

## Quantitative

1. Numerical Data - Two Types
2. Discrete - (Counting)
3. Continuous - (Measurement)



## Qualitative

1. Descriptive data based on observations
2. Involves 5 senses
3. See, feel, taste, hear, smell

## Nominal Scale Data

1. Qualitative / Categorical
2. Names, Colors, Labels, Gender, etc.
3. Order does not matter

## Ordinal Scale Data

1. Ranking / Placement
2. The order matters
3. Differences cannot be measured

## Interval Scale Data

1. The order matters
2. Differences can be measured (except ratios)
3. No True "0" Starting point

## Ratio Scale Data

1. The order matters
2. Differences are measurable (including ratios)
3. Contains a "0" Starting point

**Hypothesis** method compares two opposite statements about a population and uses sample data to decide which one is more likely to be correct. To test this assumption we first take a sample from the population and analyze it and use the results of the analysis to decide if the claim is valid or not.

### Defining Hypotheses

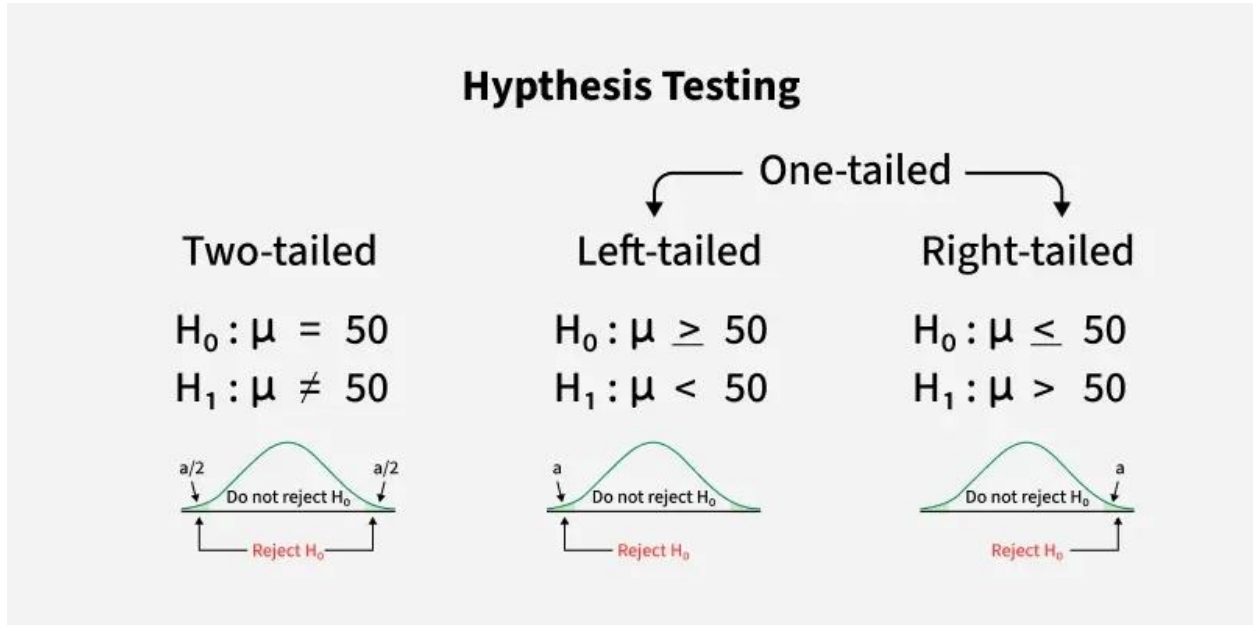
- **Null hypothesis (H<sub>0</sub>):** The null hypothesis is the starting assumption in statistics. It says there is no relationship between groups. For Example A company claims its average production is 50 units per day then here:  
Null Hypothesis:  $H_0$ : The mean number of daily visits ( $\mu$ ) = 50.
- **Alternative hypothesis (H<sub>1</sub>):** The alternative hypothesis is the opposite of the null hypothesis it suggests there is a difference between groups. **like** The company's production is not equal to 50 units per day then the alternative hypothesis would be:  
 $H_1$ : The mean number of daily visits ( $\mu$ )  $\neq$  50.

