Section 4
Introduction to Statistical Inference

4.1 – Testing With Simulation

Modeling hypotheses

In your activity this week, we examined overtime rules in the NFL and whether the coin flip gave an advantage to the team that got the ball first. One primary focus of this activity was in the _____ that were made in constructing a sampler. We simulated this scenario under the idea that there was no advantage to the team that won the flip and got the ball first.

Why would we make such an assumption? Our motivation in investigating this scenario was because we believed the overtime rules were unfair, yet we made an assumption that they were indeed fair, and that there was no advantage to either team. We did this for two reasons:

- How do you assume that there is an advantage? If we wanted to encode this into our TinkerPlots sampler, a probability model, we would need some way to encode that probability. Does the team winning with the ball first winning 55% of the time reflect an advantage? 70%? This kind of gets back to the exact question we want to answer in the first place, as we observed a win rate of about 56% (240 out of 428), and we want to know if that is evidence of an advantage in the first place.
- We can determine if there is an advantage through ruling out the possibility of no advantage, which easy to assume – this can be modeled by a win probability of 50%. If we can show that our data is either plausible to occur or very unlikely to occur under that assumption, we can rule out the idea of no advantage and conclude that there is one. That doesn't necessarily tell us how strong the advantage is in the long run, but it allows us to make a decision of advantage or no advantage!

The process we followed in that class follows three main steps:

- _____ – In class, we discussed a variety of probability models or samplers in TinkerPlots that modeled this scenario. These models should reflect the key assumptions or hypotheses that you are assuming to be true for the purposes of the simulation. In this case, this assumption was that there is no advantage.
- _____ – Once you have created an appropriate model for the scenario, we use it to simulate what happens by random chance over many, many different simulated sets of data. We can then pool and collect those results together to create a distribution to show the variability of possible results under that assumption.
- _____ – We finally then use this distribution we have created to determine where our actual, real-world results (e.g. 240 out of 428 or 56%) fall within the distribution created by our model and assumptions. Are our real-world data compatible with the assumption? We determine how likely it is to get a result like this or more extreme to measure exactly how compatible this data is with our model in order to measure the support (or lack of evidence) for our original assumption or hypothesis.

This process is a way to conduct _____ – that is, the process of generalizing a larger statistical process or population based on just a sample of data. The inferential technique we conducted with the NFL overtime data is known as _____.

Helper or hinderer?

To explore ideas of inference and hypothesis testing further, we will examine a prominent Psychology study about natural tendencies of infants, published in 2007. Infants were shown a play that showed one toy/shape playing the helping role, while a different toy was given the hindering role. Afterward, infants were given the choice to pick one of the two toys to play with.

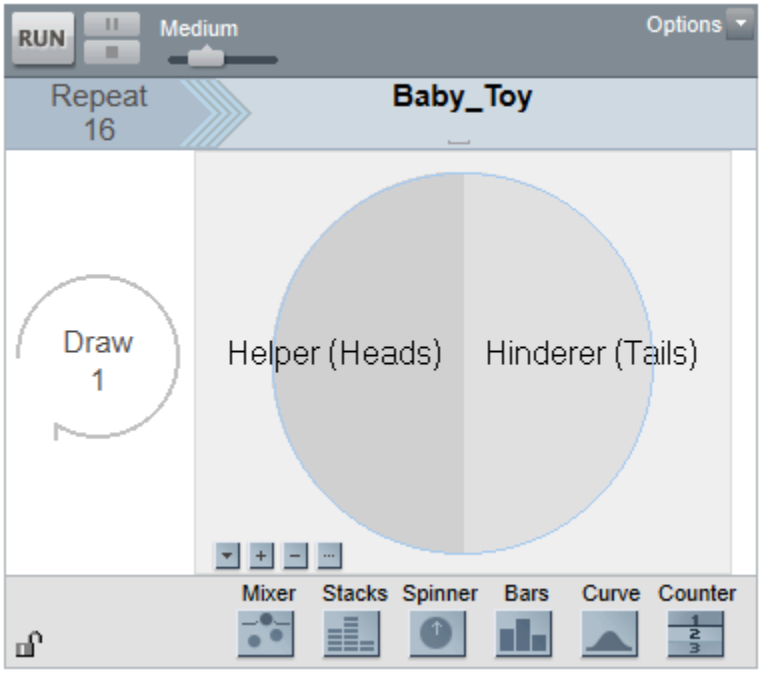
In a pilot run for this study, 16 ten-month old infants were recruited for the study, and 14 of them chose the helper toy. Is this evidence that infants have natural tendencies toward helpful actions, or was it just dumb luck? Maybe these infants didn't internalize anything from the play they were shown, and randomness alone led to 14 of the 16 infants picking the helper toy.

