



College of  
Computer Studies

GOALBASE NOTES

iN

**BUSANA**

Business Analytics

## Course Description:

The course is an introduction to Business Analytics. It covers managerial statistical tools in descriptive analytics and predictive analytics, including regression. Other topics covered include forecasting, risk analysis, simulation, and data mining, and decision analysis. This course provides students with the fundamental concepts and tools needed to understand the emerging role of business analytics in organizations and shows students how to apply basic business analytics tools in a spreadsheet environment, and how to communicate with analytics professionals to effectively use and interpret analytic models and results for making better business decision. Emphasis is placed on applications, concepts and interpretation of results, rather than theory and calculations. Students use a computer software package for data analysis.

## Learning Outcomes:

- Select, understand and apply appropriate analytical tools in the analysis of quantitative and qualitative data from a variety of business scenarios.
- Use software package for data analysis; understand data gathering and input considerations; and be able to analyze and interpret output (graphs, tables, mathematical models, etc.)
- Know considerations in collecting data and selection of appropriate analysis tools; and know how to report results in a fair, objective and unbiased manner.

## Tools or Application to Use:

- |   |   |
|---|---|
| <ul style="list-style-type: none"><li>• Student Achievement Monitoring System (SAMS)</li><li>• Facebook Messenger</li></ul> | <ul style="list-style-type: none"><li>• Google Classroom</li><li>• Google Meet</li><li>• MS Excel</li></ul> |
|---|---|

## Mode of Assessment:

- |  |  |
|--|--|
| <ul style="list-style-type: none"><li>• Online Quiz</li><li>• Activities</li></ul> | <ul style="list-style-type: none"><li>• Project</li><li>• Presentation</li></ul> |
|--|--|

## References:

- <https://catalogimages.wiley.com/images/db/pdf/9781119668015.excerpt.pdf>
- <https://www.gs.washington.edu/academics/courses/akey/56008/lecture/lecture2.pdf>

# 1

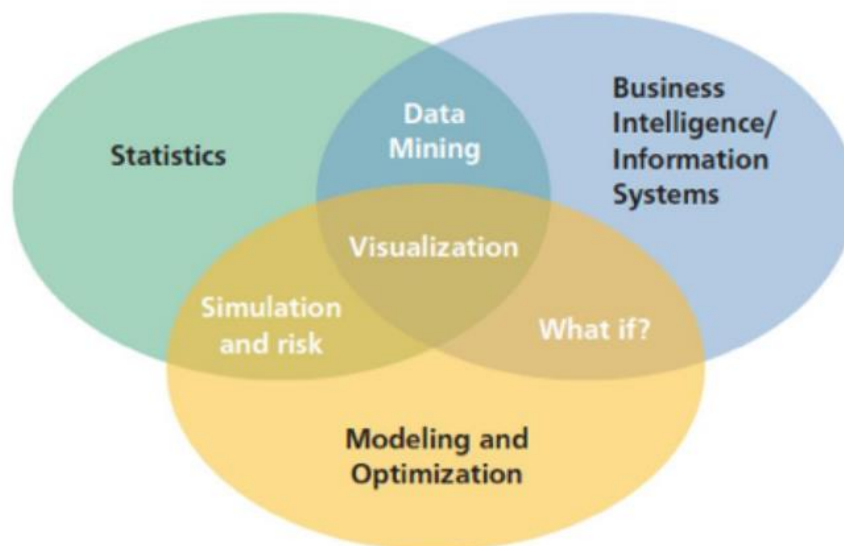
## BUSINESS ANALYTICS AND DESCRIPTIVE STATISTICAL MEASURES

### Business Analytics

**Analytics** is the use of data, information technology, statistical analysis, quantitative methods, and mathematical or computer-based models to help managers gain improved insight into their business operations and make better, fact-based decisions.

- Pricing
  - setting prices for consumer and industrial goods, government contracts, and maintenance contracts
- Customer segmentation
  - identifying and targeting key customer groups in the retail, insurance, and credit card industries
- Merchandising
  - determining brands to buy, quantities, and allocations
- Location
  - finding the best location for bank branches and ATMs, or where to service industrial equipment
- Social Media
  - understand trends and customer perceptions; assist marketing managers and product designers

### A Visual Perspective of Business Analytics



## **Impacts and Challenges**

- Benefits
  - ...reduced costs, better risk management, faster decisions, better productivity, and enhanced bottom-line performance such as profitability and customer satisfaction.
- Challenges
  - ...lack of understanding of how to use analytics, competing business priorities, insufficient analytical skills, difficulty in getting good data and sharing information, and not understanding the benefits versus perceived costs of analytics studies.