### Key Terms of Hypothesis Testing

To understand the Hypothesis testing firstly we need to understand the key terms which are given below:

- **Level of significance:** It refers to the degree of significance in which we accept or reject the null hypothesis. 100% accuracy is not possible for accepting a hypothesis so we select a level of significance. This is normally denoted with  $\alpha$  and generally it is 0.05 or 5% which means your output should be 95% confident to give a similar kind of result in each sample.
- **P-value:** When analyzing data the [p-value](#) tells you the likelihood of seeing your result if the null hypothesis is true. If your P-value is less than the chosen significance level then you reject the null hypothesis otherwise accept it.
- **Test Statistic:** Test statistic is the number that helps you decide whether your result is significant. It's calculated from the sample data you collect it could be used to test if a machine learning model performs better than a random guess.
- **Critical value:** Critical value is a boundary or threshold that helps you decide if your test statistic is enough to reject the null hypothesis
- **Degrees of freedom:** [Degrees of freedom](#) are important when we conduct statistical tests they help you understand how much data can vary.

## Types of Hypothesis:



	Null Hypothesis is True	Null Hypothesis is False
Null Hypothesis is True (Accept)	Correct Decision	Type II Error (False Negative)
Alternative Hypothesis is True (Reject)	Type I Error (False Positive)	Correct Decision

### What exactly is a *p* value?

The ***p* value**, or probability value, tells you how likely it is that your data could have occurred under the null hypothesis. It does this by calculating the likelihood of your test statistic, which is the number calculated by a statistical test using your data.

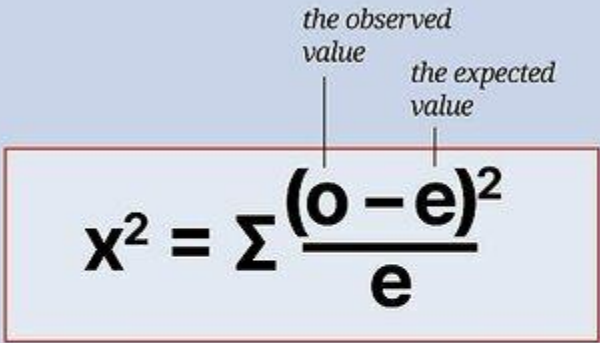
The *p* value tells you how often you would expect to see a test statistic as extreme or more extreme than the one calculated by your statistical test if the null hypothesis of that test was true. The *p* value gets smaller as the test statistic calculated from your data gets further away from the range of test statistics predicted by the null hypothesis.

The *p* value is a proportion: if your *p* value is 0.05, that means that 5% of the time you would see test statistics at least as extreme as the one you found if the null hypothesis was true.

### How to calculate *p*-value :

1. **Determine your experiment's *expected* results.** Usually, when scientists conduct an experiment and observe the results, they have an idea of what "normal" or "typical" results will look like beforehand. This can be based on past experimental results, trusted sets of observational data, scientific literature, and/or other sources. For your experiment, determine your expected results and express them as a number.
2. **Determine your experiment's *observed* results.** Now that you've determined your expected values, you can conduct your experiment and find your actual (or "observed") values. Again, express these results as numbers. If we manipulate some experimental condition and the observed results *differ* from this expected results, two possibilities are possible: either this happened by chance, or our manipulation of experimental variables *caused* the difference. The purpose of finding a *p*-value is basically to determine whether the observed results differ from the expected results to such a degree that the "null hypothesis" - the hypothesis that there is no relationship between the experimental variable(s) and the observed results - is unlikely enough to reject.
3. **Determine your experiment's *degrees of freedom*.** Degrees of freedom are a measure the amount of variability involved in the research, which is determined by the number of categories you are examining. The equation for degrees of freedom is **Degrees of freedom =  $n-1$** , where "*n*" is the number of categories or variables being analyzed in your experiment.

4. **Compare expected results to observed results with *chi square*.** Chi square (written " $\chi^2$ ") is a numerical value that measures the difference between an experiment's *expected* and *observed* values. The equation for chi square is:  $\chi^2 = \sum ((o - e)^2 / e)$ , where "o" is the observed value and "e" is the expected value.<sup>[4]</sup> Sum the results of this equation for all possible outcomes (see below).
- Note that this equation includes a  $\sum$  (sigma) operator. In other words, you'll need to calculate  $((o - e)^2 / e)$  for each possible outcome, then add the results to get your chi square value. In our example, we have two outcomes - either the car that received a ticket is red or blue. Thus, we would calculate  $((o - e)^2 / e)$  twice - once for red cars and once for blue cars.



