output	mpg
100	2.5	3.173	23.89

The logarithmic transform at the model-training stage does not not prevent the model output from being negative. We can see this by looking at the tank example:

```
mod_logmpg <- lm(log(mpg) ~ hp + wt, data = mtcars)
mod_eval(mod_logmpg, hp=10000, wt=50) %>%
  mutate(mpg = exp(model_output))
```

hp	wt	model_output	mpg
10000	50	-21.63	0

The model output is negative for the tank, but the model output corresponds to `log(mpg)`. What will keep the model from producing negative `mpg` will be the exponential transformation applied to the model output. A graph of the exponential function shows how this works.

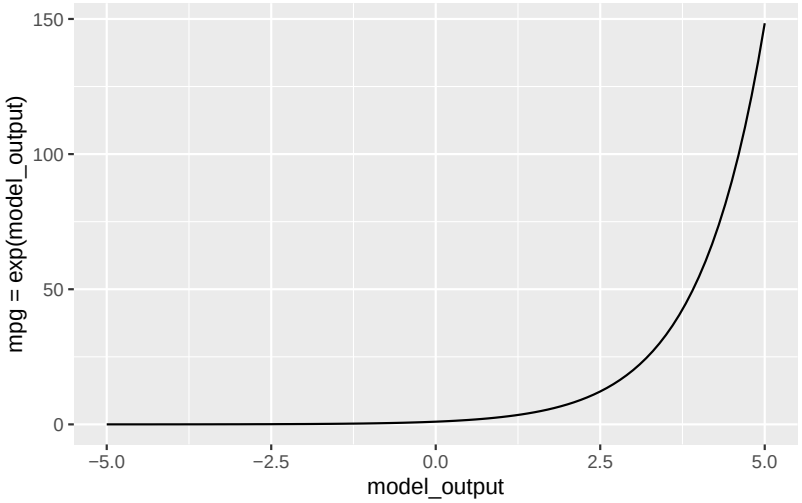


Figure 29: The exponential function translates `log(mpg)`, which can be positive or negative, into `mpg`, which can only be non-negative.

The log transform does not fix the absurdity of modeling tanks based on the fuel economy of cars. The model's prediction of `mpg` for the tank is 0.0000000004 miles/gallon, but real-world tanks do much better than that. For instance, the M1 Abrams tank is reported to get approximately 0.6 miles per gallon.

14.1 Risk

In everyday language, "risk" refers to a dangerous or unwelcome outcome. We talk about the "risk of heart disease" or the "risk of bankruptcy" or other financial loss. To apply risk to a positive outcome is non-idiomatic. For instance, for a person wanting to have a baby, we don't talk about the "risk of pregnancy," but about the "chances of becoming pregnant."

In statistics, the word "risk" is similarly used for an unwelcome outcome. However, an additional component of meaning is added. "Risk" refers to the **probability** of the unwelcome outcome. In principle, though, it would be entirely equivalent to speak of the "probability of heart disease," leaving the

phrase "heart disease" to signal that the outcome is unwanted. We talk about the "risk of death" but never the "risk of survival." Instead, we would say something like the "chances of survival."

The outcomes described by "risk" are categorical. Generically, the levels of the categorical variable might be "unwelcome" and "not unwelcome," but they might be more specifically named, say, "death" and "survival," or "lung disease" and "not."

We have been building models of such categorical output variables from the start of these Lessons. For the zero-one categorical variables we have emphasized, the model output is in the form of a probability: the probability of the outcome of the event being "one" (or whatever actual level "one" corresponds to.) If we assign one for "death" and zero for "survival," the probability which is the output of a model is a risk, but other than the choice of zero-one assignment, there is no mathematical difference (in statistics) between a risk and a probability.

It often happens that risk depends on other factors, often called "risk factors." In our modeling framework, such risk factors are merely explanatory variables. For instance, a study of the impact of smoking on health might use `outcome` represented by a categorical response variable with levels "death" or "survival."

To summarize, for statistical thinkers a model of risk takes the usual form that we have used for models of zero-one categorical models. All the same issues apply: covariates, DAGs, confidence intervals, and so on. There is, however, a slightly different style for presenting effect sizes.

Up until now, we have presented effect in terms of an arithmetic difference. As an example, we turn to the fuel-economy model introduced at the beginning of this lesson. Effect sizes are about *changes*. To look at the effect size of, say, weight (`wt`), we would calculate the model output for two cars that differ in weight (but are the same for the other explanatory variables). For instance, to know the change in fuel economy due to a 1000 pound change in weight, we can do this calculation:

```
mod_eval(logmpg_mod, hp = 100, wt = c(2.5, 3.5)) %>%
  mutate(mpg = exp(model_output))
```

hp	wt	model_output	mpg
100	2.5	3.173	23.89
100	3.5	2.973	19.55

The lighter car is predicted to get 24 mpg, the heavier car to get 19.5 mpg. The arithmetic difference in output $19.5 - 24 = -4.5$ mpg is the effect of the 1000 pound increase in weight.

There is another way to present the effect, as a **ratio** or proportion. In this style, the effect of an addition 1000 pounds is $19.5/24 = 81\%$, that is, the heavier car can go only 81% of the distance that the lighter car will travel on the same amount of gasoline. (Stating an effect as a ratio is common in some fields. For example, economists use ratios when describing prices or investment returns.)

In presenting a change in risk—that is, the change in probability resulting from a change in some explanatory variable—both the arithmetic difference and arithmetic ratio forms are used. But there is a special terminology that is used to identify the two forms. A change in the form of an arithmetic difference is called an "**absolute** change in risk." A change in the ratio form is called a "**relative** risk."

The different forms—absolute change in risk versus relative risk—both describe the same change in risk. For most decision-makers, the absolute form is most useful. To illustate, suppose exposure to a toxin increases the risk of a disease by 50%. This would be a risk ratio of 1.5. But that risk ratio might be based on an absolute change in risk from 0.00004 to 0.00006, or it might be based on an absolute change in risk from 40% to 60%. The latter is a much more substantial change in risk and ought to warrant more attention from decision makers interested.

Other ways to measure change in risk

It is important for measures of change in risk to be mathematically valid. But from among the mathematically valid measures, one wants to choose a form that will be the best for communicating with decision-makers. Those decision-makers might be the people in charge of establishing screening for diseases like breast cancer, or a judge and jury deciding the extent to which blame for an illness ought to be assigned to second-hand smoke.

Two useful ways to present a change in risk are the "**number needed to treat**" (NNT) and the "**attributable fraction**." The NNT is useful for presenting the possible benefits of a treatment or screening test. Consider these data from the US Preventive Services Task Force which take the form of the number of breast-cancer deaths in a 10-year period avoided by mammography. The confidence interval on the estimated number is also given.

Age	Deaths avoided	Conf. interval
40-49	3	0-9
50-59	8	2-17
60-69	21	11-32
70-74	13	0-32

The table does not give the risk of death, but rather the absolute risk reduction. For the 70-74 age group this risk reduction is 13/100000 with a confidence interval of 0 to 32/100000.

The NNT is well named. It gives the number of people who must receive the treatment in order to avoid one death. Arithmetically, the NNT is simply the reciprocal of the absolute risk reduction. So, for the 70-74 age group the NNT is 100000/13 or 7700 or, stated as a confidence interval, [3125 to ∞].

For a decision-maker, NNT presents the effect size in a readily understood way. For example, the 40-49 year-old group has an NTT of 33,000. The cost of the treatment could be presented in terms of anxiety prevented (mammography produces a lot of false positives) or monetary cost. The US Affordable Care Act requires health plans to fully cover the cost of a screening mammogram every one or two years for women over 40. Those mammograms each cost about $100-200. Consequently, the cost of mammography over the ten-year period (during which 5 mammograms might be performed) is roughly $5 \times \$100 \times 33000$ or about $16 million per life saved.

The attributable fraction is a way of presenting a risk ratio—in other words, a relative risk—in a way that is more concrete than the ratio itself. Consider the effect of smoking on the risk of getting lung cancer. According to the US Centers for Disease Control, "People who smoke cigarettes are 15 to 30 times more likely to get lung cancer." This statement directly gives the confidence interval on the relative risk: [15 to 30].

The attributable fraction refers to the proportion of disease in the exposed group—that is, smokers—to be attributed to expose. The general formula for attributable fraction is simple. If the risk ratio is denoted RR, the attributable fraction is

$$\text{attributable fraction} \equiv \frac{RR-1}{RR}$$

For a smoker who gets lung cancer, the confidence interval on the attributable fraction is [93% to 97%].

For second-hand smoke, the CDC estimates the risk ratio for cancer at [1.2 to 1.3]. For a person exposed to second-hand smoke who gets cancer, the attributable fraction is [17% to 23%]. Such attributions are useful for those, such as judges and juries, who need to assign a level of blame for a bad outcome.

14.2 Probability, odds, and log odds

⚠ Under construction

A probability—a number between 0 and 1—is the most used measure of the chances that something will happen, but it is not the only way nor the best for all purposes.

Also part of everyday language is the word "odds," as in, "What are the odds?" to express surprise at an unexpected event.

Odds are usually expressed in terms of two numbers, as in "3 to 2" or "100 to 1", written more compactly as 3:2 and 100:1 or even 1.5 and 100, respectively. The setting for odds is an even that might happen or not: the horse Fortune's Chance might win the race, otherwise not; it might rain today, otherwise not; the Red Sox might win the World Series, otherwise not.

The format of a *probability* assigns a number between 0 and 1 to the chances that Fortune's Chance will win, or that it will rain, or that the Red Sox will come out on top. If that number is called p, then the chances of the "otherwise outcome" must be $1 - p$. The event with probability p would be reformatted into odds as $p : (1 - p)$. No information is lost if we treat the odds as a single number, the result of the division $p/(1 - p)$. Thus, when $p = 0.25$ the corresponding odds will be $0.25/0.75$, in other words, $1/3$.

A big mathematical advantage to using odds is that the odds number can be anything from zero to infinity; it's not bounded

within 0 to 1. Even more advantageous for the purposes of accumulating risk is the logarithm of the odds, called "**log odds**." We will come back to this later.

15 Constructing a classifier

There are many yes-or-no conditions. A patient has a disease or does not. A credit-card transaction is genuine or fraudulent.

But it is not always straightforward to figure out at the time the patient comes to the clinic or the credit-card transaction is made, whether the condition is yes or no. If we could wait, the condition might reveal itself: the patient gets critically ill or the credit-hard holder complains about an unauthorized charge. But we can't wait. We want to treat the patient *before* he or she gets critically ill. We want to block the credit-card transaction before it is completed.

Instead of waiting, we measure whatever relevant variables we can when the patient arrives at the clinic or the credit-card transaction has been submitted for approval. For the patient, we might look at the concentration of specific markers for cancer in the blood. For the transaction, we might look at the shipping address to see if it matches the credit-card holder's genuine address. Such variables may provide an indication, imperfect though it may be, of whether the condition is yes or no.

A **classifier** is a statistical model used to *predict* the unknown outcome of a yes-or-no situation from information that is already available. This Lesson concerns three closely related topics about classifiers: how we collect data for training the model, how we summarize the performance of the classifier, and how we "tune" the classifier.

15.1 Identifying cases

Consider this news report and note the time lag between collection of the dietary explanatory variables and the response variable—whether the patient developed pancreatic cancer.

> Higher vitamin D intake has been associated with a significantly reduced risk of pancreatic cancer, according to a study released last week. Researchers

> combined data from two prospective studies that included 46,771 men ages 40 to 75 and 75,427 women ages 38 to 65. They identified 365 cases of pancreatic cancer over 16 years. Before their cancer was detected, subjects filled out dietary questionnaires, including information on vitamin supplements, and researchers calculated vitamin D intake. After statistically adjusting[16] for age, smoking, level of physical activity, intake of calcium and retinol and other factors, the association between vitamin D intake and reduced risk of pancreatic cancer was still significant. Compared with people who consumed less than 150 units of vitamin D a day, those who consumed more than 600 units reduced their risk by 41 percent. - *New York Times*, 19 Sept. 2006, p. D6.

