## **Summary:**



**Handling Insufficient Data in Analysis: Tips and Strategies:)**

* Use data from multiple sources to avoid limitations in analysis
* Wait for more data if time allows
* Identify trends with available data
* Look for new data set
* Stakeholders can be consulted to adjust objectives and gather more data
* Outdated or geographically-limited data should be avoided

# What to do when you find an issue with your data

# When you are getting ready for data analysis, you might realize you don’t have the data you need or you don’t have enough of it. In some cases, you can use what is known as proxy data in place of the real data. Think of it like substituting oil for butter in a recipe when you don’t have butter. In other cases, there is no reasonable substitute and your only option is to collect more data. Consider the following data issues and suggestions on how to work around them.

## **Data issue 1: no data**

| **Possible Solutions** | **Examples of solutions in real life** |
| --- | --- |
| Gather the data on a small scale to perform a preliminary analysis and then request additional time to complete the analysis after you have collected more data. | If you are surveying employees about what they think about a new performance and bonus plan, use a sample for a preliminary analysis. Then, ask for another 3 weeks to collect the data from all employees. |
| If there isn’t time to collect data, perform the analysis using proxy data from other datasets. *This is the most common workaround.* | If you are analyzing peak travel times for commuters but don’t have the data for a particular city, use the data from another city with a similar size and demographic. |

## **Data issue 2: too little data**

| **Possible Solutions** | **Examples of solutions in real life** |
| --- | --- |
| Do the analysis using proxy data along with actual data. | If you are analyzing trends for owners of golden retrievers, make your dataset larger by including the data from owners of labradors. |
| Adjust your analysis to align with the data you already have. | If you are missing data for 18- to 24-year-olds, do the analysis but note the following limitation in your report: *this conclusion applies to adults 25 years and older* *only*. |

## **Data issue 3: wrong data, including data with errors\***

| **Possible Solutions** | **Examples of solutions in real life** |
| --- | --- |
| If you have the wrong data because requirements were misunderstood, communicate the requirements again. | If you need the data for female voters and received the data for male voters, restate your needs. |
| Identify errors in the data and, if possible, correct them at the source by looking for a pattern in the errors. | If your data is in a spreadsheet and there is a conditional statement or boolean causing calculations to be wrong, change the conditional statement instead of just fixing the calculated values. |
| If you can’t correct data errors yourself, you can ignore the wrong data and go ahead with the analysis if your sample size is still large enough and ignoring the data won’t cause systematic bias. | If your dataset was translated from a different language and some of the translations don’t make sense, ignore the data with bad translation and go ahead with the analysis of the other data. |

### 

### **Use the following decision tree as a reminder of how to deal with data errors or not enough data:**

### **This illustration is a decision tree showing four possible decisions to make in order to work around data issues.**

### 

### 

# 

# 

# Calculating sample size

# Before you dig deeper into sample size, familiarize yourself with these terms and definitions:

| **Terminology** | **Definitions** |
| --- | --- |
| **Population** | The entire group that you are interested in for your study. For example, if you are surveying people in your company, the population would be all the employees in your company. |
| **Sample** | A subset of your population. Just like a food sample, it is called a sample because it is only a taste. So if your company is too large to survey every individual, you can survey a representative sample of your population. |
| **Margin of error** | Since a sample is used to represent a population, the sample’s results are expected to differ from what the result would have been if you had surveyed the entire population. This difference is called the margin of error. The smaller the margin of error, the closer the results of the sample are to what the result would have been if you had surveyed the entire population. |
| **Confidence level** | How confident you are in the survey results. For example, a 95% confidence level means that if you were to run the same survey 100 times, you would get similar results 95 of those 100 times. Confidence level is targeted before you start your study because it will affect how big your margin of error is at the end of your study. |
| **Confidence interval** | The range of possible values that the population’s result would be at the confidence level of the study. This range is the sample result +/- the margin of error. |
| **Statistical significance** | The determination of whether your result could be due to random chance or not. The greater the significance, the less due to chance. |

### 

### 

### **Why a minimum sample of 30?**

### This recommendation is based on the **Central Limit Theorem (CLT)** in the field of probability and statistics. As sample size increases, the results more closely resemble the normal (bell-shaped) distribution from a large number of samples. A sample of 30 is the smallest sample size for which the CLT is still valid. Researchers who rely on **regression analysis** –statistical methods to determine the relationships between controlled and dependent variables –also prefer a minimum sample of 30.

### Still curious? Without getting too much into the math, check out these articles:

### [Central Limit Theorem (CLT)](https://www.investopedia.com/terms/c/central_limit_theorem.asp): This article by Investopedia explains the Central Limit Theorem and briefly describes how it can apply to an analysis of a stock index.

### [Sample Size Formula](https://www.statisticssolutions.com/dissertation-resources/sample-size-calculation-and-sample-size-justification/sample-size-formula/): This article by Statistics Solutions provides a little more detail about why some researchers use 30 as a minimum sample size.

### 