Data scientists are often faced with yes-no questions about the world. You have seen some examples of such questions in this course:

* Is chocolate good for you?
* Did water from the Broad Street pump cause cholera?
* Have the demographics in California changed over the past decade?

Whether we answer questions like these depends on the data we have. Census data about California can settle questions about demographics with hardly any uncertainty about the answer. We know that Broad Street pump water was contaminated by waste from cholera victims, so we can make a pretty good guess about whether it caused cholera.

Whether chocolate or any other treatment is good for you will almost certainly have to be decided by medical experts, but an initial step consists of using data science to analyze data from studies and randomized experiments.

In this chapter, we will try to answer such yes-no questions, basing our conclusions on random samples and empirical distributions.

**Assessing Models**

In data science, a “model” is a set of assumptions about data. Often, models include assumptions about chance processes used to generate data.

Sometimes, data scientists have to decide whether or not their models are good. In this section we will discuss two examples of making such decisions. In later sections we will use the methods developed here as the building blocks of a general framework for testing hypotheses.

**U.S. Supreme Court, 1965: Swain vs. Alabama**

In the early 1960’s, in Talladega County in Alabama, a black man called Robert Swain was convicted of raping a white woman and was sentenced to death. He appealed his sentence, citing among other factors the all-white jury. At the time, only men aged 21 or older were allowed to serve on juries in Talladega County. In the county, 26% of the eligible jurors were black, but there were only 8 black men among the 100 selected for the jury panel in Swain’s trial. No black man was selected for the trial jury.

In 1965, the Supreme Court of the United States denied Swain’s appeal. In its ruling, the Court wrote “… the overall percentage disparity has been small and reflects no studied attempt to include or exclude a specified number of Negroes.”

Jury panels are supposed to be selected at random from the eligible population. Because 26% of the eligible population was black, 8 black men on a panel of 100 might seem low.

**A Model**

But one view of the data – a model, in other words – is that the panel was selected at random and ended up with a small number of black men just due to chance. This model is consistent with what the Supreme Court wrote in its ruling.

The model specifies the details of a chance process. It says the data are like a random sample from a population in which 26% of the people are black. We are in a good position to assess this model, because:

* We can simulate data based on the model. That is, we can simulate drawing at random from a population of whom 26% are black.
* Our simulation will show what a panel *would* look like *if* it were selected at random.
* We can then compare the results of the simulation with the composition of Robert Swain’s panel.
* If the results of our simulation are not consistent with the composition of Swain’s panel, that will be evidence against the model of random selection.

Let’s go through the process in detail.

**The Statistic**

First, we have to choose a statistic to simulate. The statistic has to be able to help us decide between the model and alternative views about the data. The model says the panel was drawn at random. The alternative viewpoint, suggested by Robert Swain’s appeal, is that the panel was not drawn at random because it contained too few black men. A natural statistic, then, is the number of black men in our simulated sample of 100 men representing the panel. Small values of the statistic will favor the alternative viewpoint.

**Predicting the Statistic Under the Model**

If the model were true, how big would the statistic typically be? To answer that, we have to start by working out the details of the simulation.

**Generating One Value of the Statistic**

First let’s figure out how to simulate one value of the statistic. For this, we have to sample 100 times at random from the population of eligible jurors and count the number of black men we get.

One way is to set up a table representing the eligible population and use sample as we did in the previous chapter. But there is also a quicker way, using a datascience function tailored for sampling at random from categorical distributions. We will use it several times in this chapter.

The sample\_proportions function in the datascience library takes two arguments:

* the sample size
* the distribution of the categories in the population, as a list or array of proportions that add up to 1

It returns an array containing the distribution of the categories in a random sample of the given size taken from the population. That’s an array consisting of the sample proportions in all the different categories.

To see how to use this, remember that according to our model, the panel is selected at random from a population of men among whom 26% were black and 74% were not. Thus the distribution of the two categories can be represented as the list [0.26, 0.74], which we have assigned to the name eligible\_population. Now let’s sample at random 100 times from this distribution, and see what proportions of the two categories we get in our sample.

eligible\_population **=** [0.26, 0.74]

sample\_proportions(100, eligible\_population)

array([0.27, 0.73])

That was easy! The proportion of black men in the random sample is item(0) of the output array.

Because there are 100 men in the sample, the number of men in each category is 100 times the proportion. So we can just as easily simulate counts instead of proportions, and access the count of black men only.

Run the cell a few times to see how the output varies.

*# count of black men in a simulated panel*

(100 **\*** sample\_proportions(100, eligible\_population))**.**item(0)

27.0

**Running the Simulation**

To get a sense of the variability without running the cell over and over, let’s generate 10,000 simulated values of the count.

The code follows the same steps that we have used in every simulation. First, we define a function to simulate one value of the count, using the code we wrote above.

**def** **one\_simulated\_count**():

**return** (100 **\*** sample\_proportions(100, eligible\_population))**.**item(0)

Next, we create an array of 10,000 simulated counts by using a for loop.

counts **=** make\_array()

repetitions **=** 10000

**for** i **in** np**.**arange(repetitions):

counts **=** np**.**append(counts, one\_simulated\_count())

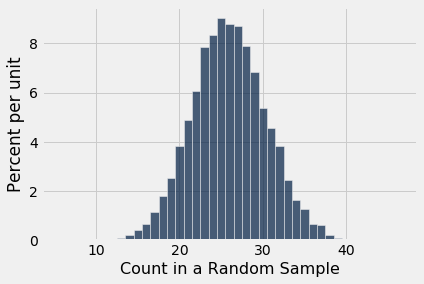
**The Prediction**

To interpret the results of our simulation, we start as usual by visualizing the results by an empirical histogram.

Table()**.**with\_column(

'Count in a Random Sample', counts

)**.**hist(bins **=** np**.**arange(5.5, 46.6, 1))



The histogram tells us what the model of random selection predicts about our statistic, the count of black men in the sample.

To generate each simulated count, we drew at 100 times at random from a population in which 26% were black. So, as you would expect, most of the simulated counts are around 26. They are not exactly 26: there is some variation. The counts range from about 10 to about 45.