## **Scope of Business Analytics**

- **Descriptive analytics:** the use of data to understand past and current business performance and make informed decisions
- **Predictive analytics:** predict the future by examining historical data, detecting patterns or relationships in these data, and then extrapolating these relationships forward in time.
- **Prescriptive analytics:** identify the best alternatives to minimize or maximize some objective

### Example 1.1: Retail Markdown Decisions

- Most department stores clear seasonal inventory by reducing prices.
- Key question: When to reduce the price and by how much to maximize revenue?
- Potential applications of analytics:
  - Descriptive analytics: examine historical data for similar products (prices, units sold, advertising, ...)
  - Predictive analytics: predict sales based on price
  - Prescriptive analytics: find the best sets of pricing and advertising to maximize sales revenue

## **Descriptive and Inferential Statistics**

When analyzing data, such as the marks achieved by 100 students for a piece of coursework, it is possible to use both descriptive and inferential statistics in your analysis of their marks. Typically, in most research conducted on groups of people, you will use both descriptive and inferential statistics to analyze your results and draw conclusions. So what are descriptive and inferential statistics? And what are their differences?

### **Descriptive Statistics**

Descriptive statistics is the term given to the analysis of data that helps describe, show or summarize data in a meaningful way such that, for example, patterns might emerge from the data. Descriptive statistics do not, however, allow us to make conclusions beyond the data we have analyzed or reach conclusions regarding any hypotheses we might have made. They are simply a way to describe our data.

Descriptive statistics are very important because if we simply presented our raw data it would be hard to visualize what the data was showing, especially if there was a lot of it. Descriptive statistics, therefore, enables us to present the data in a more meaningful way, which allows a simpler interpretation of the data. For example, if we had the results of 100 pieces of students' coursework, we may be interested in the overall performance of those students. We would also be interested in the distribution or spread of the marks. Descriptive statistics allow us to do this. How to properly describe data through statistics and graphs is an important topic and discussed in other Laerd Statistics guides. Typically, there are two general types of statistics that are used to describe data:

- **Measures of central tendency:** these are ways of describing the central position of a frequency distribution for a group of data. In this case, the frequency distribution is simply the distribution and pattern of marks scored by the 100 students from the lowest to the highest. We can describe this central position using several statistics, including the mode, median, and mean. You can learn more in our guide: Measures of Central Tendency.
- **Measures of spread:** these are ways of summarizing a group of data by describing how to spread out the scores are. For example, the mean score of our 100 students maybe 65 out of 100. However, not all students will have scored 65 marks. Rather, their scores will be spread out. Some will be lower and others higher. Measures of spread help us to summarize how spread out these scores are. To describe this spread, several statistics are available to us, including the range, quartiles, absolute deviation, variance, and standard deviation.

When we use descriptive statistics it is useful to summarize our group of data using a combination of tabulated description (i.e., tables), graphical description (i.e., graphs and charts), and statistical commentary (i.e., a discussion of the results).

## Inferential Statistics

We have seen that descriptive statistics provide information about our immediate group of data. For example, we could calculate the mean and standard deviation of the exam marks for the 100 students and this could provide valuable information about this group of 100 students. Any group of data like this, which includes all the data you are interested in, is called a population. A population can be small or large, as long as it includes all the data you are interested in. For example, if you were only interested in the exam marks of 100 students, the 100 students would represent your population. Descriptive statistics are applied to populations, and the properties of populations, like the mean or standard deviation, are called parameters as they represent the whole population (i.e., everybody you are interested in).

Often, however, you do not have access to the whole population you are interested in investigating, but only a limited number of data instead. For example, you might be interested in the exam marks of all students in the UK. It is not feasible to measure all exam marks of all students in the whole of the UK so you have to measure a smaller sample of students (e.g., 100 students), which are used to represent the larger population of all UK students. Properties of samples, such as the mean or standard deviation, are not called parameters, but statistics. Inferential statistics are techniques that allow us to use these samples to make generalizations about the populations from which the samples were drawn. It is, therefore, important that the sample accurately represents the population. The process of achieving this is called sampling (sampling strategies are discussed in detail in the section, Sampling Strategy, on our sister site). Inferential statistics arise out of the fact that sampling naturally incurs sampling error and thus a sample is not expected to perfectly represent the population. The methods of inferential statistics are (1) the estimation of the parameter(s) and (2) the testing of statistical hypotheses.

### What are the similarities between descriptive and inferential statistics?

Both descriptive and inferential statistics rely on the same set of data. Descriptive statistics rely solely on this set of data, whilst inferential statistics also rely on this data to make generalizations about a larger population.

### What are the strengths of using descriptive statistics to examine the distribution of scores?

Other than the clarity with which descriptive statistics can clarify large volumes of data, there are no uncertainties about the values you get (other than only measurement error, etc.).

### **What are the limitations of descriptive statistics?**

Descriptive statistics are limited so much that they only allow you to make summations about the people or objects that you have measured. You cannot use the data you have collected to generalize to other people or objects (i.e., using data from a sample to infer the properties/parameters of a population). For example, if you tested a drug to beat cancer and it worked in your patients, you cannot claim that it would work in other cancer patients only relying on descriptive statistics (but inferential statistics would give you this opportunity).