This was not an experiment; it was an observational study without any intervention to change anyone's diet.

15.2 The training sample

In building a classifier, we have a similar situation. Perhaps we can perform the blood test today, but that gives us only the test result, not the subject's true condition. We might have to wait years for that condition to reveal itself. Only at that point can we measure the performance of the classifier.

To picture the situation, let's imagine many people enrolled in the study, some of whom have the condition and some who don't. On Day 1 of the study, we test everyone and get raw score on a scale from 0 to 40. The results are shown in Figure 30. Each glyph is a person. The varying locations are meant to help us later on; for now, just think of them as representing where each person lives in the world. The different shapes of glyph—circle, square, triangle—are meant to remind you that people are different from one another in age, gender, risk-factors, etc.

Each person took a blood test. The raw result from that test is a score from 0 to 40. The distribution of scores is shown in

[16] That is, applying the methods of Lesson 9.

the right panel of the figure. We also show the score in the world-plot; the higher the raw score, the more blue the glyph. On Day 1, it isn't known who has the condition and who does not.

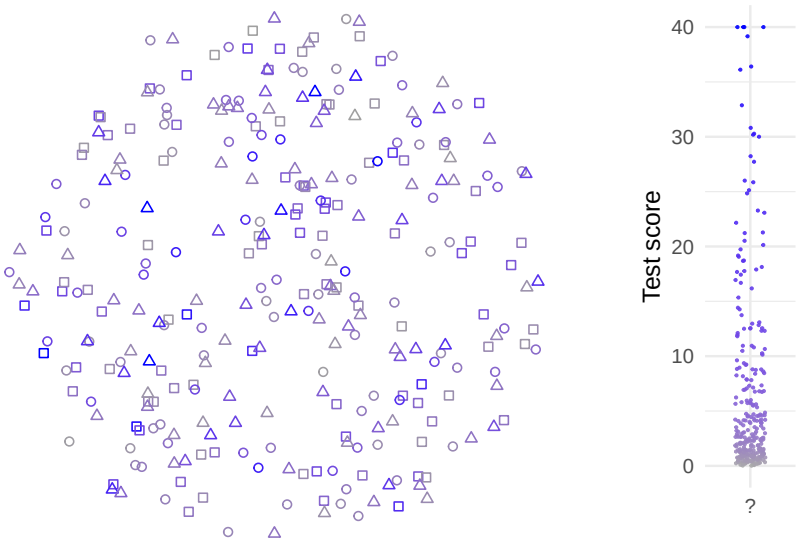


Figure 30: Day 1: The people participating in the study to develop the classifier. Each has been given a blood test which gives a score from zero (gray) to forty (blue).

Having recorded the raw test results for each person, we wait. In the pancreatic cancer study, they waited 16 years for the cancer to reveal itself.

… waiting …

After the waiting period, we can add a new column to the original data; whether the person has the condition (C) or doesn't (H).

Figure 31 shows the distribution of raw test scores for the C group and the H group. The scores are those recorded on Day 1, but after waiting to find out the patients' conditions, we can subdivide them into those who have the condition (C) and those who don't (H).

15.3 Applying a threshold

To finish the classifier, we need to identify a "**threshold score**." Raw scores above this threshold will generate a $\mathbb{P}$ test; scores below the threshold generate a $\mathbb{N}$ test.

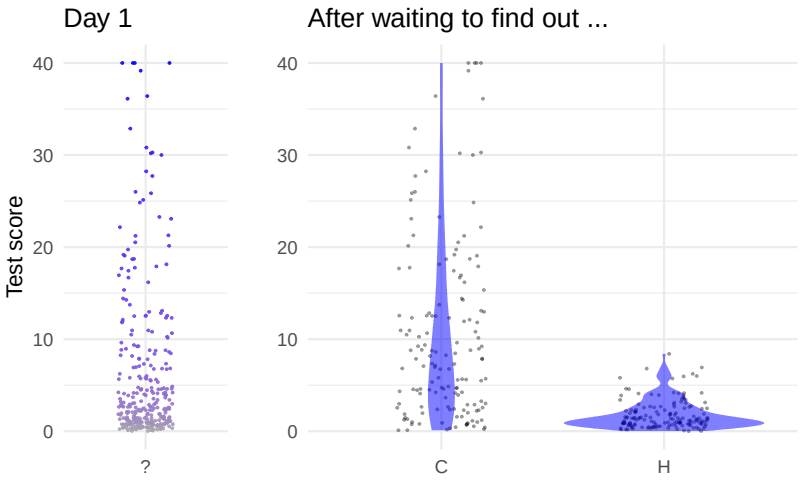


Figure 31: The distribution of raw test scores. After we know the true condition, we can break down the test scores by condition.

We can make a good guess at an appropriate threshold score from the presentation in the right panel of Figure 31. The objective in setting the threshold is to distinguish the C group from the H group. Setting the threshold at a score around 3 does a pretty good job.

It helps to give names to the two test results: $\mathbb{P}$ and $\mathbb{N}$. Anyone with a score above 3 has result $\mathbb{P}$, anyone with a score below 3 has an $\mathbb{N}$ result.

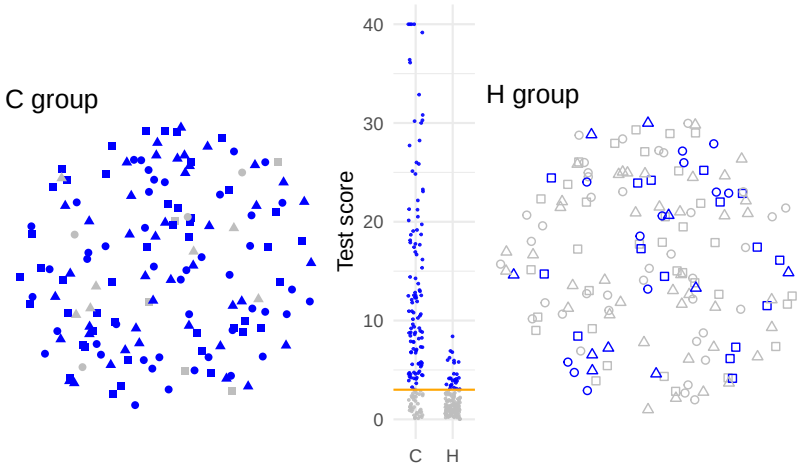


Figure 32: Blue is a $\mathbb{P}$ result, gray a $\mathbb{N}$ result.

15.4 False positives and false negatives

NARRATE Figure 32 to point out the gray dots in the C group and the blue dots in the H group. These are errors. But there are two kinds of errors.

- False-positive: blue dots in the H group. The "positive" refers to the $\mathbb{P}$ test result, the "false" simply means the test result was wrong.
- False-negative: gray dots in the C group. The "negative" refers to the $\mathbb{N}$ result. Again, the "false" means simply that the test result is out of line with the actual condition of the person.

In the training sample shown in Figure 32, there are 300 people altogether and 17 false-negatives. This gives a false-negative rate of about 6%. Similarly there are 30 false-negatives, a false-positive rate of 10%.

i Feature engineering: selling dog food

Naturally, the objective when building a classifier is to avoid errors. One way to avoid errors is by careful "**feature engineering**." Here, "features" refers to the inputs to the classifier model. Often, the designer of the classifier has multiple variables ("features") to work with. (See example.) Choosing a good set of features can be the difference between a successful classifier and one that makes so many mistakes as to be useless.

We will use the name "Bullseye" to refer to a major, national, big-box retailing chain which sells, among many other products, dog food. Sales are largely determined by customer habits; people tend to buy where and what they have previously bought. There are many places to buy dog food, for instance pet supermarkets and grocery stores.

One strategy for increasing sales involves discount coupons. A steep discount provides a consumer incentive to try something new and, maybe, leads to consumers forming new habits. But, from a sales perspective, there

is little point in providing discounts to people who already have the habit of buying dog food from the retailer. Instead, it is most efficient to provide the discount only to people who don't yet have that habit

The Bullseye marketing staff decided to build a classifier to identify pet owners who already shop at Bullseye but do not purchase dog food there. The data available, from Bullseye's "loyalty" program, consisted of individual customers' past purchases of the tens of thousands of products sold at Bullseye.

Which of these many products to use as indicators of a customer's potential to switch to Bullseye's dog food? This is where feature engineering comes in. Searching through Bullseye's huge database, the feature engineers identified that customers who buy dog food also buy carpet cleaner. But many people buy carpet cleaner who don't buy dog food. The engineers searched for purchases might distinguish dog owners from other users of carpet cleaner.

The feature engineers' conclusion: Send dog-food coupons to people who buy carpet cleaner but do not buy diapers. Admittedly, this will leave out the people who have both dogs and babies: these are false negatives. It will also lead to coupons being sent to petless, spill-prone people whose children, if any, have moved beyond diapers: false-positives.

15.5 Threshold, sensitivity and specificity

In Figure 32 the threshold between $\mathbb{P}$ and $\mathbb{N}$ is set at a score of 3. That might have been a good choice, but it pays to take a more careful look.

That graph is hard to read because the scores have a very long-tailed distribution; the large majority of scores are below 2 but the scores go up to 40. To make it easier to compare scores between the C and H groups, Figure 33 shows the scores on a nonlinear axis. Each score is marked as a letter: "P" means $\mathbb{P}$, "N" means $\mathbb{N}$. False results are colored red.

```
## PNplot(threshold=3)
knitr::include_graphics("www/PN-threshold1.png")
```

Figure 33: Redrawing the participants' scores from Figure 31 on a non-linear axis. Color marks whether the classifier gave a correct output.

Moving the threshold up would reduce the number of false-positives. At the same time, the larger threshold would *increase* the number of false-negatives. **?@fig-two-thresholds** shows what the situation would be if the threshold had been set at, say, 10 or 0.5.

By setting the threshold larger, the number of false-negatives (red Ns in **?@fig-two-thresholds**) increases, but the number of false-positives (red Ps) goes down. Setting the threshold lower reduces the number of false-negatives but increases the number of false-positives.

This trade-off between the number of false-positives and the number of false-negatives is characteristic of classifiers.

Figure 36 shows the overall pattern for false results versus threshold. At a threshold of 0, all test results are ℙ. Hence, none of the C group results are false; if there are no ℕ results, there cannot be any false-negatives. On the other hand, all of the H group are false-positives.

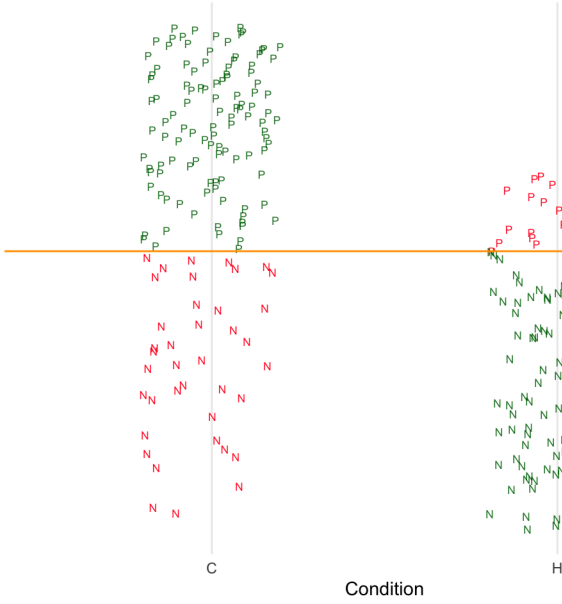
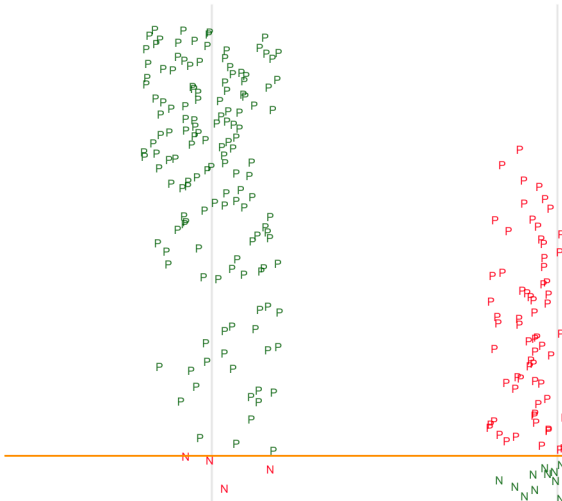


Figure 34: A **higher** threshold increases the number of false-negatives, but decreases false-positives.



