*Hypothesis test* is a structured analytical procedure used to evaluate the amount of evidence or lack thereof for a research question. A hypothesis test uses the outcome from the data obtained. The amount of evidence or lack thereof is based on probabilities and is done so in a very consistent way.

Writing the Research Question

Before a hypothesis test can be done, one needs to have a clearly stated a research question or question of interest. For instance, let’s reconsider the inherit questions from some of our previous examples.

|  |  |
| --- | --- |
| **Case Study** | **Research Question** |
| Deafness Malingering | Is there evidence that the individual being evaluated is malingering in regards to their ability to hear? |
| 1st Round Pitchers | Do pitchers taken in the 1st round of the draft have a higher risk of failure than other position players? |
| Ear Infections | Is there enough statistical evidence to say there is a difference in the duration of ear infection between the breast-fed and the bottle-fed babies? |

In general, the research question determines exactly what type of statistical analysis is appropriate. In practice, a clearly stated research question is of the utmost importance.

Parameter and Scope of Inference

In the context of hypothesis testing, the purpose of analyzing data is to answer the research question. The term, *scope of inference,* identifies to whom the conclusions of the study apply. In addition to identifying the people or objects for which the study outcomes are relevant, we need to carefully consider what exactly is being tested by the hypothesis test. The quantity being tested is called a *parameter.*

|  |
| --- |
| Definition |
| Scope of Inference: To whom the study outcomes (i.e. conclusions) apply  Parameter: The parameter is the numerical value being tested in a statistical hypothesis. This value will be represented by a Greek character in this class. |

For each of the case studies under investigation here, identify the appropriate parameter or interest.

|  |  |
| --- | --- |
| **Case Study** | **Parameter** |
| Deafness Malingering | Parameter: = the probability of identifying the correct light when deaf |
| 1st Round Pitchers | Parameter: = |
| Ear Infections | Parameter: = |

Likewise, if possible, identify the scope of inference of each case study.

|  |  |
| --- | --- |
| **Case Study** | **Scope of Inference** |
| Deafness Malingering |  |
| 1st Round Pitchers |  |
| Ear Infections |  |

Forming a Testable Hypothesis

One of the most difficult parts of a statistical analysis is to translate the posed research question into a hypothesis that can be evaluated using probabilistic statements. A statistical hypothesis has two components – a null and alternative hypothesis.

* The null hypothesis, Ho, is what the outcome from your study is being compared **against**. The null hypothesis determines exactly how the reference distribution will be setup.
* The alternative hypothesis, HA, is a restatement of the research question. The alternative hypothesis is a confirmation or endorsement of the research question.

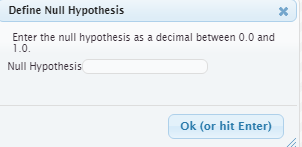
Consider some examples that we have previously discussed. For each the null and alternative hypotheses have been written out in words and with the parameters.

|  |  |  |  |
| --- | --- | --- | --- |
| **Case Study** | **Research Question** | **Hypothesis Statements** | |
| Deafness Malingering | RQ: Is there evidence that the individual being evaluated is malingering in regards to their ability to hear? | Ho: The individual is deaf; the   probability of correct light is 50%.   HA: The individual is malingering in their ability   to hear; the probability of a correct light is   *less than* 50% |  |
| 1st Round Pitchers | RQ: Do pitchers taken in the 1st round of the draft have a higher risk of failure than other position players? | Ho: 1st round pitchers have same failure rate as   other position players; the probability of   failure is 34%   HA: 1st round pitchers have a higher failure rate   than other position players; the probability   of failure is *greater than* 34% |  |
| Ear Infections | RQ: Is there enough statistical evidence to say there is a difference in the duration of ear infection between the breast-fed and the bottle-fed babies? | Ho: There is no difference in duration of   fluid between bottle- and breast-fed   babies; that is, the probability the breast-  fed baby in each pair did better is equal   to 50%.  HA: There is a difference in duration of   fluid between bottle- and breast-fed   babies; that is, the probability the breast-  fed baby in each pair did better is   *different* than 50%. |  |

Obtaining the Reference Distribution

As stated previously, the null hypothesis is used to construct the appropriate reference distribution for a given test. So far, we have constructed these reference distributions in StatKey. Specification of the null hypothesis is straight forward in StatKey.