### **What are the limitations of inferential statistics?**

There are two main limitations to the use of inferential statistics. The first, and most important limitation, which is present in all inferential statistics, is that you are providing data about a population that you have not fully measured, and therefore, cannot ever be completely sure that the values/statistics you calculate are correct. Remember, inferential statistics are based on the concept of using the values measured in a sample to estimate/infer the values that would be measured in a population; there will always be a degree of uncertainty in doing this. The second limitation is connected with the first limitation. Some, but not all, inferential tests require the user (i.e., you) to make educated guesses (based on theory) to run the inferential tests. Again, there will be some uncertainty in this process, which will have repercussions on the certainty of the results of some inferential statistics.

# 2

## EXCEL TIPS, TRICKS AND SHORTCUTS

### **What is Excel?**

Microsoft Excel is powerful data visualization and analysis software, which uses spreadsheets to store, organize, and track data sets with formulas and functions. Excel is used by marketers, accountants, data analysts, and other professionals. It's part of the Microsoft Office suite of products. Alternatives include Google Sheets and Numbers.

Excel is used to store, analyze, and report large amounts of data. It is often used by accounting teams for financial analysis but can be used by any professional to manage long and unwieldy datasets. Examples of Excel applications include balance sheets, budgets, and editorial calendars.

Excel is primarily used for creating financial documents because of its strong computational powers. You'll often find the software in accounting offices and teams because it allows accountants to automatically see sums, averages, and totals. With Excel, they can easily make sense of their business data.

While Excel is primarily known as an accounting tool, professionals in any field can use its features and formulas — especially marketers — because it can be used for tracking any type of data. It removes the need to spend hours and hours counting cells or copying and pasting performance numbers. Excel typically has a shortcut or quick fix that speeds up the process.

### **Excel Basics**

If you're just starting with Excel, there are a few basic commands that we suggest you become familiar with. These are things like:

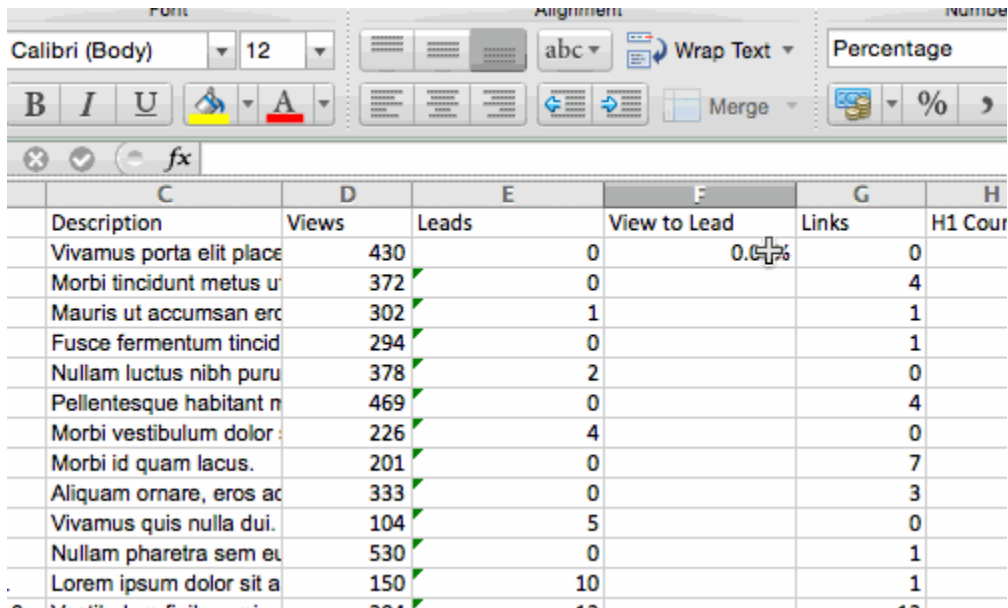
- Creating a new spreadsheet from scratch.
- Executing basic computations like adding, subtracting, multiplying, and dividing.
- Writing and formatting column text and titles.
- Using Excel's auto-fill features.
- Adding or deleting single columns, rows, and spreadsheets. (Below, we'll get into how to add things like multiple columns and rows.)
- Keeping column and row titles visible as you scroll past them in a spreadsheet, so that you know what data you're filling as you move further down the document.
- Sorting your data in alphabetical order.

Let's explore a few of these more in-depth.

For instance, why does auto-fill matter?

If you have any basic Excel knowledge, it's likely you already know this quick trick. But to cover our bases, allow me to show you the glory of autofill. This lets you quickly fill adjacent cells with several types of data, including values, series, and formulas.

