# Lab #4 – Hypothesis Testing using Randomization

Randomization is a powerful tool for determining what the sampling distribution of a statistic would look like under the null hypothesis. However, depending upon the parameter we are interested in, the way that we use the original sample to construct the randomization distribution will differ. In this lab we will use StatKey to explore the use of randomization for various hypothesis testing scenarios (ie: research questions that involve different combinations of one or two categorical or quantitative variables).

We will explore randomization by continuing to analyze the College Scorecard Dataset. Recall that this dataset contains small colleges (1000-5000 students) that primarily grant bachelor’s degrees and require the ACT for admission. The dataset contains the following variables:

* INSTNM The name of the institution
* CITY The city where the institution is located
* STABBR Abbreviation of the state where the institution is located
* REGION A categorical variable describing the region of the institution:

1 New England (CT, ME, MA, NH, RI, VT)

2 Mid East (DE, DC, MD, NJ, NY, PA)

3 Great Lakes (IL, IN, MI, OH, WI)

4 Plains (IA, KS, MN, MO, NE, ND, SD)

5 Southeast (AL, AR, FL, GA, KY, LA, MS, NC, SC, TN, VA, WV)

6 Southwest (AZ, NM, OK, TX)

7 Rocky Mountains (CO, ID, MT, UT, WY)

8 Far West (AK, CA, HI, NV, OR, WA)

9 Outlying Areas (AS, FM, GU, MH, MP, PR, PW, VI)

* LOCALE A categorical variable describing the location of the institution:

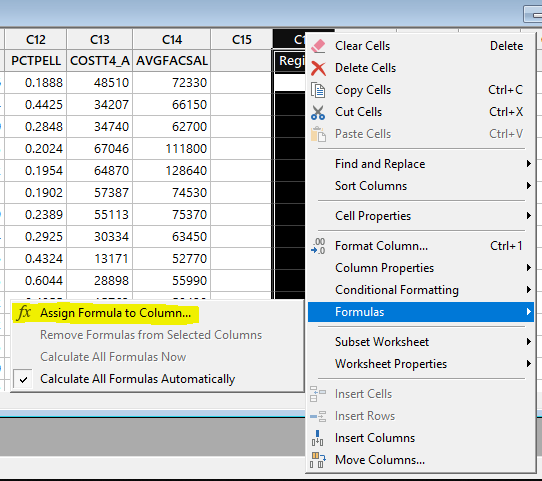
|  |  |
| --- | --- |
| 11 | City: Large (population of 250,000 or more) |
| 12 | City: Midsize (population of at least 100,000 but less than 250,000) |
| 13 | City: Small (population less than 100,000) |
| 21 | Suburb: Large (outside principal city, in urbanized area with population of 250,000 or more) |
| 22 | Suburb: Midsize (outside principal city, in urbanized area with population of at least 100,000 but less than 250,000) |
| 23 | Suburb: Small (outside principal city, in urbanized area with population less than 100,000) |
| 31 | Town: Fringe (in urban cluster up to 10 miles from an urbanized area) |
| 32 | Town: Distant (in urban cluster more than 10 miles and up to 35 miles from an urbanized area) |
| 33 | Town: Remote (in urban cluster more than 35 miles from an urbanized area) |
| 41 | Rural: Fringe (rural territory up to 5 miles from an urbanized area or up to 2.5 miles from an urban cluster) |
| 42 | Rural: Distant (rural territory more than 5 miles but up to 25 miles from an urbanized area or more than 2.5 and up to 10 miles from an urban cluster) |
| 43 | Rural: Remote (rural territory more than 25 miles from an urbanized area and more than 10 miles from an urban cluster) |

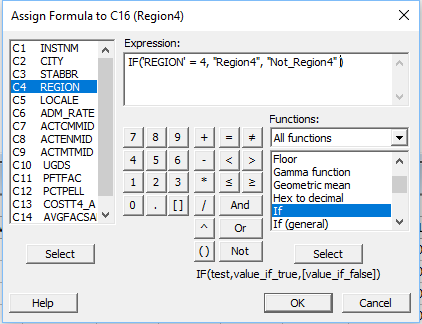
* ADM\_RATE Percentage of applicants that are admitted in to the institution
* ACTCMMID Median cumulative ACT score of enrolled students
* ACTENMID Median English ACT subscore of enrolled students
* ACTMTMID Median Math ACT subscore of enrolled students
* UGDS Total undergraduate enrollment
* PFTFAC Percent of faculty that are full time
* PCTPELL Percent of students receiving a Pell Grant
* COSTT4\_A Average yearly cost of attending
* AVGFACSAL Average faculty salary

