Making Inferences In Light of Uncertainty

Unit 7

Mathematical Concepts

- People often make claims about processes that involve chance, such as: A change in a manufacturing process leads to fewer defective parts; a change in a measurement process leads to more precise measures; a change in levels of carbon dioxide leads to global climate change. If chance is involved, there is always sample-to-sample variability. Inference about the validity of claims must be made in light of this random variability.
- To get a grip on the variability due to chance, a model is created to represent the contribution of chance to an observed value of the process.
- By running the model repeatedly, the sampling distribution of a model statistic, such as the number of defective batteries in batches of 500, can be approximated. The sampling distribution represents how the model statistic varies from sample to sample.
- The value of a statistic from a real sample, such as number of defective batteries in a real batch of 500, is compared to its model-based sampling distribution. Statistical inference asks: How likely is the value of the statistic in the real-world sample to occur just by chance? If very unlikely, we find claims about change in the process generating the distribution more credible.

Unit Overview

Building on concepts of sampling distribution and modeling introduced in the previous unit, students develop models, create model-based sampling distributions, and then make inferences by comparing the values of sample statistics to model-based sampling distributions. Investigations include inferences about claims made about (a) changes in person or improvements in methods for repeated measure of a person's arm-span, and (b) the reality of the power of illusion in a psychophysics experiment that students conduct. Extensions and formative assessments provide further opportunities for exploring model-based inference.

Measurement: Same Person? Improvement in Method? (Day 1)
A model of a sample of measurements of the arm-span of a teacher is used to generate a sampling distribution of the best guess of the real length (the median) of the teacher's arm-span and of the precision of measurement (the IQR). These model-based sampling distributions of the simulated medians and the simulated precisions are used to evaluate two possibilities about new data produced by a different class. The first is that the new

Contents

measurements are those of a different person, not the same teacher. This inference is made by comparing the sampling distribution of the model-based medians with the real-world sample median. The second is a claim that the new sample measurements reflect a more precise way of measuring arm-span, a claim that is evaluated by comparing the new sample's IQR to the sampling distribution of model-based IQR.

The Power of Illusion: Can it be Overcome? (Days 2&3)
Students participate in a psychological experiment. First, they look at a line drawn on paper and estimate its exact center (or they use a web-based program to mark the center, as described). Students make a model of their judgments based on the class data, featuring the now-familiar decomposition of variability into signal (the estimate of the midpoint of the line) and noise (the variability of individuals' estimates).

Students then estimate the center of another line drawn on paper, but one where arrowheads (< or >) are drawn at the endpoints. The arrows usually result in an illusion: The line appears shorter or longer, depending on how the arrowheads are positioned, so estimates of the center of the line, which has not changed in length, are usually biased, either to a lower or to a higher value. The illusion affects the accuracy of the estimate of the center-point. Students use TinkerPlots to model the outcomes of the experiment as a combination of signal (the true center), bias (the distortion of the illusion) and individual differences-chance error. The experiment is available at the modelingdata.org website and is more easily run from there.

The Home Run King? (Days 4&5, optional extension)
Who was a better hitter of home runs, Babe Ruth or Henry Aaron? Students build a chance model, based on resampling, of the differences between the number of home runs hit each season by Ruth and by Aaron, then use the model to evaluate a claim that one hitter was better than the other.

Formative Assessment (Days 6&7 or Days 4&5 if the optional extension is omitted)
Birth Rates invites students to first consider whether or not a birth rate of 43% boys in two different samples is reasonable given a population rate of 51%. Students then generate a model of the sampling distributions of birth rates for each sample and infer whether or not the value in either sample is unusual. (CoS 4a-d; MoV 1 3ab, 5; InI 3a 7a). For *FamilyFarms*, students use TinkerPlots to evaluate the goodness of a model of pumpkin growth and then make

an inference about the effects of an organic fertilizer, in light of the model (CoS 3d MoV 4c, 5 InI 3b 6a 7a).

Materials & Preparation — Making Inferences in Light of Uncertainty Unit 7

Read

- ☐ **Unit 7**
  Start by reading the unit to learn the content and become familiar with the activities.
- ☐ **Informal Inference** (InI) **construct map**
  Read the InI construct map and/or visit the web site (datamodeling.org) to help you discern milestones in student thinking about making inferences in light of uncertainty,

Gather

For the class

- ☐ Student worksheets
- ☐ TinkerPlots

Statistical inference is based on six big ideas. These are:

1. There is good reason to believe that the *variability in a process arises by chance*. For example, although manufacturers try very hard to reproduce objects exactly, there is always some variability that arises from random fluctuations in the production process. In nature, genetic recombination and differential access to resources results in variability among organisms of the same species. As we have found, all measurements are subject to random influences on their precision.

2. A *sample* of the random process can be obtained. For example, a sample of batteries represents the very large number of batteries produced by a manufacturing process, a sample of organisms represents a population produced by processes of reproduction and growth, and a sample of repeated measurements of an object represents a process of measurement.

3. A *statistic* measures something of interest about the distribution of values in a sample of a random process, often its center or variability. For example, the average life span of the batteries in a sample, or the precision of measure in a sample of repeated measures, or the average height of a sample of people.

4. A *model* of the random and nonrandom components of the process can be constructed, or a distribution describing the random components can be inferred by knowing something about the process. We focus on constructing a model.

