JMP032: Durability of Mobile Phone Screen - Part 1

Inferential Statistics: Confidence Intervals and Hypothesis Tests for One- and Two-Population Proportions

Produced by

Kevin Potcner, JMP Global Academic Team
kevin.potcner@jmp.com

Durability of Mobile Phone Screen - Part 1

Inferential Statistics: Confidence Intervals and Hypothesis Tests for One- and Two-Population Proportions

Key ideas:

This case study requires the use of inferential statistics, specifically hypothesis tests and confidence intervals, to evaluate the durability of mobile phone screens in a drop test. One-sample inference procedures will be used to determine if a desired level of durability is achieved for each of two types of screens. Two-sample techniques will be used, including difference in proportions, odds ratios, and relative risk, to compare performance.

Background

The durability of a product is clearly an important quality characteristic for both the end user and the manufacturer. For end users, durability is especially important for mobile phones. Dropping a phone on a hard surface, for example, can result in the screen cracking or even breaking, rendering the phone unusable. To evaluate the durability of these screens, manufacturers subject a sample of screens to a variety of tests to simulate typical wear and tear by a user, such as dropping the phone onto a concrete surface.

Material scientists for a screen manufacturer have been experimenting with two new formulations of an aluminosilicate glass. These two formulations were produced by making modifications to the material components and the curing process. For simplicity, we will refer to these two formulations as Screen Types A and B.

A sample of 10 screens of each type was developed for testing. Each screen was installed into the same style of phone. The phones were then dropped in a controlled identical manner from a height of 1 meter onto a concrete surface. A binary variable “Success” (no damage) and “Fail” (screen damage) was recorded.

One of the company’s goals is that 97% of the screens manufactured would be able to experience a drop of 1 meter without becoming damaged (i.e., the Population Success Rate).

The Task

There are two primary questions the engineering team is hoping the data can address:

1. Is there enough statistical evidence to conclude that the Population Success Rate at a 1.0m Drop Height for each of the Screen Types is at least 97%?
2. Is there statistical evidence to conclude that the Population Success Rate for one of the Screen Types is better than the other?

The Data Drop Test 1A.jmp

The data is stored in what’s referred to as Outcome/Frequency Table format.

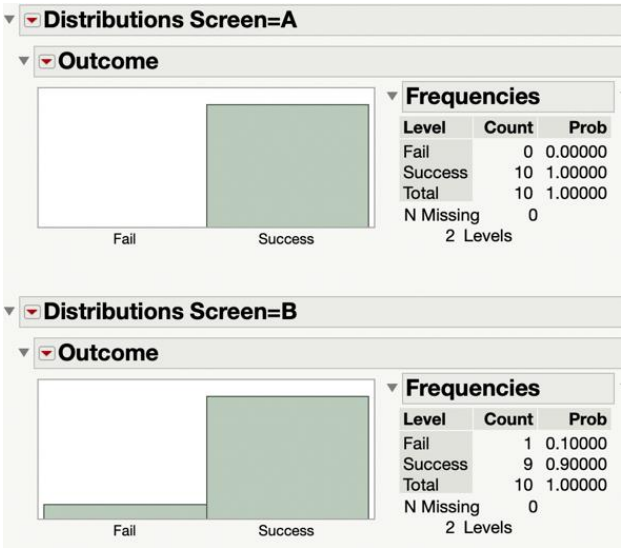
Screen	Two screen types (A, B)
Outcome	Two outcomes (Success, Fail)
Frequency	Number of phones that resulted in either Success or Fail

Analysis

Confidence interval for population proportion

We begin by summarizing the data graphically using the Distribution platform. Exhibit 1 displays bar charts, as well as counts and proportions for each possible outcome for the two Screen Types.