Increasing the threshold changes the results. At a threshold of 1, many of the H group—about 50%—are being correctly classified as ℕ. Unfortunately, the higher threshold introduces some negative results for the C group. So the fraction of correct results in the C group goes down to about 90%. This pattern continues: raising the threshold improves the fraction correct in the H group and lowers the fraction correct in the C group.

There are two names given to the fraction of correct classifications, depending on whether one is looking at the C group or the H group. The fraction correct in the C group is called the "**sensitivity**" of the test. The fraction correct in the H group is the "**specificity**" of the test.

The sensitivity and the specificity, taken together, summarize the error rates of the classifier. Note that there are two error rates: one for the C group and another for the H group. Figure 36 shows that, depending on the threshold used, the sensitivity and specificity can be very different from one another.

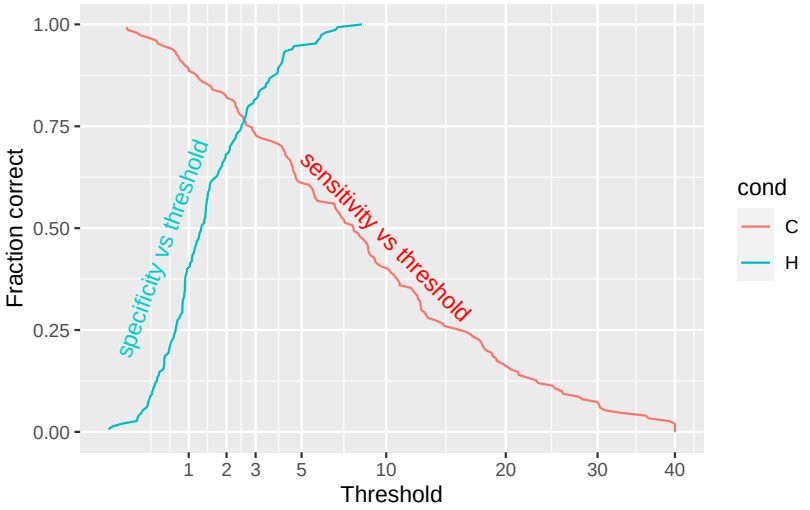


Figure 36: The choice of threshold determines the number of correct results.

Ideally, both the sensitivity and specificity would be 100%. In practice, high sensitivity means lower specificity and *vice versa*.

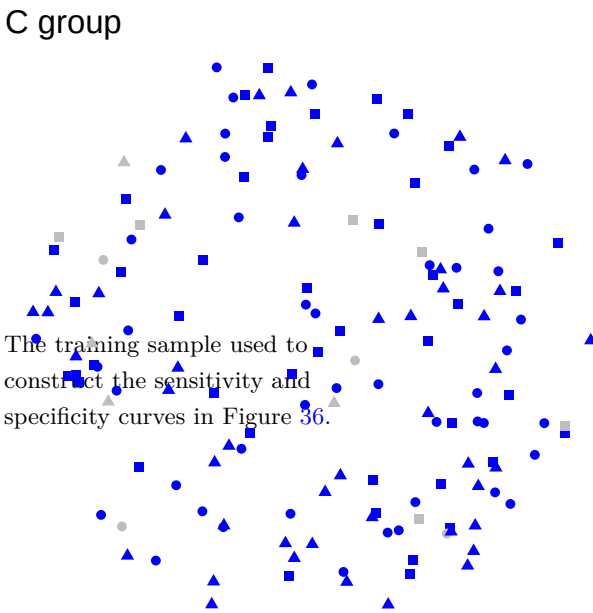
Sensitivity and specificity will be particularly important in Lesson 16 when we take into consideration the **prevalence**, that is, the fraction of the population with condition C

16 Accounting for prevalence

16.1 Prevalence

The "**prevalence**" of C is the fraction of the population who have condition C. Prevalence is an important factor in the performance of a classifier.

Lesson 15 used a **training sample**, first shown in Figure 32 and duplicated here in the margin. The training sample allowed us look at the consequences of the choice of threshold used in the test. That training sample had roughly equal numbers of people from the C and H groups. It's sensible to use such a training sample in order to make sure both the C and H groups are well represented.



The training sample used to construct the sensitivity and specificity curves in Figure 36.

The prevalence among the actual population is usually very different than in the training sample. Figure 37 illustrates the typical situation: many people in the H group and few people in the C group.

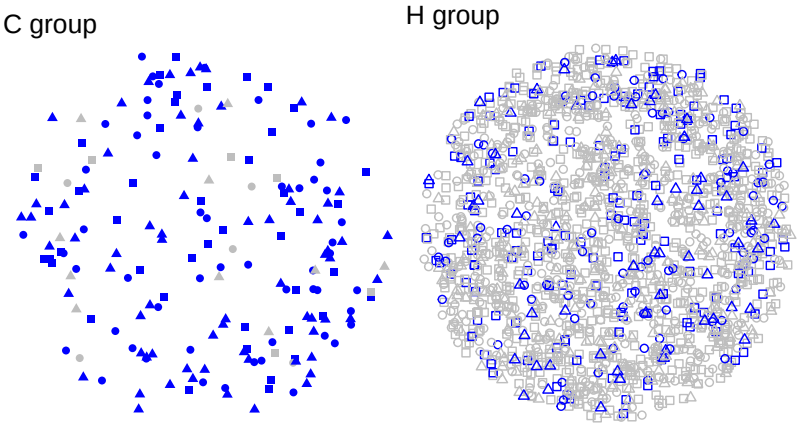


Figure 37: The population on which the classifier will be used.

The prevalence can be seen by how densely the H group is populated compared to the C group. The prevalence depicted in Figure 37 is about 10%, that is, one in ten people has condition C. In real-world conditions, prevalence is often much lower,

perhaps 0.1%. Indeed, epidemiologists often move alway from a percentage scale when quantifying prevalences, often using "cases per 100,000."

Even though the prevalence is different in Figure 32 than in Figure 32, the sensitivity is exactly the same. Likewise for the specificity.

We don't usually have comprehensive testing of a population, so drawing a picture like Figure 37 has to be done theoretically based on the limited information available: prevalence (from surveys of the population) as well as sensitivity and specificity (from the training sample). This is easy to do.

The first step is to determine the number in the C group and in the H group using the population size. If the population size is N, then the number in the C group will be $p(C)N$. We are writing the prevalence here as a probability, the probability $p(C)$ that a randomly selected person from the population has condition C. Similarly, the size of the H group is $(1-p(C))N$.

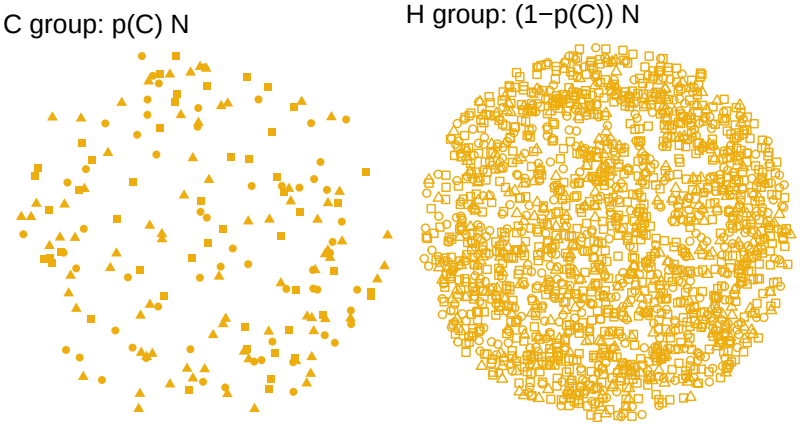


Figure 38: Knowing the prevalence allows us to determine the number of people in the C group and the number in the H group.

Consider now the sensitivity. Sensitivity is relevant only to the C group; it tells the fraction in the C group who will be correctly classified. That's enough information to know how many people in C to color blue (for $\mathbb{P}$) or gray (for $\mathbb{H}$).

Similarly, the specificity tells us what fraction among the H group to color blue and gray.

This is how Figure 37 was generated: specifying population size N, prevalence $p(C)$, and sensitivity and specificity. The false-

positives are the blue dots in the H group, the false-negatives are the gray dots in the C group.

16.2 From the patient's point of view

Figure 37 is drawn from the perspective of the epidemiologist or test developer. But it doesn't directly provide information of use to the patient, simply because the patient has only a test result ($\mathbb{P}$ or $\mathbb{H}$) but no definitive knowledge of the actual condition (C or H).

Re-organizing the epidemiologist's graph can put it in a form relevant to the patient. Instead of plotting people by C or H, we can plot them by $\mathbb{P}$ or $\mathbb{H}$. This perspective is shown in Figure 39, which is exactly the same people as in **?@fig-divided-by-condition2** but arranged differently.

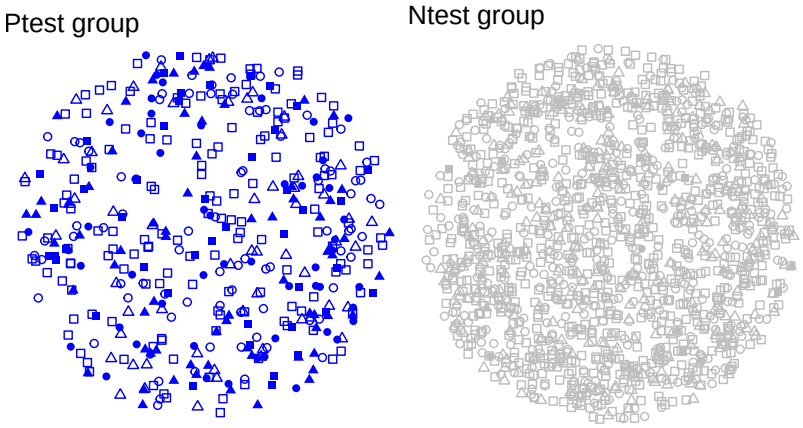


Figure 39: The population on which the classifier will be used.

For the patient who has gotten a $\mathbb{P}$ result, the left panel of Figure 39 is highly informative. The patient can see that only a small fraction of the people testing $\mathbb{P}$ actually have condition C. (The people with C are shown as filled symbols.)

The test result $\mathbb{P}$ is not definitive, it is merely a clue.

16.3 Likelihood

A "clue" is a piece of information or an observation that tells something about a mystery, but not usally everything. As an

example, consider a patient who has just woken up from a coma and doesn't know what month it is. It is a mystery. With no information at all, it is almost equally likely to be any month. So the hypotheses in contention might be labeled Jan, Feb, March, and so on.

The person looks out the window and observes snow falling. The observation of snow is a clue. It tells something about what month it might be, but not everything. For instance, the possibility that it is July becomes much less likely if snow has been observed; the possibility that it is February (or January or March) becomes more likely.

Statistical thinkers often have to make use of clues. Suppose the coma patient is a statistician. She might try to quantify the likelihood of each month given the observation of snow. Here's a reasonable try:

Month	Probability of seeing snow when looking out the window for the first time each day	Notation
January	2.3%	$p(\text{snow} \mid \text{January})$
February	3.5%	$p(\text{snow} \mid \text{February})$
March	2.1%	$p(\text{snow} \mid \text{March})$
April	1.2%	$p(\text{snow} \mid \text{April})$
May	0.5%	… and so on …
June	0.1%	
July	0	
August	0	
September	0.2%	
October	0.6%	
November	0.9%	
December	1.4%	

The table lists 12 probabilities, one for each month. For the coma patient, these probabilities let her look up which months it is likely to be. For this reason, the probabilities are called "**likelihoods**."