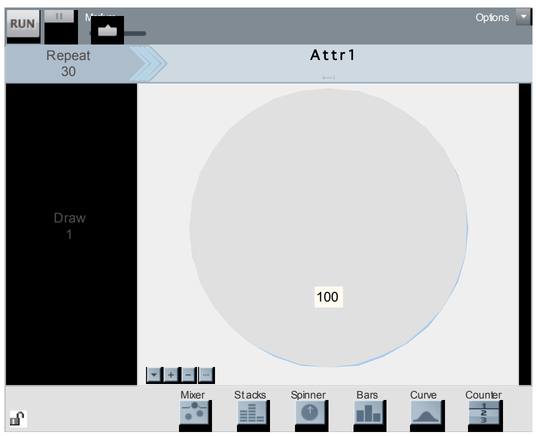
5. Assuming an adequate model of the process, the model is run repeatedly to generate the *sampling distribution* of the statistics of interest.

6. A real-world sample is collected, and the values of its statistics are compared to the values obtained for those statistics in the sampling distribution. This comparison allows for an *inference* about the likelihood of the real-world value of the statistic arising just by the chance process represented by the model.

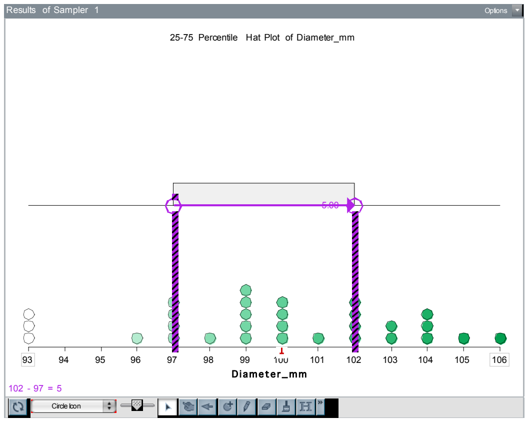
Here is an example of how these big ideas might play out to help someone make a decision. This example follows from the model constructed in Unit 6.

The Acme Cookie Company bakes circular cookies with a diameter of 100 mm. Customers can see the cookies before they buy them in a display case. People feel cheated when someone gets a bigger cookie at the same price, so Acme wants all the cookies to be exactly the same size. They form and bake the cookies in batches of 30.

Using the TinkerPlots sampler (See Acme.tp, datamodeling.org), we create a model where every batch of cookies is exactly 100 mm. in diameter. Notice that this model contains only the target value. There is no chance at work. What do you predict the median value will be when this model is run 30 times? Why?

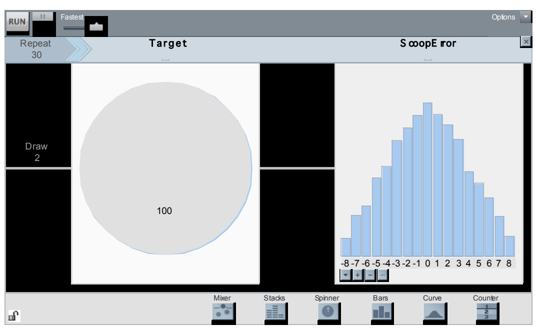


Unfortunately, perfection is not typical of the Acme Cookie Company. During production, sometimes the cookie dough sticks to the automatic scoop used to make each cookie, but with no pattern. Sometimes, the result is a bit more cookie dough, sometimes a bit less, and sometimes, the scoop has exactly the right amount. Large over- or under-fills of scoop are less common. A *sample* of a batch of 30 cookies confirms that the cookies are not very consistent. On median, diameter is 100 mm but the IQR (the inter-quartile range) is 5 mm.

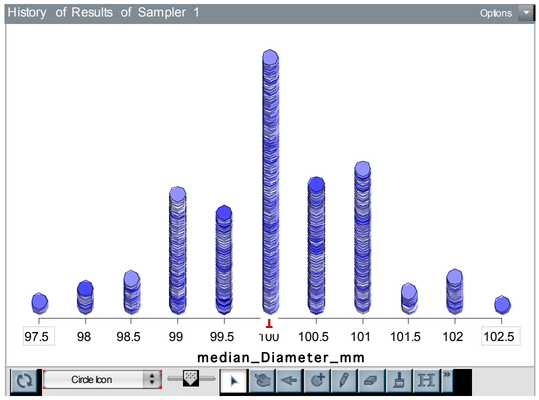


Based on this sample and a few others like it, the company builds a *model* of the process that includes the chance that the scoop will slightly over-fill or under-fill with dough. The result measured is the diameter of the cookie.

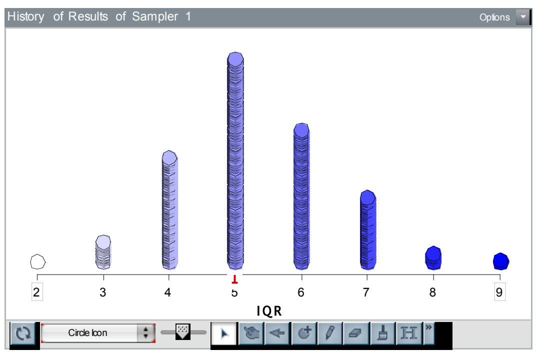
Here is their model:



To see if the model is a good account of the data that they have, they run it again and again. Repeated model runs produce an approximation to the sampling distribution of the statistics of interest: average (median) diameter of each sample of cookies and the IQR of each sample. They notice that the model median diameter and the model IQR (consistency) tend to cluster around the values that they have seen in their samples of cookies. So, they conclude that the model is a good account of the data.