The diagram shows the chi-squared test equation  $\chi^2 = \sum \frac{(o - e)^2}{e}$  enclosed in a red rectangular box. Above the box, two labels with vertical lines pointing to the variables in the equation are present: "the observed value" points to the 'o' in the numerator, and "the expected value" points to the 'e' in the numerator. Below the box, the text "chi-squared test" is written in a large, bold, black font. In the bottom right corner of the light blue background, the "wikiHow" logo is visible.

$$\chi^2 = \sum \frac{(o - e)^2}{e}$$

chi-squared test

5. **Choose a *significance level*.** Now that we know our experiment's degrees of freedom and our chi square value, there's just one last thing we need to do before we can find our p value - we need to decide on a significance level. Basically, the significance level is a measure of how certain we want to be about our results - low significance values correspond to a low probability that the experimental results happened by chance, and vice versa. Significance levels are written as a decimal (such as 0.01), which corresponds to the percent chance that random sampling would produce a difference as large as the one you observed if there was no underlying difference in the populations.

- It is a common misconception that  $p=0.01$  means that there is a 99% chance that the results were caused by the scientist's manipulation of experimental variables. This is NOT the case. If you wore your lucky pants on seven different days and the stock market went up every one of those days, you would have  $p<0.01$ , but you would still be well-justified in believing that the result had been generated by chance rather than by a connection between the market and your pants.
- By convention, scientists usually set the significance value for their experiments at 0.05, or 5 percent. This means that experimental results that meet this significance level have, at most, a 5% chance of being reproduced in a random sampling process. For most experiments, generating results that are that unlikely to be produced by a random sampling process is seen as "successfully" showing a correlation between the change in the experimental variable and the observed effect.
- Example: For our red and blue car example, let's follow scientific convention and set our significance level at **0.05**.

6. **Use a chi square distribution table to approximate your p-value.** Scientists and statisticians use large tables of values to calculate the p value for their experiment. These tables are generally set up with the vertical axis on the left corresponding to degrees of freedom and the horizontal axis on the top corresponding to p-value. Use these tables by first finding your degrees of freedom, then reading that row across from the left to the right until you find the first value *bigger* than your chi square value. Look at the corresponding p value at the top of the column - your p value is between this value and the next-largest value (the one immediately to the left of it).

- Chi square distribution tables are available from a variety of sources - they can easily be found online or in science and statistics textbooks. If you don't have one handy, use the one in the photo above or a free online table, like the one provided by [medcalc.org](http://medcalc.org) [here](#).
- Example: Our chi-square was 3. So, let's use the chi square distribution table in the photo above to find an approximate p value. Since we know our experiment has only **1** degree of freedom, we'll start in the highest row. We'll go from left to right along this row until we find a value higher than **3** - our chi square value. The first one we encounter is 3.84. Looking to the top of this column, we see that the corresponding p value is 0.05. This means that our p value is **between 0.05 and 0.1** (the next-biggest p value on the table).

7. **Decide whether to reject or keep your null hypothesis.** Since you have found an approximate p value for your experiment, you can decide whether or not to reject the null hypothesis of your experiment (as a reminder, this is the hypothesis that the experimental variables you manipulated did *not* affect the results you observed.) If your p value is lower than your significance value, congratulations - you've shown that your experimental results would be highly unlikely to occur if there was no real connection between the variables you manipulated and the effect you observed. If your p value is higher than your significance value, you can't confidently make that claim.[\[8\]](#)
- Example: Our p value is between 0.05 and 0.1 . It is not smaller than 0.05, so, unfortunately, we **can't reject our null hypothesis**. This means that we didn't reach the criterion we decided upon to be able to say that our town's police give tickets to red and blue cars at a rate that's significantly different than the national average.
  - In other words, random sampling from the national data would produce a result 10 tickets off from the national average 5-10% of the time. Since we were looking for this percentage to be less than 5%, we can't say that we're **sure** our town's police are less biased towards red cars.

### **What exactly is a confidence interval?**

A confidence interval is the [mean](#) of your estimate plus and minus the variation in that estimate. This is the range of values you expect your estimate to fall between if you redo your test, within a certain level of confidence.

**Confidence**, in statistics, is another way to describe probability. For example, if you construct a confidence interval with a 95% confidence level, you are confident that 95 out of 100 times the estimate will fall between the upper and lower values specified by the confidence interval.

Your desired confidence level is usually one minus the [alpha \( \$\alpha\$ \) value](#) you used in your [statistical test](#):

**Confidence level** =  $1 - \alpha$

So if you use an alpha value of  $p < 0.05$  for [statistical significance](#), then your confidence level would be  $1 - 0.05 = 0.95$ , or 95%.