The coma patient has 12 hypotheses for which month it is. The table as a whole is a "**likelihood function**" describing how the likelihood varies from one hypothesis to another. Think of the entries in the table as having been radioed back to Earth from the 12 hypothetical planets $\mid$ January) through $\mid$ December).

It is helpful, I think, to have a notation that reminds us when we are dealing with a likelihood and a likelihood function. We will use the fancy L to identify a quantity as a likelihood. The coma patient is interested in the likelihood of snow, which we will write L_{snow}. From the table we can see that the likelihood of snow is a function of the month, that is $L_{\text{snow}}(\text{month})$, where month can be any of January through December.

This likelihood function has a valuable purpose: It will allow the coma patient to calculate the probability of it being any of the twelve months given her observation of snow, that is $p(\text{month} \mid \text{snow})$.

In general, likelihoods are useful for converting knowledge like $L_a(b)$ into the form $p(b \mid a)$. The formula for doing the conversion is called "**Bayes' Rule**."

The form of Bayes' rule appropriate to the coma patient allows her to calculate the probability of it being any given month from the likelihoods. We also need to account for February, with only 28 days, being shorter than the other months. So we will define a probability function, $p(\text{month}) = \frac{\text{number of days in month}}{365}$

Bayes' Rule

$$p(\text{month} \mid \text{snow}) = \frac{L_{\text{snow}}(\text{month}) \cdot p(\text{month})}{L_{\text{snow}}(\text{Jan}) \cdot p(\text{Jan}) + L_{\text{snow}}(\text{Feb}) \cdot p(\text{Feb}) + \cdots + L_{\text{snow}}(\text{Dec}) \cdot p(\text{Dec})}$$

16.4 How serious is it, Doc?

Imagine a patient getting a $\mathbb{P}$ test result and wondering what the probability is of his having condition C. That is, he wants to know $p(C \mid \mathbb{P})$. This is equivalent to asking, "How serious is it, Doc?"

The doctor could point to Figure 38 as her answer. That figure was generated by creating a population with the relevant prevalence, using the sensitivity and specificity to determine the fraction of the C and H groups with $\mathbb{P}$ or $\mathbb{H}$ respectively, the *re-organizing* into new groups: the $\mathbb{P}$ group and the $\mathbb{H}$ group.

Alternatively, we can do the calculation in the same way we did for the coma patient seeing snow. There, the observation of snow was the clue. Now, the test result $\mathbb{P}$ is the clue. One of the relevant likelihoods to interpret $\mathbb{P}$ is $L_{\mathbb{P}}(C)$: the likelihood for a person who genuinely has condition C of getting a $\mathbb{P}$ result. Of course, this is just another way of writing the sensitivity.

Similarly, the specificity is $L_{\mathbb{H}}(H)$. But since our person got a $\mathbb{P}$ result, the likelihood $L_{\mathbb{H}}(H)$ is not directly relevant. (It would be relevant only to a person with a $\mathbb{H}$ result.) Fortunately, there is a simple relationship between $L_{\mathbb{P}}(H)$ and $L_{\mathbb{H}}(H)$. If we know the probability of an H person getting a $\mathbb{H}$ result we can figure out the probability of an H person getting a $\mathbb{P}$ result.

$$L_{\mathbb{P}}(H) = 1 - L_{\mathbb{H}}(H)$$

Bayes' Rule for the person with a $\mathbb{P}$ result is

$$p(C \mid \mathbb{P}) = \frac{L_{\mathbb{P}}(C) \cdot p(C)}{L_{\mathbb{P}}(C) \cdot p(C) + L_{\mathbb{P}}(H) \cdot p(H)}$$

Calculating $p(C \mid \mathbb{P})$

Suppose that $p(C) = 1\%$ for this age of patient. (Consequently, $p(H) = 99\%$.) And imagine that the test taken by the patient has a threshold score of 1. From Figure 36 we can look up the sensitivity ($L_{\mathbb{P}}(C) = 0.95$) and specificity ($L_{\mathbb{P}}(H) = 0.50$) for the test. Substituting these numerical values into Bayes' Rule gives

$$p(C \mid \mathbb{P}) = \frac{0.95 \times 0.01}{0.95 \times 0.01 + 0.50 * 0.99} = 1.9\%$$

The $\mathbb{P}$ result has changed the probability that the patient has C from 1% to 1.9%. That's big proportionally, but not so big in absolute terms.

The advantage of the Bayes' Rule form of the calculation over the $\mathbb{P}$ group in Figure 38 is that it is very easy to do the Bayes' Rule calculation for any value of prevalence $p(C)$. Why would we be interested in doing this?

Typically the prevalence of a condition is different for different groups in the population. For example, for an 80-year-old with a family history of C the prevalence might be 20% rather than the 1% that applied to the patient in the previous example. For the 80-year-old, the probability of having C given a $\mathbb{P}$ result is substantially different from the 1.9% found in the example:

$$p(C \mid \mathbb{P}) = \frac{0.95 \times 0.2}{0.95 \times 0.2 + 0.50 * 0.8} = 32\%$$

16.5 Screening tests

The reliability of a $\mathbb{P}$ result differs depending on the prevalence of C. A consequence of this is that medical screening tests are recommended for one group of people but not for another.

For instance, the US Preventative Services Task Force (USPSTF) issues recommendations about a variety of medical screening tests. According to the Centers for Disease Control (CDC) summary:

> *The USPSTF recommends that women who are 50 to 74 years old and are at average risk for breast cancer get a mammogram every two years. Women who are 40 to 49 years old should talk to their doctor or other health care provider about when to start and how often to get a mammogram.*

Recommendations such as this can be baffling. Why recommend mammograms only for people 50 to 74? Why not for older women as well? And how come women 40-49 are only told to "talk to their doctor?"

The CDC summary needs decoding. For instance, the "talk to [your] doctor" recommendation really means, "We don't think a mammogram is useful to you, but we're not going to say that straight out because you'll think we are denying you something.

We'll let your doctor take the heat, although typically if you ask for a mammogram, your doctor will order one for you. If you are a woman younger than 40, a mammogram is even less likely to give a useful result, so unlikely that we won't even hint you should talk to a doctor."

The reason mammograms are not recommended for women 40-49 is that the prevalence for breast cancer is much lower in that group of people than in the 50-74 group. The prevalence of breast cancer is even lower in women younger than 40.

So what about women 75+? The prevalence of breast cancer is high in this group, but at that age, non-treatment is likely to be the most sensible option. Cancers can take a long while to develop from the stage identified on a mammogram, and at age 75+ it's not likely to be the cause of eventual death.

The USPSTF web site goes into some detail about the reasoning for their recommendations. It's worthwhile reading to see what considerations went into their decision-making process.

16.5.1 The Loss Function

> ⚠ In Draft
>
> NEED TO FIX THIS. The prevalence wasn't included in the calculation.

In order to set the threshold at an optimal level, it is important to measure the impact of the positive or negative test result. This impact of course will depend on whether the test is right or wrong about the person's true condition. It is conventional to measure the impact as a "**loss**," that is, the amount of harm that is done.

If the test result is right, there's no loss. Of course, it's not nice that a person is C, but a $\mathbb{P}$ test result will steer our actions to treat the condition appropriately: no loss in that.

Typically, the loss stemming from a false negative is reckoned as more than the loss of a false positive. A false negative will

lead to failure to treat the person for a condition that he or she actually has.

In contrast, a false-positive will lead to unnecessary treatment. This also is a loss that includes several components that would have been avoided if the test result had been right. The cost of the treatment itself is one part of the loss. The harm that a treatment might do is another part of the loss. And the anxiety that the person and his or her family go through is still another part of the loss. These losses are not necessarily small. The woman who gets a false positive breast-cancer diagnosis will suffer from the effects of chemotherapy and the loss of breast tissue. The man who gets a false-positive prostate-cancer diagnosis may end up with urinary incontinence and impotence.

The aim in setting the threshold is to minimize the total loss. This will be the loss incurred due to false negative times the number of false negatives plus the loss incurred from a false positive times the number of false positives.

⚠ Demonstration: Setting the optimal threshold

In Lesson 15, we saw that the threshold for transforming a raw test score into a $\mathbb{P}$ or $\mathbb{H}$ result determined the sensitivity and specificity of the test. (See Figure 36.) Of course, its best if both sensitivity and specificity are as high as possible, but there is a trade-off between the two: increasing sensitivity by lowering the threshold will decrease specificity. Likewise, raising the threshold will improve specificity but lower sensitivity.

The "**loss function**" provides a way to set an optimal value for the threshold. It is a function, because the loss depends on whether the test result is a false-positive or a false-negative.

Suppose that the

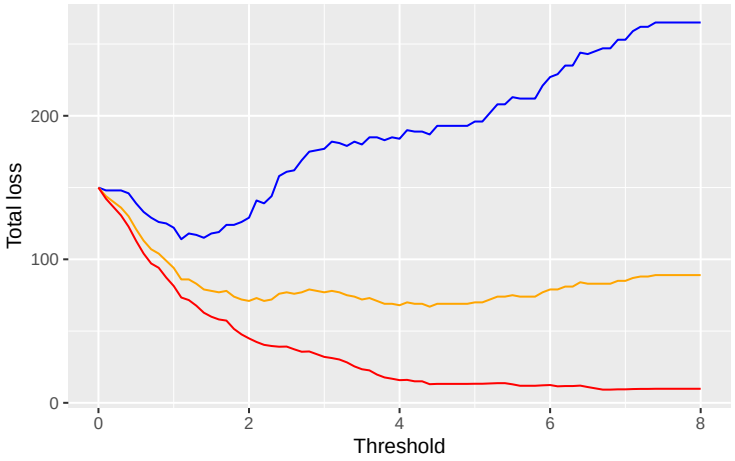


Figure 40: Total loss as a function of test threshold for the test shown in Figure 36. In the blue curve, a false-negative is 3 times more costly than a false-positive. In the orange curve they are equally costly. In the red curve, a false-positive is 10 times more costly than a false-negative.

17 Hypothesis testing

In Lesson 16, we looked at the accounting process that is used for building a classifier and interpreting the results. That accounting process is also applicable, in whole or part, to a more general procedure in statistics called, variously, "**hypothesis testing**" or "**Null hypothesis testing**" (NHT), or "**Null hypothesis significance testing**" (NHST). Textbooks tend to use the shorter name: "hypothesis testing." But these Lessons will use NHT as the name, because that is a more complete description of the actual process.

The word "test" is familiar to all who have ever been students, but it will still be useful to have a definition. This one seems reasonable:

> "*A procedure intended to establish the quality, performance, or reliability of something, especially before it is taken into widespread use.*" – Oxford Languages

Based on this general definition, one would expect that "hypothesis test" will be "a procedure intended to establish the correctness or applicability of a hypothesis, especially before relying on that hypothesis to guide action in the world." Regrettably, this definition does not align well with the statistical procedure called NHT. This lack of alignment causes confusion and error. It has also led to controversy about the use of the procedure.

The educator's response to the controversy is often, and reasonably, "We need to teach NHT, whatever its flaws, because that is the procedure that everyone follows." Educational reformer George Cobb followed up with a question: Why does everyone use it? His answer: "Because it's the centerpiece of the introductory statistics curriculum."

In this Lesson, we will introduce the logic of Null hypothesis testing (NHT) and compare it to other procedures that also have a claim to be the way to test hypotheses. In Lesson 37,

we will show how to do the computations involved to create the number—called the **p-value**—that is usually taken as the end-result of NHT. And in Lesson 38, we will return to the controversy, with the goal of helping the reader avoid pitfalls in interpretation of the p-value.

17.1 Tests, generally

In Lessons 34 and 35, the classifiers we built had two possible outputs, $\mathbb{P}$ or $\mathbb{N}$. The classifier is only part of a bigger procedure which we called a "test." To avoid unnecessary abstraction, our examples featured medical tests.