**Comparing the Prediction and the Data**

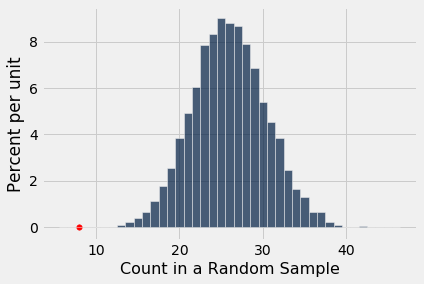
Though the simulated counts are quite varied, very few of them came out to be eight or less. The value eight is far out in the left hand tail of the histogram. It’s the red dot on the horizontal axis of the histogram.

Table()**.**with\_column(

'Count in a Random Sample', counts

)**.**hist(bins **=** np**.**arange(5.5, 46.6, 1))

plots**.**scatter(8, 0, color**=**'red', s**=**30);



The simulation shows that if we select a panel of 100 jurors at random from the eligible population, we are very unlikely to get counts of black men as low as the eight that were in Swain’s jury panel. This is evidence that the model of random selection of the jurors in the panel is not consistent with the data from the panel.

When the data and a model are inconsistent, the model is hard to justify. After all, the data are real. The model is just a set of assumptions. When assumptions are at odds with reality, we have to question those assumptions.

While it is *possible* that a panel like Robert Swain’s could have been generated by chance, our simulation demonstrates that it is very unlikely. Thus our assessment is that the model of random draws is not supported by the evidence. Swain’s jury panel does not look like the result of random sampling from the population of eligible jurors.

This method of assessing models is very general. Here is an example in which we use it to assess a model in a completely different setting.

**Mendel’s Pea Flowers**

[Gregor Mendel](https://en.wikipedia.org/wiki/Gregor_Mendel) (1822-1884) was an Austrian monk who is widely recognized as the founder of the modern field of genetics. Mendel performed careful and large-scale experiments on plants to come up with fundamental laws of genetics.

Many of his experiments were on varieties of pea plants. He formulated sets of assumptions about each variety; these were his models. He then tested the validity of his models by growing the plants and gathering data.

Let’s analyze the data from one such experiment to see if Mendel’s model was good.

In a particular variety, each plant has either purple flowers or white. The color in each plant is unaffected by the colors in other plants. Mendel hypothesized that the plants should bear purple or white flowers at random, in the ratio 3:1.

**Mendel’s Model**

For every plant, there is a 75% chance that it will have purple flowers, and a 25% chance that the flowers will be white, regardless of the colors in all the other plants.

**Approach to Assessment**

To go about assessing Mendel’s model, we can simulate plants under the assumptions of the model and see what it predicts. Then we will be able to compare the predictions with the data that Mendel recorded.

**The Statistic**

Our goal is to see whether or not Mendel’s model is good. We need to simulate a statistic that will help us make this decision.

If the model is good, the percent of purple-flowering plants in the sample should be close to 75%. If the model is not good, the percent purple-flowering will be away from 75%. It may be higher, or lower; the direction doesn’t matter.

The key for us is the *distance* between 75% and the percent of purple-flowering plants in the sample. Big distances are evidence that the model isn’t good.

Our statistic, therefore, is the **distance between the sample percent and 75%**:

∣∣sample percent of purple-flowering plants−75∣∣|sample percent of purple-flowering plants−75|

**Predicting the Statistic Under the Model**

To see how big the distance would be if Mendel’s model were true, we can use sample\_proportions to simulate the distance under the assumptions of the model.

First, we have to figure out how many times to sample. To do this, remember that we are going to compare our simulation with Mendel’s plants. So we should simulate the same number of plants that he had.

Mendel grew a lot of plants. There were 929 plants of the variety corresponding to this model. So we have to sample 929 times.

**Generating One Value of the Statistic**

The steps in the calculation:

* Sample 929 times at random from the distribution specified by the model and find the sample proportion in the purple-flowering category.
* Multiply the proportion by 100 to get a pecent.
* Subtract 75 and take the absolute value of the difference.

That’s the statistic: the distance between the sample percent and 75.

We will start by defining a function that takes a proportion and returns the absolute difference between the corresponding percent and 75.

**def** **distance\_from\_75**(p):

**return** abs(100**\***p **-** 75)

To simulate one value of the distance between the sample percent of purple-flowering plants and 75%, under the assumptions of Mendel’s model, we have to first simulate the proportion of purple-flowering plants among 929 plants under the assumption of the model, and then calculate the discrepancy from 75%.

model\_proportions **=** [0.75, 0.25]

proportion\_purple\_in\_sample **=** sample\_proportions(929, model\_proportions)**.**item(0)

distance\_from\_75(proportion\_purple\_in\_sample)

1.7491926803014053

That’s one simulated value of the distance between the sample percent of purple-flowering plants and 75% as predicted by Mendel’s model.

**Running the Simulation**

To get a sense of how variable the distance could be, we have to simulate it many more times.

We will generate 10,000 values of the distance. As before, we will first use the code we developed above to define a function that returns one simulated value Mendel’s hypothesis.

**def** **one\_simulated\_distance**():

proportion\_purple\_in\_sample **=** sample\_proportions(929, model\_proportions)**.**item(0)

**return** distance\_from\_75(proportion\_purple\_in\_sample)

Next, we will use a for loop to create 10,000 such simulated distances.

distances **=** make\_array()

repetitions **=** 10000

**for** i **in** np**.**arange(repetitions):

distances **=** np**.**append(distances, one\_simulated\_distance())

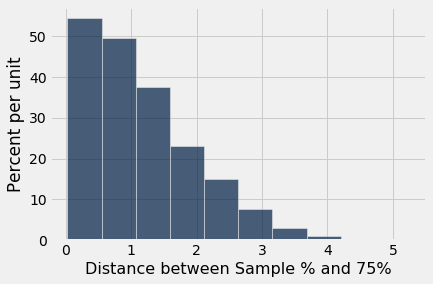
**The Prediction**

The empirical histogram of the simulated values shows the distribution of the distance as predicted by Mendel’s model.

Table()**.**with\_column(

'Distance between Sample % and 75%', distances

)**.**hist()



Look on the horizontal axis to see the typical values of the distance, as predicted by the model. They are rather small. For example, a high proportion of the distances are in the range 0 to 1, meaning that for a high proportion of the samples, the percent of purple-flowering plants is within 1% of 75%, that is, the sample percent is in the range 74% to 76%.