There are multiple ways to deploy this feature, but the fill handle is among the easiest. Select the cells you want to be the source, locate the fill handle in the lower-right corner of the cell, and either drag the fill handle to cover cells you want to fill or just double click:



The screenshot shows the Microsoft Excel interface. The ribbon at the top includes the 'Font' tab (Calibri, size 12), the 'Alignment' tab (left-align, center-align, right-align, justify-align, wrap text, merge), and the 'Number' tab (percentage, currency, decimal). Below the ribbon is a formula bar with 'fx'. The main area displays a table with 6 columns (C-H) and 12 rows. The first row is a header, and the subsequent rows contain data. The 'View to Lead' column (F) has a value of 0.04% in the first row, and a small crosshair cursor is visible over it.

Description	Views	Leads	View to Lead	Links	H1 Cour
Vivamus porta elit place	430	0	0.04%	0	
Morbi tincidunt metus u	372	0		4	
Mauris ut accumsan erc	302	1		1	
Fusce fermentum tincid	294	0		1	
Nullam luctus nibh puru	378	2		0	
Pellentesque habitant n	469	0		4	
Morbi vestibulum dolor	226	4		0	
Morbi id quam lacus.	201	0		7	
Aliquam ornare, eros ac	333	0		3	
Vivamus quis nulla dui.	104	5		0	
Nullam pharetra sem eu	530	0		1	
Lorem ipsum dolor sit a	150	10		1	

excel autofill Similarly, sorting is an important feature you'll want to know when organizing your data in Excel.

Sometimes you may have a list of data that has no organization whatsoever. Maybe you exported a list of your marketing contacts or blog posts. Whatever the case may be, Excel's sort feature will help you alphabetize any list.

Click on the data in the column you want to sort. Then click on the "Data" tab in your toolbar and look for the "Sort" option on the left. If the "A" is on top of the "Z," you can just click on that button once. If the "Z" is on top of the "A," click on the button twice. When the "A" is on top of the "Z," that means your list will be sorted in alphabetical order. However, when the "Z" is on top of the "A," that means your list will be sorted in reverse alphabetical order.

Let's explore more of the basics of Excel (along with advanced features) next.

### **How to Use Excel**

To use Excel, you only need to input the data into the rows and columns. And then you'll use formulas and functions to turn that data into insights.

We're going to go over the best formulas and functions you need to know. But first, let's take a look at the types of documents you can create using the software. That way, you have an overarching understanding of how you can use Excel in your day-to-day.



## Documents You Can Create in Excel

Not sure how you can use Excel in your team? Here is a list of documents you can create:

- **Income Statements:** You can use an Excel spreadsheet to track a company's sales activity and financial health.
- **Balance Sheets:** Balance sheets are among the most common types of documents you can create with Excel. It allows you to get a holistic view of a company's financial standing.
- **Calendar:** You can easily create a spreadsheet monthly calendar to track events or other date-sensitive information.

Here are some documents you can create specifically for marketers.

- **Marketing Budgets:** Excel is a strong budget-keeping tool. You can create and track marketing budgets, as well as spending, using Excel. If you don't want to create a document from scratch.
- **Marketing Reports:** If you don't use a marketing tool such as Marketing Hub, you might find yourself in need of a dashboard with all of your reports. Excel is an excellent tool to create marketing reports.
- **Editorial Calendars:** You can create editorial calendars in Excel. The tab format makes it extremely easy to track your content creation efforts for custom time ranges.
- **Traffic and Leads Calculator:** Because of its strong computational powers, Excel is an excellent tool to create all sorts of calculators — including one for tracking leads and traffic.

This is only a small sampling of the types of marketing and business documents you can create in Excel. We've created an extensive list of Excel templates you can use right now for marketing, invoicing, project management, budgeting, and more.

In the spirit of working more efficiently and avoiding tedious, manual work, here are a few Excel formulas and functions you'll need to know.

## Excel Formulas

It's easy to get overwhelmed by the wide range of Excel formulas that you can use to make sense of your data. If you're just getting started using Excel, you can rely on the following formulas to carry out some complex functions — without adding to the complexity of your learning path.

- **Equal sign:** Before creating any formula, you'll need to write an equal sign (=) in the cell where you want the result to appear.
- **Addition:** To add the values of two or more cells, use the + sign. Example: =C5+D3.
- **Subtraction:** To subtract the values of two or more cells, use the - sign. Example: =C5-D3.
- **Multiplication:** To multiply the values of two or more cells, use the \* sign. Example: =C5\*D3.
- **Division:** To divide the values of two or more cells, use the / sign. Example: =C5/D3.

Putting all of these together, you can create a formula that adds, subtracts, multiplies, and divides all in one cell. Example: =(C5-D3)/((A5+B6)\*3).

For more complex formulas, you'll need to use parentheses around the expressions to avoid accidentally using the PEMDAS order of operations. Keep in mind that you can use plain numbers in your formulas.

## Excel Functions

Excel functions automate some of the tasks you would use in a typical formula. For instance, instead of using the + sign to add up a range of cells, you'd use the SUM function. Let's look at a few more functions that will help automate calculations and tasks.