Medical tests often have three or more outcomes such as "high," "normal," or "low." This does not fundamentally change the situation from our $\mathbb{P}$ or $\mathbb{N}$ paradigm.

In a medical test, the first step in the procedure usually involves a measurement procedure for an individual, for instance, counting white blood cells or measuring the concentration of prostate-specific antigen (PSA).

The second step is purely arithmetical, comparing the measurement result to a threshold, thereby determining if the output should be $\mathbb{P}$ or $\mathbb{N}$. Lessons 34 and 35 were largely about how to set a good value for the threshold and involved stating a *loss function* and considering the prevalence of the condition involved.

A more familiar kind of test is the one taken by students in school, for instance, an algebra test. As the reader knows, in an algebra test, the subject is made to answer questions, the number of correct answers counted, and that count applied to a threshold to determine whether the overall result is "pass" or "fail." There is not an explicit underlying condition, say, "expert" or "dilettante." Consequently, there is no such thing as a false-negative (an expert who fails the test) or a false-positive (a dilettante who passes the test). All that matters is the test result itself

Unlike medical tests, academic tests are generally not the product of the sort of careful development phase described in Lesson 16. In an algebra test, there is no training data, that is, no sample of subjects who have been observed definitively to be "experts" or "dilettantes" and who take the test to be classified as "pass" or "fail." There is no calculation of a sensitivity or

specificity to characterize the test and no use of a loss function to bend sensitivity, specificity, and prevalence into a threshold dividing "pass" from "fail." Academicians never tell their students what the false positive and false negative rates are.

Students and teachers think of an academic test as that part of the overall procedure where questions are asked and answered. Of course, there is a follow-up procedure that we call "grading," where the correct answers are counted and the count converted to a pass/fail result. Sometimes the threshold between pass and fail is not fixed, but is set by the instructor to achieve a desired pass rate. This is called "grading on a curve," the threshold depends on the observed counts.

NHT is like "grading on a curve." The data collection and summary of the data is **not part of the procedure**. Like "grading on a curve", NHT starts at the point where the data have already been recorded and summarized. (Typically, the summary is the coefficient from a regression model or some other statistical measure like R^2.)

NHT is only about grading the summary, which is done on a curve. The grading is accomplished by examining a likelihood calculated from the summary. As notation for this likelihood, we will write $L_{\mathbb{S}}$(Null hypothesis).

The phrase "Null hypothesis" is too long to make for pleasant mathematical reading. Consequently, a shorter symbol, H_0_, is often used. The subscript *0* is a shorthand for "null." The "H" indicates a hypothesis.

17.2 The Null hypothesis

The key to understanding Null hypothesis testing is to know what the Null hypothesis is claiming.

In a technical sense, it would suffice to say that the Null hypothesis is a claim that the effect size is zero (or that R^2 is zero) except for sampling variation. Making sense of this requires that one know what an "effect size" (or "R^2") is and what "sampling variation" means. At this point in these Lessons, you ought to

know all of these things. But how to talk about hypothesis testing to a general audience?

One strategy is to describe the Null as the "absence of any effect" or "relationship." Another common strategy is to avoid mentioning the Null at all and use alternates such as "the result is significant" or "not due to chance."

Another way to think about the Null hypothesis is algorithmically. A Null-hypothesis relationship—really, a lack of relationship—can be created between two variables by **shuffling** one of them. Shuffling was introduced briefly in Lesson 10, where it was used to simulate an explanatory variable unrelated to the response.

To illustrate, consider the `Galton` data about the heights of adult children and their parents. We will make a simple model of `height` as a function of `sex`—everyday experience suggests a relationship between the two variables.

```
lm(height ~ sex, data=Galton) %>% confint()
```

	lwr	upr
(Intercept)	64.0	64.0
sexM	4.8	5.4

If the Null hypothesis were true, that is, if `sex` were unrelated to `height`, the `sexM` coefficient ought to be close to zero and a confidence interval on the coefficient will usually include zero. But for `height ~ sex` the confidence interval does not include zero.

Now consider what happens if we shuffle one or both of the variables.

```
lm(height ~ shuffle(sex), data=Galton) %>% confint()
```

	lwr	upr
(Intercept)	66.00	67.00
shuffle(sex)M	-0.35	0.59

```
lm(shuffle(height) ~ sex, data=Galton) %>% confint()
```

	lwr	upr
(Intercept)	66.00	67.00
sexM	-0.47	0.47

```
lm(shuffle(height) ~ shuffle(sex), data=Galton) %>% confint()
```

	lwr	upr
(Intercept)	67.00	67.00
shuffle(sex)M	-0.75	0.19

All these confidence intervals include zero, as expected.

Figure 41: Planet Null, known symbolically as $|H_0)$. Any pattern on Planet Null is attributed to chance, in the form of sampling variation.

In terms of the planet metaphor for hypotheses, the Null hypothesis $|\ H_0)$ is a planet. Variables on this planet are always unrelated to one another. The possible indication of a pattern is due to sampling variation, not a genuine relationship.

Riddle: How do we get to Planet Null?

```
Take the space shuffle.
```

Before the advent of ubiquitous computing, the Null hypothesis was implemented using algebra and probability theory. An example of such theory appeared in Lesson 10. The blue diagonal line in Figure 15 reflects what the average value of R^2 would be if a large number of trials were run on the model `y ~ x1 + x2 + ... + xk` where the `x`'s are unrelated to the `y`. Another part of the theory has to do with the "distribution" of the F statistic from Lesson 29, which we will discuss in Lesson 37.

17.3 The Alternative hypothesis

In Null hypothesis testing, there is only the one hypothesis—the Null—under consideration. Since the world $\mid H_0$) can be created by shuffling, the computations for NHT can be done pretty easily even without the probability theory just mentioned.

There has been a controversy since the 1930s about whether hypothesis testing—in the broad sense—should involve two (or more) competing hypotheses. One of these could be the Null hypothesis, the other, which we call the "Alternative hypothesis" (H_a) a statement of a specific non-null relationship.

Figure 42: Planet Alt, that is, $\mid H_a$) might look like this. We draw it as a cartoon planet, since any particular hypothesis is a product of the imagination.

The situation with two hypotheses would be very similar to that presented in Lessons 34 and 35. In those lessons, the two hypotheses were C and H. In developing a classifier, one starts by collecting a training sample which is a mixture of cases of C and H. But, in general, with a competition of hypothesis—H_0 and H_a—we don't have any real-world objects to sample that are known to be examples of the two hypotheses. Instead, we have to create them computationally. Instances of H_0 can be made by data shuffling. But instances of H_a need to be generated by some other mechanism, perhaps one akin to the DAGs we have used in these lessons.

With mechanisms to generate data from both the Null and Alternative hypotheses, we would take the statistical summary $\mathbb{S}$ of the actual data, and compute the likelihoods for each hypothesis: $L_{\mathbb{S}}(H_0)$ and $L_{\mathbb{S}}(H_a)$. It should not be too controversial in a practical process to set the prior probability for each hypothesis at the same value: $p(H_0) = p(H_a) = \frac{1}{2}$. Then, turn the crank of Bayes' Rule (Section 16.4) to compute the posterior probabilities. If the posterior of one or the other hypothesis is **much greater** than $\frac{1}{2}$, we would have compelling evidence in favor of that hypothesis.[17]

There are specialized methods of Bayesian statistics and whole courses on the topic. An excellent online course is *Statistical Rethinking*.

[17] In the same spirit, we might simply look at the likelihood ratio, $L_{\mathbb{S}}(H_a) \div L_{\mathbb{S}}(H_0)$ and draw a confident conclusion only when the ratio turns out to be much greater than 1, say, 5 or 10.

Before the widespread acceptance of the Bayesian approach, statisticians Jerzy Neyman and Egon Pearson proposed a two-hypothesis framework in 1933. We will discuss this in **?@sec-power**.

> ⚠ Not an alternative!
>
> If you have studied statistics before, you likely have been exposed to NHT. Many textbook descriptions of NHT appear to make use of an "alternative hypothesis" within NHT. This style is traditional and so common in textbooks that it seems disrepectful to state plainly that it is wrong. There is only one hypothesis being tested in NHT: the Null.
> In the textbook presentation of NHT, the "alternative" hypothesis is not a specific claim—for instance, "the drug reduces blood pressure by 10 mmHg". Instead, the student is given a pointless choice of three versions of the alternative. These are usually written $H_a \neq H_0$ or as $H_a < H_0$ or as $H_a > H_0$, and amount to saying "the effect size is non-zero," "the effect size is negative," or "the effect size is positive."
> Outside of textbooks, only $H_a \neq H_0$ is properly used. The other two textbook choices provide, at best, variations on exam questions. At worst, they are a way to put a thumb on the scale to disadvantage the Null.

17.4 "Under the Null"

Using the shuffling algorithm, the computation underlying NHT is very much like the technique introduced in Lesson 5: create a set of trials to represent the sampling distribution. The twist in NHT is that the sampling distribution is calculated on Planet Null. Or, in conventional statistical language, what's computed is the "sampling distribution under the Null."

The phrase "under the Null" is often described as being shorthand for "assuming the Null to be true." When doing a calculation, "assuming the Null to be true" might be better expressed

more emphatically: "arranging things so that the Null is true" or "enforcing the Null."

We will use the `Galton` data, and the model `height ~ mother`, to illustrate computing the "sampling distribution under the Null." The relationship being expressed by `height ~ mother` is that the adult child's height is related at least in part to the mother's height. As always, the effect size is a good way to quantify that relationship. Here it is:

```
mod <- lm(height ~ mother, data=Galton)
mod %>% coef()
```

```
(Intercept)      mother
    46.6908      0.3132
```

```
mod %>% confint()
```

	lwr	upr
(Intercept)	40.2951	53.0864
mother	0.2134	0.4129

These reports says that changing the model input by one inch will translate to about 0.3 inches in the output. The uncertainly due to sampling variation broadens the 0.3 into an interval, 0.2 to 0.4 inches of height per inch of mother.

Now we rocket off to Planet Null. On Planet Null, children are not related to their mothers. Or, put another way, the unique relationship between a mother and her child is a Planet-Earth concept. On Planet Null, each mothers and children are paired at random.

To calculate the `mother` coefficient on Planet Null, use `shuffle()`. Here's the result from one sample:

```
lm(height ~ shuffle(mother), data=Galton) %>% coef_summary()
```

term	coefficient
(Intercept)	67.1960
shuffle(mother)	-0.0068

Repeating the above generates another sample from Planet Null:

```
lm(height ~ shuffle(mother), data=Galton) %>% coef_summary()
```

term	coefficient
(Intercept)	62.791
shuffle(mother)	0.062

From these two Planet Null samples, it is already evident that the `shuffle(mother)` coefficient on Planet Null tends to be closer to zero than the `mother` coefficient from Planet Earth.

We can make a more compelling and precise statement if we look at a large number of Planet-Null samples:

```
Trials <- do(1000) * {
  lm(height ~ shuffle(mother), data=Galton) %>% coef()
}
```

```
Trials %>%
  ggplot(aes(y = mother, x = "All")) +
  geom_jitter(alpha = 0.2, width=.15) +
  geom_violin(alpha = .2, fill = "blue", color = NA) +
  geom_point(aes(y = 0.31, x = 1), color="red") +
  xlab("")
```

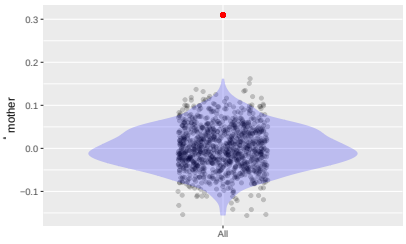


Figure 43: Values of the `shuffle(mother)` coefficient. Each gray dot is the result from one shuffling trial. The violin shows the sampling distribution under the null. The red dot is the `mother` coefficient found without shuffling.