**Comparing the Prediction and the Data**

To assess the model, we have to compare this prediction with the data. Mendel recorded the number of purple and white flowering plants. Among the 929 plants that he grew, 705 were purple flowering. That’s just about 75.89%.

705 **/** 929

0.7588805166846071

So the observed value of our statistic – the distance between Mendel’s sample percent and 75 – is about 0.89:

observed\_statistic **=** distance\_from\_75(705**/**929)

observed\_statistic

0.8880516684607045

Just by eye, locate roughly where 0.89 is on the horizontal axis of the histogram. You will see that it is clearly in the heart of the distribution predicted by Mendel’s model.

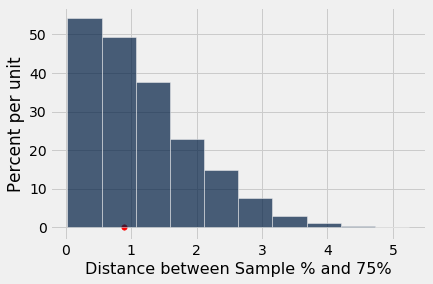
The cell below redraws the histogram with the observed value plotted on the horizontal axis.

Table()**.**with\_column(

'Distance between Sample % and 75%', distances

)**.**hist()

plots**.**scatter(observed\_statistic, 0, color**=**'red', s**=**30);



The observed statistic is like a typical distance predicted by the model. By this measure, the data are consistent with the histogram that we generated under the assumptions of Mendel’s model. This is evidence in favor of the model.

**Multiple Categories**

We have developed a way of assessing models about chance processes that generate data in two categories. The method extends to models involving data in multiple categories. The process of assessment is the same as before, the only difference being that we have to come up with a new statistic to simulate.

Let’s do this in an example that addresses the same kind of question that was raised in the case of Robert Swain’s jury panel. This time, the data are more recent.

**Jury Selection in Alameda County**

In 2010, the American Civil Liberties Union (ACLU) of Northern California presented a [report](https://www.aclunc.org/sites/default/files/racial_and_ethnic_disparities_in_alameda_county_jury_pools.pdf) on jury selection in Alameda County, California. The report concluded that certain ethnic groups are underrepresented among jury panelists in Alameda County, and suggested some reforms of the process by which eligible jurors are assigned to panels. In this section, we will perform our own analysis of the data and examine some questions that arise as a result.

Some details about jury panels and juries will be helpful in interpreting the results of our analysis.

**Jury Panels**

A jury panel is a group of people chosen to be prospective jurors; the final trial jury is selected from among them. Jury panels can consist of a few dozen people or several thousand, depending on the trial. By law, a jury panel is supposed to be representative of the community in which the trial is taking place. Section 197 of California’s Code of Civil Procedure says, “All persons selected for jury service shall be selected at random, from a source or sources inclusive of a representative cross section of the population of the area served by the court.”

The final jury is selected from the panel by deliberate inclusion or exclusion. The law allows potential jurors to be excused for medical reasons; lawyers on both sides may strike a certain number of potential jurors from the list in what are called “peremptory challenges”; the trial judge might make a selection based on questionnaires filled out by the panel; and so on. But the initial panel is supposed to resemble a random sample of the population of eligible jurors.

**Composition of Panels in Alameda County**

The focus of the study by the ACLU of Northern California was the ethnic composition of jury panels in Alameda County. The ACLU compiled data on the ethnic composition of the jury panels in 11 felony trials in Alameda County in the years 2009 and 2010. In those panels, the total number of people who reported for jury service was 1,453. The ACLU gathered demographic data on all of these prosepctive jurors, and compared those data with the composition of all eligible jurors in the county.

The data are tabulated below in a table called jury. For each ethnicity, the first value is the proportion of all eligible juror candidates of that ethnicity. The second value is the proportion of people of that ethnicity among those who appeared for the process of selection into the jury.

jury **=** Table()**.**with\_columns(

'Ethnicity', make\_array('Asian', 'Black', 'Latino', 'White', 'Other'),

'Eligible', make\_array(0.15, 0.18, 0.12, 0.54, 0.01),

'Panels', make\_array(0.26, 0.08, 0.08, 0.54, 0.04)

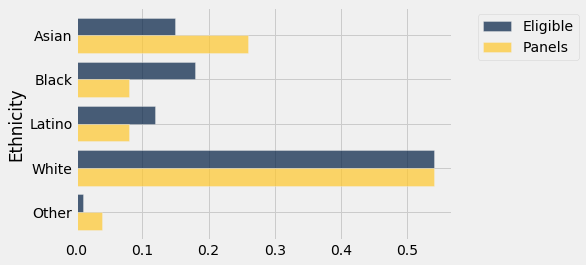
)

jury

| **Ethnicity** | **Eligible** | **Panels** |
| --- | --- | --- |
| Asian | 0.15 | 0.26 |
| Black | 0.18 | 0.08 |
| Latino | 0.12 | 0.08 |
| White | 0.54 | 0.54 |
| Other | 0.01 | 0.04 |

Some ethnicities are overrepresented and some are underrepresented on the jury panels in the study. A bar chart is helpful for visualizing the differences.

jury**.**barh('Ethnicity')



**Comparison with Panels Selected at Random**

What if we select a random sample of 1,453 people from the population of eligible jurors? Will the distribution of their ethnicities look like the distribution of the panels above?

We can answer these questions by using sample\_proportions and augmenting the jury table with a column of the proportions in our sample.

**Technical note.** Random samples of prospective jurors would be selected without replacement. However, when the size of a sample is small relative to the size of the population, sampling without replacement resembles sampling with replacement; the proportions in the population don’t change much between draws. The population of eligible jurors in Alameda County is over a million, and compared to that, a sample size of about 1500 is quite small. We will therefore sample with replacement.