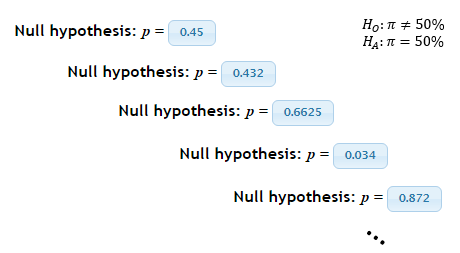


|  |  |  |  |
| --- | --- | --- | --- |
| **Case Study** | **Research Question** | **Null Hypothesis Setup** | |
| Deafness Malingering | RQ: Is there evidence that the individual being evaluated is malingering in regards to their ability to hear? |  |  |
| 1st Round Pitchers | RQ: Do pitchers taken in the 1st round of the draft have a higher risk of failure than other position players? |  |  |
| Ear Infections | RQ: Is there enough statistical evidence to say there is a difference in the duration of ear infection between the breast-fed and the bottle-fed babies? |  |  |

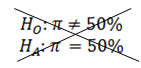
Comment: The research question must be stated in such a way that a spinner can be set up. Some hypothesis are \*not\* testable. For example, suppose for the ear infection example the research question was stated as follows.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Research Question** | **Hypothesis** |  |
| **A** | Is there enough statistical evidence to say there is a **difference** in the duration of ear infection between the breast-fed and the bottle-fed babies? |  | A hypothesis that can be tested |
| **B** | Is there enough statistical evidence to say there is **no difference** in the duration of ear infection between the breast-fed and the bottle-fed babies? |  | This hypothesis CANNOT be tested |

The reason research question B is \*not\* valid is that it is not possible to set up a single null hypothesis for the data to be compared against. There are an infinite number of possible values under the null that can be used to show a difference between breast-fed and bottle-fed. Sketch some example below.



Different parameter values in the null will produce different reference distribution. It is not clear which reference distribution should be used to compare our data against. Thus, it is not possible to statistically test research question B.



For each of the above case studies, 1000 trials of the simulation were run. These 1000 trials represent the reference distribution and are shown here.

|  |  |
| --- | --- |
| Setting up the  Null Hypothesis | Reference Distributions  (under the null hypothesis) |
|  |  |
|  |  |
|  |  |

|  |
| --- |
| Definition |
| Reference Distribution: A graph of the outcomes from many repeated trials. These outcomes are used to measure and evaluate the amount of evidence the data provides for the research question. |

Making a Formal Decision

The determination of what constitutes an outlier in our reference distribution has been somewhat ambiguous up to this point. Our discussion here will eliminate this ambiguity.

R.A. Fisher, one of the founding fathers of modern statistics, says that 1 in 20 is convenient to take as a limit in judging whether a deviation ought to be considered (statistically) significant or not. (Source: R.A Fisher (1925), *Statistical Methods for Research Methods*, p44.) For better or worse, the 1 in 20 or 5% has become the standard by which statistical significance is determined.

|  |
| --- |
| Statistical Significance |
| * 1 in 20 (i.e. 5%) is the standard by which statistical significance is determined |

Measuring the amount of evidence a set of data provides for a research question requires one to first consider which of the possible values provide evidence for the stated research question. The possible values that are said to provide enough evidence for the research question will certainly depend on how the research question is stated. Once again, consider our three case studies.

|  |  |  |
| --- | --- | --- |
| **Case Study** | **Research Question** | **Evidence for the research question** |
| Deafness Malingering | RQ: Is there evidence that the individual being evaluated is malingering in regards to their ability to hear? |  |
| 1st Round Pitchers | RQ: Do pitchers taken in the 1st round of the draft have a higher risk of failure than other position players? |  |
| Ear Infections | RQ: Is there enough statistical evidence to say there is a difference in the duration of ear infection between the breast-fed and the bottle-fed babies? |  |

|  |
| --- |
| Type of Test |
| * Left-tailed test: When interest lies in the left tail of the reference distribution * Right-tailed test: When Interest lies in the right tail of the reference distribution * Two-tailed test: When interest lies in both tails of the reference distribution |