Exhibit 1 Distribution

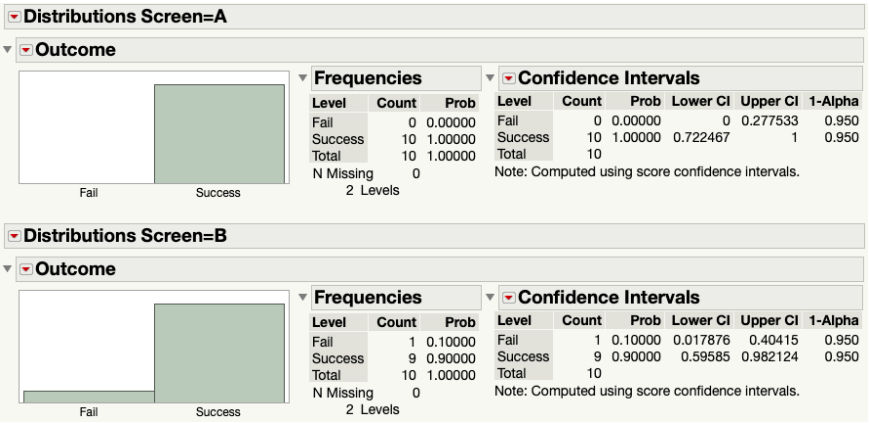


To create, Analyze>Distribution. Select Outcome as the Y Variable, Frequency as the Freq Variable, and Screen as the By Variable.

All of the 10 phones (100%) with Screen Type A did not experience any screen damage, while one of the 10 phones (10%) with Screen Type B did. As this is just the result of a test performed on 20 phones, it's important that we don't immediately conclude that Screen A is better than B in general without a more formal statistical analysis. It's possible that these results could occur in such a test, if, in fact, the Population Success Rates between the two Screen Types are the same. Our analyses will allow us to quantify this probability. We will also be able to estimate the Population Success Rates for each Screen Type accompanied by a measure of statistical uncertainty.

We will begin that analysis by first calculating 95% confidence interval estimates of the Population Success Rate for each Screen Type. Confidence intervals estimate a population parameter by providing a range of plausible values for the parameter based on just a sample. Exhibit 2 displays confidence intervals based upon these data.

Exhibit 2 Distribution with Confidence Intervals



To create, choose Confidence Interval > 0.95 from the red triangle next to the bar chart for each Screen Type.

These intervals are interpreted as follows: We are 95% confident that the Population Success Rate for Screen Type A at the 1m Drop Height is between 72.2% and 100%, and we are 95% confident that the Population Success Rate for Screen Type B at the 1m Drop Height is between 59.6% and 98.2%. The confidence intervals quantify the statistical uncertainty in an estimate. Due to the small sample sizes (10 phones for each Screen Type), there is quite a bit of uncertainty, which is reflected in the width of the confidence intervals (100 – 72.7 = 27.3 for Screen Type A and 98.2 – 59.6 = 38.6 for Screen Type B).

Technical Note: The confidence intervals shown in Exhibit 2 and the hypothesis test conducted in the next section are based upon a statistical procedure that uses the binomial distribution. Another common method used to estimate a population proportion is based upon the normal distribution. Results using that technique will be not identical to those shown here.

To statistically demonstrate that the desired 97% Population Success Rate at the 95% confidence level, the lower bound of the confidence interval would need to exceed 97%, which translates to any plausible value that we estimate the Success Rate to be is greater than 97%. In our data, the lower bound is 72.2% for Screen A and 59.6% for Screen B. Clearly, we have not produced the statistical evidence to demonstrate the Success Rate is above 97%. In fact, the margin of error in these confidence intervals reveals that it would not be possible to demonstrate that level of a Success Rate at the 95% confidence level based upon testing only 10 phones. Even with the perfect results for Screen Type A, the best we can estimate the Population Success Rate to be is 72.7%. We would need to test enough phones to be able to produce a confidence interval with a much smaller margin of error such as, for example, the confidence interval [97.5% - 100%].

In the exercises, you will perform analyses based upon more testing.

Hypothesis test for one proportion

Since the confidence intervals demonstrated that the desired criteria of 97% Population Success Rate was not met, it is not necessary to perform a hypothesis test as it will reach the same conclusion. We will still conduct a hypothesis test to illustrate how to do so and then interpret the output since you will be asked to perform this analysis in the exercises.