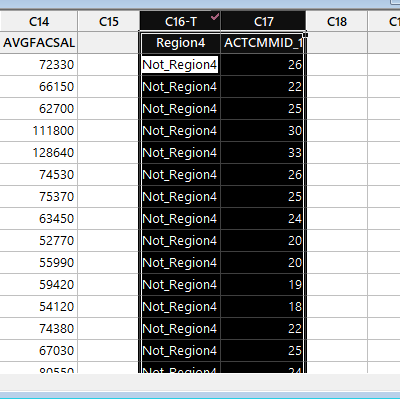
## An Example of Randomization Testing in StatKey

Most often we are interested in performing two-sided hypothesis tests. This example will walk through a randomization test for a difference in means using the College Scorecard Data. Recall that testing for a difference in means actually requires two variables: one variable defining the groups (ie: which cases contribute to which mean) and one quantitative variable from which the means are calculated.

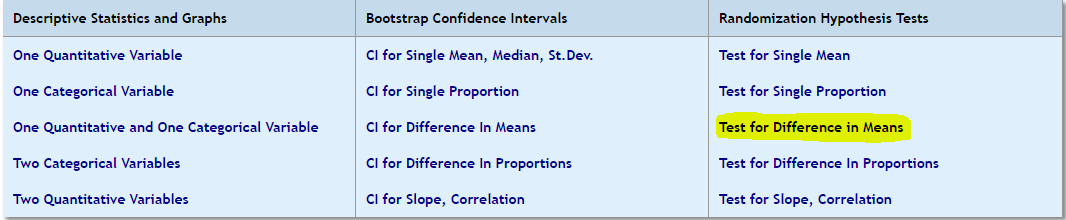
In our example we will determine whether the mean cumulative ACT score is different for small colleges in the Plains Region than it is for small colleges in other regions. The first step in answering this question is to format the data appropriately, this requires us to create a new variable describing whether or not a college is in Region 4 (Plains). A refresher on how to create this variable is shown below:

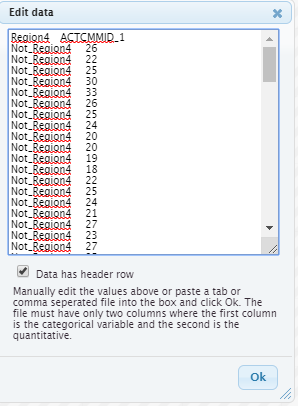




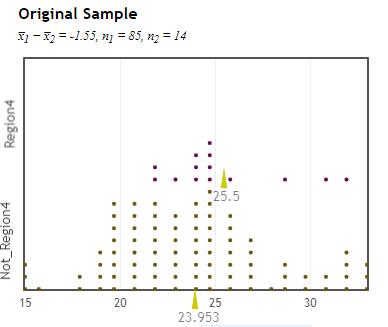


After we have the data formatted properly, we need to enter it into StatKey:

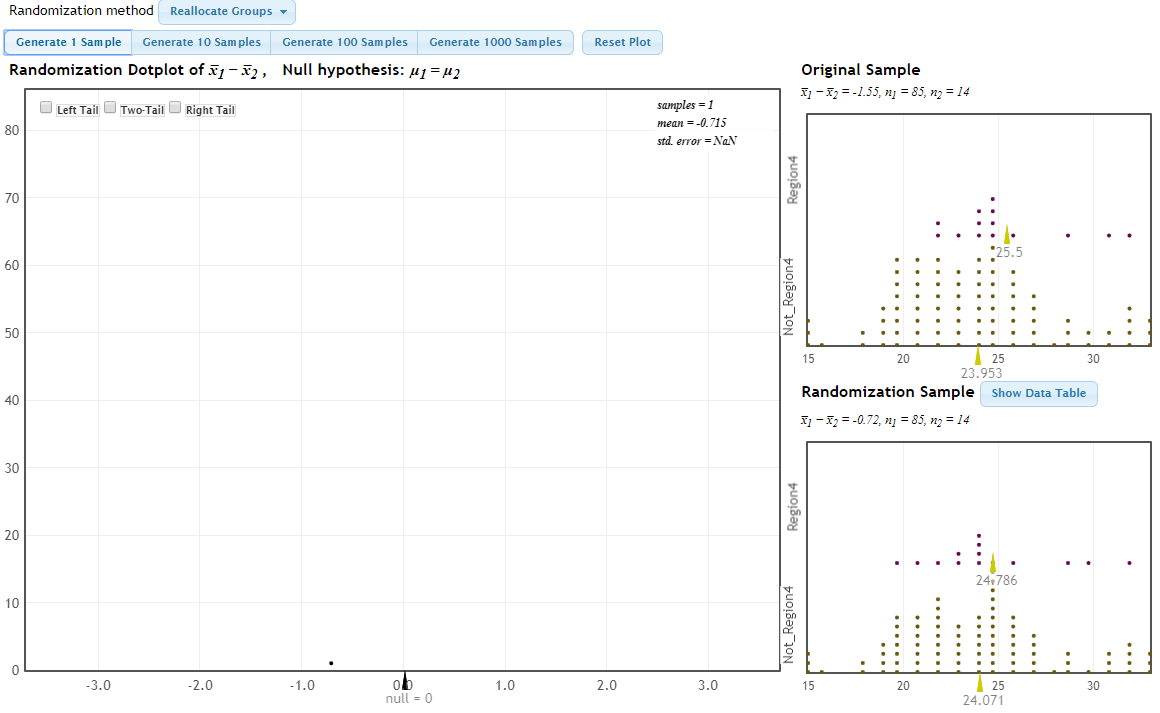


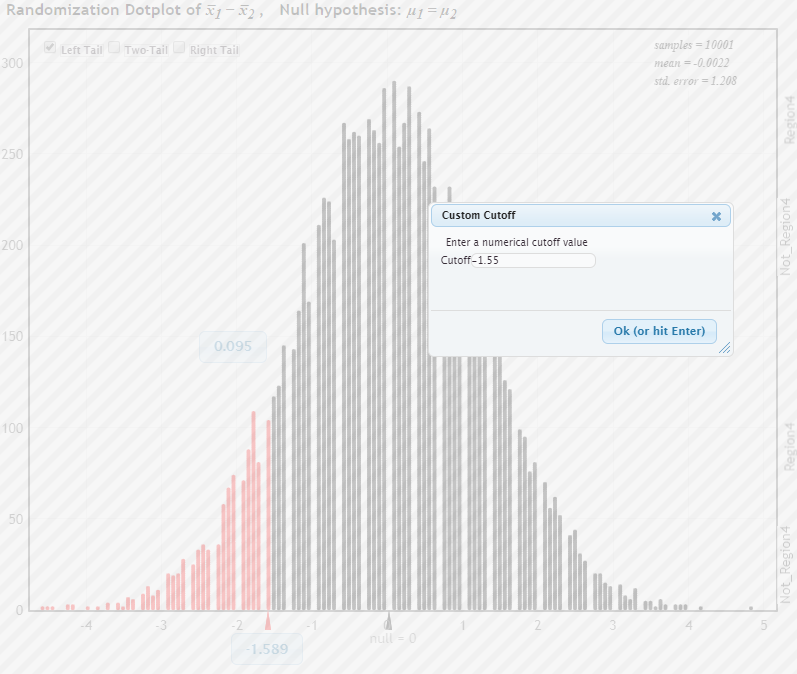


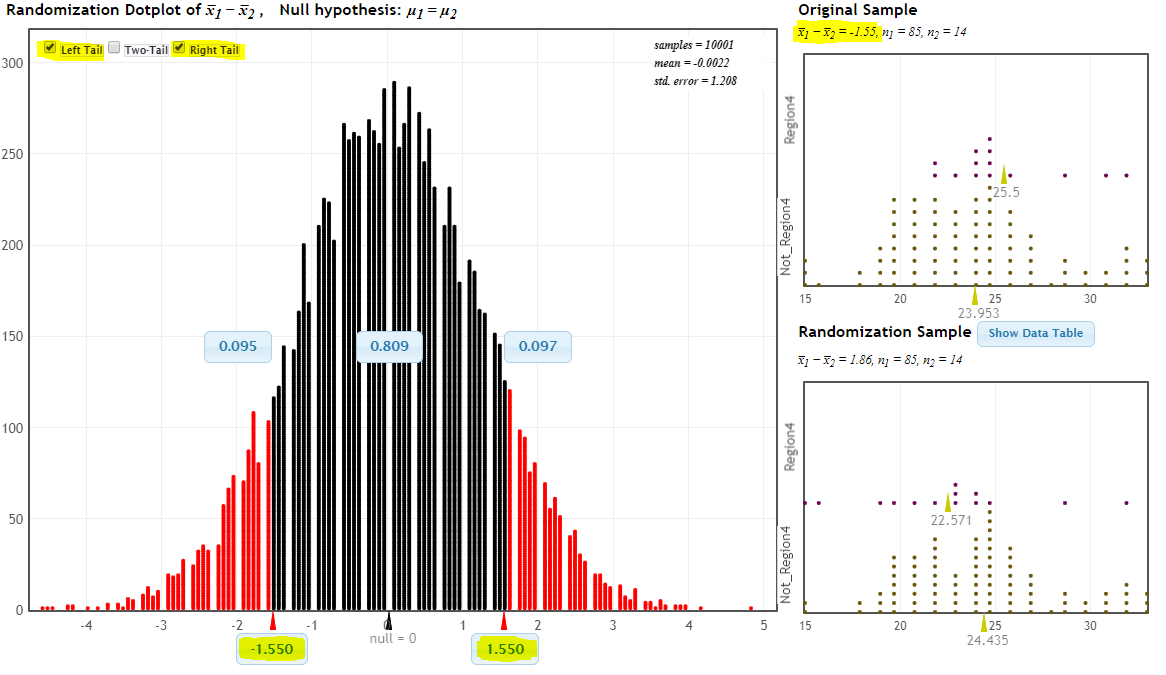
In StatKey, we can view the data we input in the top right corner in the section “Original Sample”. Notice how StatKey reports that the observed difference in means is -1.55. To make the right interpretations, we need to be careful in recognizing that   is the mean of “Not\_Region4” and is the mean of “Region4”. So, a negative difference indicates a higher mean ACT for Plains region colleges.



Now we are ready to construct the randomization distribution and perform a hypothesis test:







Here we see that the two-sided p-value is approximately 0.19. Please note that a formal hypothesis test requires you to clearly state the null and alternative hypotheses, state your observed test statistic, and provide your p-value with an in-context conclusion. These images are merely an example of how to do things in StatKey, NOT a template for hypothesis testing!

## Part 1: Randomization Testing for a Single Mean

The premise of randomization distributions are to simulate the frequency of values of a particular statistic that might be expected if the null hypothesis is true. We saw the example of shuffling the group labels when testing a difference in means of two groups, but this isn’t the only statistic we might base a hypothesis test on. Each section of this lab will address a different type of test, as well as how the test can be executed using randomization.