In the cell below, we sample at random 1453 times from the distribution of eligible jurors, and display the distribution of the random sample along with the distributions of the eligible jurors and the panel in the data.

eligible\_population **=** jury**.**column('Eligible')

sample\_distribution **=** sample\_proportions(1453, eligible\_population)

panels\_and\_sample **=** jury**.**with\_column('Random Sample', sample\_distribution)

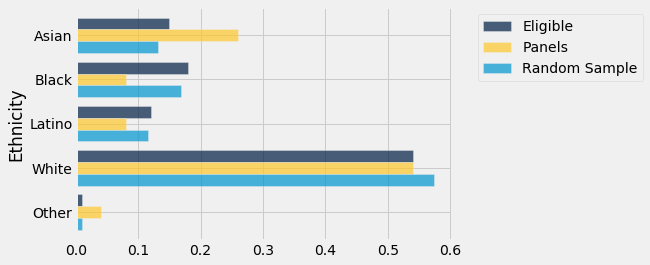
panels\_and\_sample

| **Ethnicity** | **Eligible** | **Panels** | **Random Sample** |
| --- | --- | --- | --- |
| Asian | 0.15 | 0.26 | 0.13214 |
| Black | 0.18 | 0.08 | 0.168617 |
| Latino | 0.12 | 0.08 | 0.115623 |
| White | 0.54 | 0.54 | 0.573985 |
| Other | 0.01 | 0.04 | 0.00963524 |

The distribution of the random sample is quite close to the distribution of the eligible population, unlike the distribution of the panels.

As always, it helps to visualize.

panels\_and\_sample**.**barh('Ethnicity')



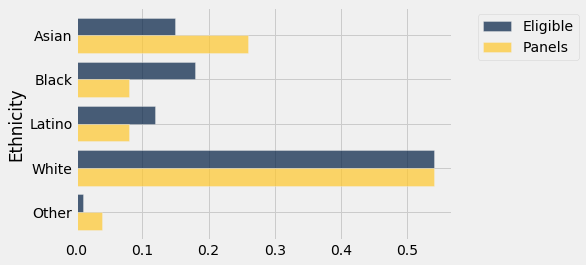
The bar chart shows that the distribution of the random sample resembles the eligible population but the distribution of the panels does not.

To assess whether this observation is particular to one random sample or more general, we can simulate multiple panels under the model of random selection and see what the simulations predict. But we won’t be able to look at thousands of bar charts like the one above. We need a statistic that will help us assess whether or not the model or random selection is supported by the data.

**A New Statistic: The Distance between Two Distributions**

We know how to measure how different two numbers are – if the numbers are xx and yy, the distance between them is |x−y||x−y|. Now we have to quantify the distance between two distributions. For example, we have to measure the distance between the blue and gold distributions below.

jury**.**barh('Ethnicity')



For this we will compute a quantity called the *total variation distance* between two distributions. The calculation is as an extension of the calculation of the distance between two numbers.

To compute the total variation distance, we first take the difference between the two proportions in each category.

*# Augment the table with a column of differences between proportions*

jury\_with\_diffs **=** jury**.**with\_column(

'Difference', jury**.**column('Panels') **-** jury**.**column('Eligible')

)

jury\_with\_diffs

| **Ethnicity** | **Eligible** | **Panels** | **Difference** |
| --- | --- | --- | --- |
| Asian | 0.15 | 0.26 | 0.11 |
| Black | 0.18 | 0.08 | -0.1 |
| Latino | 0.12 | 0.08 | -0.04 |
| White | 0.54 | 0.54 | 0 |
| Other | 0.01 | 0.04 | 0.03 |

Take a look at the column Difference and notice that the sum of its entries is 0: the positive entries add up to 0.14, exactly canceling the total of the negative entries which is -0.14.

This is numerical evidence of the fact that in the bar chart, the gold bars exceed the blue bars by exactly as much as the blue bars exceed the gold. The proportions in each of the two columns Panels and Eligible add up to 1, and so the give-and-take between their entries must add up to 0.

To avoid the cancellation, we drop the negative signs and then add all the entries. But this gives us two times the total of the positive entries (equivalently, two times the total of the negative entries, with the sign removed). So we divide the sum by 2.

jury\_with\_diffs **=** jury\_with\_diffs**.**with\_column(

'Absolute Difference', np**.**abs(jury\_with\_diffs**.**column('Difference'))

)

jury\_with\_diffs

| **Ethnicity** | **Eligible** | **Panels** | **Difference** | **Absolute Difference** |
| --- | --- | --- | --- | --- |
| Asian | 0.15 | 0.26 | 0.11 | 0.11 |
| Black | 0.18 | 0.08 | -0.1 | 0.1 |
| Latino | 0.12 | 0.08 | -0.04 | 0.04 |
| White | 0.54 | 0.54 | 0 | 0 |
| Other | 0.01 | 0.04 | 0.03 | 0.03 |

jury\_with\_diffs**.**column('Absolute Difference')**.**sum() **/** 2

0.14

This quantity 0.14 is the *total variation distance* (TVD) between the distribution of ethnicities in the eligible juror population and the distribution in the panels.

We could have obtained the same result by just adding the positive differences. But our method of including all the absolute differences eliminates the need to keep track of which differences are positive and which are not.

**Simulating One Value of the Statistic**

We will use the total variation distance between distributions as the statistic to simulate. It will help us decide whether the model of random selection is good, because large values of the distance will be evidence against the model.

Keep in mind that **the observed value of our statistic is 0.14**, calculated above.

Since we are going to be computing total variation distance repeatedly, we will write a function to compute it.

The function total\_variation\_distance returns the TVD between distributions in two arrays.

**def** **total\_variation\_distance**(distribution\_1, distribution\_2):

**return** sum(np**.**abs(distribution\_1 **-** distribution\_2)) **/** 2

This function will help us calculate our statistic in each repetition of the simulation. But first, let’s check that it gives the right answer when we use it to compute the distance between the blue (eligible) and gold (panels) distributions above.

total\_variation\_distance(jury**.**column('Panels'), jury**.**column('Eligible'))

0.14

This agrees with the value that we computed directly without using the function.

In the cell below we use the function to compute the TVD between the distributions of the eligible jurors and one random sample. This is the code for simulating one value of our statistic. Recall that eligible\_population is the array containing the distribution of the eligible jurors.

sample\_distribution **=** sample\_proportions(1453, eligible\_population)

total\_variation\_distance(sample\_distribution, eligible\_population)

0.0202202339986235

Notice that the distance is quite a bit smaller than 0.14, the distance between the distribution of the panels and the eligible jurors.