We can state the hypothesis of interest as:

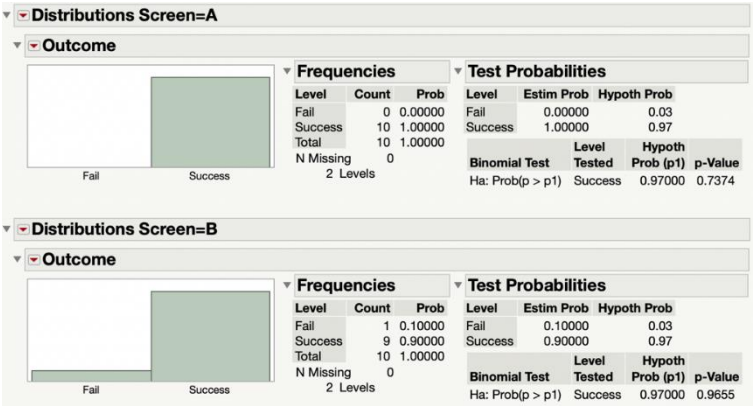
$$H_O : p_A \leq 0.97$$
$$H_A : p_A > 0.97$$

$$H_O : p_B \leq 0.97$$
$$H_A : p_B > 0.97$$

where p_A and p_B are the Population Success Rates for each Screen Type.

Exhibit 3 shows the results of the hypothesis test added to the distribution output.

Exhibit 3 Distribution with Hypothesis Tests



To create, choose Test Probabilities from the red triangle next to the bar chart for each screen type. Type in the value 0.97 for the Hypothesis Probability for Success. Choose Probability greater than hypothesized value (exact one-sided binomial test).

The p-values for both tests are large (0.7374 for Screen A and 0.9655 for Screen B) indicating, as we already determined, that there is no statistical evidence to suggest the desired 97% Success Rate was met at our chosen significance (0.05).

It is important to note that we should not conclude that our analyses have produced statistical evidence that the screens do not meet the criteria. Rather, we haven't produced statistical evidence to conclude that they do. Perhaps we would be able to produce the necessary evidence if a larger number of screens were tested.

Confidence interval and hypothesis test for difference in two proportions

Another objective of the Drop Test was to evaluate if there was statistical evidence to conclude that the two Screen Types have different Success Rates. Notice how much the confidence intervals overlap [72.2% - 100%] for Screen A and [59.6% - 98.2%] for Screen B. From this we can see that there isn't sufficient statistical evidence to conclude a difference in the Success Rates between the two Screens Types. Again, it may be the result of a small number of screens tested; a larger number tested could potentially produce the necessary statistical evidence to demonstrate a difference.

Nonetheless, we'll demonstrate the types of statistical analyses that are commonly done when comparing two population proportions, since you'll be asked to perform these analyses in the exercises with a larger number of phones tested. Five different statistical techniques will be demonstrated:

1. A contingency table analysis.
2. Hypothesis test for the difference in two proportions.
3. Confidence interval estimate for the difference in two proportions.
4. Confidence interval estimate for the relative risk.
5. Confidence interval estimate for the odds ratio.

Each analysis will result in the same conclusion that there isn't enough statistical evidence to conclude a difference in the Success Rate between the two Screen Types.

The hypothesis test of interest for the first three techniques can be written as:

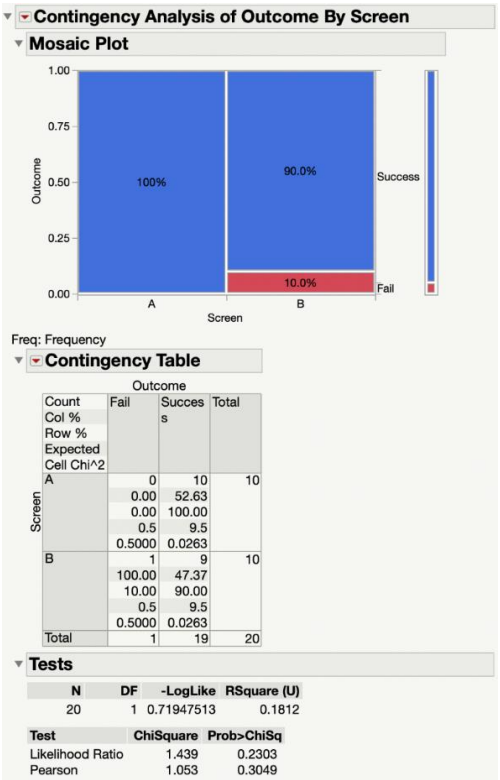
$H_O : p_A = p_B$ or equivalently $H_O : p_A - p_B = 0$

$H_A : p_A \neq p_B$ $H_A : p_A - p_B \neq 0$

where p_A and p_B are the Population Success Rates for Screen Types A and B respectively.

