

Let's consider the probability experiment in which we roll a fair six-sided die.

**Question #1** What is the sample space,  $S$ , for our experiment? That is, list all the possible outcomes.

$S = \{1, 2, 3, 4, 5, 6\}$

**Question #2** What is the probability associated with each outcome?

Each outcome is equally likely so each has probability  $1/6$

**Question #3** What is the probability that you roll an even number?

There are 3 even numbers (2, 4, 6) out of 6 numbers in the sample space, so the probability is  $3/6 = 1/2$ .

**Question #4** What is the probability that you roll a number less than 5?

There are 4 numbers less than 5 (1, 2, 3, 4) out of 6 numbers in the sample space, so the probability is  $4/6 = 2/3$ .

**Question #5** What is the probability that you roll an even number AND a number less than 5?

There are 2 even numbers less than 5 (2, 4) out of 6 numbers in the sample space, so the probability is  $2/6 = 1/3$ .

**Question #6** What is the probability that you roll an even number OR a number less than 5?

There are 3 even numbers and 4 numbers less than 5 in the sample space, so you might think the probability is  $7/6$ . However, that list of 7 numbers contains two numbers (2 and 4) twice, so it's really  $7/6 - 2/6 = 5/6$ .

Alternatively: There are 5 numbers in the sample space (1, 2, 3, 4, 6) that are either even or less than 5 or both, so the probability is  $5/6$ .

**Question #7** Are the events "roll an even number" and "roll a number less than 5" disjoint? Why or why not?

No, they are not disjoint, because there is at least one outcome in both events (in fact, there are two – 2 and 4).

**Question #8** Are the events "roll an even number" and "roll a number less than 5" independent? Why or why not?

The probability of both events happening =  $2/6 = 1/3$  from Question 5.

The product of the individual probabilities of the two events is  $(1/2)(2/3) = 1/3$

Since the two probabilities are equal, the events are independent.

**Alternatively: The probability of rolling an even number is  $1/2$ .**

**Suppose I know that the number is less than 5. This means that the number must be 1, 2, 3, or 4.**

**Two of those numbers are even, so the probability of rolling an even number is  $2/4 = 1/2$ .**

**Since the probability of rolling an even number stays at  $1/2$  even after knowing the number is less than 5, the events are independent.**

**Alternatively: The probability of rolling a number less than 5 is  $2/3$ .**

**Suppose I know that the number is even. This means that the number must be 2, 4, or 6.**

**Two of those numbers are less than 5, so the probability of rolling a number less than 5 is  $2/3$ .**

**Since the probability of rolling a number less than 5 stays at  $2/3$  even after knowing the number is even, the events are independent.**

Now let  $X$  be the discrete random variable representing the outcome of rolling a fair six-sided die; that is,  $X = 1$  if you roll a 1,  $X = 2$  if you roll a 2, etc.

**Question #9** Fill in the table to describe the Probability Distribution of  $X$ :

$X = x$	$P(X = x)$
<b>1</b>	<b><math>1/6</math></b>
<b>2</b>	<b><math>1/6</math></b>
<b>3</b>	<b><math>1/6</math></b>
<b>4</b>	<b><math>1/6</math></b>
<b>5</b>	<b><math>1/6</math></b>
<b>6</b>	<b><math>1/6</math></b>

**Question #10** Use the table to find  $P(1 < X \leq 4)$ .

**The values of  $x$  inside the interval  $1 < x \leq 4$  are 2, 3, and 4. Therefore,**

**$P(1 < X \leq 4) = P(X = 2) + P(X = 3) + P(X = 4) = 1/6 + 1/6 + 1/6 = 3/6 = 1/2$ .**

**Alternatively: The values of  $x$  not in the interval  $1 < x \leq 4$  are 1, 5, and 6. Therefore**

**$P(1 < X \leq 4) = P(X \text{ is not } 1, 5, \text{ or } 6) = 1 - P(X \text{ is } 1, 5, \text{ or } 6) = 1 - 1/2 = 1/2$ .**



Let  $X$  be the random variable representing the outcome of rolling a fair six-sided die; that is,  $X = 1$  if you roll a 1,  $X = 2$  if you roll a 2, etc.

Recall that in Lab 2, you found the probability mass function of  $X$  to be:

$X = x$	$P(X = x)$
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

**Question #1** Compute the expected value (mean) of  $X$ .

$$E(X) = (1)(1/6) + (2)(1/6) + (3)(1/6) + (4)(1/6) + (5)(1/6) + (6)(1/6)$$

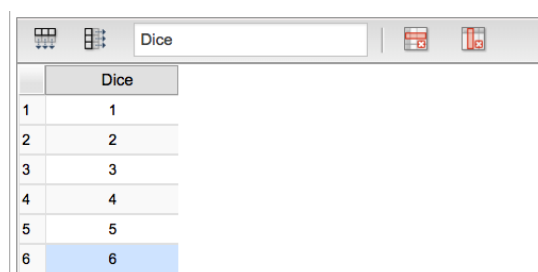
$$= (1 + 2 + 3 + 4 + 5 + 6)/6 = 21/6 = 3.5$$

**Question #2** Compute the variance and standard deviation of  $X$ .

$$\text{Var}(X) = (1 - 3.5)^2 (1/6) + (2 - 3.5)^2 (1/6) + (3 - 3.5)^2 (1/6) + (4 - 3.5)^2 (1/6) + (5 - 3.5)^2 (1/6) + (6 - 3.5)^2 (1/6) = (6.25 + 2.25 + 0.25 + 0.25 + 2.25 + 6.25)/6 = 17.5/6 = 2.92$$


$$\text{SD}(X) = \sqrt{\text{Var}(X)} = 1.71$$

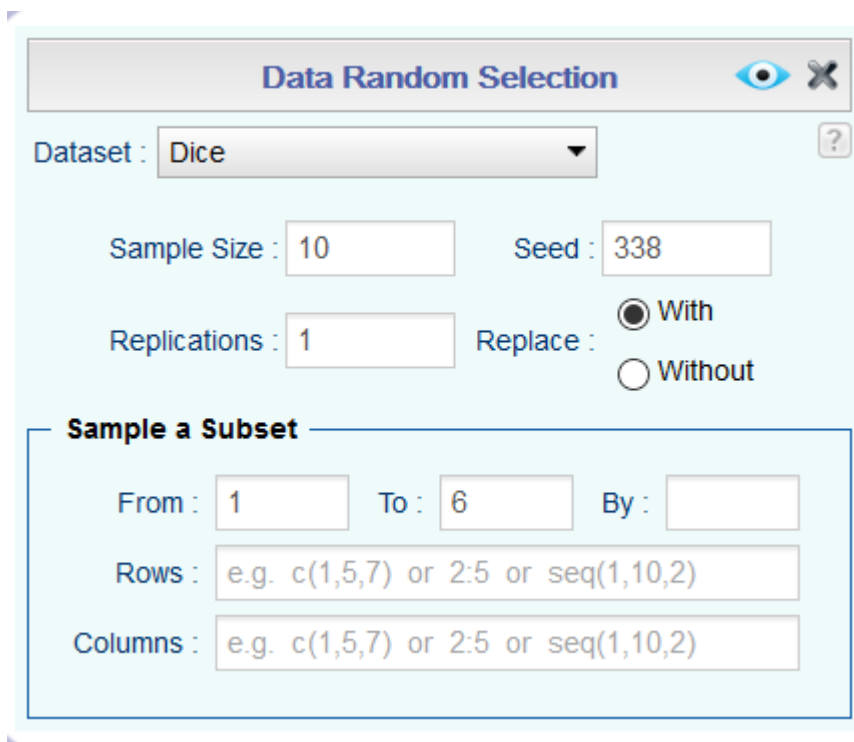
Now we are going to simulate the act of rolling a die in Rguroo. Use Rguroo's [Data Editor](#) to make a dataset that has values 1 through 6, as shown below. Call your dataset **Dice**.



	Dice
1	1
2	2
3	3
4	4
5	5
6	6

Since the die is fair, every outcome is equally likely, and we can simulate directly from our dataset.

Go to [Probability-Simulation](#)  [Random Selection](#) and take a sample of size 10 from the **Dice** dataset, using seed 338, as shown below. View the resulting set of 10 simulated rolls of a die.



The dialog box is titled "Data Random Selection" and has a close button (X) and a help button (?). It contains the following fields and options:

- Dataset: **Dice** (dropdown menu)
- Sample Size: **10** (text box)
- Seed: **338** (text box)
- Replications: **1** (text box)
- Replace: ☒ **With** (radio button), ☐ **Without** (radio button)
- Sample a Subset** (section header)
  - From: **1** (text box)
  - To: **6** (text box)
  - By: (text box)
  - Rows: **e.g. c(1,5,7) or 2:5 or seq(1,10,2)** (text box)
  - Columns: **e.g. c(1,5,7) or 2:5 or seq(1,10,2)** (text box)

**Question #3** What is the relative frequency of the number 6 (the proportion of the time you got 6)? Is it close to the probability you computed earlier?

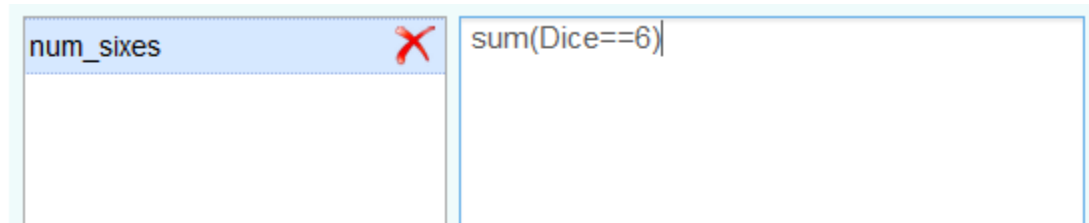
**You should have gotten 2 sixes, for a relative frequency of 2/10 or 0.2. This is as close to the probability we computed as we can get in 10 rolls.**

**Question #4** What is the mean value rolled? Is it close to the mean you computed earlier?

**The mean we get with seed 338 is 3.9, which is close but not exactly equal to 3.5.**

If we do much more than 10 simulated rolls, it's going to be difficult to obtain the relative frequency of 6's and the mean value rolled by hand. So we're going to use the [Statistic](#) menu to obtain the frequency of 6's and the mean value.

Create two variables, **num\_sixes** and **mean\_roll**, and in the center box type the formulas as shown below.



The interface shows a table with two columns. The first column contains the variable name **num\_sixes** and a red X icon. The second column contains the formula **sum(Dice==6)**.

num_sixes	sum(Dice==6)
-----------	--------------

num_sixes		mean(Dice)
mean_roll		

Go back into the sample menu and change the sample size of 10 to 100; 1000; 10,000; and 100,000.

**Question #5** Convert each value of **num\_sixes** to a relative frequency and fill in table below:

Number of Rolls	Relative Frequency (Proportion of 6's)	Mean
10	2/10 = 0.2	3.9
100	14/100 = 0.14	3.49
1,000	163/1000 = 0.163	3.536
10,000	1650/10000 = 0.165	3.4994
100,000	16842/100000 = 0.16842	3.50323

**Question #6** Describe what happens to the relative frequency of the occurrence of observing a 6 as the number of rolls increases from 10 to 100,000.

**Ideally, the relative frequency should be approaching our probability of 1/6 as the number of rolls increases. However, because of some randomness, we don't exactly get there smoothly.**

**Question #7** Describe what happens to the mean value of the rolls as the number of rolls increases from 10 to 100,000.

**Ideally, the mean should be approaching our computed mean of 3.5 as the number of rolls increases. However, because of some randomness, once we get around 3.5 we don't exactly get closer every time.**

On Titanium, you will see a dataset called *HairEyeColor*. Download the dataset and then import it to Rguroo following the same procedure you used in Lab 1.

**Question #1** Copy or screenshot the Summary table and paste it below.

*Categorical Variables*

Variable	Level 1	Level 2	Level 3	Level 4
Hair	Black:108	Blond:127	Brown:286	Red:71
Eye	Blue:215	Brown:220	Green:64	Hazel:93
Sex	Female:313	Male:279		

**Question #2** How many variables are there in this dataset? Are the variables numerical or categorical? Specifically name one of the categorical variables and state its levels.

**All three variables are categorical. The variables are:**

**Hair: takes values Black, Blond, Brown, and Red.**

**Eye: takes values Blue, Brown, Green, and Hazel.**

**Sex: takes values Female and Male.**

Right click on the data set (on the left panel) and select [View](#) to observe the raw data in the interface. Upon doing so, you will see that there are 4 columns: Case no., Hair, Eye, and Sex. The [View](#) interface only allows you to see 25 rows of the dataset at a time, but you can use the buttons and text boxes at the bottom of the interface to change which rows are viewed.

**Question #3** How many cases are in the dataset? (Hint: look at the overall number of rows)

**592 cases are in this dataset.**

Now let's graphically depict the eye colors using a bar graph. Click on the [Plots](#) section, then go to the drop menu for [Create Plot](#) and select [Barplot](#). A window will pop up that looks like the following:

Barplot

Dataset :
Select a Dataset

Numerical / Freq
Categorical

Numerical / Freq
Categorical

Var. Filter...

No items to show.

Selected

No items to show.

☐ Numericals on Axis

Factor 1 :

Factor 2 :

Function : Mean

☐ Confidence Bar

☐ Add Value Labels

Label

Title :

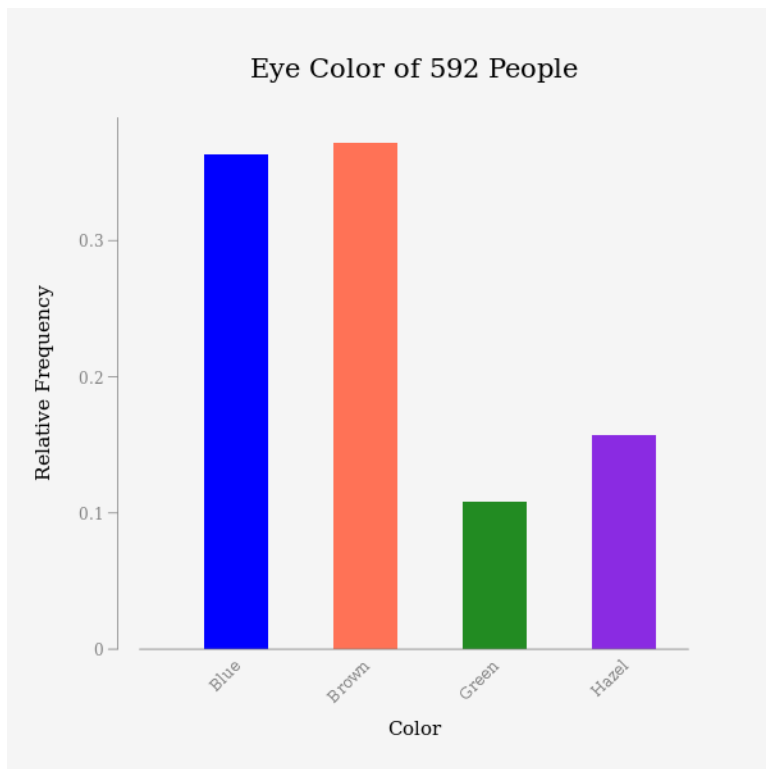
X-Axis :

Y-Axis :

We first need to select the dataset by clicking the drop-down menu currently showing *Select a Dataset*; choose *HairEyeColor*. Now click the *Categorical* tab, select the *Factor 1* drop down menu and click on *Eye*. Click on *Relative Frequency* to change the numbers shown in the plot from frequency to relative frequency. Fill in an appropriate *Label* for the *Title*, *X-Axis*, and *Y-Axis*. Click on the eye icon to view the bar graph.

**Question #4** Copy the graph and paste it below.





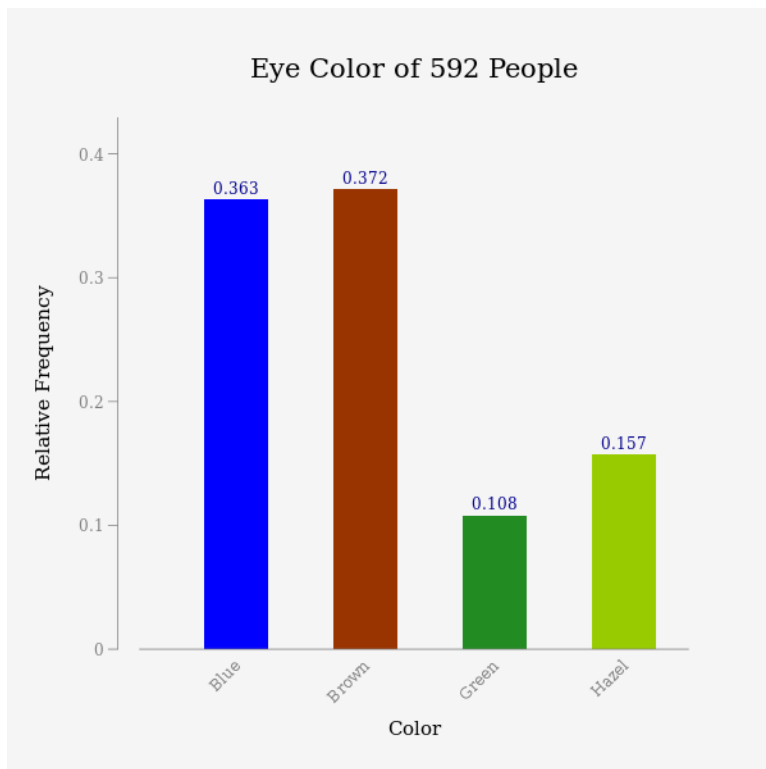
**Question #5** Which category has the most people? Which has the least?

**Brown eyes are the most common and green eyes are the least common.**

You can add the specific proportion of each category as well as other features by clicking on the [Details](#) tab. To add the proportions, go to [Bars](#), [Value Labels](#), [Error Bars](#). In the right side of the [Bars](#) tab you will see a section called [Value Labels](#); click the [Add Value Labels](#) box. Press the eye icon to see the change.

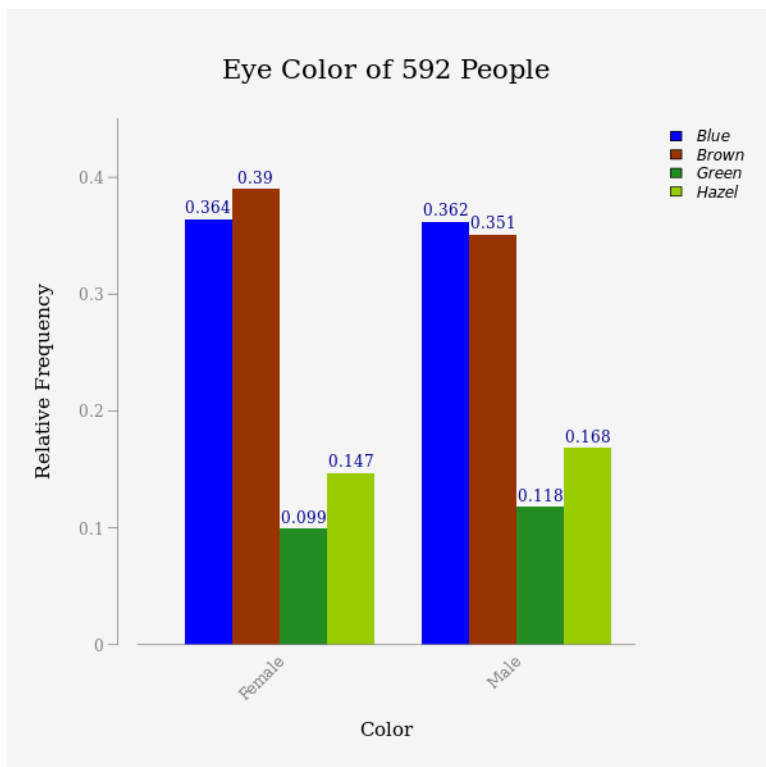
Now, let's change the bar colors from the default. Click on the [Level Editor](#) button. Select *Eye* from [Factor](#) list on the left side. The list of levels of *Eye* now appears in the middle column. For one eye color at a time, click on the level, then in the right column, find the [Color](#) option. You can either type a color name (for instance, brown) in the text box, or click on the color palette to the right, and select a color.

**Question #6** Copy the new graph and paste it below.



We can also visually compare the distribution of Eye Color in males and females. To do this, click on the [Basics](#) button and select Sex from the [Factor 2](#) menu. Select the eye icon to view the new graph. The legend may appear horizontal – to fix this click the [Details](#) button, find the [Legend and Grid](#) menu, and in the [Properties](#) tab, select the [Vertical](#) orientation.

**Question #7** Copy the new graph and paste it below.



**Question #8** Which color is most prevalent for females; which color for males?

**Brown is most prevalent for females (39%) and blue is for males (36.2%).**

Suppose that we are randomly guessing on a 30-question multiple-choice exam. We're guessing a little better than chance alone would predict – we estimate we have a 35% chance of getting each question correct.



**Question #1** Let  $X$  be the number of questions we get correct. Evaluate the BINS assumptions for this scenario to explain why  $X$  can be modeled as a binomial random variable. As part of your evaluation, identify the parameters  $n$  and  $p$ .

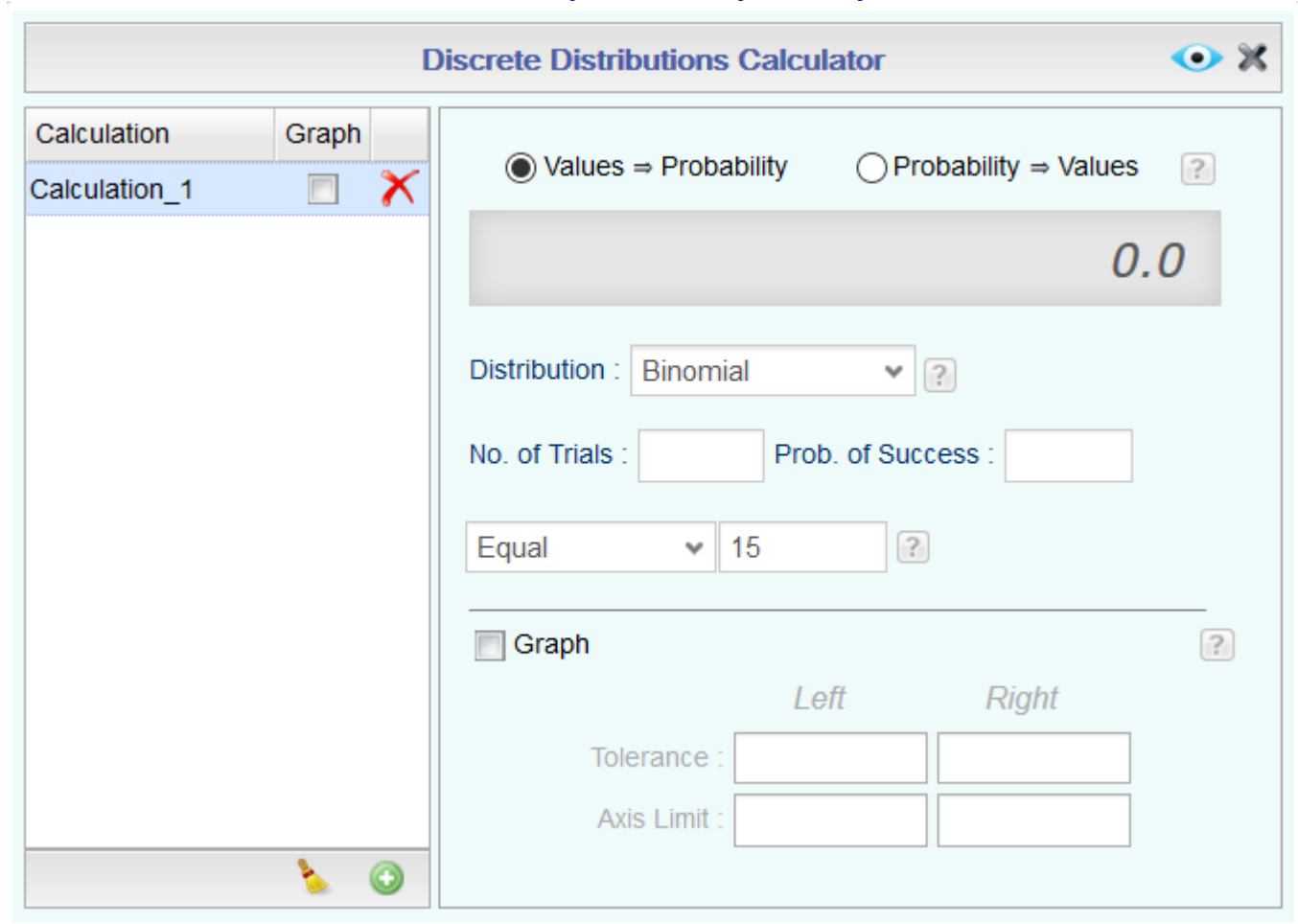
**B:** we have a binary response on each question, either we get it correct or we don't

**I:** we are assumed to get each question right/wrong independently of the other questions

**N:** there are  $n = 30$  questions

**S:** the probability of getting any given question correct is  $p = 0.35$

In Rguroo, click on [Probability-Simulation](#)  [Distribution Calculator](#)  [Discrete](#). Select the [Binomial](#) Distribution and fill in the values of  $n$  (*No. of Trials*) and  $p$  (*Prob. of Success*).



Discrete Distributions Calculator

Calculation Graph

Calculation\_1

☒ Values  $\Rightarrow$  Probability ☐ Probability  $\Rightarrow$  Values

0.0

Distribution : Binomial

No. of Trials : 15 Prob. of Success :

Equal 15

☐ Graph

Left Right

Tolerance : Axis Limit :

**Question #2** Fill in the rest of the diagram as shown above. What is the probability of getting exactly 15 questions correct?

**For questions 2-6, let  $X$  be the number of questions answered correctly.**

$$P(X=15) \approx 0.035$$

For **Questions #3-6**, keep the distribution the same, but change the *Equal* and **15** to new values reflective of the new question. Note that the checkbox to the right of the text box means to include the value; e.g., *Below* 15 with the checkbox checked means less than or equal to 15, while *Below* 15 without the checkbox checked means strictly less than 15.

**Question #3** What is the probability of getting at most 10 questions correct?

$$P(X \leq 10) \approx 0.508$$

**Question #4** What is the probability of getting 18 or more questions correct?

$$P(X \geq 18) \approx 0.0045$$

**Question #5** What is the probability of getting more than 10 questions correct, but at most 18?

$$P(10 < X \leq 18) \approx 0.491$$

**Question #6** What is the probability of getting strictly fewer than 5 questions correct?

$$P(X < 5) \approx 0.0075$$

We can also think of this problem in terms of sample proportions. Remember that the sample proportions do not have a binomial distribution, but we can use the formula  $\hat{p} = X/n$  to convert between sample proportions  $\hat{p}$  and sample counts of success  $X$ .

**Question #7** What is the probability of getting exactly 60% on this 30-question test?

**For questions 7-10, let  $\hat{p}$  be the sample proportion of questions correct (out of 30).**

$$P(\hat{p} = 0.6) = P(X = 18) \approx 0.0031$$

**Question #8** What is the probability of getting a score of 45% or lower?

$$P(\hat{p} \leq 0.45) = P(X \leq 13.5) = P(X \leq 13) \approx 0.873$$

**Question #9** What is the probability of getting a score lower than 50%?

$$P(\hat{p} < 0.5) = P(X < 15) \approx 0.935$$

**Question #10** Suppose that we pass the test (70% or above). Is this consistent with our assumption that we are randomly guessing with a 35% chance on each question? Why or why not?

**No, this is not consistent (unless you believe you are unusually lucky!). We find that  $P(\hat{p} \geq 0.7) = P(X \geq 21) = 0.0001$ . In other words, if we are randomly guessing with probability 0.35 of getting any given question right, there is a 0.01% chance that we pass. This is extremely unlikely and it suggests that our assumption is wrong.**

**Clinical trials** are studies that provide intervention to human subjects and attempt to determine whether the intervention is safe and effective. Investigate the clinical trial, “A Behavioral Intervention to Improve Hypertension Control in Veterans,” which can be found online at

<https://clinicaltrials.gov/ct2/show/study/NCT00286754>

**Question #1** For this study, what is the real-world question of interest?

**There are several slightly different real-world questions of interest that could be asked, given the purpose of the study. For instance: Can we improve adherence to treatment for high blood pressure? Does a particular treatment (stage-matched intervention) do a better job of lowering veterans' blood pressure?**

**Question #2** What is the population of interest in this study? What was the sample being studied (what are the experimental units, and how many are there)?

**The population of interest in this study is United States veterans, age 21 years or older, with hypertension, on anti-hypertensive medication for 1 year, and with uncontrolled blood pressure (you might not have all of these inclusion criteria, and that's okay). The sample being studied was 533 people from this population.**

**Question #3** What was the factor (experimental/explanatory) variable in this study? What were the primary outcome (response) variables? Classify each variable as categorical (qualitative) or numerical (quantitative).

**The factor in this study was the type of treatment given to the veterans. This variable is categorical.**

**One primary outcome variable was whether the blood pressure was under control or not under control. This variable is categorical.**

**The other primary outcome variable was systolic blood pressure. This variable is numerical.**

**Question #4** What was the control group in this study? What was/were the treatment group(s)?

**The control group was the UC group: patients who received their usual care.**

**The treatment groups were the groups who got a different intervention than usual care: stage-matched intervention (SMI) or health education intervention (HEI).**

**Question #5** Did the experimenters avoid confounding due to the placebo effect? If so, how? If not, why not?

**The experimenters were not able to completely avoid confounding due to the placebo effect, since subjects in the Usual Care group knew they were getting their usual care. However, the subjects in the other two groups only knew they were getting a different treatment; they did not know which**

group they were in. Therefore, the placebo effect would be the same in both the SMI and HEI group.

**Question #6** The experimenters included a number of “exclusion criteria.” People meeting one of these criteria would not be included in the study. For one of those criteria, explain why including people meeting that criterion might bias the results.

**Limited life expectancy:** Patients not expected to live beyond the duration of the study due to something other than hypertension would perhaps not have the ability to adhere to a long-term treatment.

**Inability to understand English:** patients who do not understand English would not be able to effectively converse with the English-speaking social workers and might unintentionally follow the wrong directions.

**Lack of a landline telephone:** Most of these kinds of studies prefer landline telephones because cell phone coverage can vary quite a bit from person to person. Without a landline telephone, people may not be reached in a timely manner and may not accurately follow the intervention.

**Unable to follow the study protocol:** Patients who are unable to comply with the treatment may deviate from the intended interventions.

**Recent major surgery:** There may be complications from the surgery that could bias the blood pressure in unknown ways.

**Not living in the study area for duration of the study:** If they are not available for follow-up then some of their responses will be missing. The experimenters want to remove these people from the study before they drop out on their own.

**Unable to provide informed consent:** This is not a source of bias. Enrolling subjects who do not give informed consent is a major ethical lapse and could result in the entire study being retracted.

“A/B testing” is fancy name for randomized controlled experiments that manipulate only a single factor. Typically, most A/B testing is now done via the Internet, including websites, apps, e-mail marketing, etc. Open the paper “Controlled Experiments on the Web: Survey and Practical Guide,” which can be found at <http://www.exp-platform.com/Documents/controlledExperimentDMKD.pdf> and read sections 2 and 3.

**Question #7** Why is an A/B test called an A/B test?

**An A/B test refers to a test that manipulates a single experimental variable, which has a control (“A”) and treatment (“B”) value, and seeks to compare a response under variant (treatment) A to the response under variant B.**

**Question #8** What is an A/A test and why would researchers use it?

**An A/A test is an A/B test in which both variants are the control. This test would be used to obtain an estimate of the variability in the response variable, or to make sure that the system used to run the experiments is working correctly.**



**Question #9** For one of the examples discussed in section 2 (pages 143-148 of the document), explain what the treatment and control groups are.

**Doctor FootCare:** The control group is the visitors to the website who see the old checkout screen. The treatment group is the visitors to the website who see the new checkout screen.

**Microsoft Office:** The control group is the users who saw the initial implementation of the feedback widget. The treatment group is the users who saw the new implementation with the five-star rating.

**MSN Home Page:** The control group is the users who saw the shopping page without ads. The treatment group is the users who saw the shopping page with the 3 ads.

**Behavior-Based Search:** The control group is the users whose results were returned using Amazon's old recommendation algorithm. The treatment group is the users whose results were returned using the new, Behavior-Based Search algorithm.

**Question #10** For the example you chose in **Question #9**, what was the Overall Evaluation Criterion (OEC)? Is the Overall Evaluation Criterion based on an experimental (explanatory) variable or an outcome (response)?

**The OEC is almost always a statistic computed from the values of an outcome/response variable.**

**Doctor FootCare:** Overall, they want to measure the conversion rate, or the proportion of visitors who purchase something, between the two conditions. The OEC here would be computed from a categorical response variable (purchased/did not purchase).

**Microsoft Office:** Overall, they want to measure the response rate, or the proportion of visitors who actually leave a response, between the two conditions. The OEC here would be computed from a categorical response variable (left feedback/did not leave feedback). Alternatively, you could argue that the OEC is the actual rating on the 1- to 5-star scale (the number of stars left by the users who did respond), which could be viewed as categorical or numerical.

**MSN Home Page:** Overall, they want to compare the net amount of monetary income generated between the two conditions. This is based on a numerical response variable.

**Behavior-Based Search:** Overall, they want to compare the amount of money users spent on Amazon between the two conditions. This is based on a numerical response variable.

The ELISA test was an early test used to screen blood donations for antibodies to HIV. A study (Weiss et. al. 1985) found that the conditional probability that a person would test positive given that they had HIV was 0.97, and the conditional probability that a person would test negative given that they did NOT have HIV was 0.926. The World Almanac gives an estimate of the probability of a person in the USA of having HIV to be 0.0026.

**Question #1** Given the numbers in the paragraph above, identify the base rate (prevalence), sensitivity, and specificity of the test.

**Prevalence = 0.0026**

**Sensitivity =  $P(\text{test positive} \mid \text{disease}) = 0.97$**

**Specificity =  $P(\text{test negative} \mid \text{no disease}) = 0.926$**

**Question #2** Suppose 10000 random people are tested. How many of them do you expect to actually have HIV? How many do you expect not to have HIV?

**Out of 10,000 people, we expect  $10,000 \times 0.0026 = 26$  people to have HIV and the remaining 9974 to not have HIV.**

**Question #3** Of those with HIV, how many do you expect to test positive?

**Of the 26 people with HIV, we expect 97% to test positive. This is 25.22 people or, rounded, about 25 people.**

**Question #4** Of those without HIV, how many do you expect to test negative?

**Of the 9974 people without HIV, we expect 92.6% to test negative. This is 9235.924 people, or rounded, about 9236 people.**

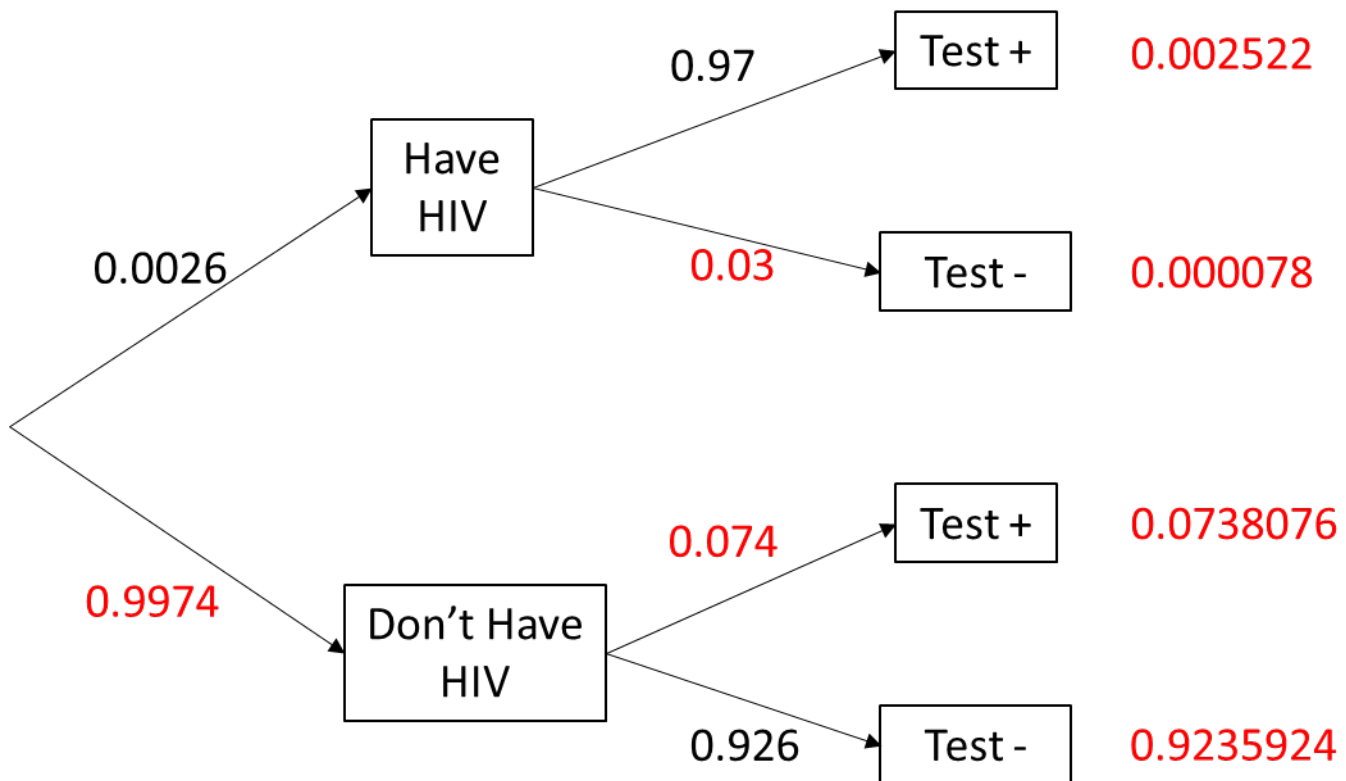
**Question #5** Draw a two-way table to represent this situation. Fill in your answers to Questions #2-4 in the appropriate cells, then fill in the rest of the table. Do not solve for a probability yet.

	Test Positive	Test Negative	Total
Have HIV	25.22	0.78	26
Don't Have HIV	738.076	9,235.924	9,974
Total	763.296	9,236.704	10,000

**Alternatively, using whole numbers:**

	Test Positive	Test Negative	Total
Have HIV	25	1	26
Don't Have HIV	738	9,236	9,974
Total	763	9,237	10,000

**Question #6** Draw a tree diagram to represent this situation. Fill in the probabilities on each branch of the tree. Do not solve for a probability yet.



**In the diagram above, the numbers in black are given and the numbers in red are computed.**

Recall that Bayes' Theorem says:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

**Question #7** In the situation of testing for HIV, what should Event A be? What should Event B be? How do you know?

**Here Event A should be "have HIV" and Event B should be "test positive for HIV." There are a variety of ways of figuring this out, but the easiest is to note that we want to know the probability of having HIV given a positive test.**

**Question #8** Rewrite Bayes's Theorem for this situation, plugging in the correct numbers for each probability. Do not solve for a probability yet.

$$\frac{P(\text{HIV} + | \text{test} +)}{P(\text{HIV} + | \text{test} +) + P(\text{HIV} - | \text{test} +)}$$

**Question #9** Suppose a random person is tested and they test positive. Using the method (two-way table, tree diagram, Bayes's Theorem) that makes the most sense to you, find the conditional probability that this person has HIV given that they test positive.

**Using any of the three methods, it can be found that the conditional probability that this person has HIV, given that they test positive, is about 0.033 or 3.3%.**

**Question #10** Based on your results, would you recommend that this be the first and only test used to detect HIV? Why or why not?

**This test should not be the first and only test used to detect HIV. Only 3.3% of people who test positive will actually have HIV. We probably need additional tests (with preferably higher specificity) to detect whether these people actually have HIV.**

**This counterintuitive result (even if you test positive you are still very unlikely to have a disease) is why there is a lot of disagreement about the proper procedures for different kinds of cancer screenings. If the base rate is very low, then the vast majority of people who test positive will be undergoing unnecessary, expensive, potentially harmful treatment.**

[Tan et al. \(2018\)](#) were interested in how pet dogs bond with complete strangers. In one of their studies, they placed a food treat under one of two cups (the dog did not know which bowl had the food) and had a stranger point the dog toward the cup with the food. Although the researchers did many trials to see how trust evolved, one of the results they analyzed was whether dogs would pick up the cue from the stranger on the very first trial. A total of 53 dogs were tested.

**Question #1** Assuming that the tested dogs were independent (i.e., the dogs did not communicate anything to each other), check the other three assumptions of a binomial setting.

**B condition: MET.** Binary outcome with success = dog got the food; failure = dog did not get the food

**N condition: MET.** The researchers observed  $n = 53$  dogs without this number depending on the result from any of the dogs.

**S condition: KIND OF MET.** We will assume that all the dogs have the same probability of getting the food cup (even though we don't know that that probability is). This may not be a completely realistic assumption, but we don't know enough about the dogs to say anything about what affects their probability of getting the food cup.

**Question #2** What is the sample size in this study?

**Sample size:  $n = 53$  dogs**

**Question #3** What is the parameter  $p$  about which we would like to make inference?

**The parameter  $p$  represents the probability that a dog will “listen” to the stranger and pick the food cup the stranger is pointing to; or, equivalently, it represents the population proportion of dogs that will pick the “correct” cup when a stranger points to it.**

According to the researchers, if dogs do not trust strangers, they would be just as likely to pick the “correct” bowl as the “incorrect” bowl. Let's make this our “nothing unexpected is happening” condition.




**Question #4** What is the value of the parameter  $p$  under the null hypothesis?

**Under the null hypothesis, a dog is equally likely to pick each bowl and  $p = 0.5$ .**

**Question #5** What type of “test statistic” can we compute from sample data in the binomial setting? What type of sampling distribution does it have, and what are the parameters of that distribution under our null hypothesis?

**The test statistic we can compute from sample data in the binomial setting is the number of successes. Here we define success as “dog got the food on the first try.” It is known to have a binomial distribution.**

**Under our null hypothesis, the parameters of the binomial distribution are  $n = 53$  and  $p = 0.5$ .**

Now we will use Rguroo to set up the distribution of our test statistic under the null hypothesis. Click on [Probability-Simulation](#)  [Distribution Calculator](#)  [Discrete](#). Select the [Binomial](#) Distribution and fill in the values of  $n$  ([No. of Trials](#)) and  $p$  ([Prob. of Success](#)). Next, click the [Probability](#)  [Values](#) button at the top of the dialog, since we will be giving a probability.

Suppose that under our “something unexpected is happening” condition, where dogs *do* trust strangers, a “practically significant” value of  $p$  is assumed to be 0.7. Since the value under  $H_1$  is higher than the value under  $H_0$ , we will look in the [Upper Tail](#) of the distribution.

**Question #6** According to convention, what is our desired maximum probability of committing a Type I Error?

**$\alpha = 0.05$**

Enter your answer from **Question #6** in the appropriate text box, check the [Graph](#) box, and preview.

**Question #7** In the text at the top of the graph, Rguroo gives two possible critical regions of  $X$ :  $X \geq 32$  and  $X \geq 33$ . Which one should you use and why?

**We should use  $X \geq 33$  because  $P(X \geq 32) = 0.08449 \geq 0.05$ , while  $P(X \geq 33) = 0.04919 \leq 0.05$ , and we want to keep our  $\alpha$  value below 0.05.**

Now go back into the probability calculator ([Basics](#)) and click the [+](#) button to add another calculation. Enter the same value of  $n$ , but now in the [Prob. of Success](#) box, enter the value of  $p$  under the *alternative* hypothesis  $H_1$ .

Finally, fill in the critical region (from **Question #7**), check the [Graph](#) box, and preview.

**Question #8** According to Rguroo, what is the power of our test at our sample size and specific alternative hypothesis?

**The power of our test at  $n = 53$  and specific alternative  $H_1$ :  $p = 0.7$  is 0.9138 or 91.38%.**

**Question #9** What is the probability of committing a Type II Error, given our sample size, alternative hypothesis, and  $\alpha$  value?

**The probability of committing a Type II Error is  $\beta = 1 - \text{power} = 0.0862$  or 8.62%.**

[Save](#) the calculator to Rguroo – you will need them for Lab 9!

**Question #10** Based on the work you have done in this lab, do you believe that the researchers have a high enough power to detect a practically significant effect of “dog trust in strangers”?

**By convention, we need power to be at least 80% to believe we have high enough power to detect a practically significant effect. Since our power is 91.38%, we do believe we have high enough power.**

[Tan et al. \(2018\)](#) were interested in how pet dogs bond with complete strangers. In one of their studies, they placed a food treat under one of two cups (the dog did not know which bowl had the food) and had a stranger point the dog toward the cup with the food. Although the researchers did many trials to see how trust evolved, one of the results they analyzed was whether dogs would pick up the cue from the stranger on the very first trial. A total of 53 dogs were tested.

Recall from Lab #8 that under the following conditions:

$$H_0: p = 0.5$$

$$H_1: p = 0.7$$

$$\alpha = 0.05$$

$$n = 53$$

our binomial test had a power of 0.9138. In this lab we will explore the effect of changing the sample size, effect size, and  $\alpha$  value on power.

Open up your saved Probability Calculator from Lab 8. First, we will explore changing the value of  $p$  under  $H_1$ . Because this change does not affect the sampling distribution under  $H_0$ , we have the same critical region, and so we only need to worry about the second calculation we saved.

**Question #1** What is the power under the *new* alternative  $H_1: p = 0.6$ ?

**The new power is 0.4259.**

**Question #2** What is the power under the *new* alternative  $H_1: p = 0.8$ ?

**The new power is 0.9992.**

**Question #3** How does the power change as the alternative value of  $p$  gets further from the null value of 0.5?

**As the alternative value of  $p$  gets further from the null value of 0.5, the power increases.**

Now we will explore the effect of changing the desired maximum  $\alpha$ -value. Although this change does not affect the sampling distribution under  $H_0$ , it *does* affect the critical region, so we need to change *both* calculations. First, change the first calculation to use a *new* value  $\alpha = 0.01$ .

**Question #4** What is the critical region for  $\alpha = 0.01$ ? Is it a larger or smaller critical region compared to  $\alpha = 0.05$ ?

**The critical region is now  $X \geq 36$ . This is a smaller critical region compared to  $\alpha = 0.05$ .**

**Question #5** Enter this new critical region in the second calculation and reset  $p$  to its original alternative value of 0.7. What is the new power?

**The new power is 0.6895.**



**Question #6** Repeat the steps using  $\alpha = 0.10$ . What is the power of the test at this new  $\alpha$  value?

**At  $\alpha = 0.10$  the new critical region is  $X \geq 32$  and the power is 0.9505.**

**Question #7** How does the power change as the probability of Type I Error increases? Why do you suspect it changes in that direction?

**As the probability of Type I Error increases, the power also increases because the critical region gets bigger, so under the same alternative distribution the probability of being in the critical region increases.**

Finally, we will explore the effect of changing the sample size  $n$ . This change affects the sampling distribution under both  $H_0$  and  $H_1$ , so we need to change *both* calculations. First, change the first calculation to use  $n = 30$  and reset to  $\alpha = 0.05$ .

**Question #8** What is the critical region for  $n = 30$ ?

**The new critical region is  $X \geq 20$ .**

**Question #9** Change the second calculation to use  $n = 30$  and enter the new critical region. What is the new power?

**The new power is 0.7304.**

**Question #10** Repeat the steps using  $n = 100$ . What is the power of the test using this new sample size?

**The new critical region is  $X \geq 59$  and the new power is 0.9928.**

**Question #11** How does the power change as sample size increases? Why do you suspect it changes in that direction? (Hint: think about the critical regions in terms of sample proportions!)

**As the sample size increases, the power generally also increases (there is some weird behavior due to rounding, which we will ignore in this class). When we had  $n = 30$ , our critical region in terms of proportions was  $\hat{p} \geq 0.667$ . When the sample size increased to  $n = 53$ , the critical region in terms of proportions got larger ( $\hat{p} \geq 0.623$ ), and when the sample size increased to  $n = 100$ , the critical region got even larger ( $\hat{p} \geq 0.59$ ). Again, as the critical region gets larger, if the parameters of the distribution stay the same, the probability of being in the critical region increases.**

**The best way to think about this “intuitively” is to try to visualize the two distributions being on the same x-axis. As the sample size increases, the two distributions look “further and further apart.” As they separate, it starts becoming more and more likely to find a test statistic in the Critical Region under the alternative distribution.**

[Tan et al. \(2018\)](#) were interested in how pet dogs bond with complete strangers. In one of their studies, they placed a food treat under one of two cups (the dog did not know which bowl had the food) and had a stranger point the dog toward the cup with the food. Although the researchers did many trials to see how trust evolved, one of the results they analyzed was whether dogs would pick up the cue from the stranger on the very first trial. A total of 53 dogs were tested.

In the study, 27 of the 53 dogs picked the cup with the food on the first trial.

**Question #1** According to the Neyman-Pearson paradigm we have been working with in Labs 8 and 9, should we accept the null hypothesis  $p = 0.5$  or the alternative hypothesis  $p = 0.7$ ? Why?



**We should accept the null hypothesis  $p = 0.5$  because our observed value of the test statistic ( $X = 27$ ) falls outside the critical region.**

If dogs do not trust strangers initially, they would be expected to choose between the two cups more-or-less at random on the first trial. However, if dogs tend to trust strangers, they would be expected to choose the cup with food more often than would be suggested by chance alone.

**Question #2** Using the parameter  $p$ , write the null and alternative hypothesis according to the “null hypothesis significance testing” paradigm. Label which one is the null hypothesis ( $H_0$ ) and which is the alternative ( $H_a$ ).

**$H_0: p = 0.5$**

**$H_a: p > 0.5$**

Now we will use Rguroo to set up the distribution of our test statistic under the null hypothesis. Click on [Probability-Simulation](#)  [Distribution Calculator](#)  [Discrete](#). Select the [Binomial](#) Distribution and fill in the values of  $n$  ([No. of Trials](#)) and  $p$  ([Prob. of Success](#)) under  $H_0$ .

**Question #3** To compute the p-value according to the “null hypothesis significance testing” paradigm, which of the events below should we find the probability of? Justify your answer.

- a) Exactly 27 dogs getting the cup with food
- b) 27 or fewer dogs getting the cup with food
- c) 27 or more dogs getting the cup with food

**c) 27 or more dogs getting the cup with food. Since our alternative hypothesis has a  $>$  sign, outcomes that are “more favorable” to the alternative hypothesis are outcomes in which more dogs got the food cup.**

Fill in the rest of the dialog based on your answer to **Question #3**. Then click the eye button ([Preview](#)) to find that probability.

**Question #4** According to Rguroo, what is the p-value?

0.5

**Question #5** Using a significance level of 0.05, should you reject the null hypothesis or fail to reject the null hypothesis? Justify your answer.

**We should fail to reject the null hypothesis because the p-value of 0.5 is greater than the significance level of 0.05.**

**Question #6** Can you conclude that dogs do not trust strangers? Can you conclude that dogs do trust strangers? Or can you not make a conclusion either way? Justify your answer.

**Since we fail to reject the null hypothesis, we cannot conclude that the alternative hypothesis is correct – we cannot conclude that dogs trust strangers. However, we started out by assuming that our null hypothesis is correct and dogs don't trust strangers, and we cannot make conclusions of the form “we assumed something was true, we didn't find any evidence to the contrary, so it must be true.” Therefore, we cannot make a conclusion about whether dogs trust strangers based on this data.**

**Question #7** Would you make the same conclusion using the Neyman-Pearson testing paradigm? Why or why not?

**No. Under the Neyman-Pearson testing paradigm, we must either conclude that dogs trust strangers ( $H_0$ ) or that they do not ( $H_1$ ). There is no middle ground for not making a conclusion. If we rejected the null hypothesis, however, we would come to the same conclusion using both paradigms.**

Now we will use the [Proportion Inference](#)  [One Proportion](#) module in [Analytics](#) to “automatically” do a binomial hypothesis test.

In the [Factor Label](#) box, type a name for the response variable, and in the boxes below it, type names for your [Success](#) and [Failure](#) levels of that variable. In the top right, enter the appropriate values of [Sample Size](#) and [# of Successes](#). In the bottom right, enter your alternative hypothesis using the corresponding drop-down menu and text box, then uncheck [Large sample z](#) and check [Binomial](#) instead.

**Question #8** Paste the “Test of Hypothesis” table output from Rguroo below.

*Test of Hypothesis: Food*  
*Method: Binomial Exact Test*

Alternative Hypothesis  $H_a$ : Proportion of 'Food' is greater than 0.5

Sample Size	No. of Successes	Sample Proportion	P-value	BFB	95% Lower CL	95% Upper CL
53	27	0.50943	0.5	1	0.38897	1

*Test is not significant at 5% level.*

**Question #9** Identify the value of the test statistic and the p-value from the table. Are these values the same as you got using the probability calculator?

**The value of the test statistic is No. of successes = 27.**

**The p-value of 0.5 is found in the P-value column.**

**These are the same values we found using the probability calculator.**

For Lab 11 and 12, you will collect your own data. For Lab 11, we will analyze the colors of Starburst. For Lab 12, we will analyze the personality types of statistics students, but we will collect the data now.

To collect the personality type data, please open the following link in your browser window: <http://www.16personalities.com/free-personality-test>, and complete the 10-15 minute personality test. Once you complete the test, your personality will be classified into one of 16 “types,” but we will consider only the 4 main personality categories: Analysts, Diplomats, Sentinels, and Explorers. The figure below shows someone classified as an “Architect,” which is one of the Analyst types as shown by the highlighting. Note that you may have to click on “Start Reading” to find this type. While you are waiting for the rest of the class to finish, feel free to peruse your “results,” or to click on the Personality Types tab to look at all of the possible types.



**Question #1** Once you obtain your individual results, complete the Questionnaire on Titanium for your personality type, major, and the number of Starbursts in each color you received.

In this lab, we will test the claim that all four main colors are equally likely in the population of all Starburst.

**Question #2** Write out the null hypothesis for this goodness-of-fit test.

**$H_0: p_{\text{Red}} = 0.25, p_{\text{Orange}} = 0.25, p_{\text{Pink}} = 0.25, p_{\text{Yellow}} = 0.25$**

**It's okay to say all four colors are equally likely, but remember that we want to create a model, so we should specify the actual values of all four proportions.**

**Question #3** If the null hypothesis is true, how many of each color Starburst would we expect to see in our sample?

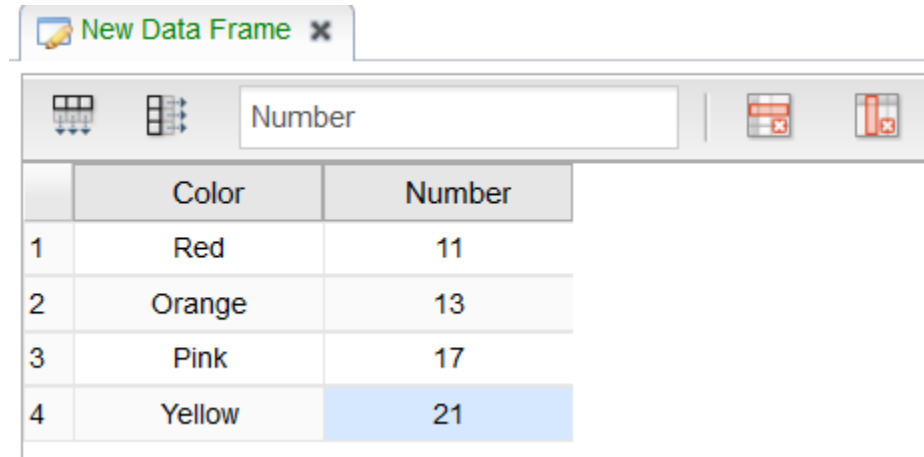
**Our sample had 88 candies in it, so if the null hypothesis is true, we expect 22 red Starburst, 22 orange Starburst, 22 pink, and 22 yellow.**

**Question #4** If the sample size assumptions are met, what would be the sampling distribution of your test statistic (i.e., what is the type of distribution and the degrees of freedom)?

**If the sample size assumptions are met, our test statistic would have a  $\chi^2$  distribution with 3 degrees of freedom.**

Let's put in our data and running the goodness-of-fit test in Rguroo.

Click on [Data](#) → [Data Import](#) → [Create New Data Frame](#). Create the data frame as shown in the screenshot below, except put in the correct counts as indicated by the Titanium survey.



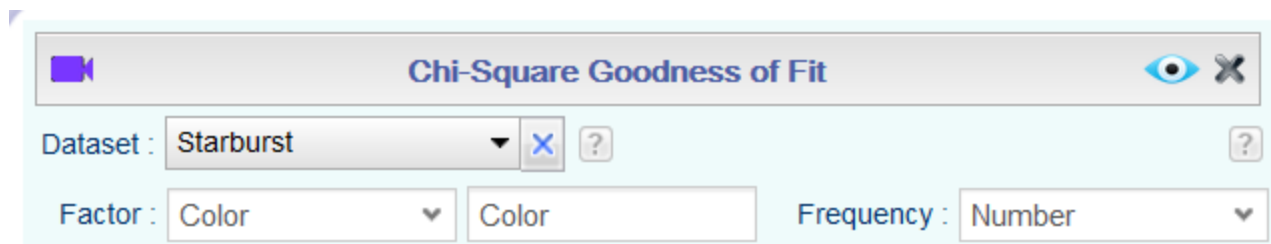
	Color	Number
1	Red	11
2	Orange	13
3	Pink	17
4	Yellow	21

[Save](#) the dataset as the Rguroo dataset *Starburst*.

**Question #5** Are the sample size assumptions are met for a chi-square goodness of fit test? Show how you checked the assumptions.

**Yes, the sample size assumptions are met. The expected counts in all categories are  $22 \geq 5$ .**

Click on [Analytics](#) → [Goodness of Fit](#) and fill out the dialog as shown below.



Chi-Square Goodness of Fit

Dataset : Starburst X ?

Factor : Color Color Frequency : Number

The observed counts should now automatically pop up in the table in the dialog. Fill in the expected probability of each color.

If the sample size assumptions are met, check the [Chi-Square](#) box. Otherwise, we will create a simulated chi-square distribution; check the [Simulation](#) box.

**Question #6** Copy the table output by Rguroo below.

## Chi-Squared Goodness of Fit Test

Research Hypothesis  $H_a$ : Population proportions of Color are different from the expected distribution

Observed Test Statistic	Degrees of Freedom	p-value
7.45455	3	0.05874

*Test is not significant at the 5% significance level*

**Question #7** What is the value of the chi-square test statistic as computed by Rguroo?

**7.45455 (you should probably round this to 7.45 or 7.455)**

**Question #8** What is the p-value for this test?

**0.05874 (it's okay to round this to 0.059 or 0.06)**

**Question #9** Using a 5% significance level, what can you conclude about the distribution of Starburst?

**We fail to reject the null hypothesis. We conclude that our model (which assumes each color has a 25% chance of showing up) is a reasonable approximation of the true/real distribution of Starburst colors.**

**Question #10** Do you believe that your conclusion (from **Question #9**) actually applies to the population of all Starburst? (HINT: Think about the sample we used and the way we collected the data)

**Answers will vary. Essentially, I'm looking for one of two things:**

- 1) No, the conclusion does not apply, for any number of reasons why the bag of Starburst might not actually represent a "random sample" of Starburst (the bagging process may not be random; different manufacturing plants may have different distributions; etc.)**
- 2) Yes, the conclusion does apply, because there is no obvious way in which this sample of 88 Starburst would look meaningfully different from any other sample of 88 Starburst (except because of random luck).**

The two-way table below shows the result of the survey asking you for your personality type and major in two sections of Math 338.

OBSERVED COUNTS	Biology	Computer Science	Other	TOTAL
Analyst	2	5	0	7
Diplomat	13	10	5	28
Sentinel	3	1	2	6
Explorer	3	4	0	7
<b>TOTAL</b>	<b>21</b>	<b>20</b>	<b>7</b>	<b>48</b>

Let's investigate whether personality type affects the choice of major.

**Question #1** Write the null hypothesis for this test of independence.

**$H_0$ : the personality type does not affect the choice of major. OR, there is no relationship/association between the personality type and the major.**

**Question #2** If the sample size assumptions are met (all expected counts  $\geq 5$ ), what would be the sampling distribution of your test statistic (i.e., what is the type of distribution and the degrees of freedom)?

**Our test statistic would come from a  $\chi^2$  distribution with  $(r-1)(c-1) = (4-1)(3-1) = 6$  df.**

**Question #3** If the null hypothesis is true, calculate the expected number of analysts who are Computer Science majors. If your number is not an integer, round it to at least one decimal place.

**If the null hypothesis is true, we would expect  $(7)(20)/48 = 2.92$  analysts who are Computer Science majors.**

**Question #4** If the null hypothesis is true, calculate the Pearson residual and contribution to the chi-squared statistic for analysts who are Computer Science majors.

**Pearson residual:**  $\frac{5-2.92}{\sqrt{2.92}} = 1.22$

**Contribution to chi-squared statistic:  $(1.22)^2 = 1.49$**

Now, let's put in our data and run the test of independence in Rguroo. Import the majors.csv data file from Titanium into Rguroo.

Click on [Analytics](#) → [Contingency Table](#), select your newly imported dataset, and fill out the dialog as shown below.



* Factor 1 :	Personality.Type ▼	Personality.Type
* Factor 2 :	Major ▼	Major
Frequency :	Num. Variable... ▼	

**Question #5** To obtain the p-value, can we use the sampling distribution from **Question #2**, or do we have to simulate a sampling distribution? Explain your reasoning. (HINT: look at your answer to **Question #3**)

**Clearly we do not have 5 expected counts in every cell of the table. Therefore we cannot use a  $\chi^2$  distribution with 6 df to estimate our p-value, and we must simulate the sampling distribution.**

If you answered “yes” to **Question #5**, check the *Chi-Square* box. Otherwise, we will create a simulated chi-square distribution; check the *Simulation* box.

**Question #6** Copy the table with a p-value below.

### *Chi-Squared Test of Independence by Simulation*

Random generator seed = 100

Observed Test Statistic	p-value	Number of Simulations
7.19184	0.3142	10000

*Test is not significant at the 5% significance level*

**Question #7** What is the value of the observed chi-square test statistic as computed by Rguroo?

**7.19184 (best to round to 7.19)**

**Question #8** What is the p-value for this test?

**0.3142 (can round to 0.31)**

**Question #9** Using a 5% significance level, can you conclude that people’s personality type affects their choice of major?

**Since  $0.31 > 0.05$ , we fail to reject our hypothesis ( $H_0$ ). It appears reasonable that people’s personality type does not affect their choice of major.**

**Question #10** Do you believe that your answer (from **Question #9**) applies to all students at Cal State Fullerton? (HINT: Think about the sample we used and the way we collected the data)

**If we wanted to model all CSUF students, we would have problems with both the sample we used and the data collection method. First, we only looked at (mostly) biology and computer science majors, so it’s hard to say whether our results would apply to people in other majors. Second, these personality tests can give pretty different results even for the same person taking the test an hour later, so it’s not clear whether people were classified accurately.**