Model Simulation of the Cookie Making Process: Average Diameter

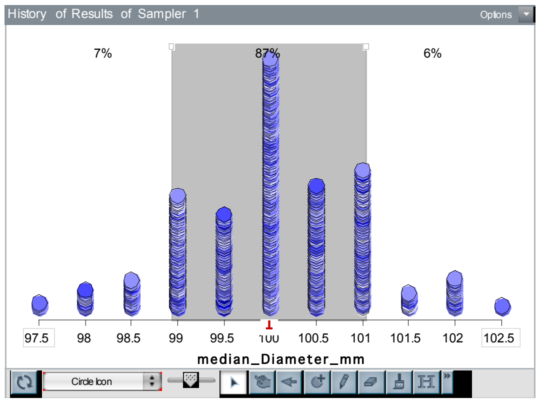


Model Simulation of the Consistency of the Size of the Cookies from Batch to Batch: IQR

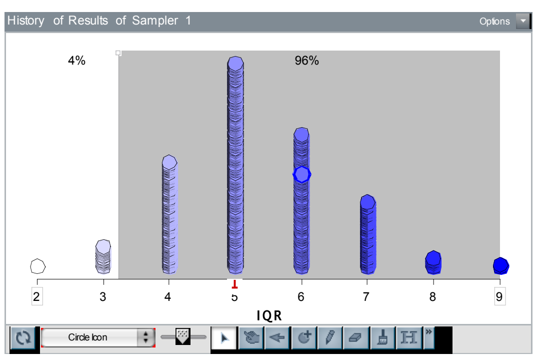
The model and the data on which it is based indicate that sharp-eyed customers may notice the differences in the size of the cookies! Once a company loses its reputation, it can take years before customers will trust it again. Acme will not rest until production can be made more consistent.

Company engineers propose coating the automatic scoop with water before dipping it into the batter. They say that will help make the process more consistent. Acme Company tries this new water-process method with a <u>real</u> sample of 30 cookies, and they find that the median diameter is 99 and the IQR is 3. Have the engineers really solved the problem?

Although the new cookie batch median of 99 is not equal to the target value of 100, looking at the sampling distribution of the medians, one would expect this result just by chance. Notice that about 87% of the median values of the diameters of the cookies in all the simulated batches of cookies fall between 99 and 101 mm., based on the model of the old production process. So far, the new process is like the old process, because the scoop size is what determines diameter of the cookie.



Now, let's look at the new, improved water-process value for consistency. The IQR of the sample of cookies produced by the water-process method is 3. If we look at the sampling distribution of the IQR's, a value of 3 or less occurs just by chance about 4% of the time. So, we are likely safe to conclude that the engineers got it right! The water-process method really helps Acme make more consistent cookies. This means that the diameter of the cookies produced will tend to cluster more tightly around the target diameter of 100 mm. More consistency = greater customer satisfaction.



But, even a strong likelihood is no guarantee of certainty. We are not saying that the engineers <u>must</u> be right. There is about a 4% chance that, in fact, they just got lucky. Perhaps, the water-process method is really no better than the old method. However, the magnitude of the difference in the precision of the first and second samples favors the inference that the water-process method is better.

To explore this model of statistical inference, see AcmeModel.tp, on the modelingdata.org website, Unit 7.

Making Decisions about New Measurements

In this activity, students use a TinkerPlots file, LucasArmSpan(u7).tp, that contains a display of the measurements of a teacher's arm-span produced by members of her class. (The file is found on modelingdata.org, Unit 7). Students first revisit what the statistics of median and IQR measure about the distribution of this batch of data. The median represents the best guess of the real length of the teacher's arm-span, and the IQR represents the precision of the measurements, the tendency of the measurements to agree. Students next examine a model of the measurement process developed by the class, again using TinkerPlots (MLArmSpanModel.tp). The students examine the model and during whole-class discussion, the nonrandom and random components of the model are described. The file also contains a sample of one run of the model, with the sample median and sample IQR displayed. The sampling distributions of the sample medians and of the sample IQR's are also displayed, and students describe what they notice about each sampling distribution. After discussion of each sampling distribution, students are in a position to make statistical inference. Using the Divider tool of TP, students compare the values of the median and IQR from a second real sample of measurements to answer two questions. The first addresses whether or not the second median value of 159 cm. came from the measurements of the first teacher, or perhaps someone else. The second addresses whether or not the measurements in the second real sample reflect a better method of measurement, or is the second sample's IQR of 10 just due to chance?

Note. This first activity is a review of the process of model building and model fit developed in Unit 6. After deciding that the model fits, its sampling distribution is used to guide inference.

Whole Group

1. **Introduce the activity: making decisions**

   a. Ask students to open LucasArmSpan(u7) and mention that the file contains the measurements of a teacher's arm-span by her students.

   b. Ask questions like the ones below to remind students about the meaning of the statistics in the data display.

Q: What does the median measure? (The best guess of the real length of the teacher's arm-span)
Q: What does the distance measured by the ruler represent? (The IQR).
Q: What does the IQR measure? (The tendency for the measurements to agree, an indicator of the precision of the measurements.)

2. Ask students to open MLArmspanModel.tp

a. Ask students to look at the model displayed in the file and to take a moment to think about the components of the model and how it works.

b. Ask questions, like the ones below, to ensure that all students understand that the model combines nonrandom and random components.

Q: What are the meanings of the different spinners? How about Mdn (the best guess, sample median, of the real length)?
Q: How about Gap? Why are the values negative (underestimates)? What kinds of values are most likely?
Q: How about Lap? Why are the values positive? (over-estimates)
Q: How about Slack? (Errors that result when the teacher's arm droops as she gets tired.)
Q: How about CalcE? (Errors students make when they add or subtract measurement values as they measure.)
Q: Which parts model chance? Which do not?

c. Ask questions like the ones below to ensure that students understand that the "Results of Sampler 1" show a simulation of a sample of 30 measurements. The median is the best guess in that sample and the IQR is the estimate of the precision of measure in that sample.

Q: What does each value in "Results of Sampler 1" represent? How was it made? (by summing the outcomes of each part of the model: Mdn + Gap + Lap + Slack + CalcE)
Q: What does the Mdn measure? The IQR?

d. Ask questions like the ones below to ensure that students understand that SimIQR and SimMdn displays are the sampling distributions of the median and of the IQR, based on the model developed by the class.