**How it works:** To test a single mean against a null value, StatKey takes the original data points and shifts their location such that the shifted mean is the value that is hypothesized under the null. The shifted data points are then sampled with replacement in order to introduce the appropriate amount of variability.

### Question #1:

Why are the shifted data re-sampled with replacement? What would happen if these data were sampled without replacement?

### Question #2:

In 2015-16, the federal government issued Pell grants to 53% of college students across the United States. The next few questions will look at the research question: “Does the average small college enroll the same proportion of students with Pell grants as the national rate?”

1. Decide upon the appropriate null and alternative hypotheseses and state them using proper statistical notation in the table below
2. Input the appropriate data into StatKey, make sure that StatKey knows your null hypothesis
3. Generate 10 randomized samples one-at-a-time, tracking the standard deviation of each in the table below

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Null Hypothesis: | | | | | | | | | | |
| Alternative Hypothesis: | | | | | | | | | | |
| Randomized Sample # | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Standard Deviation |  |  |  |  |  |  |  |  |  |  |

### Question #3:

How do the standard deviations of the randomized samples compare with that of the original sample? Why might this relationship be important? (Think about the goal of creating the sampling distribution under the null hypothesis)

### Question #4:

Conduct a *two-sided randomization test* at the level, summarize your findings including: your p-value, your decision with regards to the null hypothesis, and what it means in the context of the research question.