- **SUM:** The SUM function automatically adds up a range of cells or numbers. To complete a sum, you would input the starting cell and the final cell with a colon in between. Here's what that looks like SUM(Cell1:Cell2). Example: =SUM(C5:C30).
- **AVERAGE:** The AVERAGE function averages out the values of a range of cells. The syntax is the same as the SUM function: AVERAGE(Cell1:Cell2). Example: =AVERAGE(C5:C30).
- **IF:** The IF function allows you to return values based on a logical test. The syntax is as follows: IF(logical\_test, value\_if\_true, [value\_if\_false]). Example: =IF(A2>B2,"Over Budget","OK").
- **VLOOKUP:** The VLOOKUP function helps you search for anything on your sheet's rows. The syntax is VLOOKUP(lookup value, table array, column number, Approximate match (TRUE), or Exact match (FALSE)). Example: =VLOOKUP([@Attorney],tbl\_Attorneys,4,FALSE).
- **INDEX:** The INDEX function returns a value from within a range. The syntax is as follows: INDEX(array, row\_num, [column\_num]).
- **MATCH:** The MATCH function looks for a certain item in a range of cells and returns the position of that item. It can be used in tandem with the INDEX function. The syntax is: MATCH(lookup\_value, lookup\_array, [match\_type]).
- **COUNTIF:** The COUNTIF function returns the number of cells that meet certain criteria or have a certain value. The syntax is COUNTIF(range, criteria). Example: =COUNTIF(A2:A5,"London").

## Excel Tips

### Use Pivot tables to recognize and make sense of data.

Pivot tables are used to reorganize data in a spreadsheet. They won't change the data that you have, but they can sum up values and compare different information in your spreadsheet, depending on what you'd like them to do.

Let's take a look at an example. Let's say I want to take a look at how many people are in each house at Hogwarts. You may be thinking that I don't have too much data, but for longer data sets, this will come in handy.

To create the Pivot Table, I go to Data > Pivot Table. If you're using the most recent version of Excel, you'd go to Insert > Pivot Table. Excel will automatically populate your Pivot Table, but you can always change around the order of the data. Then, you have four options to choose from.

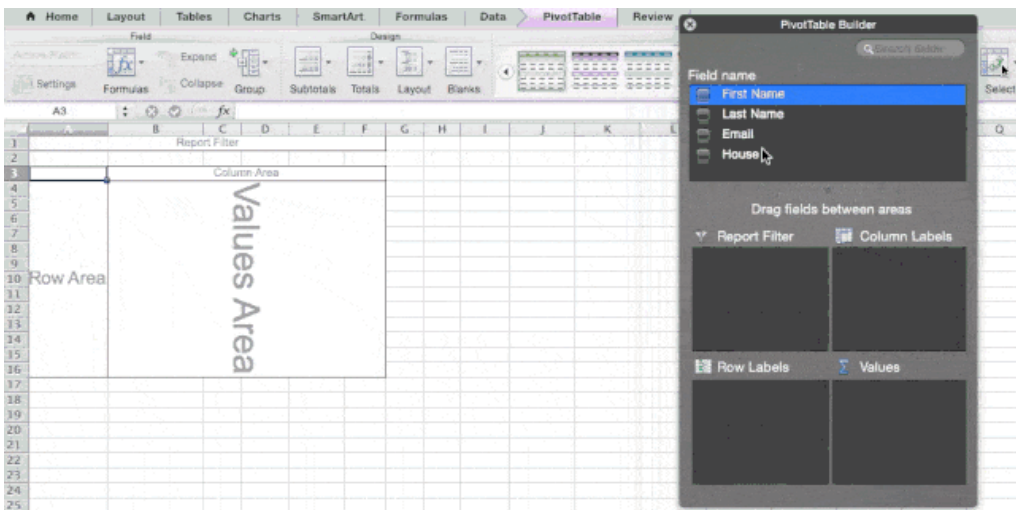
**Report Filter:** This allows you to only look at certain rows in your dataset. For example, if I wanted to create a filter by house, I could choose to only include students in Gryffindor instead of all students.

**Column Labels:** These would be your headers in the dataset.

**Row Labels:** These could be your rows in the dataset. Both Row and Column labels can contain data from your columns (e.g. First Name can be dragged to either the Row or Column label — it just depends on how you want to see the data.)

**Value:** This section allows you to look at your data differently. Instead of just pulling in any numeric value, you can sum, count, average, max, min, count numbers, or do a few other manipulations with your data. In fact, by default, when you drag a field to Value, it always does a count.

Since I want to count the number of students in each house, I'll go to the Pivot table builder and drag the House column to both the Row Labels and the Values. This will sum up the number of students associated with each house.



### Add more than one row or column.

As you play around with your data, you might find you're constantly needing to add more rows and columns. Sometimes, you may even need to add hundreds of rows. Doing this one-by-one would be super tedious. Luckily, there's always an easier way.

To add multiple rows or columns in a spreadsheet, highlight the same number of preexisting rows or columns that you want to add. Then, right-click and select "Insert."