Q: Look at SimMdn. What does each value represent? (The median of a simulated sample of 30). About how many sample medians are there? (200) What do you notice about them?
Q: Look at SimIQR. What does each value represent? (The IQR of a simulated sample of 30). What do you notice about them?

Whole Group

3. **Ask students to work in pairs to consider two claims about a new sample of real measurements made by another class.**

   a. The first claim: In the new sample, the median is 159 cm. Henry claims that it is different than the 157 cm. of the first sample, so it must be of a different teacher. Melissa disagrees. She says that the median of the second sample is of the same teacher, because it is just the kind of difference in a sample's median that you could get, just by chance.

   Q: If you use the divider tool with 2 divisions and turn on percent, how likely is a median of 159 or more? (24% or about 1 out of every 4 samples)

   b. The second claim: In the new sample, the IQR is 10. George claims that this is very unlikely just by chance, so the second class had a better method of measurement. Their way is more precise, because than the IQR in the first sample was 14.5. Amani disagrees. She says that the IQR of 10 could happen just by chance, so the second class did not really have a better method of measurement.

   Q: If you use the divider tool, Split Dividers with 2 divisions and turn on percent, how likely is an IQR of 10 or less, according to the model? (5% or about 1 out of every 20 samples)

Whole Group

4. **Ask students to share their conclusions and rationale for each claim.**

a. Help students understand that statistical inference is never certain. We can only make statements about how likely something is or is not. Statistics cannot be used to prove anything. All we can do is make a model of a chance process and then make decisions. Conventionally, we establish a threshold and say that if we can expect the value of a statistic to happen just by chance about 5% or less of the time, then we tend to accept this level of uncertainty and make the inference that values of the statistic in this region may very well have resulted from a different process. To avoid mistakes, we can raise the threshold to 1% or even .01%, but if we do, we tend to mistakenly consign to chance that which is not. However, we can never be entirely certain!

Q: Let's say that we accept the second class's conclusion that they had a better method for measuring. After all, their class IQR was 10. Could we be wrong? How likely is it that we are wrong? How can the sampling distribution help us know how likely it is that we are wrong?

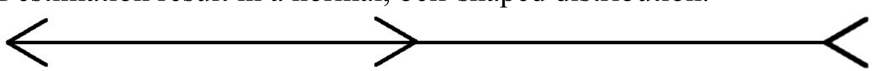
Student Thinking: Uncertain inference

Students often treat anything outside of the center clump as indicating a real difference. So, if the value of a sample statistic is not in the mid-50 percent of the sampling distribution, many students will say that it is unlikely to arise just by chance. Conventionally, we are much more cautious, and typically do not accept a difference as due to anything other than chance unless the value of the statistic is in the lower or upper 5% or less, of the distribution expected by chance alone. We do not believe that it is important at this time for students to accept convention, but rather that they understand that when chance is involved, decisions have to made in light of uncertainty. It is very productive for students to think about a neighborhood of potential values of a statistic that could arise just by chance.

The Power of Illusion

In this activity, students estimate and mark the center of a line segment, such as the one displayed below. People are usually good at this, and although there is chance error in estimation, under-estimates are as likely as over-estimates. There are many sources of chance error, resulting in "normally" distributed individual difference variation.

Students perform the same task with a congruent line segment that has a "closed" arrow at its left endpoint (the line segment terminates in an arrowhead) and an "open" arrow (angled segments) that goes beyond the right endpoint. The task is to move a closed arrow from left to right until its head reaches the center. The closed arrows result in an illusion that the distance is shorter than it really is, and so people tend to overcompensate, thus overshooting the mid-point of the line segment. This results in overestimates of center (shifts to the right). There is some debate about whether or not the variability of estimates around the shifted center tends to increase compared to the unmarked condition. But the sum of the errors of estimation result in a normal, bell-shaped distribution.



The third condition of this experiment is to mark the center of another congruent line segment that has an open arrow at its left endpoint and a closed arrow at its right endpoint. The task is to move an open arrow from the left to right until its head reaches the center. The open arrows create an illusion of longer length, so people tend to under-compensate for this illusion, resulting in undershooting the true center. Compared to the previous condition, there is a shift in center, but in the opposite direction (to the left). This activity can be conducted either with paper-and-pencil or on-line at modelingdata.org.

Paper and pencil version: students estimate and mark the center of three line segments (same line length, different configurations of arrows) using arrowheads printed on transparencies.

Note: If you prefer not to conduct the experiment with your class, results from a sample of 82 sixth grade students is located in eyeillusion.tp, and a model of the Normal condition is in EyeNormalModel.tp on the modelingdata.org website.

Whole group

1. **Introduce the experiment.**

   Today, we are going to try out an experiment to see if we can overcome the power of a visual illusion. In each condition of the experiment, you will try to mark the exact center of a line without using any tools, like rulers, or anything other than your eyes.

Individual

2. **Conduct the experiment.**

   a. For the line with no arrows: Look at the line and estimate its middle. Mark that point. Use just your eyes. Don't fold the paper or use a ruler.
   b. For the line with the closed arrows: Look at the line and move the arrow from the left side of the segment until the arrow tip is at the middle of the line. Mark that point. Just use your eyes. Don't fold the paper or use a ruler.

   c. For the line with open arrows: Look at the line and move the arrow from the left side of the segment until the arrow tip is at the middle of the line. Mark that point. Just use your eyes. Don't fold the paper or use a ruler.

      Note: If you like, you can vary the order randomly for each student and discuss with the class why varying the order might make the experiment better. Or, you can ask how many different orders are possible and then have each student use one of these different orders.