Making a statistical decision is straight forward once the appropriate critical region has been identified.

|  |
| --- |
| Making a Formal Decision |
| * If the outcome from the observed data is within the bottom 5%, top %, or outside 5%, then we can say that the data provides enough (statistical) evidence for the research question. * If the outcome from the observed data is not the bottom 5%, top %, or outside 5% , then we must conclude that the data does not provide enough (statistical) evidence for the research question. |

Measuring Evidence

The **p-value** is our method of determining whether or not you are in the bottom, top, or outside 5% of the reference distribution. The p-value clearly measures and identifies the *amount* of evidence that an observed outcome from a set of data provides for the research question.

|  |
| --- |
| Definition |
| P-Value: the probability of observing an outcome as extreme or more extreme than the observed outcome |

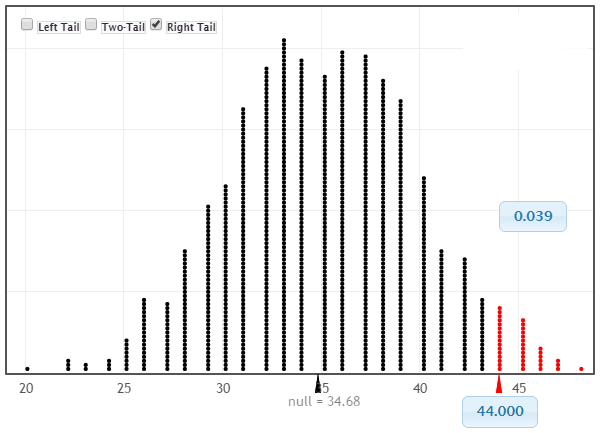
Comments:

* The “p” in p-value stands for probability
* “As extreme or more extreme” implies the possible outcomes that would provide additional support for the research question.

Example 4.1: Compute the p-value for the 1st Round Pitchers case study.

|  |  |
| --- | --- |
| **1st Round Pitchers Case Study** | |
| Research Question | Do pitchers taken in the 1st round of the draft have a higher risk of failure than other position players? |
| Testable Hypothesis | Ho: 1st round pitchers have same failure rate as other position players; the   probability of failure is 34%   HA: 1st round pitchers have a higher failure rate than other position   players; the probability of failure is *greater than* 34% |
| Parameter | = the probability of 1st round pitcher not making it to the majors |
| Rewrite of Hypotheses |  |
| Type of Test | Right-tailed |
| Observed Outcome from Study | 44 correct guesses out of 102 |
| Identify the values for computing the p-value | What values are as extreme as or more extreme than the observed outcome? |

The reference distribution with 1000 repeated trials. StatKey can be used to easily compute the p-value.

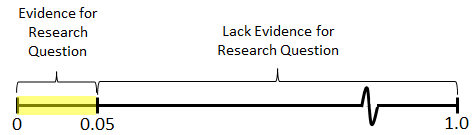


|  |  |
| --- | --- |
| P-Value | The probability of observing an outcome as extreme or more extreme than the observed outcome  P-Value = \_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

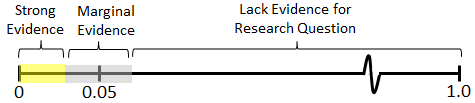
The decision rule for p-values requires us to compare the p-value to the standard guideline proposed by R. A. Fisher for determining statistical significance.

|  |
| --- |
| Making a Decision with P-Values |
| * If the p-value is less than 0.05, then we can say that the data provides enough (statistical) evidence for the research question. * If the p-value is not less than 0.05, then we must conclude that the data does not provide enough (statistical) evidence for the research question. |

This decision rule for better or worse has been widely accepted as appropriate for determining statistical significance.