**When do you use confidence intervals?**

You can calculate confidence intervals for many kinds of statistical estimates, including:

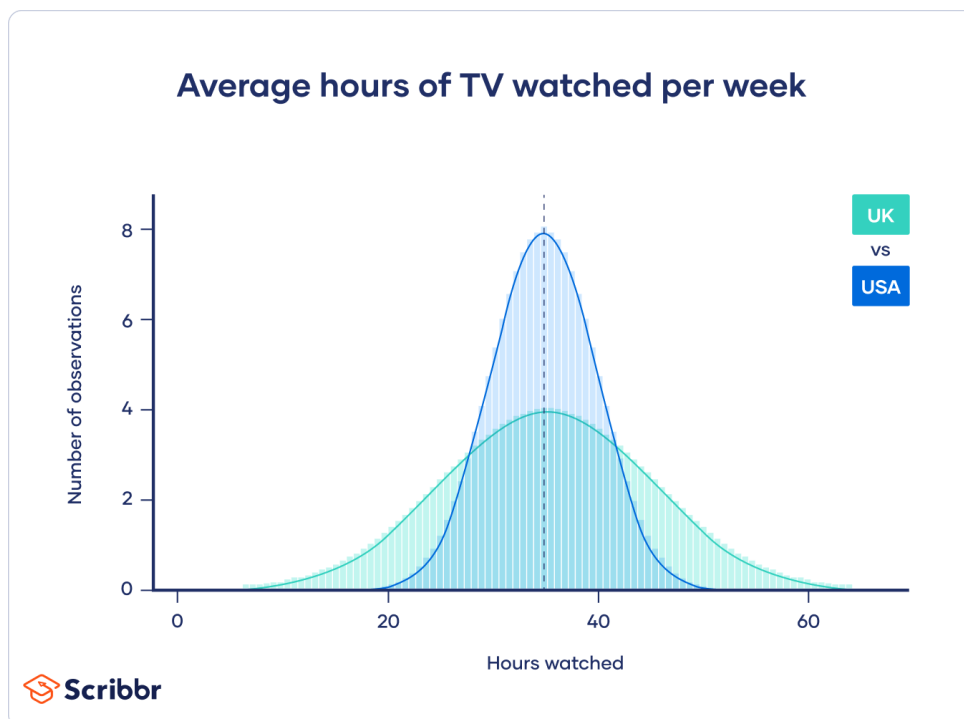
- Proportions
- Population means
- Differences between population means or proportions
- Estimates of variation among groups

These are all point estimates, and don't give any information about the variation around the number. Confidence intervals are useful for communicating the variation around a point estimate.

Example: Variation around an estimate You survey 100 Brits and 100 Americans about their television-watching habits, and find that both groups watch an average of 35 hours of television per week.

However, the British people [surveyed](#) had a wide variation in the number of hours watched, while the Americans all watched similar amounts.

Even though both groups have the same point estimate (average number of hours watched), the British estimate will have a wider confidence interval than the American estimate because there is more variation in the data.





## **How Does Regression Analysis Work?**

Regression analysis works by constructing a mathematical model that represents the relationships among the variables in question. This model is expressed as an equation that captures the expected influence of each independent variable on the dependent variable.

End-to-end, the regression analysis process consists of data collection and preparation, model selection, parameter estimation, and model evaluation.

### **Step 1: Data Collection and Preparation**

The first step in regression analysis involves gathering and preparing the data. As with any data analytics, data quality is imperative—in this context, preparation includes identifying all dependent and independent variables, cleaning the data, handling missing values, and transforming variables as needed.

### **Step 2: Model Selection**

In this step, the appropriate regression model is selected based on the nature of the data and the research question. For example, a simple linear regression is suitable when exploring a single predictor, while multiple linear regression is better for use cases with multiple predictors. Polynomial regression, logistic regression, and other specialized forms can be employed for various other use cases.

### **Step 3: Parameter Estimation**

The next step is to estimate the model parameters. For linear regression, this involves finding the coefficients (slopes and intercepts) that best fit the data. This is more often accomplished using techniques like the least squares method, which minimizes the sum of squared differences between observed and predicted values.

### **Step 4: Model Evaluation**

Model evaluation is critical for determining the model's goodness of fit and predictive accuracy. This process involves assessing such metrics as the coefficient of determination (R-squared), mean squared error (MSE), and others. Visualization tools—scatter plots and residual plots, for example—can aid in understanding how well the model captures the data's patterns.