The set of `mother` coefficients in these trials reflects us the "sampling distribution under the Null." (Figure 43) The "sampling distribution under the Null" is represented by the gray dots. The violin gives an easier to read representation. For this model, the `mother` coefficient from the sample from Planet Earth is so far different that it cannot pretend to be an instance from Planet Null.

18 Calculating a p-value

This Lesson continues the discussion of Null hypothesis testing (NHP) started in Lesson 17. Recall from that lesson that the Null hypothesis is a statement in line with the claim that "there is no relationship between these variables" or "nothing is going on." For example, in a study about the effectiveness of a new drug, the Null hypothesis will be that the drug has no effect at all. Another example: In an economics study about the possible relationship between a country's "corruption index" and interest rates, the Null hypothesis would be "corruption is unrelated to interest rates."

The work that *precedes* an NHP involves acquiring data, modeling it in a way that illuminates the *relationship of interest*, then summarizing that model. Often, the summary takes the form of a model coefficient, but it might be the model's R^2 or the incremental R^2 from a nested set of models. (See Section 10.5.) Whatever the details, we will call the summary $\mathbb{Sreal}$.

The primary calculation of an NHP, described in Section 17.4, is to generate simulated data frames, each involving shuffling to annul the *relationship of interest*. Then, model and summarize the simulated data frames in exactly the same way as for the actual data. The result for the first simulated data frame is $\mathbb{Snull}_1$, for the second $\mathbb{Snull}_2$, and up to, say, $\mathbb{Snull}_{1000}$.

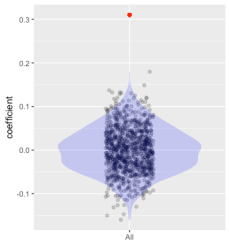
The exact number of simulated data frames does not matter. We will use 1000 to simplify the arithmetic.

At this point, there are 1001 summaries: one $\mathbb{Sreal}$ and 1000 $\mathbb{Snull}_i$, where $i = 1, 2, \ldots, 1000$. These become the raw material for calculating numerical culmination of the NHT: the **p-value**.

18.1 The p-value

Figure 43 (reproduced in the margin) shows an example of the raw material: 1000 values of $\mathbb{Snull}_i$ and the single, unique value of $\mathbb{Sreal}$ from the model `lm(height ~ mother, data=Galton)`.

```
knitr::include_graphics("www/fig-shuffle-mother.png")
```



The p-value comes from the fraction of the $\mathbb{Null}_i$ that are larger in value than the $\mathbb{Real}$. In Figure 43, *none* of the 1000 $\mathbb{Null}$ are larger than $\mathbb{Real}$. Therefore, we write $p < 1/1000$ or $p < 0.001$.

The p-value is the final result of an NHT. TALK ABOUT how to write about the p-value LATER ON.

NHT is such a popular technique that statistical software will find the p-values from regression models for you. Consequently, in practice the shuffling technique is reserved for more specialized situations. For these Lessons, the specialized "situation" is pedagogical; we are trying to help you understand the *concept* of p-values and shuffling provides a concrete way to do this. In most modeling work, however, pedagogy is not an issue. So expect to use software to calculate p-values.

The p-value software we use in these Lessons is the two summary functions, `regression_summary()` and `anova_summary()`.

```
lm(height ~ mother, data = Galton) %>% regression_summary()
```

term	estimate	std.error	statistic	p.value
(Intercept)	46.6908	3.2587	14.328	0
mother	0.3132	0.0508	6.163	0

The regression model `height ~ mother` has two coefficients (the "estimate" column from `regression_summary()`). Insofar as our concern is the relationship between child's `height` and `mother`, only the p-value for `mother` is of interest.

> **Software for P-values**
>
> In these Lessons, we use the `regression_summary()` and `anova_summary()` R functions to calculate p-values from models. These two functions come from the `{math300}` package, which was written specially for these Lessons. In

standard R, the equivalents are `summary()` and `anova()`. In fact, `regression_summary()` and `anova_summary()` are merely *wrappers* around `summary()` and `anova()`. The wrappers make sure that the output is in the form of a data frame and therefore suitable for data wrangling. The output from `summary()` and `anova()`, however, are not in a data-frame format.

18.2 Basic interpretation of p-values

When the p-value is small, the likelihood of $\mathbb{S}\mathtt{real}$ under the Null hypothesis, that is, $L_{\mathbb{S}\mathtt{real}}$(Null hypothesis) is also small. A small likelihood of a given hypothesis means that the hypothesis is not a compelling explanation for the observed $\mathbb{S}\mathtt{real}$.

There is a formal vocabulary for NHT. Instead of saying, casually, "The Null hypothesis is not a compelling explanation for $\mathbb{S}\mathtt{real}$, the formal NHT statement is,"The Null hypothesis is **rejected**."

Commonly, the threshold 0.05 is given as the numerical definition of "small" in "small likelihood." For example, the p-value on the `mother` coefficient in the model `lm(height ~ mother, data=Galton)` is $p < 0.01$. This is obviously less than 0.05, so the outcome of the NHT is to "reject the Null hypothesis."

Suppose, on the other hand, that the p-value had been "large," that is $0.05 < p$. What phrase should we use to summarize this situation. It is tempting—but wrong—to think that this leads to "accepting" the Null hypothesis. Instead the proper NHT phrase to use is "fail to reject the Null."

We will return in Lesson 19 to the question of whether 0.05 is a good threshold to use. That's part of a broader controversy.

Another part of the NHT formal vocabulary is the phrase "statistically significant." In the everyday sense of the word, "significant" suggests "important,", "worthy of attention," or "noteworthy." In NHT speak, "statistically significant" is a synonym for "reject the Null." The NHT meaning of "statistically significant" has nothing at all to do with utility of the result.

Different fields have different standards for defining small. For instance, it's common in psychology to consider $p < 0.10$ as fairly small, while in physics, "small" means perhaps $p < 0.001$ or even $p < 0.000001$.

It may seem odd that there is no universal agreement about "small." The reason is that p-values are part of a *standard operating procedure* for evaluating research results to know if they are worthy of publication.

In physics, laws and models are meant to be exact or close to exact. Lord Rutherford (1871-1935), an important physicist who won the Nobel prize in 1908, famously disparaged the use of statistics, reportedly saying, "If your experiment needs statistics, you should have done a better experiment." This was in an era where the p-value *standard operating procedure* had not yet been invented. Today, when p-values are common in most fields, Rutherford's distaste for statistical method is reflected in p-value thresholds like $p < 0.000001$.

In other fields such as economics or psychology or clinical medicine, models are sought that are *useful* but without any expectation that they be exact. (In the 19th and early 20th century, psychologists and economists sometimes used the vocabulary of "law" to describe their findings, but "model" is more appropriate, because, unlike physics, the laws are not strictly enforced!) Often, in economics or psychology or medicine, the size of a sample used to train a model is less than, say, $n = 100$. And the units of observation—people or countries, for instance—are different one from the other, quite unlike, say, electrons, which are all the same. Consequently, sampling variation is often an important source of noise, obscuring relationships or even suggesting relationships that are not really there. (See Lesson 12.) This situation—small sample size, variation in observational units, and large sampling variation—would cause many useful findings to go unreported, as would happen if $p < 0.000001$ were the standard. So a less stringent threshold for publication is used, most commonly 0.05.

18.3 P-values for coefficients

For any regression model, the "regression report" contains one row for every coefficient for the model. Each of these rows will have a coefficient value ("`estimate`") and a p-value. There are two additional columns: a standard error for the coefficient ("`std.error`") and a value (labeled "`statistic`") that is *always* just the estimate divided by the standard error.

The idea of the standard error was introduced in Lesson 5 (Section 5.1). The point of the standard error is to summarize the amount of sampling variation in the coefficient. But we standardized on the confidence interval format to summarize sampling variation.

Many statisticians think there is little point to calculating a p-value on a model coefficient because the confidence interval contains all the information needed. Importantly, model coefficients and their confidence intervals come with units; the units are the connection between the number and the real world. P-values are without units, and their value depends strongly on the sample size. Thus, they mix together relevant information about the magnitude of the effect and incidental information about the size of the sample used for training.

18.4 P-values for F

[[Say how a model coefficient is different from a set of terms. The question is whether the new term adds information on top of the existing terms.]]

Sometimes the interest is more general: Do any of these terms contribute to explaining variation in the response variable? In such situations, the appropriate p-value is one that compares one model to another. This style of p-value—not on the individual coefficients but on model terms—comes from a calculation called "analysis of variance."

18.5 Traditional names for tests

18.6 Tests in textbooks

Statistics textbooks usually include several different settings for "hypothesis tests." I've just pulled a best-selling book off my shelf and find listed the following tests spread across eight chapters occupying about 250 pages.

- hypothesis test on a single proportion
- hypothesis test on the mean of a variable
- hypothesis test on the difference in mean between two groups (with 3 test varieties in this category)
- hypothesis test on the paired difference (meaning, for example, measurements made both before and after)
- hypothesis test on counts of a single categorical variable
- hypothesis test on independence between two categorical variables
- hypothesis test on the slope of a regression line
- hypothesis test on differences among several groups
- hypothesis test on R^2

As statistics developed, early in the 20th century, distinct tests were developed for different kinds of situations. Each such test was given its own name, for example, a "t-test" or a "chi-squared test." Honoring this history, statistics textbooks present hypothesis testing as if each test were a new and novel kind of animal.

In fact, almost all the different tests named in introductory statistics books are really just different manifestations of regression. Regression is to "animal" the way t-test is to "elephant." An important theme in the history of statistics is that out of the diversity of statistical methods, almost all of them are encompassed by one method: regression modeling.

In these Lessons, we've focussed on that one method, rather than introducing all sorts of different formulas and calculations which, in the end, are just special cases of regression. Nonetheless, most people who are taught statistics were never told that the different methods fit into a single unified framework. Consequently, they use different names for the different methods.

Communicating in a world where people learned the traditional names, you have to be able to recognize those names know which regression model they refer to. In the table below, we will use different letters to refer to different kinds of explanatory and response variables.

- `x` and `y`: quantitative variables
- `group`: a categorical variable with multiple (≥ 3) levels.
- `yesno`: a categorical variable with exactly two levels (which can always be encoded as a zero-one quantitative variable)

Model formula	traditional name
`y ~ 1`	**t-test** on a single mean
`yesno ~ 1`	**p-test** on a single proportion.
`y ~ yesno`	**t-test** on the difference between two means
`yesno1 ~ yesno2`	**p-test** on the difference between two proportions
`y ~ x`	**t-test** on a slope
`y ~ group`	**ANOVA** test on the difference among the means of multiple groups
`y ~ group1 * group2`	**Two-way ANOVA**
`y ~ x * yesno`	**t-test** on the difference between two slopes. (Note the `*`, indicating interaction)

Another named test, the **z-test**, is a special kind of t-test where you know the variance of a variable without having to calculate it from data. This situation hardly every arises in practice, and mostly it is used as a soft introduction to the t-test.

18.7 P-values and covariates

Use cancer/grass-treatment example from Lesson 11 to illustrate how failing to think about covariates *before* the study analysis can lead to false discovery.

Use age in marriage data.

So, standard operating procedures were based on the tools at hand. We will return to the mismatch between hypothesis testing and the contemporary world in Lesson 19.

Make this table nicer by constructing it in some other system.

.	do not reject H_0_	reject H_0_ in favor of H_A_
H_0_ true	Correct decision	Type 1 error
H_A_ true	Type 2 error	Correct decision

> A **Type 1 error**, also called a **false positive**, is rejecting the null hypothesis when H_0 is actually true. Since we rejected the null hypothesis in the gender discrimination (from the Case Study) and the commercial length studies, it is possible that we made a Type 1 error in one or both of those studies. A **Type 2 error**, also called a **false negative**, is failing to reject the null hypothesis when the alternative is actually true. A Type 2 error was not possible in the gender discrimination or commercial length studies because we rejected the null hypothesis.