One may choose to use some slight variations of this rule. For example, it may be desirable to differentiate between a p-value of 0.01 vs. 0.04 because a p-value of 0.01 provides stronger evidence (i.e. the observed outcome is more of an outlier) than 0.04. However, this is typically not done and both are said to provide evidence for the research question. One may consider a more flexible rule; such as the one provided here.



Bross (1971) suggests that such modifications would be detrimental in evaluating evidence.

“Anyone familiar with certain areas of the scientific literature will be well aware of the need for curtailing language-games. Thus if there were no 5% level firmly established, then some persons would stretch the level to 6% or 7% to prove their point. Soon others would be stretching to 10% and 15% and the jargon would become meaningless. Whereas nowadays a phrase such as *statistically significant difference* provides some assurance that the results are not merely a manifestation of sampling variation, the phrase would mean very little if everyone played language-games. To be sure, there are always a few folks who fiddle with significance levels--who will switch from two-tailed to one-tailed tests or from one significance test to another in an effort to get positive results. However such gamesmanship is severely frowned upon.”

*Source*: Bross IDJ (1971), "Critical Levels, Statistical Language and Scientific Inference," in *Foundations of Statistical Inference*.

Use the aforementioned decision rule to make a decision regarding the research question. Also, write a conclusion without the use of statistical jargon, i.e. using laymen’s language.

|  |  |
| --- | --- |
| **1st Round Pitchers Case Study** | |
| Research Question | Do pitchers taken in the 1st round of the draft have a higher risk of failure than other position players? |
| Testable Hypothesis | Ho: 1st round pitchers have same failure rate as other position players; the   probability of failure is 34%  HA: 1st round pitchers have a higher failure rate than other position   players; the probability of failure is *greater than* 34% |
| P-Value | P-Value = 0.039 or 3.9% |
| Decision | Is the p-value less than 0.05? **Yes**   * If “Yes”, then data is said to provide enough evidence for the research question * If “No”, then data does not provide enough evidence for research question |
| Conclusion | Write a conclusion in laymen’s terms.  The data does provide enough statistical evidence to say pitchers taken in the 1st round of the draft have a higher risk of failure than other position players (p-value = 0.039). |

A well-written conclusion should consist of three parts i) decision, ii) restatement of question and iii) strength of evidence.

|  |  |
| --- | --- |
| Part | Conclusion |
| Conclusion |  |
| Part I:  Decision | *Enough evidence* |
| Part II:  Restate question | *Research Question* |
| Part III:  Report p-value | *Strength of evidence* |

Example 4.2: Consider next the Ear Infection case study. Use the p-value approach to make a decision for the stated research question. Write a final conclusion in laymen’s terms as well.

|  |  |
| --- | --- |
| **Ear Infection Case Study** | |
| Research Question | Is there enough statistical evidence to say there is a difference in the duration of ear infection between the breast-fed and the bottle-fed babies? |
| Testable Hypothesis | Ho: There is no difference in duration of fluid between bottle- and breast-fed   babies; that is, the probability the breast-fed baby in each pair did better   is equal to 50%.  HA: There is a difference in duration of fluid between bottle- and breast-fed   babies; that is, the probability the breast-fed baby in each pair did better   is *different* than 50%. |
| Parameter | = the probability of a breast-fed baby doing better |
| Rewrite of Hypotheses |  |
| Type of Test | Two-tailed |
| Observed Outcome from Study | For 16 out of the 23 pairs, breast-fed baby did better |
| Identify the values for computing the p-value | What values are as extreme as or more extreme than the observed outcome? |

Distributions from StatKey can be used to estimate the p-value for this investigation.

|  |  |
| --- | --- |
| Probability of observing 7 or less | Probability of observing 16 or more |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| P-Value | The probability of observing an outcome as extreme or more extreme than the observed outcome   |  |  |  |  | | --- | --- | --- | --- | |  | Probability of observing 16 or more | : | 0.044 | | + | Probability of observing 7 or less | : | 0.046 | |  | P-Value | = |  | | |
| Decision | Is the p-value less than 0.05?   * If “Yes”, then data is said to provide enough evidence for the research question * If “No”, then data does not provide enough evidence for research question |
| Conclusion | Write a conclusion in laymen’s terms. |