3. Measure the distance from the left end of the segment to the estimate.

Exchange your paper with a neighbor. Record the number of mm from the left end of the line to the center mark for each line. Write this down. This is the distance from the left end of the line segment to its center mark.

Whole Group

4. Create a TinkerPlots file of the estimates.

Be sure that students think about the data structure that will be necessary to distinguish among the conditions of the experiment.

Q: How should we record the data in TinkerPlots?
Q: How many attributes do we need? What should they be called?

Note: Two attributes are most economical. The first, Estimate, records the length in mm from the left end of the line to the center mark for each line. The second, Condition, indicates whether the line segment was normal, open-arrow, or closed-arrow. Your class might decide upon other names. An optional third attribute might be Error, representing the difference between the true center and the estimated center.

5. Look at the data, focusing on the Normal condition.

Ask students which statistics might be worth knowing and to interpret their meaning.

Q: How can we display the data so that we can see the estimates of the center of the line for each condition separately?
Q: What statistic will be the best for estimating the real center of the line? Why do you think so? Will it still be the best if I drag this person's estimate way off to here (drag to create extreme value)? Why do you think so?
Q: What statistic will be best for estimating how much our estimates tended to agree? Why do you think so? Will it still be the best if I drag this person's estimate way off to here (drag to create extreme value)? Why do you think so?

Pairs

6. Create a model that accounts for the data in the Normal condition.

Students create a model of the data estimates in the Normal condition. Encourage students to select statistics and develop sampling distributions of those statistics.

Q: Is the model you have developed a good one? Why do you think so?

Whole group

7. Use the sampling distribution of the Normal condition to make inference.

Ask students to consider: In light of the sampling distribution of the median for the model of the Normal condition, was your class able to overcome the illusion in either of the other two conditions? Why do you think so? In light of the sampling distribution of the IQR or average deviation, did the variability of people's estimates of center increase with the Arrow conditions or was it about the same?

Pairs

8. Revise the model for the Normal condition to account for one of the other conditions.

Ask students to revise the model for the Normal condition to account for the estimates in either the Open-Arrow or Closed-Arrow conditions.

Note: One way to revise the model is to add a constant, nonrandom term to it. This will affect the center, but the not variability.

The Home Run King (Optional, Advanced Topic)

This activity extends inference via a sampling distribution, not by developing a model of a random process, but by treating a sample of data as a population and then randomly sampling, with replacement, its values to create new, random samples. This is called bootstrapping. Here it is applied to a debate about baseball: Should Henry Aaron or should Babe Ruth be considered the best home run hitter in the history of the game? In this activity, the sample consists of the number of home runs that Ruth hit for every one of the 15 years that he played, and the number of home runs that Aaron hit for every one of the 21 years that he played. Neither player hit the same number of home runs every year.

The variability from year-to-year in the number of home runs hit by any batter is likely due to chance. For example, sometimes a ball hit with the same force will travel out of the ballpark with a wind behind it but will stay in the park when the wind is blowing in. Some years may be a bit windier than others, resulting in random differences in the total number of home runs hit each year by the same batter. We don't know all the sources of chance in baseball, but we believe that there are many of them. Not all of us can play baseball, and even among professional baseball players, some are clearly more skilled than others. So, hitting baseballs is clearly more than just due to chance. On average, Ruth hit more home runs than Aaron, but is the difference just about what one would expect just by chance? Or, was Ruth more skilled?

To find out, we put all the home runs each player hit for every year that they played into a mixer (Ruth's 15 years and Aaron's 21 years). But, we pretend that we don't know which player goes with each year's total home runs. If the difference between them is just due to chance, then it does not matter. It is equally likely that it is Ruth or it is Aaron.

Because Ruth played 15 years, we randomly draw a year's total 15 times. Each time, we throw the total back into the mixer, so it always has a 1/36 chance of being drawn. Because Aaron played for 21 years, we randomly draw a year's total 21 times, also throwing the total back into the mixer (Ruth's 15 years + Aaron's 21 years = 36 years). Then, for each random sample of 15, we find the average and say that it is Ruth's. And for each random sample of 21, we find the average and say that it is Aaron's. We find the difference between the averages, and because we chose at random, the difference is just due to chance or luck. We repeat this process a large number of times to get the sampling distribution of the differences. Then, we compare the real difference between Ruth and Aaron to the values in the sampling distribution to see how likely it is that differences of that size occurred just by chance. If the likelihood is low, then we might decide that the difference is due to skill of the players and/or to other nonrandom factors.

Whole Group

1. Noticing the difference in number of home runs hit every year by Ruth and by Aaron.

a. Ask students to open homeruns.tp (download from modelingdata.org) and inform them that that it contains the number of home runs hit by two different baseball players: Babe Ruth and Henry Aaron.

Q: How should we display the data?

b. Ask students what they notice. The following questions may help:

Q: What statistics might help us compare Ruth and Aaron? Why do you think so? What does the statistic measure?
Q: Why might the number of home runs hit by each player not be the same every year? What might chance have to do with it?
Q: What random things could be why a player hits different numbers of home runs from year-to-year? What are some nonrandom things that might influence home run totals?

2. Modeling differences between Ruth and Aaron as due to chance.

a. Ask students to inspect the model in homerunmodel.tp. Help students interpret the model.

Q: What is in the first mixer? (home runs for both players)
Q: What does the second, Year, counter do? (It runs a counter from 1 to 36, sending the first 15 random draws to the mixer called Ruth and the remaining 21 random draws to the mixer called Aaron)
Q: What is in the mixer called Ruth? Aaron?

b. Help students interpret the display of the Sampler.