Just like we did with the NFL activity, we can test this idea statistically by asking a probability question: if the infants truly were picking these toys at random and were not influenced by the play, how likely was it for at least 14 of these infants to pick that toy? Using this probability, we can evaluate whether the naïve hypothesis we made that infants pick the toys with equal probability is potentially valid. A larger probability would seem to show that this data could plausibly occur under random chance, where a small probability would indicate that maybe this assumption is not a good one.

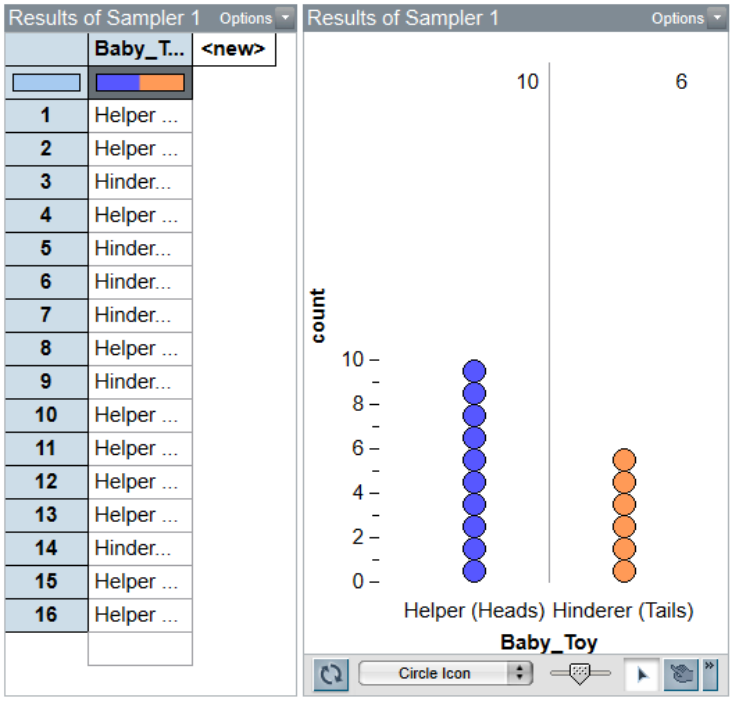
To address this kind of hypothesis test, one could imagine simulating the babies' outcomes using a coin, let's say heads represents a baby picking the helper toy, and tails represents the hinderer toy. This would accurately reflect our assumption of no preference toward either toy. We could imagine flipping 16 coins now to simulate what one repetition of this study might look like under this assumption of no preference. If we do this many times, we can build those results for the number of heads (helper toys chosen by babies) we observed into a large distribution of results. But instead of doing this with a physical coin, let's try doing this in TinkerPlots!

Simulating the coin experiment for the helper or hinderer study

To be able to collect more sets of 16 coin flips to emulate what babies with no toy preference might do, we could use a spinner device like the one shown to the right. When we click run, we will simulate one possible repetition of the study. This reflects our assumption, and represents the _____ step of the hypothesis testing process.