Exhibit 4 is the output from a contingency table analysis.

Exhibit 4 Mosaic Plot and Contingency Table



To create, choose Analyze>Fit Y by X. Place Outcome in the Y Response Category, Screen in the X Grouping Category, and Frequency in the Freq field. Right-click on the mosaic plot and choose Label by Percent under Cell Labeling. The red triangle in the Contingency Table output provides choices for the information to display in the table cells.

The p-values for the likelihood ratio (0.2303) and Pearson (0.3049) show that, as expected, there is not enough statistical evidence to suggest a difference in Population Success Rates between the two Screen Types.

Exhibit 5 contains the results of a two-sample proportions tests, which is another method to test the hypothesis. The Chi-square test from the contingency table analysis can be used if there are more than two categories (e.g., Screen Type C, D, etc.) while the two-sample proportions test is only available when there are two groups, as is the case here.

Exhibit 5 Two-Sample Proportions Tests

Two Sample Test for Proportions

Description	Proportion Difference	Lower 95%	Upper 95%
P(Success|A)-P(Success|B)	0.1	-0.17918	0.34585

Adjusted Wald Test (Null Hypothesis)	Prob
P(Success|A)-P(Success|B) ≤ 0	0.2669
P(Success|A)-P(Success|B) ≥ 0	0.7331
P(Success|A)-P(Success|B) = 0	0.5338

To create, choose Two Sample Test for Proportions from the top red triangle.

As was expected, the p-values for this statistical analysis technique are large (0.5338), demonstrating that there is no statistical evidence to suggest a difference in the Success Rate between the two Screen Types. Also note that the confidence interval for $p_A - p_B$ (the difference in the Population Success Rates) is [-.018, 0.346]. This is interpreted as: We estimate with 95% confidence that the Population Success Rate for Screen Type A is -0.18 units smaller than and up to 0.346 units larger than the Population Success Rate for Screen Type B. That is, we estimate that the Population Success Rate for Screen Type A could be smaller than and also could potentially be larger than the Population Success Rate for Screen Type B. Statistically significant evidence at the chosen level of confidence would only be present if the interval is entirely >0 or <0. This confidence interval shows that 0 is a plausible value for $p_A - p_B$. In other words, there is no difference in the Success Rates.

Confidence interval for relative risk

The fourth approach that can be used to conduct a two-sample comparative analysis is by examining the ratio of the Success Rates known as the relative risk. A hypothesis test framework for this ratio can be written as:

$$H_0: \frac{p_A}{p_B} = 1$$

$$H_A: \frac{p_A}{p_B} \neq 1$$

where p_A and p_B are the Population Success Rates for Screen Types A and B respectively.

This approach phrases the parameter of interest as a percent increase or decrease between the two Success Rates versus the difference between them. It can be a useful way to describe the difference between two proportions. Exhibit 6 shows the results for that analysis.

Exhibit 6 Relative Risk

Relative Risk

Description	Relative Risk	Lower 95%	Upper 95%
P(Success|A)/P(Success|B)	1.111111	0.903718	1.366098

To create, choose Relative Risk from the top red triangle. Choose desired response outcome and category in the numerator.

The confidence interval [0.904, 1.366] is interpreted as: We estimate with 95% confidence that the Population Success Rate for Screen Type A is between 90.4% the size of and up to 36.6% larger than the Population Success Rate for Screen Type A. That is, the Population Success Rate for Screen Type A could be smaller than and also could potentially be larger than the Population Success Rate for Screen Type B. Statistically significant evidence of a difference at the chosen level of confidence would only be present if the confidence interval estimate for the relative risk is entirely >1 or <1. This confidence interval shows that 1 is a plausible value for p_A/p_B. In other words, there is no difference in the Success Rates.

Confidence interval for odds ratio

The last approach we'll illustrate is to evaluate the odds ratio. The odds for an outcome is the probability of the outcome occurring divided by the probability of it not occurring; it is written mathematically as:

$$\frac{p_A}{(1 - p_A)}$$

In this case, p_A would be the probability that Screen Type A would not be damaged in the drop test and $(1 - p_A)$ is the probability that it would. For example, if the probability of a screen not being damaged is 0.90, then the probability of the screen being damaged is 0.10 and the odds would be 0.90/0.10 = 9. It is interpreted as the probability of a screen not being damaged is 9 times more likely than the probability of it being damaged.