### Question #5:

Now conduct a *one-sided randomization test* at the level evaluating whether small colleges issue Pell grants to a larger proportion of students than the national average. State the appropriate null and alternative hypotheses and summarize your findings including: your p-value, your decision with regards to the null hypothesis, and what it means in the context of the research question.

## Part 2: Randomization Testing for a Single Proportion

**How it works:** StatKey simulates weighted coin flips using the proportion specified in the null hypothesis. For example, if the null hypothesis specifies, StatKey simulates coin flips with a 30% chance of coming up “heads”, which refers to the outcome category that you input.

### Question #6:

The next few questions will look at the research question: “Does a given state have its share of small colleges? “

1. For a state of your choosing, create a new binary categorical variable (using an IF formula in Minitab) to identify which cases are located that state. Include the observed statistic in the table below.
2. Decide upon an appropriate null hypothesis and state it using proper notation (you may want to use outside information to inform your null value). Include your null hypothesis using proper statistical notation in the table below.

|  |  |
| --- | --- |
| Null Hypothesis: | Observed Statistic: |
|  |  |

### Question #7:

Input the appropriate data into StatKey and conduct a two-sided randomization test at the level, summarize your findings including: your p-value, your decision with regards to the null hypothesis, and what it means in the context of the research question.

## Part 3: Randomization Testing for a Difference in Proportions

**How it works:** We will only use StatKey’s default “reallocation” method, this approach shuffles the group labels for each case in the same way that we saw in the notes when testing for a difference in means in the “LightatNight” example.

### Question #8:

The next few questions will look at the research question: “The most common locale for small colleges are small towns (LOCALE = 32), but does the proportion of small colleges located in small towns differ by region?” To make this question fit within the scope of “testing for a difference in proportions”, I’d like you to identify two regions of your choosing that have at least 1 small college in a small town locale. Use the information from a two-way frequency table in Minitab to fill out the following table”

|  |  |  |  |
| --- | --- | --- | --- |
|  | Small Colleges in Small Towns | Small Colleges not in Small Towns | Proportion of the Region in Small Towns |
| Region \_\_\_(fill in)\_\_\_\_ |  |  |  |
| Region \_\_\_(fill in)\_\_\_\_ |  |  |  |

### Question #9:

Determine the appropriate null hypothesis for this research question and fill out the following table:

|  |  |
| --- | --- |
| Null Hypothesis: | Observed Statistic: |
|  |  |

### Question #10:

Input the appropriate data into StatKey and conduct a two-sided randomization test at the level, summarize your findings including: your p-value, your decision with regards to the null hypothesis, and what it means in the context of the research question.

## Part 4: Randomization Testing for Slope/Correlation

**How it works:** The Y variable (the response) is shuffled so that each value of X (the explanatory variable) ends up paired with a different, randomly chosen value of Y in each randomized sample

### Question #11:

The next few questions will look at the research question: “Are tuition costs predictive of faculty salaries at small colleges?” To begin, state the relevant explanatory and response variables in the table below:

|  |  |
| --- | --- |
| Explanatory Variable: | Response Variable: |
|  |  |

Question #12:

Recall that both correlation and regression can be used to evaluate the relationship between two quantitative variables. For each of these methods, determine the appropriate null hypothesis for this research question and fill out the following table: (Hint: for regression, think about how much you’d expect the average faculty salary to increase per unit increase in tuition if the two variables were unrelated)

|  |  |  |
| --- | --- | --- |
|  | Null Hypothesis: | Observed Statistic: |
| Using the correlation coefficient |  |  |
| Using the regression line |  |  |

Question #13:

Input the appropriate data into StatKey and conduct a two-sided randomization test at the level, summarize your findings including: your p-value, your decision with regards to the null hypothesis, and what it means in the context of the research question. Provide your testing results for both the correlation coefficient and the regression line, you only need to interpret these results once.

### Question #14:

Thinking about Parts 1-4, compare and contrast the similarities and differences in how the randomization distribution was constructed for each of these different types of hypothesis tests. What do you think are the key attributes of the original sample that must be maintained?