We are now ready to run a simulation to assess the model of random selection.

**Predicting the Statistic Under the Model of Random Selection**

The total variation distance between the distributions of the random sample and the eligible jurors is the statistic that we are using to measure the distance between the two distributions. By repeating the process of sampling, we can see how much the statistic varies across different random samples.

The code below simulates the statistic based on a large number of replications of the random sampling process, following our usual sequence of steps for simulation. We first define a function that returns one simulated value of the total variation distance under the hypothesis of random selection. Then we use our function in a for loop to create an array tvds consisting of 5,000 such distances.

*# Simulate one simulated value of*

*# the total variation distance between*

*# the distribution of a sample selected at random*

*# and the distribution of the eligible population*

**def** **one\_simulated\_tvd**():

sample\_distribution **=** sample\_proportions(1453, eligible\_population)

**return** total\_variation\_distance(sample\_distribution, eligible\_population)

tvds **=** make\_array()

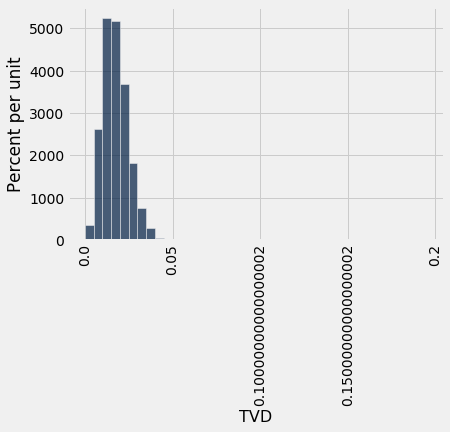
repetitions **=** 5000

**for** i **in** np**.**arange(repetitions):

tvds **=** np**.**append(tvds, one\_simulated\_tvd())

The empirical histogram of the simulated distances shows that drawing 1453 jurors at random from the pool of eligible candidates results in a distribution that rarely deviates from the eligible jurors’ race distribution by more than about 0.05.

Table()**.**with\_column('TVD', tvds)**.**hist(bins**=**np**.**arange(0, 0.2, 0.005))



**Assessing the Model of Random Selection**

The panels in the study, however, were not quite so similar to the eligible population. The total variation distance between the panels and the population was 0.14, which is far out in the tail of the histogram above. It does not look at all like a typical distance between a random sample and the eligible population.

The data in the panels is not consistent with the predicted values of the statistic based on the model of random selection. So our analysis supports the ACLU’s calculation that the panels were not representative of the distribution provided for the eligible jurors.

**Some Possible Explanations for the Differences**

As with most such analyses, however, our analysis does not say *why* the distributions are different or what the difference might imply.

The ACLU report discusses several possible reasons for the discrepancies. For example, some minority groups were underrepresented on the records of voter registration and of the Department of Motor Vehicles, the two main sources from which jurors are selected. At the time of the study, the county did not have an effective process for following up on prospective jurors who had been called but had failed to appear. The ACLU listed several other reasons as well. Whatever the reasons, it seems clear that the composition of the jury panels was different from what we would have expected in a random sample from the distribution in the Eligible column.

**Questions about the Data**

We have developed a powerful technique that helps decide whether one distribution looks like a random sample from another. But data science is about more than techniques. In particular, data science always involves a thoughtful examination of how the data were gathered.

**Eligible Jurors.** First, it is important to remember that not everyone is eligible to serve on a jury. On its [website](http://www.alameda.courts.ca.gov/pages.aspx/jury-duty-overview), the Superior Court of Alameda County says, “You may be called to serve if you are 18 years old, a U.S. citizen and a resident of the county or district where summoned. You must be able to understand English, and be physically and mentally capable of serving. In addition, you must not have served as any kind of juror in the past 12 months, nor have been convicted of a felony.”

The Census doesn’t maintain records of the populations in all these categories. Thus the ACLU had to obtain the demographics of eligible jurors in some other way. Here is their own description of the process they followed and some flaws that it might contain.

“For the purpose of determining the demographics of Alameda County’s jury eligible population, we used a declaration that was prepared for the Alameda County trial of People v. Stuart Alexander in 2002. In the declaration, Professor Weeks, a demographer at San Diego State University, estimated the jury eligible population for Alameda County, using the 2000 Census data. To arrive at this estimate, Professor Weeks took into account the number of people who are not eligible for jury services because they do not speak English, are not citizens, are under 18, or have a felony conviction. Thus, his estimate should be an accurate assessment of who is actually eligible for jury service in Alameda County, much more so than simply reviewing the Census report of the race and ethnicity of all people living in Alameda County. It should be noted, however, that the Census data on which Professor Weeks relied is now ten years old and the demographics of the county may have changed by two or three percent in some categories.”

Thus the distribution of ethnicities of eligible jurors used in the analysis is itself an estimate and might be somewhat out of date.

**Panels.** In addition, panels aren’t selected from the entire eligible population. The Superior Court of Alameda County says, “The objective of the court is to provide an accurate cross-section of the county’s population. The names of jurors are selected at random from everyone who is a registered voter and/or has a driver’s license or identification card issued by the Department of Motor Vehicles.”

All of this raises complex questions about how to accurately estimate the ethnic composition of eligible jurors in Alameda County.

It is not clear exactly how the 1453 panelists were classified into the different ethnic categories (the ACLU report says that “attorneys … cooperated in collecting jury pool data”). There are serious social, cultural, and political factors that affect who gets classified or self-classifies into each ethnic category. We also don’t know whether the definitions of those categories in the panels are the same as those used by Professor Weeks who in turn used Census categories in his estimation process. Thus there are also questions about the correspondence between the two distributions being compared.

Thus, while we have a clear conclusion about the data in our table – the panels do not look like a random sample from the distribution provided for eligible jurors – questions about the nature of the data prevent us from concluding anything broader.

**Decisions and Uncertainty**

We have seen several examples of assessing models that involve chance, by comparing observed data to the predictions made by the models. In all of our examples, there has been no doubt about whether the data were consistent with the model’s predictions. The data were either very far away from the predictions, or very close to them.

But outcomes are not always so clear cut. How far is “far”? Exactly what does “close” mean? While these questions don’t have universal answers, there are guidelines and conventions that you can follow. In this section we will describe some of them.