Example 4.3: Helper vs. Hinder Case Study

In a study reported in a November 2007 issue of *Nature*, researchers investigated whether infants take into account an individual’s actions towards others in evaluating that individual as appealing or aversive, perhaps laying the foundation for social interaction (Hamlin, Wynn, and Bloom, 2007). In one component of the study, sixteen 10-month-old infants were shown a “climber” character (a piece of wood with “google” eyes glued onto it) that could not make it up a hill in two tries. Then they were shown two scenarios for the climber’s next try, one where the climber was pushed to the top of the hill by another character (“helper”) and one where the climber was pushed back down the hill by another character (“hinderer”). The infant was alternately shown these two scenarios several times. Then the child was presented with both pieces of wood (the helper and the hinderer) and asked to pick one to play with. The color and shape and order (left/right) of the toys were varied and balanced out among the 16 infants.

*References:*

* *Hamlin, J. Kiley, Karen Wynn, and Paul Bloom. “*[*Social evaluation by preverbal infants*](http://course1.winona.edu/cmalone/stat110/HelperHinderer.pdf)*.” November 22, 2007. Nature, Volume 150.*
* *Introducing Concepts of Statistical Inference. Rossman, Chance, Cobb, and Holcomb. NSF/DUE/CCLI # 0633349.*
* [*Video*](http://www.yale.edu/infantlab/socialevaluation/Helper-Hinderer.html) *showing experimental setup.*

Research Question: Do 10-month old infants prefer the Helper toy over the Hinderer toy?

The authors of this work have provided videos to help explain their experimental setup.

[*Video Link*](http://www.yale.edu/infantlab/socialevaluation/Helper-Hinderer.html)

Essential portions of these videos are provided here.

|  |  |
| --- | --- |
| Infants watched a video of Helper  toy several times | Infants watched a video of Hinder  toy several times |
|  |  |

|  |
| --- |
| After watching the videos, both toys were presented to the infant.    Outcome: The toy first selected by the infant |

Questions

1. In the screen shots provided above, the yellow triangle is the Helper toy and the Blue square is the Hinderer toy. Why might it be important to change the color and shape of the Helper and Hinderer toys throughout the experiment? Explain.
2. The last sentence in the case study description mentions that color and shape were varied among the 16 infants in this study. Your friend makes the following false statement, “The results from this study should not be trusted because the experimental setting (i.e. color and shape of the two toys) was not exactly the same for each infant.” Explain why this statement is false.
3. What statistical concept may present itself if color and shape are not changed throughout the experiment? A statistician would randomly assign the color and shape to the various roles. Provide a brief synopsis for how you would proceed with the random assignment for this experiment.

|  |  |
| --- | --- |
| **Helper vs. Hinder Case Study** | |
| Research Question | Do 10-month old infants prefer the Helper toy over the Hinderer toy? |
| Testable Hypothesis | Ho:  HA: Infants have a preference for the Helper toy; that is, the probability of selecting   the Helper toy is *greater than* 50%. |
| Parameter | = the probability a 10-month old will select the Helper toy |
| Rewrite of Hypotheses |  |
| Type of Test | Circle one: Left-tailed | Right-tailed | Two-tailed |
| Observed Outcome from Study | 14 out of the 16 10-month old infants selected the Helper toy |
| Identify the appropriate values for computing the p-value | What values are as extreme as or more extreme than the observed outcome? |

Use StatKey to obtain he reference distribution.



|  |  |
| --- | --- |
| P-Value | P-Value = the probability of observing an outcome as extreme or more extreme than the observed outcome  P-Value = \_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| Decision | Is the p-value less than 0.05?   * If “Yes”, then data is said to provide enough evidence for the research question * If “No”, then data does not provide enough evidence for research question |
| Conclusion | Write a conclusion in laymen’s terms. |

Obtaining an Exact Distribution

There is one caveat regarding to our current approach to obtaining a p-value. Certainly, different simulations will produce slightly different reference distributions. As we’ve discussed previously, the same general pattern will be the same, but variations do exist.