⚠ The chi-squared test

Most statistics books include two versions of a test invented around 1900 that deals with counts at different levels of a categorical variable. This chi-squared test is genuinely different from regression. And, in theoretical statistics the chi-squared distribution has an important role to play.
The chi-squared test of independence could be written, in regression notation, as `group1 ~ group2`. But regression does not handle the case of a categorical variable with multiple levels.
However, in practice the chi-squared test of independence

is very hard to interpret except when one or both of the variables has two levels. This is because there is nothing analogous to model coefficients or effect size that comes from the chi-squared test.
The tendency in research, even when `group1` has more than two levels, is to combine groups to produce a `yesno` variable. Chi-squared can be used with the response variable being `yesno` and almost all textbook examples are of this nature.
But for a `yesno` response variable, a superior, more flexible and more informative method is logistic regression.

ANOVA, which is always a comparison of two models, say `y~1` versus `y~group` involves something called an F-test. For the simpler setting of the t-test, the model `y~yesno`, an F-test can also be done. Which to do, t or F? It turns out that t^2 is exactly the same as F.

19 False discovery

19.1 Avoid bad habits

NEEDS RE-ORGANIZATION

Sometimes the interest is more general: Do any of these terms contribute to explaining variation in the response variable? In such situations, the appropriate p-value is one that compares one model to another. This style of p-value—not on the individual coefficients but on model terms—comes from a calculation called "analysis of variance."

Nobody likes to summarize their work with the word "fail." And so, when "fail to reject the Null hypothesis" is the correct conclusion, people express this in softer ways.

It's very common for the conclusion "fail to reject the Null" simply not to be reported at all. Historically, and even today, some journals will not accept for publication a scientific article with the conclusion "fail to reject the Null."

Consider the situation of a researcher whose years-long project has led to a p-value of 0.07. To soften the blow of "fail to reject," the researcher will report the p-value itself so that the reader can see how close it is to small. In some literatures, you will see language like "tending to significance" instead of "fail to reject." In some fields, research publications will show the notation $p < 0.1$. This also indicates failure to reject the null hypothesis.

Journalists eager to publish reports about scientific work, but facing a p-value that is a little too large, will occasionally qualify their report with this phrase: "… although the work did not reach the rigorous scientific standard for statistical significance."

All of these are dodges. There's nothing "rigorous" about $p < 0.05$ although seems unfair that a researcher who had a plausible idea and did the work to test it honestly does not get to publish that work and receive acknowledgement that they are a hard-working part of the overall scientific enterprise.

Another problem with p-values stems from misinterpretation of the admittedly difficult logic that underlies them. The misinterpretations are encouraged by the use of the term "**tests of significance**" to the p-value method. Particularly galling is the use of the description "**statistically significant**" to describe a result where $p < 0.05$. The everyday meaning of "significant" as something of importance is in no way justified by $p < 0.05$. Instead, the practical importance or not is more clearly signaled by examining an effect size. (It's extremely disappointing that journalists, who are writing for an audience that for the most part has no understanding of p-value methodology, use "significant" when reporting on the statistics of research findings. It would be more honest to use a neutral term such as "null-validated" or "p-validated" which does not confuse the statistical result with actual practical importance.)

This example of a regression table shows that p-values can sometimes be very, very small. Such smallness is often misinterpreted as indicating that a very powerful result has been found. This is simply nonsense, which is why the more dignified notation $p < 0.05$ or `*` is to be preferred.

> **i** "Significance" and significant digits
>
> In the regression summary of the `height ~ mother` model, the p-value on the `mother` coefficient was reported as `1.079105e-09`. This is a symptom of the choice by software designers to report more digits than are genuinely useful.
> Statistician Jeffrey Witmer, in an editorial in the *Journal*

of Statistics Education, distinguishes between the "mathematical" information in a p-value and the "statistical" information. Mathematically, the p-value is the result of a calculation. Statistically, the p-value is used as a symbol to indicate whether the Null hypothesis is a plausible explanation for a statistical result.

Witmer proposes a simple rule for printing p-values: Round to 1 significant digit. This means that a p-value computed to be 0.382 would be reported as 0.4. A p-value of 0.0079 would be reported as 0.008. The justification for this rule is that there is no information in the second non-zero digit of a p-value that can meaningfully guide a conclusion about whether the Null hypothesis is a plausible explanation for a statistical result. There may be a mathematical difference between 0.0079 and 0.008, but there is no meaningful statistical difference.

Witmer also offers a simple solution to the problem of people misinterpreting "statistically significant" as related to the everyday meaning of "significant." Replace the term "statistically significant" with "statistically discernible." There is no difference between the everyday sense of "discernible"—able to be perceived—and the statistical implications. In conveying statistical information, "discernible" is more descriptive than "significant." For example, it would be appropriate to describe the implications of a p-value $p < 0.03$ as, "the relationship is barely discernible from the sampling variation."

Use the confidence interval instead

NHT applied to an effect size is intended to demonstrate whether the effect size is sufficiently far from zero that we can reasonably conclude that it is non-zero. There is a simpler way to do this: look at the confidence interval on the effect size.

But, for those journals that (unwisely) require p-values, you'll have to use NHT to generate them.

Since "fail" and "reject" are unattractive words, in practice other expressions are used. One of the notations is $p < 0.05$,

another is to put a asterisk (*) next to the value of the effect size or R^2. Both of these correspond to "reject the Null." The notation used for "fail to reject" is to put nothing next to the effect size or R^2, but it would be more appropriate simply to list the effect size as "n.s." to stand for "not significant."

> **In Draft: Power**
>
> Mention the idea of power and why it's helpful to look at power when interpreting a "failure to reject."

There is, I think, a helpful analogy to be made between hypothesis testing and the familiar ways that we try to avoid information overload on the Internet.

"Internet protocols" organize communication into standard format "packets" that are easily and rapidly transmitted, routed, and received. These packets make possible the vast web of connections that is the Internet. Anyone can put any digital content they like *inside* a packets; the protocols are neutral in this regard. The Internet protocols were *not* designed to determine what content is worth transmitting and what is worth receiving. We rely on other systems for that, mostly at the receiving end. There are spam filters to avoid email accounts being flooded with worthless or harmful messages. There are recommender systems that compare your history of music or movie streaming to that of others in order to identify what new content you might like. Search engines look inside web pages to identify connections and rank highly those pages that are linked to by other highly ranked pages. These systems leave creators free to follow their interests, ideas, and imaginations, while providing a little guidance to people who want to access some content but avoid being overcrowded by other content that is not worthwhile.

Historically, there were earlier waves of technology that increased the ability to communicate. Printing and postal systems emerged in the 13th and following systems. Before those

innovations, communication was outrageously expensive, requiring hand-copying of manuscripts, couriers, and camel trains. Content was controlled to some extent by authorities: government censorship; church "indices" and spritual authorities; and often the authorities of those famous classical philosophers and poets whose work and thought was promulgated by early universities.

About four centuries ago, such authorities were being challenged. It slowly became accepted to make judgements based on observations and to disregard antique authorities. Enlightened "scientists" communicated their discoveries in hand-written letters to one another.[18] In the late 1600s, another, possibly more efficient means of communication was developed: scientific societies where members met and read aloud their work to an audience, and the journals of such societies which enabled mass communication to those scientists distant from the society's meetings in time or space.

Early scientific journals are delightful collations on diverse and miscellaneous subjects. Everything seems to have been of interest to everyone. Publication was regulated by the recommendations of "members" of the society; new members were admitted by the consensus of earlier members.

Over the centuries, the growth of scientific content and the specialization of methodology called for research findings to be sorted by area. But there was still need to regulate publication, to avoid distracting readers with worthless information.

Hand-in-hand with the scientific revolution's reliance on observation and data rather than authority came the need to standardize methods for summarizing data. This might be called a "statistical protocol" by analogy to Internet protocols, but there is no wise governing body, only consensus and "accepted practice."

[18]The word "science" comes from the Latin for "to know." But a dictionary definition of "science" makes clear the "scientific" style for gaining knowledge: "The systematic study of the structure and behavior of the physical and natural world through observation, experimentation, and the testing of theories against the evidence obtained." (Source: Oxford Languages)

The data from a bench-top experiment might consist of, say, six numbers: three from the treatment and three from controls. The arithmetic means of these two groups is practically certain to be non-zero, even if the the treatment had no effect. This meant that a means was needed to establish when the difference in means was large enough to suggest the two groups might be genuinely different and that the treatment did have an effect. The statistical protocol to decide such things needed to be simple: computers weren't available and there were no courses to teach statistical method until the 1960s. In the 1930s, prominent statistical pioneer Ronald Fisher published a slim volume, *Statistical Methods for Research Workers* which laid out methods for managing and standardizing the calculations. Fisher's authority was substantial but not absolute. Differing philosophical views also came to influence "accepted practice."

Early statistics books and courses codified "accepted practice." What emerged is the system of calculations that we call "hypothesis testing" and the ubiquitous p-value. Still, this was rooted in the need to avoid journals wasting library-shelf-space and reader time with experiments that produced arithmetic differences between groups that were accidental and not genuinely "significant.'

The now-codified accepted practice was in many ways similar to the protocols used by search engines and social media to direct our eyes and ears to content that might, possibly, be worthwhile. These systems are far from perfect, sometimes hiding good content or promoting worthless content. And, of course, the worth of content is a matter of personal interests and values, something that computer algorithms can mimic only imperfectly.

"Hypothesis testing" is an ad hoc set of not always consistent concepts cobbled together by a unorganized community of independent researchers, steered perhaps by the perceived authority of one statistical celebrity or another. It is not a mathematically derived, highly optimized calculation of objective worth, just a simple means to deal with the fact that arithmetic differences are influenced by sampling variation and noise, and that a detected difference might not reliably point to a genuine difference between groups.

It is simply not possible to understand hypothesis testing in the same way you can understand differentiation or data wrangling.

Hypothesis testing emerged in an era of bench-top and agricultural experiments conducted by a small community of self-identified scientists working without central control. It might have been a practicable solution to the problem of information overload in that era of small data. But the protocol has been frozen in place by textbooks; each generation passing it along to the next as received wisdom, in much the same way as the views of classical philosophers and poets were passed down to later generations as authoritative and unchallengeable.

So lets step back from this frozen statistical protocol of hypothesis testing and point out inconsistencies and peculiarities that make it hard to make sense of and perhaps unsuited to the needs of handling information overload in todays world of big data and huge scientific enterprise.

p-value is an inseparable tangle of the amount of data available and the effect size. With enough data, practically everything has a small p-value.

Hundreds of thousands (perhaps millions) of scientists churning out research results. A filter that eliminates 95% of the nonsense still lets through an unfathomable mass of content.

So many choices in research and analysis methods—which covariates to include, whether to exclude an inconvenient point as an outlier, multiple choices for the response variable, all combined with a professional priority to "publish or perish."

19.2 False discovery

SIMPLIFY THIS. Make a DAG with hundreds of explanatory variables, none of which is connected to the response variable.

19.3 Sources of false discovery

[NEEDS STREAMLINING and take out references to examples like Potomac/Austin]

How did the coupon classifier system identify so many accidental patterns, patterns that existed in the training data but not in the testing data?

One source of false discovery stems from having multiple potential response variables. In the Potomac/Austin example, there were ten different classifiers at work, one for each of the ten Austin products. Even if the probability of finding an accidental pattern in one classifier is small, looking in ten different places dramatically increases the odds of finding something.

Similarly, having a large number of explanatory variables – we had 100 in the coupon classifier – provides many opportunities for false discovery. The probability of an accidental pattern between one outcome and one explanatory variable is small, but with many explanatory variables each being considered it's much more likely to find something.

A third source of false discovery at work in the coupon classifier relates to the family of models selected to implement the classifier. We used a tree model classifier capable of searching through the (many) explanatory variables to find ones that are associated with the response outcome. Unbridled, the tree model is capable of very fine stratification. Each coupon classifiers stratified the customers into about 200 levels. On average, then, there were about 50 customers in each strata. But there is variation, so many of the strata are much smaller, with ten or fewer customers. The small groups were constructed by the tree-building algorithm to have similar outcomes among the members, so it's not surprising to see a very strong pattern in each group. For each classifier, about 15% of all customers fall into a strata with 20 or fewer customers.

19.4 Identifying false discovery

We use data to build statistical models and systems such as the coupon-assignment machine. False discovery occurs when a pattern or model performance seen with one set of data does not generalize to other potential data sets.

The basic technique to avoid false discovery is called **cross validation**. One simple approach to cross validation splits the

data frame into two randomly selected non-overlapping sets of rows: one for training and the other for testing. Use the training data to build the system. Use the *testing* data to evaluate the system's performance.

Most often, cross validation is used to test model prediction performance such as the root-mean-square error or the sensitivity and specificity of a classifier. This can be accomplished by taking the trained model and providing as input the explanatory variables from the testing data, then comparing the model output to the actual response variable values in the testing data. Note that using testing data in this way does not involve retraining the model on the testing data.

How big should the training set be compared to the testing set? For now, we'll keep things simple and encourage use of a 50:50 split or something very close to that.

This is a simple and reliable approach that should always be used.

19.5 False discovery and multiple testing

When the main interest is in an effect size, standard procedure calls for calculating a confidence interval on the effect. For example, a 2008 study examined the possible relationship between a woman's diet before conception and the sex of the conceived child. The popular press was particularly taken by this result from the study:

> Women producing male infants consumed more breakfast cereal than those with female infants. The odds ratio for a male infant was 1.87 (95% CI 1.31, 2.65) for women who consumed at least one bowl of breakfast cereal daily compared with those who ate less than or equal to one bowlful per week. [@fetal-sex-2008]

The model here is a classifier of the sex of the baby based on the amount of breakfast cereal eaten. The effect size tells the change in the odds of a male when the explanatory variable changes from one bowlful of cereal per week to one bowl per

day (or more). This effect size is sensibly reported as a ratio of the two odds. A ratio bigger than one means that boys are more likely outcomes for the one-bowl-a-day potential mother than the one-bowl-a-week potential mother. The 95% confidence interval is given as 1.31 to 2.65. This confidence interval does not contain 1. In a conventional interpretation, this provides compelling evidence that the relationship between cereal consumption and sex is not a false pattern.

But the confidence interval is not the complete story. The authors are clear in stating their methodology: "Data of the 133 food items from our food frequency questionnaire were analysed, and we also performed additional analyses using broader food groups." In other words, the authors had available more than 133 potential explanatory variables. For each of these explanatory variables, the study's authors constructed a confidence interval on the odds ratio. Most of the confidence intervals included 1, providing no compelling evidence of a relationship between that food item and the sex of the conceived child. As it happens, breakfast cereal produced the confidence interval that was the most distant from an odds ratio of 1.

Let's look at the range of confidence intervals that can be found from studying 100 potential random variables that are each unrelated to the response variable. We'll simulate a response randomly generated "sex" G and B where the odds of G is 1. Similarly, each explanatory variable will be a randomly generated "consumption" high or low where the odds of high is 1. A simple stratification of sex by consumption will generate the odds of G for those cases with consumption Y and also the odds of G for those cases with consumption N. Taking the ratio of these odds gives, naturally enough, the odds ratio. We can also calculate from the stratified data a 95% confidence interval on the odds ratio.

So that the results will be somewhat comparable to the results in @fetal-sex-2008, we'll use a similar sample size, that is, n = 740. Table @ref(tab:sex-consumption-1) shows one trial of the simulation.

(ref:sex-consumption-1-cap) A stratification of sex outcome (B or G) on consumption (high or low) for one trial of the simulation described in the text.

Table 11: (ref:sex-consumption-1-cap)

	high	low
B	165	182
G	211	182

Referring to Table @ref(tab:sex-consumption-1), you can see that the odds of G when consumption is low is 182 / 182 = 1. The odds of G when consumption is high is 211/165 = 1.28. The 95% confidence interval on the odds ratio can be calculated. It is 0.95 to 1.73. Since that includes 1, the data underlying Table @ref(tab:sex-consumption-1) provide little or no evidence for a relationship between sex and consumption. This is exactly what we expect, since the simulation involves entirely random data.

Figure 44 shows the 95% confidence interval on the odds ratio for 133 trials like that in Table @ref(tab:sex-consumption-1). The confidence interval from each trial is shown as a horizontal line. The large majority of them include 1. That's to be expected because the data have been generated so that sex and consumption have no relationship except those arising by chance.

```
Warning: geom_vline(): Ignoring `mapping` because `xintercept` was provided.
```

Nonetheless, out of 133 simulations there are six where the confidence interval does not include 1. These are shown in red. By necessity, one of the intervals will be the most extreme. If instead of numbering the simulations, we had labelled them with food items – e.g. grapefruit, breakfast cereal, toast – we would have a situation very similar to what seems to have happened in the sex-vs-food study. (For a more detailed analysis of the impact of multiple testing in @fetal-sex-2008, see @young-2009.)

Suppose now that half of the data used in @fetal-sex-2008 had been held back as testing data. Using the training data, it would be an entirely legitimate practice to generate hypotheses about which specific food items might be related to the sex of the baby. The validity of any one selected hypothesis could then

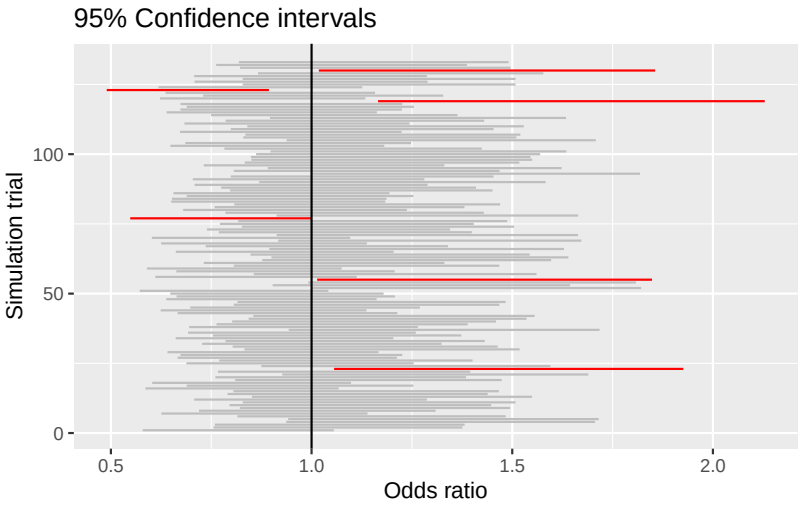


Figure 44: Confidence intervals on the odds ratio comparing female and male birth rates for many trials of simulated data with no genuine relationship between the explanatory and response variables.

be established using the testing data without the ambiguity introduced by multiple testing. The testing data confidence interval can be taken at face value; the training data confidence interval cannot.

19.6 Example: Organic discovery?

It's easy to find organic foods in many large grocery stores. Advocates of an organic diet are attracted by a view that it is sustainable, promotes small farms, and helps avoid contact with pesticides. There are also nay-sayers who make valid points, but that is not our purpose here. Informally, I find that many people and news reports point to the health benefits of an organic diet. Usually they believe that these benefits are an established fact.

A 2018 New York *Times* article observed:

> *People who buy organic food are usually convinced it's better for their health, and they're willing to pay dearly for it. But until now, evidence of the benefits*

> *of eating organic has been lacking.* [@NYT-2018-10-23-Rabin]

The new evidence of health benefits is reported in an article in the *Journal of the American Medical Association: Internal Medicine* [@baudry-2018]

Describing the findings of the research, the *Times* article continued:

> *Even after these adjustments [for covariates], the most frequent consumers of organic food had 76 percent fewer lymphomas, with 86 percent fewer non-Hodgkin's lymphomas, and a 34 percent reduction in breast cancers that develop after menopause.*

The study warrants being taken seriously: it involved about 70,000 French adults among whom 1340 cancers were noted. The summary of organic foot consumption was a scale from 0 to 32 and included 16 labeled products including dairy, meat and fish, eggs, coffee and tea, wine, vegetable oils, and sweets such as chocolate. Adjustment was made for a substantial number of covariates: age, sex, educational level, marital status, income, physical activity, smoking, alcohol intake, family history of cancer, body mass index, hormonal treatment for menopause, and others.

Yet … the research displays many of the features that can lead to false discovery. For instance, results were reported for four different types of cancer: breast, prostate, skin, lymphomas. The study reports p-values and hazard ratios[19] comparing cancer rates among the four quartiles of the organic consumption index.

Comparing the most organic (average organic index 19.36/32) and the least organic (average index 0.72/32) groups, the 95% confidence interval on the relative risk and p-values given in the study's Table 4 are:

- Breast cancer: 0.66 - 1.16 (p = 0.38)
- Prostate cancer: 0.61- 1.73 (p = 0.39)
- Skin cancer: 0.49 - 1.28 (p = 0.11)

[19] Hazard ratios are analogous to risk ratios.

- Lymphomas: 0.07 - 0.69 (p = 0.05)

You might be surprised to see that the confidence interval on the relative risk for breast cancer includes 1.0, which suggests no evidence for an effect. As clearly stated in the report, the risk reduction for breast cancer is seen only in a subgroup of study participants: those who are postmenopausal. And even then, the confidence intervals continue to include 1.0:

- Breast cancer pre-menopausal: 0.67 - 1.52 (p = 0.85)
- Breast cancer post-menopausal: 0.53 - 1.18 (p = 0.18)

So where is the claimed 34% reduction in breast cancer cited in the New York Times article. It turns out the the study used two different indices of organic food consumption. The 0 to 32 scale which includes many items for which the amount consumed is very small (e.g., coffee, chocolate) and a "simplified, plant derived organic food score." It's only when you look at the full 0 to 32 scale that you see the reduction in post-menopausal breast cancer: the confidence interval is 0.45 to 0.96 (p = 0.03).

What about cancer rates overall? For the 0 to 32 scale the risk ratio was 0.58 - 1.01 (p = 0.10). To see the claimed reduction clearly you need to look at the simplified food score which gives 0.63 - 0.89 ($p < 0.005$). And it's only in comparing the highest-index quarter of participants with the low

```
Warning: geom_hline(): Ignoring `mapping` because `yintercept` was provided.
```

19.7 NOTES IN DRAFT

"Statistical crisis" in science

https://www.americanscientist.org/article/the-statistical-crisis-in-science

Garden of the Forking Paths

Ionedes

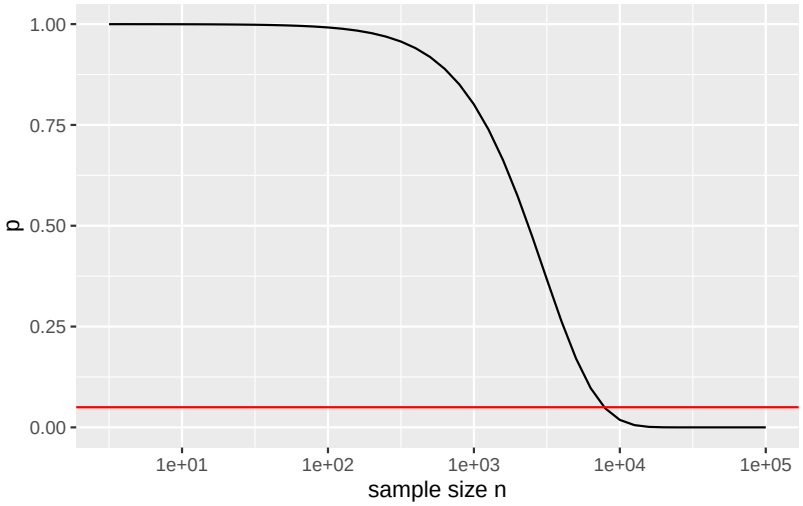


Figure 45: The p-value as a function of sample size n when the test statistic R-squared has the trivial value 0.001. The horizontal line shows the usual threshold for "significance" of $p < 0.05$.