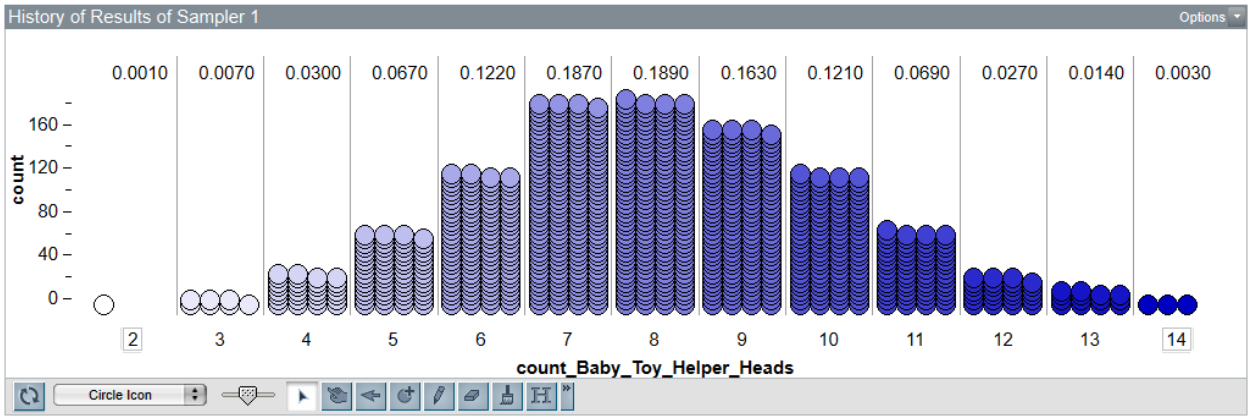


With an appropriate model created, we can now move on to the _____ step of the hypothesis test. We would start this by running our sampler, which would produce one possible



outcome of the helper or hinderer study. The resulting table and plot of one such trial is given to the left. Based on our assumption, we would have expected to see 8 babies pick each toy, but due to this process being subject to random chance, we actually observed 10 babies choose the helper toy and 6 choose the hinderer toy. We can continue to click run to see more possible results, but we really would like to keep track of the number of helper toys each time we do that.

Thus, we complete the simulation process by right clicking on our number of helpers and choosing collect statistics. This creates a new table that will keep track of this statistic you chose every time the sampler is run. We can then run this a large number of times (say, 1000) and create a new plot of these results, which is shown below. Over many trials, we can now see that our original expectation per trial of 8 helper toys is confirmed, with 8 being at the center of the distribution. However, there is quite a bit of variability about that expected value!



Finally, with the simulation complete, we are now at the _____ step of the hypothesis test. To do this, we now want to consider how likely it is to get a result of 14 helper toys chosen or something larger than that. In our simulation of 1000 trials of this study, we only had 3 times where 14 even occurred, a probability of 0.003 or 0.3%. This probability is known as a _____.

> *Example*: If the original study had found only 11 out of the 16 babies chose the helper toy, what would the resulting *p*-value be based on the simulation?

Conducting the TinkerPlots simulation in R

We could also carry out a similar simulation to this in R. One way we could consider this is through taking a sample of values, which we learned how to do a few sections ago with the sample function.

```
sample(c(0,1), 16, replace=TRUE)
```

This code above functions very similarly to our spinner we created – we are sampling from two possible outcomes 16 times, setting with replacement to true. But why represent the outcomes as 0 and 1? If we allow 1 to represent the helper toy and 0 the hinderer toy, then counting the number of helper toys is as simple as taking the sum of all outcomes in the vector:

```
sum(sample(c(0,1), 16, replace=TRUE))
```

This now gives us a code to replicate one trial of the helper or hinderer study. But how do we do this many times? Let's start by creating a placeholder vector for us to store 1000 of these values. But rather than write out a really long vector, we can use the replicate function:

```
helpers = rep(0, 1000)
```

This creates a vector of 1000 entries, all 0 for now, but we will replace them shortly! We now need a way to have each entry become one of those sum/sample functions we wrote above. To do this, we use a control structure called a **for loop**.

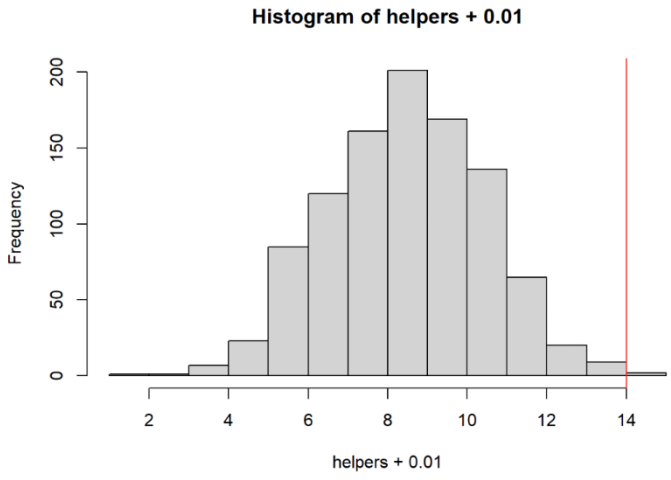
```
for (i in 1:1000) {
  helpers[i] = sum(sample(c(0,1), 16, replace=TRUE))
}
```