In the example below, I want to add three rows. By highlighting three rows and then clicking insert, I'm able to add three blank rows into my spreadsheet quickly and easily.

	A	B	C	D	E
1	First Name	Last Name	Email	House	
2	Harry	Potter	hpotter@hogwarts.edu	Gryffindor	
3	Hermione	Granger	hgranger@hogwarts.edu	Gryffindor	
4			rweasley@hogwarts.edu	Gryffindor	
5			dmalfoy@hogwarts.edu	Slytherin	
6			cchang@hogwarts.edu	Ravenclaw	
7			llovegood@hogwarts.edu	Ravenclaw	
8			ntonks@hogwarts.edu	Hufflepuff	
9			habbott@hogwarts.edu	Hufflepuff	
10					
11					
12					
13					
14					

### Use filters to simplify your data.

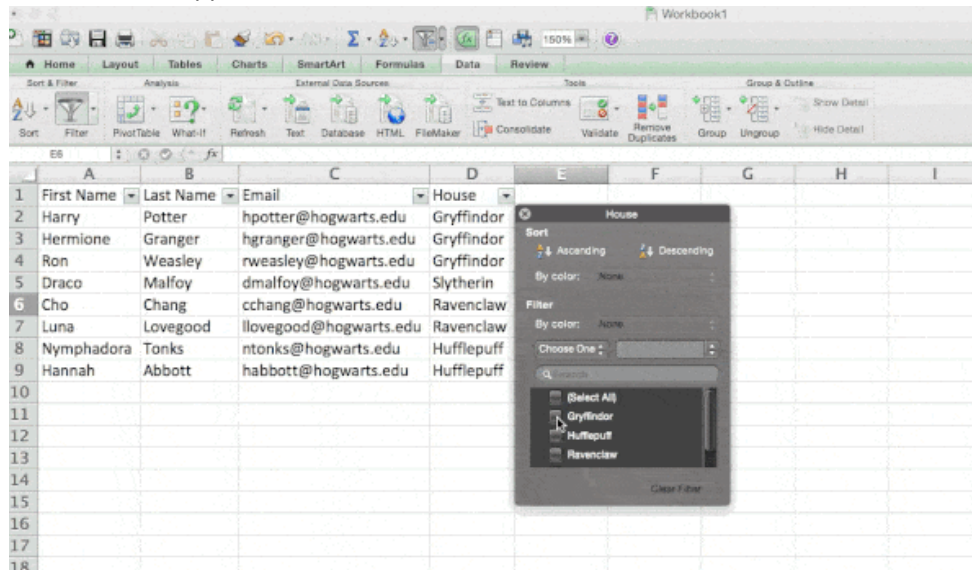
When you're looking at very large data sets, you don't usually need to be looking at every single row at the same time. Sometimes, you only want to look at data that fit into certain criteria.

That's where filters come in.

Filters allow you to pare down your data to only look at certain rows at one time. In Excel, a filter can be added to each column in your data — and from there, you can then choose which cells you want to view at once.

Let's take a look at the example below. Add a filter by clicking the Data tab and selecting "Filter." Clicking the arrow next to the column headers and you'll be able to choose whether you want your data to be organized in ascending or descending order, as well as which specific rows you want to show.

In my Harry Potter example, let's say I only want to see the students in Gryffindor. By selecting the Gryffindor filter, the other rows disappear.

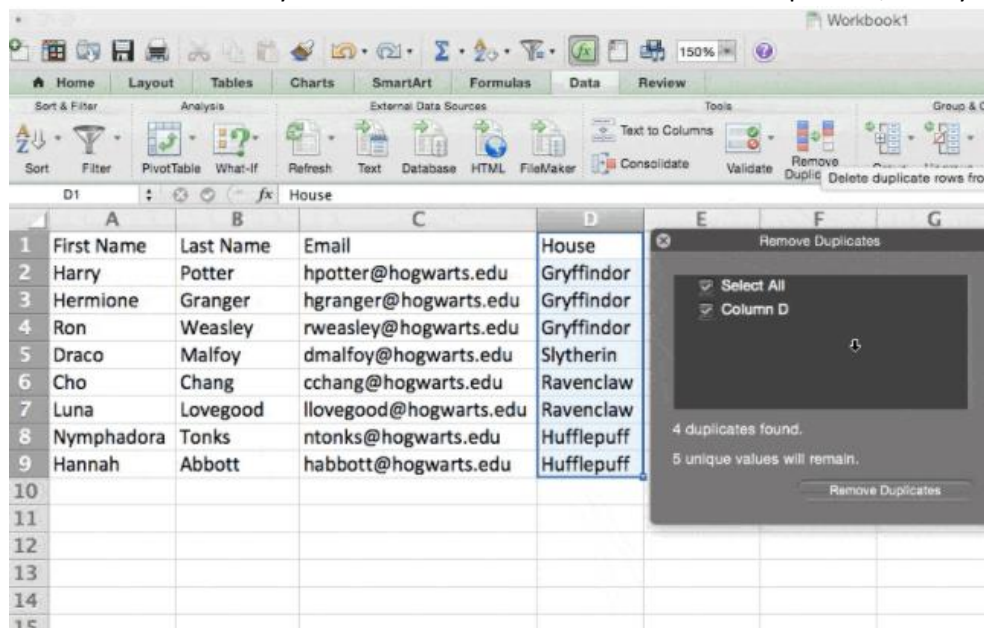


Pro Tip: Copy and paste the values in the spreadsheet when a Filter is on to do additional analysis in another spreadsheet.

### Remove duplicate data points or sets.

Larger data sets tend to have duplicate content. You may have a list of multiple contacts in a company and only want to see the number of companies you have. In situations like this, removing the duplicates comes in quite handy.

To remove your duplicates, highlight the row or column that you want to remove duplicates of. Then, go to the Data tab and select "Remove Duplicates" (which is under the Tools subheader in the older version of Excel). A pop-up will appear to confirm which data you want to work with. Select "Remove Duplicates," and you're good to go.

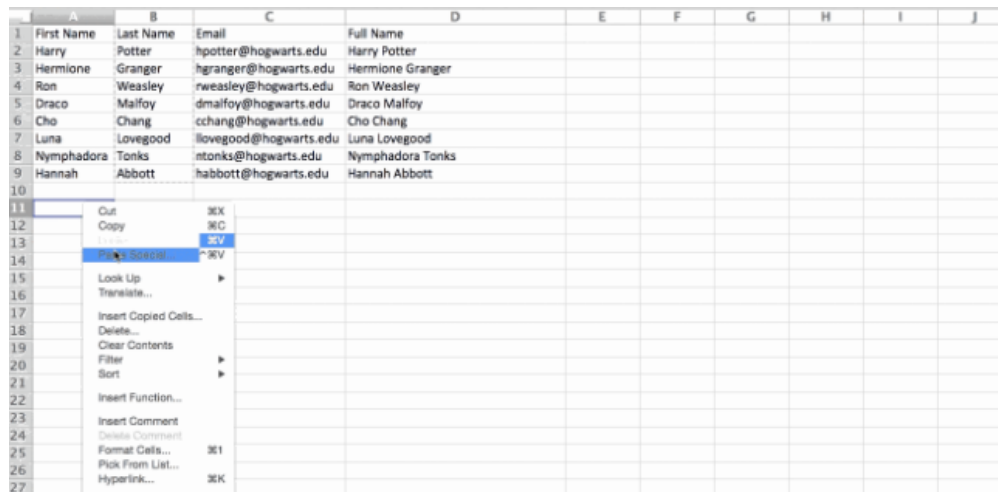


You can also use this feature to remove an entire row based on a duplicate column value. So if you have three rows with Harry Potter's information and you only need to see one, then you can select the whole dataset and then remove duplicates based on email. Your resulting list will have unique names without any duplicates.

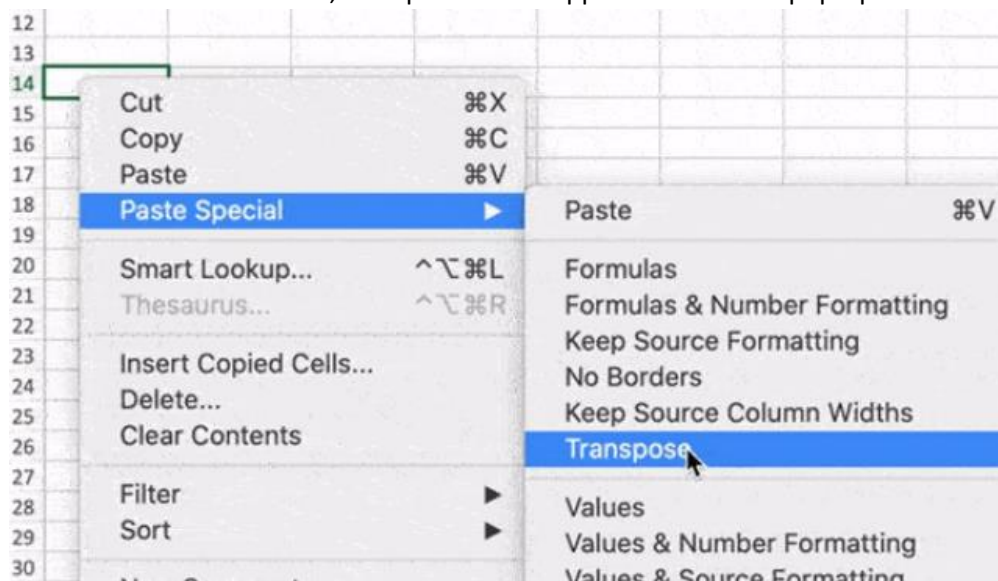
### Transpose rows into columns.