But first let us develop a general framework of decision making, into which all our examples will fit.

What we have developed while assessing models are some of the fundamental concepts of statistical tests of hypotheses. Using statistical tests as a way of making decisions is standard in many fields and has a standard terminology. Here is the sequence of the steps in most statistical tests, along with some terminology and examples. You will see that they are consistent with the sequence of steps we have used for assessing models.

**Step 1: The Hypotheses**

All statistical tests attempt to choose between two views of the world. Specifically, the choice is between two views about how the data were generated. These two views are called *hypotheses*.

**The null hypothesis.** This is a clearly defined model about chances. It says that the data were generated at random under clearly specified assumptions about the randomness. The word “null” reinforces the idea that if the data look different from what the null hypothesis predicts, the difference is due to *nothing* but chance.

From a practical perspective, **the null hypothesis is a hypothesis under which you can simulate data.**

In the example about Mendel’s model for the colors of pea plants, the null hypothesis is that the assumptions of his model are good: each plant has a 75% chance of having purple flowers, independent of all other plants.

Under this hypothesis, we were able to simulate random samples, by using sample\_proportions(929, [0.75, 0.25]). We used a sample size of 929 because that’s the number of plants Mendel grew.

**The alternative hypothesis.** This says that some reason other than chance made the data differ from the predictions of the model in the null hypothesis.

In the example about Mendel’s plants, the alternative hypothesis is simply that his model isn’t good.

**Step 2: The Test Statistic**

In order to decide between the two hypothesis, we must choose a statistic that we can use to make the decision. This is called the **test statistic**.

In the example of Mendel’s plants, our statistic was the absolute difference between the sample percent and 75% which was predicted by his model.

∣∣sample percent of purple-flowering plants−75∣∣|sample percent of purple-flowering plants−75|

To see how to make the choice in general, look at the alternative hypothesis. What values of the statistic will make you think that the alternative hypothesis is a better choice than the null?

* If the answer is “big values,” you might have a good choice of statistic.
* So also if the answer is “small values.”
* But if the answer is “both big values and small values,” we recommend that you look again at your statistic and see if taking an absolute value can change the answer to just “big values”.

In the case of the pea plants, a sample percent of around 75% will be consistent with the model, but percents much bigger or much less than 75 will make you think that the model isn’t good. This indicates that the statistic should be the *distance* between the sample percent and 75, that is, the absolute value of the difference between them. Big values of the distance will make you lean towards the alternative.

The **observed value of the test statistic** is the value of the statistic you get from the data in the study, not a simulated value. Among Mendel’s 929 plants, 705 had purple flowers. The observed value of the test statistic was therefore

abs ( 100 **\*** (705 **/** 929) **-** 75)

0.8880516684607045

**Step 3: The Distribution of the Test Statistic, Under the Null Hypothesis**

The main computational aspect of a test of hypotheses is figuring out *what the values of the test statistic might be if the null hypothesis were true*.

The test statistic is simulated based on the assumptions of the model in the null hypothesis. That model involves chance, so the statistic comes out differently when you simulate it multiple times.

By simulating the statistic repeatedly, we get a good sense of its possible values and which ones are more likely than others. In other words, we get a good approximation to the probability distribution of the statistic, as predicted by the model in the null hypothesis.

As with all distributions, it is very useful to visualize this distribution by a histogram. We have done so in all our examples.

**Step 4. The Conclusion of the Test**

The choice between the null and alternative hypotheses depends on the comparison between what you computed in Steps 2 and 3: the observed value of the test statistic and its distribution as predicted by the null hypothesis.

If the two are consistent with each other, then the observed test statistic is in line with what the null hypothesis predicts. In other words, the test does not point towards the alternative hypothesis; the null hypothesis is better supported by the data. This was the case with the assessment of Mendel’s model.

But if the two are not consistent with each other, as is the case in our example about Alameda County jury panels, then the data do not support the null hypothesis. That is why we concluded that the jury panels were not selected at random. Something other than chance affected their composition.

If the data do not support the null hypothesis, we say that the test *rejects* the null hypothesis.

**The Meaning of “Consistent”**

In the example about Alameda County juries, it was apparent that our observed test statistic was far from what was predicted by the null hypothesis. In the example about pea flowers, it is just as clear that the observed statistic is consistent with the distribution that the null predicts. So in both of the examples, it is clear which hypothesis to choose.

But sometimes the decision is not so clear. Whether the observed test statistic is consistent with its predicted distribution under the null hypothesis is a matter of judgment. We recommend that you provide your judgment along with the value of the test statistic and a graph of its predicted distribution under the null. That will allow your reader to make his or her own judgment about whether the two are consistent.

Here is an example where the decision requires judgment.

**The GSI’s Defense**

A Berkeley Statistics class of about 350 students was divided into 12 discussion sections led by Graduate Student Instructors (GSIs). After the midterm, students in Section 3 noticed that their scores were on average lower than the rest of the class.

In such situations, students tend to grumble about the section’s GSI. Surely, they feel, there must have been something wrong with the GSI’s teaching. Or else why would their section have done worse than others?

The GSI, typically more experienced about statistical variation, often has a different perspective: if you simply draw a section of students at random from the whole class, their average score could resemble the score that the students are unhappy about, just by chance.

The GSI’s position is a clearly stated chance model. We can simulate data under this model. Let’s test it out.

**Null Hypothesis.** The average score of the students in Section 3 is like the average score of the same number of students picked at random from the class.

**Alternative Hypothesis.** No, it’s too low.

A natural statistic here is the average of the scores. Low values of the average will make us lean towards the alternative.

Let’s take a look at the data.

The table scores contains the section number and midterm score for each student in the class. The midterm scores were integers in the range 0 through 25; 0 means that the student didn’t take the test.

scores **=** Table**.**read\_table(path\_data **+** 'scores\_by\_section.csv')

scores

| **Section** | **Midterm** |
| --- | --- |
| 1 | 22 |
| 2 | 12 |
| 2 | 23 |
| 2 | 14 |
| 1 | 20 |
| 3 | 25 |
| 4 | 19 |
| 1 | 24 |
| 5 | 8 |
| 6 | 14 |

... (349 rows omitted)

To find the average score in each section, we will use group.

section\_averages **=** scores**.**group('Section', np**.**average)

section\_averages**.**show()

| **Section** | **Midterm average** |
| --- | --- |
| 1 | 15.5938 |
| 2 | 15.125 |
| 3 | 13.6667 |
| 4 | 14.7667 |
| 5 | 17.4545 |
| 6 | 15.0312 |
| 7 | 16.625 |
| 8 | 16.3103 |
| 9 | 14.5667 |
| 10 | 15.2353 |
| 11 | 15.8077 |
| 12 | 15.7333 |

The average score of Section 3 is 13.667, which does look low compared to the other section averages. But is it lower than the average of a section of the same size selected at random from the class?

To answer this, we can select a section at random from the class and find its average. To select a section at random to we need to know how big Section 3 is, which we can by once again using group.

scores**.**group('Section')

| **Section** | **count** |
| --- | --- |
| 1 | 32 |
| 2 | 32 |
| 3 | 27 |
| 4 | 30 |
| 5 | 33 |
| 6 | 32 |
| 7 | 24 |
| 8 | 29 |
| 9 | 30 |
| 10 | 34 |

... (2 rows omitted)

Section 3 had 27 students.

Now we can figure out how to create one simulated value of our test statistic, the random sample average.

First we have to select 27 scores at random without replacement. Since the data are already in a table, we will use the Table method sample.

Remember that by default, sample draws with replacement. The optional argument with\_replacement = False produces a random sample drawn without replacement.

scores\_only **=** scores**.**drop('Section')

random\_sample **=** scores\_only**.**sample(27, with\_replacement**=**False)

random\_sample

| **Midterm** |
| --- |
| 20 |
| 14 |
| 14 |
| 25 |
| 14 |
| 12 |
| 10 |
| 22 |
| 16 |
| 15 |

... (17 rows omitted)

The average of these 27 randomly selected scores is

np**.**average(random\_sample**.**column('Midterm'))

16.814814814814813

That’s the average of 27 randomly selected scores. The cell below collects the code necessary for generating this random average.

Now we can simulate the random sample average by repeating the calculation multple times.

**def** **random\_sample\_average**():

random\_sample **=** scores\_only**.**sample(27, with\_replacement**=**False)

**return** np**.**average(random\_sample**.**column('Midterm'))

sample\_averages **=** make\_array()

repetitions **=** 10000

**for** i **in** np**.**arange(repetitions):

sample\_averages **=** np**.**append(sample\_averages, random\_sample\_average())

Here is the histogram of the simulated averages. It shows the distribution of what the Section 3 average might have been, if Section 3 had been selected at random from the class.

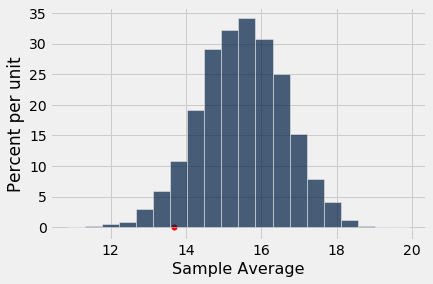
The observed Section 3 average score of 13.667 is shown as a red dot on the horizontal axis. You can ignore the last line of code; it just draws the dot.

averages\_tbl **=** Table()**.**with\_column('Sample Average', sample\_averages)

averages\_tbl**.**hist(bins**=**20)

observed\_statistic **=** 13.667

plots**.**scatter(observed\_statistic, 0, color**=**'red', s**=**30);



As we said earlier, small values of the test statistic will make us lean towards the alternative hypothesis, that the average score in the section is too low for it to look like a random sample from the class.

Is the observed statistic of 13.667 “too low” in relation to this distribution? In other words, is the red far enough out into the left hand tail of the histogram for you to think that it is “too far”?

It’s up to you to decide! Use your judgment. Go ahead – it’s OK to do so.

**Conventional Cut-offs and the P-value**

If you don’t want to make your own judgment, there are conventions that you can follow. These conventions tell us how far out into the tails is considered “too far”.

The conventions are based on the area in the tail, starting at the observed statistic (the red dot) and looking in the direction that makes us lean toward the alternative (the left side, in this example). If the area of the tail is small, the observed statistic is far away from the values most commonly predicted by the null hypothesis.

Remember that in a histogram, area represents percent. To find the area in the tail, we have to find the percent of sample averages that were less than or equal to the average score of Section 3, where the red dot is. The array sample\_averages contains the averages for all 10,000 repetitions of the random sampling, and observed\_statistic is 13.667, the average score of Section 3.

np**.**count\_nonzero(sample\_averages **<=** observed\_statistic) **/** repetitions

0.0564

About 5.7% of the simulated random sample averages were 3.667 or below. If we had drawn the students of Section 3 at random from the whole class, the chance that their average would be 13.667 or lower is about 5.7%.

This chance has an impressive name. It is called the *observed significance level* of the test. That’s a mouthful, and so it is commonly called the *P-value* of the test.

**Definition:** The P-value of a test is the chance, based on the model in the null hypothesis, that the test statistic will be equal to the observed value in the sample or even further in the direction that supports the alternative.\*\*

If a P-value is small, that means the tail beyond the observed statistic is small and so the observed statistic is far away from what the null predicts. This implies that the data support the alternative hypothesis better than they support the null.

How small is “small”? According to the conventions:

* If the P-value is less than 5%, it is considered small and the result is called “statistically significant.”
* If the P-value is even smaller – less than 1% – the result is called “highly statistically significant.”

By this convention, our P-value of 5.7% is not considered small. So we have to conclude that the GSI’s defense holds good – the average score of Section 3 is like those generated by random chance. Formally, the result of the test is not statistically significant.

When you make a conclusion in this way, we recommend that you don’t just say whether or not the result is statistically significant. Along with your conclusion, provide the observed statistic and the P-value as well, so that readers can use their own judgment.

**Historical Note on the Conventions**

The determination of statistical significance, as defined above, has become standard in statistical analyses in all fields of application. When a convention is so universally followed, it is interesting to examine how it arose.

The method of statistical testing – choosing between hypotheses based on data in random samples – was developed by Sir Ronald Fisher in the early 20th century. Sir Ronald might have set the convention for statistical significance somewhat unwittingly, in the following statement in his 1925 book *Statistical Methods for Research Workers*. About the 5% level, he wrote, “It is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not.”