The main parts of this for loop are defined by what's in the parentheses and what's in the curly brackets. The parentheses defines what variable you are looping over, and what values that variable will take in each iteration. We want to store a value in each entry of the `helpers` vector eventually, so we want to put something in all entries 1 through 1000. Finally, we write in square brackets what we want to do on each value of our variable `i`. Thus, this for loop is effectively running the following 1000 lines of code:

```
helpers[1] = sum(sample(c(0,1), 16, replace=TRUE))
helpers[2] = sum(sample(c(0,1), 16, replace=TRUE))
...
helpers[1000] = sum(sample(c(0,1), 16, replace=TRUE))
```

The only thing that changes in the square brackets is all of the values of `i`, which are as defined in the parentheses above. Thus, we've achieved our goal, and created a vector of simulated numbers of helpers! We can finally visualize this by creating a histogram.

```
hist(helpers)
abline(v=14, col="red")
mean(helpers >= 14)
```



Using mean here to find a percentage seems a bit counterintuitive – but using the

">=" operator is effectively checking all 1000 differences to see if they are at least 14 or not, and assigning 1 (True) or 0 (False). Since there are only 1's and 0's here, finding the mean is the equivalent as the percentage of 1's present – both are found by adding up 1s and dividing by the sample size. Thus, we have the *p*-value!

R is also much more efficient than TinkerPlots, so running a simulation of 10,000 or even a million repetitions is quite easy! Try adjusting the number of repetitions to something higher yourself, and you'll notice that the shape of the histogram becomes much smoother and more consistent.

4.2 – Formalizing the Hypothesis Test

Defining hypotheses

When conducting hypothesis tests, researchers will often try to formalize their assumptions or hypotheses in terms of parameters. We know that the researchers hypothesized that babies had helpful tendencies, but in statistical testing, we usually write out these hypotheses in terms of a population proportion of babies that would choose the helper toy, so we typically write out our hypothesis in terms of this proportion, *p*.

We don't just write out one hypothesis – we write out two! One hypothesis represents the assumption that we made in carrying out the simulation: that the babies had no preference for either toy. This is referred to as the _____ hypothesis. The other hypothesis reflects the researcher's hypothesis and basis for doing the statistical test. This is the _____ hypothesis. Using appropriate notation, we would write out these hypotheses as follows:

H_0: _____

H_a: _____

Based on these hypothesis and our simulation, we found a *p*-value of 0.3%. Generally, the *p*-value is defined as the probability of getting a result as extreme or more, assuming your null hypothesis is true. This matches what we did, as we assumed the babies had no preference for either toy to be true, and found the probability of getting 14 or more babies to pick the helper toy.

This measure represents the level of evidence that we have against our null hypothesis. The lower the probability, the less plausible it is to get our data if this hypothesis were to be true, giving evidence that the null hypothesis is not a good hypothesis. But the higher the probability, the more plausible it is to get our data under this hypothesis, giving credibility to the null hypothesis. It is often difficult to determine what constitutes "enough" evidence to go against a null hypothesis. Let's try carrying out another example setting up hypothesis and using TinkerPlots to illustrate this idea.

Example: The scenario from your nightmares has come true – you're taking a multiple choice exam but are very unprepared to take the test. There are 50 questions on the exam and each question has four choices. Your score comes back and you get 17 questions correct. While the grade for this exam is not ideal, this is better than 25% of the questions which is better than what you would expect from random guessing. But could you have obtained a score like this even if you were randomly guessing? Write out your hypotheses for this test and use TinkerPlots to find the *p*-value in this scenario. (collecting 250 samples should be enough – this takes some time!)

H_0: _____

H_a: _____

Measuring evidence

The results of this test bring up an important question: what *p*-values constitute enough evidence to go against your null hypothesis? A 10% chance isn't super likely, but it does happen 1 in 10 times, which shows that it is still somewhat plausible to occur. 10% is not a particularly high probability, but I'm sure we've all experienced weather forecasts with a 10% chance of rain where it rains all day.