The odds ratio is the ratio of two different odds. If, for example, the probability of Screen Type B not being damaged is 0.80, the odds of the screen not being damaged are 0.80/0.20 = 4. The odds ratio between the two Screen Types would be 9/4 = 2.25. It is interpreted as the odds of Screen Type A not being damaged are 2.25 the size of the odds of Screen Type B not being damaged.

An odds ratio equal to 1 would occur when the odds of the two outcomes are the same. In other words, it occurs when the Success Rates are the same.

A hypothesis test for the odds ratio can be written as:

$$H_0: \frac{p_A/(1-p_A)}{p_B/(1-p_B)} = 1$$

$$H_A: \frac{p_A/(1-p_A)}{p_B/(1-p_B)} \neq 1$$

where p_A and p_B are the Population Success Rates for Screen Types A and B respectively.

The odds ratio for our data would be based upon this calculation:

$$\frac{1/0}{0.9/0.10}$$

The 0 in the denominator part of the numerator cannot be computed, thus we are unable to calculate the odds ratio for these specific data. The data you'll be analyzing in the exercises does not have this result so the odds ratio can be calculated.

Technical Note: It is not uncommon to have data that does not allow a certain statistical procedure to be performed or the results of an analysis are not very reliable. It is often the case with data that is based on very rare events.

Summary

Statistical insights

Our analyses have shown that there is no statistical evidence in the data needed to demonstrate the desired 97% Population Success Rates. We also did not produce any statistical evidence to conclude a difference in the Population Success Rates between the two screens. As was stated earlier, we should not conclude that our analyses have produced statistical evidence that the screens do not meet the desired Success Rate; rather, it's that we haven't produced statistical evidence to conclude they do. Similarly, we have not produced statistical evidence to conclude that the two Screen Types have the same Success Rate, simply that we haven't produced the statistical evidence to say they are different. More tests may be able to produce that evidence.

Implications and next steps

Based upon the inconclusive results, the engineers decide to expand upon the study and test 40 more phones of each Screen Type. The exercises will ask you to perform the same analyses as demonstrated here on this larger sample size. In addition, you'll be asked to find the confidence level at which it can be concluded that the desired Success Rate is met. You will also evaluate 95% confidence intervals for different "what-if" scenarios to determine how many phones would need to be tested and the results of those tests to demonstrate the desired Success Rate at that level of confidence.

Exercises

Use the Drop Test 1B.jmp data set to answer the following questions:

1. Analyze the results of all 50 drop tests for each Screen Type via hypothesis tests and confidence intervals to determine if there is statistical evidence to demonstrate the desired 97% Population Success Rate. Based upon the confidence intervals, what are estimates of the lowest and highest possible Population Success Rates? Comment on the uncertainty in estimating the Population Success Rates.

2. Evaluate different confidence levels:
    a. Create a new data table in the same table/frequency format but enter the values 0 for the failures and 50 for the successes.
    b. Create a 95% confidence interval for the Population Success Rate for these fabricated data. Did this result in a confidence interval that demonstrates the desired 97% Success Rate?
    c. Create a 90%, 80%, and 75% confidence interval from this scenario. At what level of confidence did it derive an interval that provides the statistical evidence to demonstrate 97% Success Rate?

3. How many screens would need to be tested with none of them being damaged to generate the statistical evidence required to demonstrate the 97% Population Success Rate at 95% confidence? To answer this, continue to change the results of this fabricated scenario by increasing the number of tests with all tests resulting in a success and no failures. For each, examine the 95% confidence interval.

4. Perform the five different analyses illustrated to determine if statistical evidence exists at the 95% confidence level to conclude a difference in the Success Rate between the two Screen Types:
    a. Contingency table analysis.
    b. Two-sample proportions test.
    c. Confidence intervals for the difference of two proportions.
    d. Confidence intervals for relative risk.
    e. Confidence intervals for the odds ratio.

    Provide an interpretation for the confidence intervals.

5. What are your ideas and recommendations for further study to improve upon evaluating the performance of these two Screen Types in a drop test?