Example 4.4: Consider the Staring case study in which two different reference distributions were obtained.

|  |  |
| --- | --- |
| **Staring Case Study** | |
| Research Question | Is there enough statistical evidence to say the individual in this study has the ability to correctly identify when someone is starring at them? |
| Testable Hypothesis | Ho: The individual is just guessing; that is, the probability of a correct guess is 50%. HA: The individual is answering correctly more often than not; that is, the   probability of a correct guess is *greater than* 50%. |
| Observed Outcome from Study | 14 correct guesses out of 20 |

The possible effect of the different simulation on the calculation of a p-value.

|  |  |
| --- | --- |
| Simulation #1 | Simulation #2 |
|  |  |
| P-Value = 64/1000 = 0.064 | P-Value = 58/1000 = 0.058 |
| Decision: Is p-value less than 0.05? No | Decision: Is p-value less than 0.05? No |
| Conclusion: The data does not provide enough support to say this to say the individual in this study has the ability to correctly identify when someone is starring at them (p-value = 0.064). | Conclusion: The data does not provide enough support to say this to say the individual in this study has the ability to correctly identify when someone is starring at them (p-value = 0.053). |

Fortunately, in the two cases presented above, the final conclusion is the same and the discrepancy between the two plots is minimal. The amount of discrepancy between these two reference distributions is reduced when a larger number of trails are used, i.e. a larger number of samples.

The **binomial probability distribution** can be used instead of the obtaining a reference distribution via simulation. The binomial probability distribution is based on an **infinite number of trials**. This has two advantages: 1) as the number of trials increase, the pattern in our reference distribution is more exact, and 2) prevents different statisticians from getting slightly different p-values.

Conditions for a Binomial Probability Model:

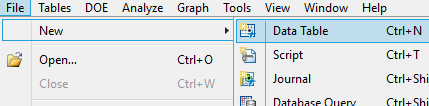
|  |  |
| --- | --- |
| assume | A binomial probability model can be used if:   1. There are a fixed number of observations under study. 2. There are only two possible outcomes. Historically, one outcome is generically labeled a “Success” and the other a “Failure”. 3. The probability of a “Success” remains constant. 4. The observations under study are independent. |

Discuss whether or not these conditions are reasonable for the Staring case study.

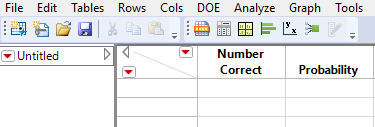
* Fixed Number of Trials
* Two Outcomes
* Probability of “Success” is constant
* Observations are independent

Getting Binomial Probabilities in JMP

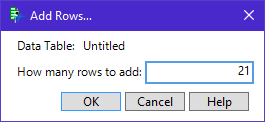
Step 1: Obtain an empty data table, i.e. worksheet.



Step 2: Create two new columns – the first will contain the possible outcomes and second will contain the associated probability for each outcome.



Step 3: Select Rows > Add Rows… Specify the number of rows required for your analysis.



Step 4: Autofill the Number Correct column

|  |  |
| --- | --- |
| Right click on column, select Formula… | Select Row > Sequence |

|  |
| --- |
| Specify the smallest and largest possible values |

Step 5: Obtaining the probabilities

|  |  |  |
| --- | --- | --- |
| Right click on column, select Formula… | Select Discrete Probability > Binomial Probability | |
| Specification of the quantities used to obtain the binomial probabilities in JMP.    What are these values?   * p (or ): this value is the specified parameter for the analysis, i.e. the location of the reference distribution on the number line * n: this value is the sample size for the analysis, i.e. maximum value on the number line * k: the possible outcomes, i.e. values on the number line   Correct specification for the Staring case study. | |

|  |  |
| --- | --- |
| Binomial Probabilities | To create the graph, select Graph > Chart. Place Number Correct in the X, Levels box. Select Probability > Statistics > Data to specify that the probabilities are to be plotted. A needle chart has been specified under Options.    The resulting binomial distribution |