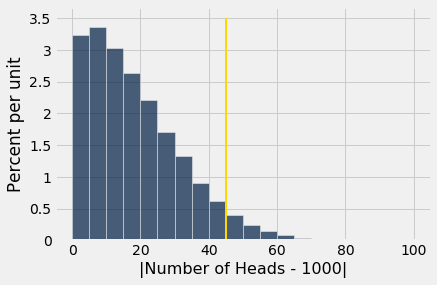
What was “convenient” for Sir Ronald became a cutoff that has acquired the status of a universal constant. No matter that Sir Ronald himself made the point that the value was his personal choice from among many: in an article in 1926, he wrote, “If one in twenty does not seem high enough odds, we may, if we prefer it draw the line at one in fifty (the 2 percent point), or one in a hundred (the 1 percent point). Personally, the author prefers to set a low standard of significance at the 5 percent point …”

Fisher knew that “low” is a matter of judgment and has no unique definition. We suggest that you follow his excellent example. Provide your data, make your judgment, and explain why you made it.

Whether you use a conventional cutoff or your own judgment, it is important to keep the following points in mind.

* Always provide the observed value of the test statistic and the P-value, so that readers can decide whether or not they think the P-value is small.
* Don’t look to defy convention only when the conventionally derived result is not to your liking.
* Even if a test concludes that the data don’t support the chance model in the null hypothesis, it typically doesn’t explain *why* the model doesn’t work.
* n the process by which we decide which of two hypotheses is better supported by our data, the final step involves a judgment about the consistency of the data and the null hypothesis. While this step results in a good decision a vast majority of the time, it can sometimes lead us astray. The reason is chance variation. For example, even when the null hypothesis is true, chance variation might cause the sample to look quite different from what the null hypothesis predicts.
* **Wrong Conclusions**
* If you are testing a null hypothesis against the alternative that the null hypothesis isn’t true, then there are four ways of classifying reality and the result of the test.

|  | **Null is True** | **Alternative is True** |
| --- | --- | --- |
| **Test Favors the Null** | Correct result | Error |
| **Test Favors the Alternative** | Error | Correct result |

* In two out of the four cells of this table, the test result is wrong. One type of error occurs if the test favors the alternative hypothesis when in fact the null hypothesis is true. The other type of error occurs if the test favors the null hypothesis when in fact the alternative hypothesis is true.
* Since the null hypothesis is a completely specified chance model, we can estimate the chance of the first type of error. The answer turns out to be essentially the cutoff that we use for the P-value. Let’s see how.
* **The Chance of an Error**
* Suppose you want to test whether a coin is fair or not. Then the hypotheses are:
* **Null:** The coin is fair. That is, the results are like draws made at random with replacement from *Heads, Tails*.
* **Alternative:** The coin is not fair.
* Suppose you are going to test this hypothesis based on 2000 tosses of the coin. You would expect a fair coin to land heads 1000 times out of 2000, so a reasonable test statistic to use is
* test statistic = ∣∣ number of heads−1000 ∣∣test statistic = | number of heads−1000 |
* Small values of this statistic favor the null hypothesis, and large values favor the alternative.
* We have simulated this statistic under the null hypothesis many times, and drawn its empirical distribution.
* 
* The area to the right of 45 (where the gold line is) is about 5%. Large values of the test statistic favor the alternative. So if the test statistic comes out to be 45 or more, the test will conclude that the coin is unfair.
* However, as the figure shows, a fair coin can produce test statistics with values 45 or more. In fact it does so with chance about 5%.
* So *if the coin is fair* and our test uses a 5% cutoff for deciding whether it is fair or not, then there is about a 5% chance that the test will wrongly conclude that the coin is unfair.
* **The Cutoff for the P-value is an Error Probability**
* The example above is a special case of a general fact:
* **If you use a**pp**% cutoff for the P-value, and the null hypothesis happens to be true, then there is about a**pp**% chance that your test will conclude that the alternative is true.**
* The 1% cutoff is therefore more conservative than 5%. There is less chance of concluding “alternative” if the null happens to be true. For this reason, randomized controlled trials of medical treatments usually use 1% as the cutoff for deciding between the following two hypotheses:
* **Null Hypothesis.** The treatment has no effect; observed differences between the outcomes of the treatment and control groups of patients are due to randomization.
* **Alternative Hypothesis.** The treatment has an effect.
* The idea is to control the chance of concluding the treatment does something if in fact it does nothing. This reduces the risk of giving patients a useless treatment.
* Still, even if you set the cutoff to be as low as 1%, and the treatment does nothing, there is about a 1% chance of concluding that the treatment does something. This is due to chance variation. There is a small chance that data from random samples end up leading you astray.
* **Data Snooping and P-Hacking**
* The discussion above implies that if each of 100 different research groups runs a separate randomized controlled experiment about the effect of a treatment that in fact has no effect, and each experiment uses a 1% cutoff for the P-value, then by chance variation, one of the experiments is expected to wrongly conclude that the treatment does have an effect.
* Unfortunately, that could be the one that gets published. This is why it is important that experiments be *replicated*. That is, other researchers ought to be able to carry out the experiment and see if they get similar results.
* It is not uncommon for researchers to test multiple hypotheses using the same data. For example, in a randomized controlled trial about the effect of a drug, researchers might test whether the drug has an effect on various different diseases.
* Now suppose the drug has no effect on anything. Just by chance variation, a small percent of the tests might conclude that it does have an effect. So, when you read a study that uses tests of hypotheses and concludes that a treatment has an effect, always ask how many different effects were tested before the researchers found the one that was reported.
* If the researchers ran multiple different tests before finding one that gave a “highly statistically significant” result, use the result with caution. The study could be marred by *data snooping*, which essentially means torturing the data into making a false confession. This is sometimes also called *p-hacking*.
* In such a situation, one way to validate the reported result is by replicating the experiment and testing for that particular effect alone. If it comes out significant again, that will validate the original conclusion.
* **Technical Note: The Other Kind of Error**
* There is, of course, another kind of error: concluding that the treatment does nothing when in fact it does something. Discussions of that error are outside the scope of this course. Just be aware that life isn’t easy: if you set up your test to reduce one of the two errors, you almost always increase the other one.