To make decisions for a hypothesis test, we need to determine what constitutes enough evidence to go against our null hypothesis. What we determine as "enough" can change depending on the context of our test and how important the decision we make is. But as a rough guideline, this is how statisticians typically interpret the strength of their evidence in testing:

***p*-value**	**Interpretation**
> 0.1	No evidence against the null hypothesis
0.05 – 0.1	Some/weak evidence against the null hypothesis
0.01 – 0.05	Moderate evidence against the null hypothesis
0.001 – 0.01	Strong evidence against the null hypothesis
< 0.001	Very strong evidence against the null hypothesis

When conducting a hypothesis test, researchers often set a cutoff value for what constitutes enough evidence for their purposes. This value, denoted α, is referred to as the _____. Based on the *p*-value we get, we would make decisions about our test as follows:

If _____, Reject the null hypothesis (H_0)

If _____, Fail to reject the null hypothesis (H_0)

Typically, α is often set to 0.05 as a "default" value. This is a standard originally set by statistician RA Fisher nearly 100 years ago, and was done so rather arbitrarily. Yet even today, many research journals that accept quantitative work use 0.05 as the default level of significance for tests. This creates a bit of a false dichotomy – 0.051 and 0.049 are very similar levels of evidence for *p*-values, but we would interpret them very differently according to the way we have defined the decision of a test. If we conduct a hypothesis test and get a *p*-value that is very close to our significance level, rather than make a broad decision based on a borderline result, we should try to to replicate the study to see if how our results might differ with a new sample.

Question: Why do we use the terms "reject" and "fail to reject" for hypothesis testing? Why wouldn't we "accept" our null hypothesis based upon the results of a test?

Using R to conduct a hypothesis test

At this point, you might have realized that all of the probabilities that we have computed thus far are just like binomial probabilities we did last section, as they are independent, have two outcomes, and have a fixed sample size/probability. So why did we go through the simulations? To see the probability models that we use to generate them and emphasize that *they assume a null hypothesis is true!* But rather than carry tests out using these simulations every time, we can also just compute a binomial probability. Remember from last section that we can use the `pbinom` function in R to compute these probabilities.

Example: Using the `pbinom` function, find the *p*-value for the helper-hinderer scenario where 14 of the 16 babies used the helper toy.

Another way we can compute these probabilities quickly is using the binom.test function.

```
binom.test(x, n, p=p0, alternative=ALT)
```

The values of `x` and `n` are your count and sample size, and `p0` is the null value. For the alternative (`ALT`), specify one of three options: "`less`", "`greater`", or "`two.sided`" depending on the direction of your test. We'll discuss the two-sided option next section!

Example: Using the `binom.test` function, find the *p*-value for the helper-hinderer scenario where 14 of the 16 babies used the helper toy.

To wrap up this section, let's formally conduct the NFL hypothesis test we did in our last activity, going through all the steps and interpretations done in a hypothesis test.

> *Example*: The National Football League (NFL) uses an overtime period to determine a winner for games that are tied at the end of regulation time. Between 1974 and 2009, the overtime period started with a coin flip that determined which team got the ball first in overtime, and then the team that scored first won the game. Rules were changed after 2009 because fans and players both believed that these rules were unfair for the team that lost the coin flip. That is, they believed that the team that won the coin flip and got the ball first in overtime had an advantage at winning the game. Between 1974 and 2009, there were 428 games that went to overtime and 240 of them were won by the team that won the coin toss. Conduct a hypothesis test to address the following research question: Did the NFL's old overtime rules give the team that won the coin toss an advantage at winning the game?

Write out hypotheses:

H_0: _____

H_a: _____

Conduct the test and evaluate the evidence:

Interpret the *p*-value:

Make a decision for the test:

Draw conclusions:

4.3 – Additional Practice

Example: In Western Countries, only about 12% of the population identifies as "left-handed." While part of hand preference may be environmentally conditioned, genetics may also play a role. One theory posits that red-headed people are more likely to be left-handed! Based on genetic theory, we'd like to see if red-headed people might be more likely to be left-handed than the general population. Let's say that we take a random sample of 125 red-headed people. We found that 40 of them had a preference for their left-hand. Conduct a full hypothesis test for this scenario using a significance level of $\alpha = 0.01$:

Write out hypotheses:

H_0: _____

H_a: _____

Conduct the test and evaluate the evidence:

Interpret the *p*-value:

Make a decision for the test:

Draw conclusions: