The Commencement Candy Paradox: Individual Samples vs. Population Truth

A custom confectionary factory designs specialized candy party favors for a college graduation. The packaging machine distributes candies into small favors bags featuring the university's school colors: White and Royal Blue. The factory claims the machine is configured for a perfectly even 50/50 color split. You are assisting with setup for the ceremony and open a single favor bag to check its contents. When you spill the pieces out onto the table, you are immediately confused by why the colors don't have the same amount. You carefully separate the colors and count them up:

Your First Bag: 31 White Candies and 21 Royal Blue Candies

1. Calculate the proportion of white candies found within this first bag. Express your answer as a fraction and as a decimal rounded to three places.

2. Seeking more clarity, you open a second favor bag from the shipment. This bag contains 27 white candies and 24 royal blue candies. Calculate the proportion of white candies for this second bag.

Sally the Statypus says: Hold on a second! Look at that second bag. It has a total of 51 candies. If the factory is aiming for a clean half-and-half split, how can a bag with an odd number of candies ever deliver exactly half? Is it physically possible?

Statypus Insight: The Constraint of Integer Atoms

Consider a machine configured to yield a specific candy color proportion of exactly 0.65. Written as a reduced fraction, $0.65 = \frac{13}{20}$. This structural property means that an individual sample proportion can only hit the true target value perfectly if the sample size (n) is a multiple of 20 (e.g., $n = 20, 40, 60$). Since our proportion was $0.5 = \frac{1}{2}$, we can only have a sample proportion match this if n is a multiple of 2, i.e. an even number. Thus, if $n = 51$, hitting the target exactly is a physical impossibility.

Given that both random bags show distinct variation, you grow curious about how much empirical evidence these two wonky bags actually offer about the machine's calibration. Instead of guessing, we need to lean directly into this variation and inspect it across a macro scale.

Zooming Out: Uncovering the Macro-Level Truth

To conduct a thorough audit of the machinery, you spend the afternoon opening a total of **100 graduation favor bags**. Instead of trusting guesses, we will use our computing environment to audit the true output of the factory.

Coding Corner: Acquiring the Population Audit

Navigate to academy.statypus.org and download the sister script `ASweetCelebration.r` and open it in RStudio. Run the command block under **Part 1: Macro Population Audit** to automatically download the hosted dataset containing the records of all 100 bags. Look closely at the dataframe in your Environment pane.

3. Using the `sum()` function on the appropriate columns in the `candy_data` dataframe, calculate the grand total of White candies and Royal Blue candies produced across the entire afternoon's run.

4. Using these grand totals, calculate the proportion of white candies for the massive aggregate batch. While we cannot assume this value is the absolute population parameter (p), explain why this aggregate calculation is our "best" guess of the machine's true calibration.

The Bracket Operator: Manual Row Extraction

We know the overall batch is highly balanced. But remember, the commencement guests do not see the massive aggregate pile; they only open a single bag. We need to verify our handwritten math from Page 1 against the dataframe.

In R, we can extract specific values from a column vector using the bracket operator `[ ]`. For example, `candy_data$white_count[1]` will extract the number of white candies found in the bag we started with.

5. Run the commands in **Part 2: Manual Row Extraction** of your script. Use the bracket operator to extract the values and manually calculate the proportion of white candies for **Bag 3**. Write out the fraction and the resulting decimal below.

Mapping Guest Experiences: The Birth of a Distribution

Bill the Statypus says: Stop this nonsense right now! Manually extracting rows with brackets to calculate proportions one-by-one is what accountants do with paper ledgers. We are data scientists. We have a computer. We are going to calculate the proportions for all 100 bags simultaneously using vector math! Run **Part 3: Vectorization** in your script.

Sally the Statypus says: Now that we have all 100 guest experiences calculated instantly, let's draw a picture! Run the `hist()` function provided in the script to see what all these different proportions look like when stacked together.

6. Sketch the histogram produced by your R script. Describe its shape as best as you can

Statypus Insight: Continuous Models for Discrete Realities

Originally, individuals were individual candies that weree either white or royal blue. If we focus on a fixed sample size, our individuals are now the bags of candies and the variable we are looking at now is a sample proportion.

🏛 Seneca the Statypus: The Sampling Distribution of $\hat{p}$

Up until this point, a variable represented a trait measured on an individual subject (like the color of a single candy). Now, we are treating an entire sample as a single entity, calculating a summary measurement that changes from sample to sample. We define this new variable using special notation:

$\hat{p}$ (pronounced **"p-hat"**) = The proportion of successes found in a single random sample.

The histogram you just built is called a **Sampling Distribution**. It tracks the behavior and variation of $\hat{p}$ across many replicated samples.

Diving into the Book: The Theoretical Blueprint

We do not want to physically unbox hundreds of bags or run computer simulations every single time we need to evaluate a sample proportion. Probability theory provides an elegant mathematical shortcut to predict the behavior of $\hat{p}$, provided our samples meet a few strict conditions.

Task: Open Chapter 9 of the Statypus textbook (`https://r.statypus.org/`) and navigate to **Section 9.1**. Use the text to answer the following questions.

7. According to the textbook, we can model the sampling distribution of $\hat{p}$ using a Normal curve. What are the mean and standard deviation of this Normal model?

8. **The 5% Rule:** Explain the 5% Rule in your own words. Why must our population be sufficiently large compared to our sample size (n) when we are sampling without replacement? Review Example 9.2 to help understand the concept.

9. **The Success/Failure Condition:** Write down the specific mathematical inequalities required to satisfy the Success/Failure condition.

Bill the Statypus says: This condition is in place to ensure we are safe to use a symmetric model. Having too few successes or failures would lead to a skewed distribution that the Normal distribution cannot capture.

The Mathematical Shortcut & Model Breakdown

Seneca the Statypus: Theorem 9.3: Central Limit Theorem for Proportions

If the 5% Rule and Success/Failure conditions are met, then the sampling distribution of $\hat{p}$ (for a fixed sample size of n) can be modeled by a Normal distribution with mean and standard deviation given below.

$$\hat{p} \sim \text{Norm}\left(\mu_{\hat{p}} = p,\ \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}\right)$$

Sally the Statypus says: It is cool that we are using the Normal distribution to model the proportions that come from a sample that was 50/50!

10. Let us assume the factory machine is operating exactly at its design configuration ($p = 0.50$), and we are looking at an average bag capacity of $n = 50$. Calculate the theoretical mean ($\mu_{\hat{p}}$) and theoretical standard deviation ($\sigma_{\hat{p}}$) for this sampling distribution by hand.

Reflection: Exploring Model Breakdown via Desmos

Scan the attached QR code to open a Desmos exploration. Play with the sliders for "Sample Size," "p," and "# of samples." Play around with each of them until you have a feel for what each one does. When you feel you understand it, explain what each of the three does below. Tap the downward wedges to collapse the unnecessary components.

Chapter 9

The Analytical Audit

Bill the Statypus says: We now have the complete picture. The first bag you opened on Page 1 had an observed proportion of $\hat{p} = 0.596$. We know the theoretical center is 0.50 and the theoretical spread is roughly 0.0707. It is time to use our Normal distribution tools from Chapter 7 to find out how rare a result like that actually is!

Coding Corner: Theoretical Tail Verification

Navigate to **Part 4: Theoretical Tail Verification** in your R script. We will complete the `pnorm()` command to find the mathematical probability of a guest opening a bag with a white candy proportion of 0.596 or higher, assuming the machine is perfectly calibrated.

11. Complete the missing arguments inside the template code in Part 4 of your script. Based on this, what proportion of bags do we expect to have at least 59.6% white candies?

The Reality Check: Empirical Shuffling

Sally the Statypus says: Theoretical models are beautiful, but let's see how our real data stacked up! Run the final line of code in Part 4. We are going to pass a logical condition directly into the `sum()` function to search our 100 actual bags and count exactly how many generated a proportion of 0.596 or greater.

12. Record the exact integer output generated by executing the last command in the sister script.

Reflection: Is this acceptable?

Do you think that opening a bag with nearly 60% white candies represents an anomaly that points to a broken factory calibration, or if it represents an expected manifestation of simple sampling variation.