Q: What does the display show? (the results of 36 repetitions)

Q: Why is the median shown? (average number of homeruns)
Q: What does the 10 mean? (the difference between Ruth Mdn and Aaron Mdn, Ruth-Aaron)
Q: What might a -10 mean? (Ruth hit on average 10 fewer homeruns than Aaron)

c. Help students interpret the sampling distribution.

Q: What does each value on the display Diff_Homers represent? (the difference in medians between the chance sample of Ruth homers and the chance sample of Aaron homers)
Q: Why is the sampling distribution centered at 0?
Q: Why is the sampling distribution symmetric?

3. Making inference.

Help students evaluate the claim that because on average Ruth hit 8 more home runs than Aaron, he was the better home run hitter.

Q: Comparing the medians, Ruth averaged 8 more home runs than Aaron. How likely is this just by chance? Use the sampling distribution to find out.
Q: What other statistics might you look at if you wanted to compare the players? Why might you do so?

Note: At your option, students could explore measures of consistency of home run production via average deviation or IQR. They could also explore other measures of center, such as the mean.

Formative Assessment

The first formative assessment problem considers birth rates in two counties in North Carolina. (Thanks to Professor Hollylynne Lee for suggesting this problem.) The overall birth rate in the state is 51% male (there is usually a slight bias for male births in the world). But in two counties, the birth rate for males was 43% during that year's census. Students consider whether or not this birth rate is unusual for each county, assuming a chance process. One county had a total of 56 births that year and the other had 314 births. After initial consideration, students generate TinkerPlots models for each county and reconsider their initial decision.

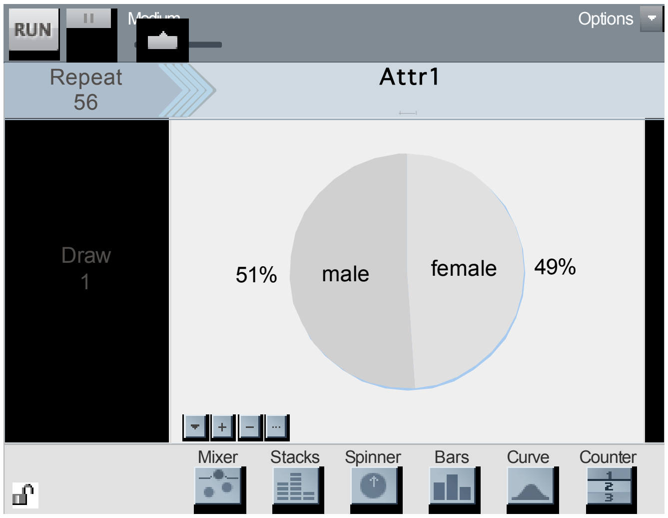
Note: Recall that sample size influences chance variability. The number of births in one county is approximately six times that of the other county. Hence, one would expect less sampling variability for the larger number of births than for the smaller number of births. That would make the percent male births observed in the larger county (43%) less likely to occur just by chance, indicating that perhaps in that county, factors other than chance are at work.

1. **Pose the problem to students** and ask them to decide about whether or not they agree with each statement.

2. **Elicit student reasoning about each statement and solicit justifications.** Keep a running record of the number of students agreeing with each statement and some of the reasons why.

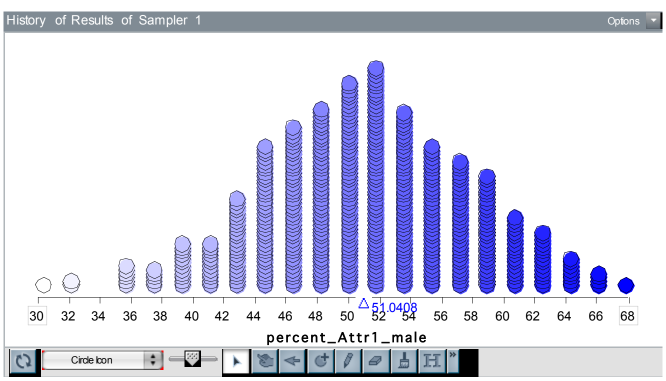
3. **Have students build models with TinkerPlots for each county and run them repeatedly** to generate sampling distributions of the percent male birth, assuming 51% of the births will be male, and that the sex of a baby occurs by chance.

4. **After surveying students' models, select models that reflect different levels of the Modeling Variability (MoV) construct.** Conduct a conversation aimed at helping all students understand the rationale for models such as these:



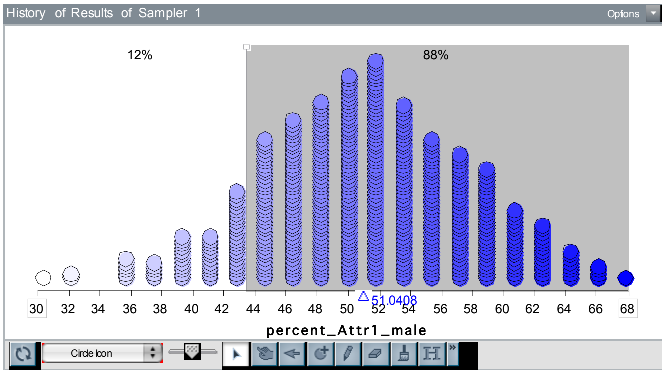
5. Check on student interpretations of the resulting sampling distribution by asking questions such as:

Q: What does each dot represent? (the percent of male births for a sample of 56 births)
Q: Why isn't each dot exactly the same?
Q: What is the mean percent of male births and why might it be that value?



An empirical sampling distribution (n = 500 samples of size 56) for percent male births assuming that the probability of a male birth is 0.51

Q: What percent of the samples had percentages of male births of 43% or less? (This may vary slightly depending on the size of the sampling distribution.)



6. **Have students again indicate which statements they agree with.** This time, students should justify their statements by referring to the models and results of the sampling distributions.

If time permits, the second assessment item, *Family Farms Pumpkins,* poses a question about the effects of fertilizer on pumpkin growth. The item includes a model that students can run. The model fits the growth of pumpkins without fertilizer. Students can generate a sampling distribution of model parameters to guide inference about the potential effects of fertilizer.

To guide the formative assessment conversation, consider:

1. Focus on the nature of the center clump. An IQR of 1 and a median of 18 lbs. suggests that 50% of the data ranges between 17.5 lbs. and 18.5 lbs. if the data are symmetric about 18 lbs., or some other range, such as 17.2 to 18.2. But whatever the exact boundaries, it suggests a tight center clump.

2. Elicit a range of responses to focus on different ways of thinking about model fit. Include examination of the model-based sampling distribution of the IQR and the median, in comparison to the real data sample.

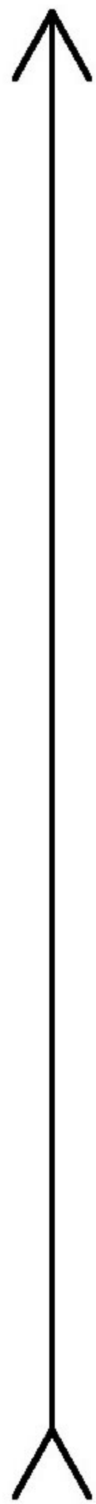
3. Help students use the sampling distribution of the model to evaluate the claim of improved growth with organic fertilizer. Use the TinkerPlots divider and percent tools to consider how likely the new sample median would arise just by chance, if the model of unfertilized pumpkin growth was adequate.

4. Guide students to develop a new model of growth by modifying the model for unfertilized growth. One way is to add a constant term representing the effect of fertilizer as the difference between the fertilized and unfertilized medians.

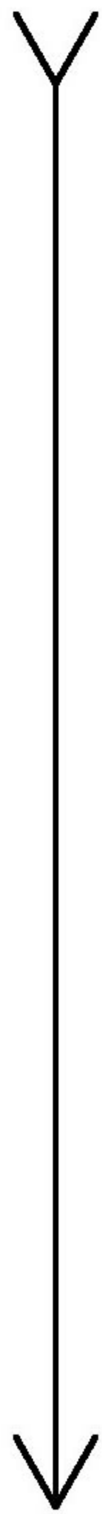
Investigating Inference

1. Mark a middle point by estimation.
2. Measure the distance from the beginning of the line to the middle point you marked

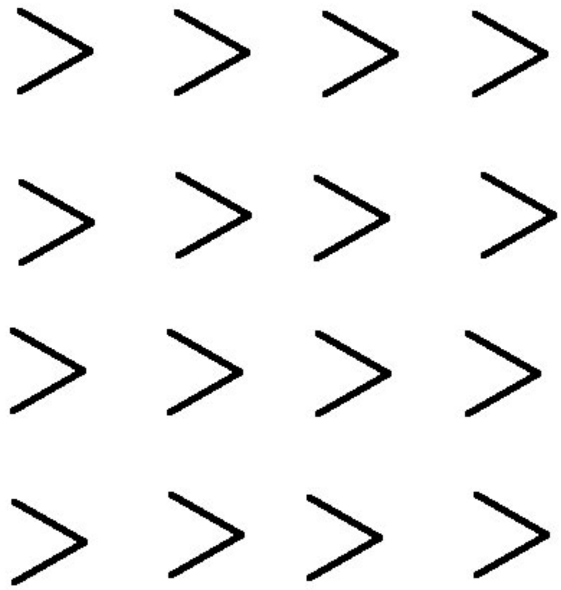
Your measurement: ___________mm



Your measurement: ________mm



Your measurement: ________ mm



Birth Rates

The state of North Carolina records the number of male babies and the number of female babies in every county in the state. Across all counties, 51% of the births are male, but biologists working for the state health department noticed that in two counties, 43% of the births were male. In Hyde County there were 56 births that year, and in Martin County, there were 314 births.

1. Circle the statement that comes closest to what you think is true.

   a. Both Hyde and Martin counties are very unusual, because the birth rate for boys in the state is 51% in the state.

   b. Neither Hyde and Martin counties are very unusual, because 43% could just happen by chance.

   c. The birth rate for Martin county is more unusual than the birth rate for Hyde county, even though it is 43% for each, because there were more births in Martin county.

   d. The birth rate for Hyde county is more unusual than the birth rate for Martin county, even though it is 43% for each, because there were fewer births in Hyde county.

2. Use TinkerPlots to build a model of the birth rate in Hyde County. According to your model, how likely is a birth rate for males of 43% or less?

3. Use TinkerPlots to build a model of the birth rate in Martin County. According to your model, how likely is a birth rate of 43% or less?

4. Considering your models, circle the statement that comes closest to what you think is true.

    a. Both Hyde and Martin counties are very unusual, because the birth rate for boys in the state is 51% in the state.

    b. Neither Hyde and Martin counties are very unusual, because 43% could just happen by chance.

    c. The birth rate for Martin county is more unusual than the birth rate for Hyde county, even though it is 43% for each, because there were more births in Martin county.

    d. The birth rate for Hyde county is more unusual than the birth rate for Martin county, even though it is 43% for each, because there are fewer births in Hyde county.

    Explain your choice:

Unit Quiz — Making Inferences in Light of Uncertainty Unit 7

Family Farms Pumpkins

Family Farms (FF) grows pumpkins and cans the pumpkin for pie makers. It has developed models of the variability of the weights of the pumpkins in its fields. Usually, pumpkins range from 1 to 35 pounds. Collecting samples from many fields (n =100 randomly selected pumpkins from each field), the median pumpkin weight is 18 pounds and the IQR is 1.

1. If the IQR is 1, then 50% of the pumpkins range in weight between about ____ pounds and about ____ pounds.

2. FF has developed a model of the weights of the pumpkins by combining the effects of chance differences in amount of light, moisture, and the weights of pumpkin's parents on the pumpkin's weight after it is done growing. The model is in FamilyFarms.tp. Run the model. Do you think it is a good model? Why or why not?

3. FF tries out an organic fertilizer to increase growth. The median weight of the pumpkins in the test field is 19.5 pounds. Some of the testers are very happy with the results. An increase in a 1.5 pounds is a lot. Other testers say that the result is just due to chance. Organic fertilizer really doesn't do much for pumpkins. It just smells bad. If you assume that the FF model of growth without fertilizer is an adequate model, what would you say about this difference with fertilizer?

4. How would you modify the FF model to account for the effect of the fertilizer, if there is one?

Birth Rates

Question 1: Birth Rates and CoS		
Level	**Performance**	**Example**
CoS 4D Predict and justify changes in a sampling distribution based on changes in properties of a sample.	Chooses 1c.	
CoS 4B Recognize that the sample-to-sample variation in a statistic is due to chance	Chooses 1b.	
CoS 3B A statistic measures a characteristic of a sample distribution.	Chooses 1a or 1d.	

Question 2: Birth Rates and Modeling Variability (MoV)		
Level	**Performance**	**Example**
MoV(5)	Judge model fit in light of variability across repeated samples with same model.	Builds model as in MoV3B, creates sampling distribution statistic of percent male births and creates a sampling distribution of this statistic.
MoV(3b)	Evaluate fit of chance device by appealing to relations between simulated and observed values.	Builds model with spinner divided into regions of 51% (male) and 49% (female). Sets repetitions for 56, the sample size. Runs it a few times. Notices that sometimes the statistic is close to 43%.
MoV(3a-)	Uses chance device to represent variability.	Builds model with spinner divided into regions of 43% (male) and 57% (female). Or does not set the repetitions for 56, the sample size.
MoV(1)	Attribute variability to specific sources or causes.	Says that there is about an equal chance of having a male or a female birth.

Question 3: Birth Rates and Modeling Variability (MoV)		
Level	**Performance**	**Example**
MoV(5)	Judge model fit in light of variability across repeated samples with same model.	Builds model as in MoV3B, collects statistic of percent male births and creates a sampling distribution of this statistic.
MoV(3b)	Evaluate fit of chance device by appealing to relations between simulated and observed values.	Builds model with spinner divided into regions of 51% (male) and 49% (female). Sets repetitions for 314, the sample size.
MoV(3a-)	Uses chance device to represent variability.	Builds model with spinner divided into regions of 43% (male) and 57% (female). Or does not set the repetitions for 314, the sample size.
MoV(1)	Attribute variability to specific sources or causes.	Says that there is about an equal chance of having a male or a female birth.

Question 4: Birth Rates and Informal Inference (InI)		
Level	**Performance**	**Example**
InI(7a)	Sampling-distribution guided inference.	Refers to sampling distributions to suggest that 43% or less occurs in X percent of the samples for Martin county, but in Y percent of the samples for Hyde county. X << Y, so Martin county is less likely if only chance is at work.
InI (6a)	Consider sample in light of sample-to-sample variability.	Notices that there is sample to sample variability about 51% to justify choice, but does not make reference to sampling distribution, even if constructed previously.
InI(3b)	Compare 2 distributions based on specific values.	Judges based on the equal percent of male births in each county.

Family Farms Pumpkins

Question 1: Family Farms Pumpkins and Conceptions of Statistics (CoS)		
Level	**Performance**	**Example**
CoS(3d)	Demonstrate knowledge of relations among components of a statistic.	Says about 17.5 to 18.5 (this need not be symmetric about 18, so a response like 17.2 to 18.2 would be fine as well).

Question 2: Model Fit Family Farms Pumpkins and Modeling Variability (MoV)		
Level	**Performance**	**Example**
MoV(5)	Judge model fit in light of variability across repeated samples with same model.	Runs model repeatedly to develop sampling distribution. Notes correspondence between statistics of sampling distribution and of field data. May critique model because it collapses many sources of variability
MoV(4c)	Compare model outputs to data and judge accuracy.	Runs model a few times. Notes that its median and IQR are about right.

Question 3: Family Farms Pumpkins and Informal Inference (InI)		
Level	**Performance**	**Example**
InI 7b	Sampling-distribution guided inference	Compares fertilized sample median to sampling distribution of model of unfertilized growth, quantifies chance by proportion or percent to argue that the fertilized sample median is unlikely.
InI(6a)	Expect particular regions of distributions to have sample-to-sample variability.	Anticipates that unfertilized sample medians will vary but that the fertilized sample median is more than might be expected. Does not relate to sampling distribution. Or says that because there is sample to sample variability, the fertilized sample median may just be due to chance.
InI(3b)	Compare 2 distributions based on specific values.	The medians are different. So it must not be due to chance.

Question 4: Family Farms Pumpkins and Modeling Variability (MoV)		
Level	**Performance**	**Example**
MoV(5)	Judge model fit in light of variability across repeated simulation with the same model.	Adjusts model by adding a constant term of 1.5 (or a value in a reasonable neighborhood around 1.5) and then creates corresponding sampling distribution.
MoV4(b)	Create model of total variability as composition of chance and constant devices.	Adjusts model by adding a constant term of 1.5 (or a reasonable value in that neighborhood) but does not consider sampling distribution.