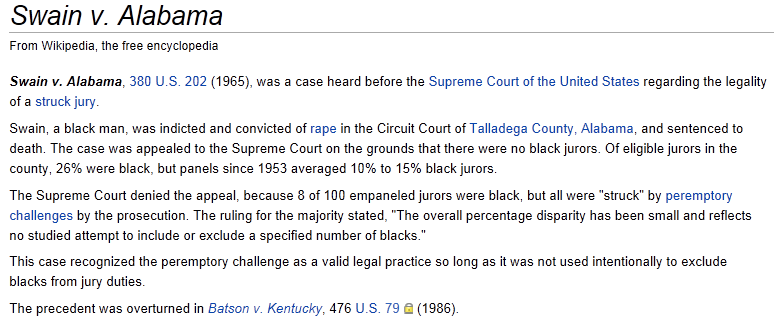
Questions

|  |  |
| --- | --- |
| 1. The p-value for the Staring case study would requires us to add up the probability values for 14, 15, 16, 17, 18, 19, and 20. Compute this value.   P-Value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |

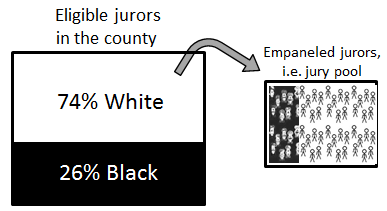
1. What is the interpretation of this value? That is, explicitly communicate what this value is telling us.

Example 4.5: Consider the following Wiki entry for the Swain vs. Alabama Supreme Court case.

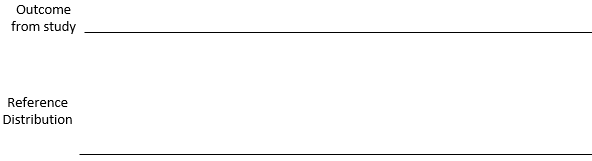
<http://en.wikipedia.org/wiki/Swain_v._Alabama>



A schematic for picking a jury pool for the Swain vs. Alabama case is provided below. The information for this schematic was provided by Wikipedia.



The Wiki article above states that there were 8 (out of 100) empaneled jurors that were black. Specify the setup that would allow us to investigate the fairness of Swain’s jury?



# Blacks in Empaneled Jury Pool

Next, we will evaluate the following statement which was made by the Supreme Court through the following research question.

|  |  |
| --- | --- |
| Statement made by Supreme Court | *“The overall percentage disparity has been small and reflects no studied attempt to include or exclude a specified number of blacks.”* |
| Research Question | Is there evidence of racial discrimination (against blacks) in the selection of Swains’ empaneled jury? |

|  |  |
| --- | --- |
| **Swain vs. Alabama Case Study** | |
| Research Question | Is there evidence of racial discrimination (against blacks) in the selection of Swains’ empaneled jury? |
| Testable Hypothesis | Ho:  HA: |
| Parameter | = the probability that |
| Rewrite of Hypotheses |  |
| Type of Test | Circle one: Left-tailed | Right-tailed | Two-tailed |
| Observed Outcome from Study | 8 black out of 100 empaneled jurors |
| Identify the appropriate values for computing the p-value | What values are as extreme as or more extreme than the observed outcome? |

Getting the p-value in JMP.

|  |  |
| --- | --- |
| Setting up the data table    Creating the list of possible outcomes    Getting the binomial probabilities | The Binomial Distribution Probabilities |

|  |  |
| --- | --- |
| P-Value | P-Value = the probability of observing an outcome as extreme or more extreme than the observed outcome  P-Value = \_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| Decision | Is the p-value less than 0.05?   * If “Yes”, then data is said to provide enough evidence for the research question * If “No”, then data does not provide enough evidence for research question |
| Conclusion | Write a conclusion in laymen’s terms. |

Getting cumulative probabilities in JMP

|  |  |
| --- | --- |
| Individual Probabilities    Cumulative Probabilities |  |