When you have rows of data in your spreadsheet, you might decide you want to transform the items in one of those rows into columns (or vice versa). It would take a lot of time to copy and paste each header — but what the transpose feature allows you to do is simply move your row data into columns, or the other way around.

Start by highlighting the column that you want to transpose into rows. Right-click it, and then select "Copy." Next, select the cells on your spreadsheet where you want your first row or column to begin. Right-click on the cell, and then select "Paste Special." A module will appear — at the bottom, you'll see an option to transpose. Check that box and select OK. Your column will now be transferred to a row or vice-versa.



On newer versions of Excel, a drop-down will appear instead of a pop-up.



## Split up text information between columns.

What if you want to split out information that's in one cell into two different cells? For example, maybe you want to pull out someone's company name through their email address. Or perhaps you want to separate someone's full name into a first and last name for your email marketing templates.

Thanks to Excel, both are possible. First, highlight the column that you want to split up. Next, go to the Data tab and select "Text to Columns." A module will appear with additional information.

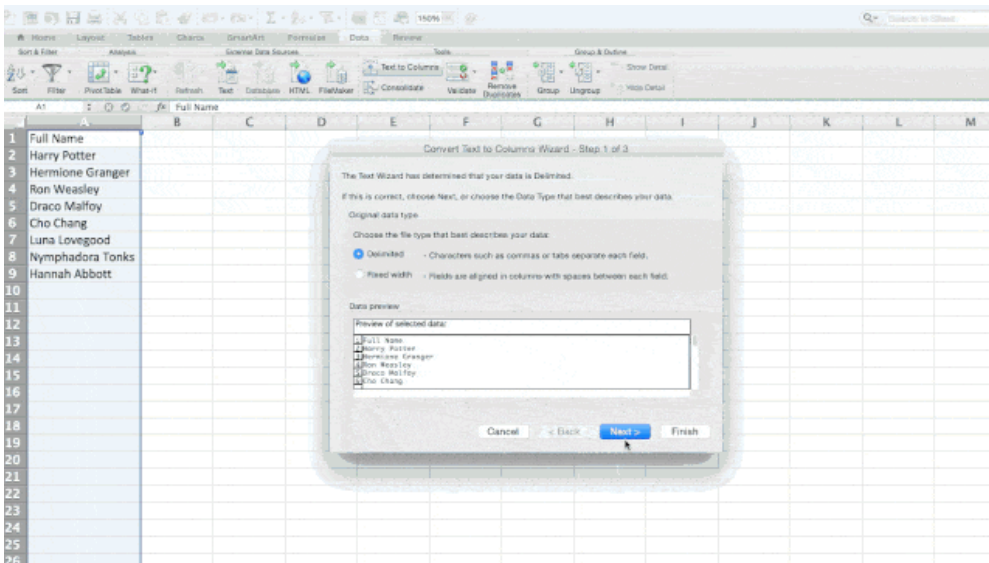
First, you need to select either "Delimited" or "Fixed Width."

- "Delimited" means you want to break up the column based on characters such as commas, spaces, or tabs.
- "Fixed Width" means you want to select the exact location on all the columns that you want the split to occur.

In the example case below, let's select "Delimited" so we can separate the full name into first name and last name.

Then, it's time to choose the Delimiters. This could be a tab, semi-colon, comma, space, or something else. ("Something else" could be the "@" sign used in an email address, for example.) In our example, let's choose the space. Excel will then show you a preview of what your new columns will look like.

When you're happy with the preview, press "Next." This page will allow you to select Advanced Formats if you choose to. When you're done, click "Finish."



## Use formulas for simple calculations.

In addition to doing pretty complex calculations, Excel can help you do simple arithmetic like adding, subtracting, multiplying, or dividing any of your data.

- To add, use the + sign.
- To subtract, use the - sign.
- To multiply, use the \* sign.
- To divide, use the / sign.

You can also use parentheses to ensure certain calculations are done first. In the example below  $(10+10*10)$ , the second and third 10 were multiplied together before adding the additional 10. However, if we made it  $(10+10)*10$ , the first and second 10 would be added together first.



	A	B	C	D	E
1	First Name	Last Name	Email	House	House Points
2	Harry	Potter	hpotter@hogwarts.edu	Gryffindor	10
3	Hermione	Granger	hgranger@hogwarts.edu	Gryffindor	10
4	Ron	Weasley	rweasley@hogwarts.edu	Gryffindor	10
5	Draco	Malfoy	dmalfoy@hogwarts.edu	Slytherin	0
6	Cho	Chang	cchang@hogwarts.edu	Ravenclaw	0
7	Luna	Lovegood	llovegood@hogwarts.edu	Ravenclaw	0
8	Nymphadora	Tonks	ntonks@hogwarts.edu	Hufflepuff	0
9	Hannah	Abbott	habbott@hogwarts.edu	Hufflepuff	0
10					
11	Gryffindor	=E2+E3*			
12	Slytherin				
13	Ravenclaw				
14	Hufflepuff				

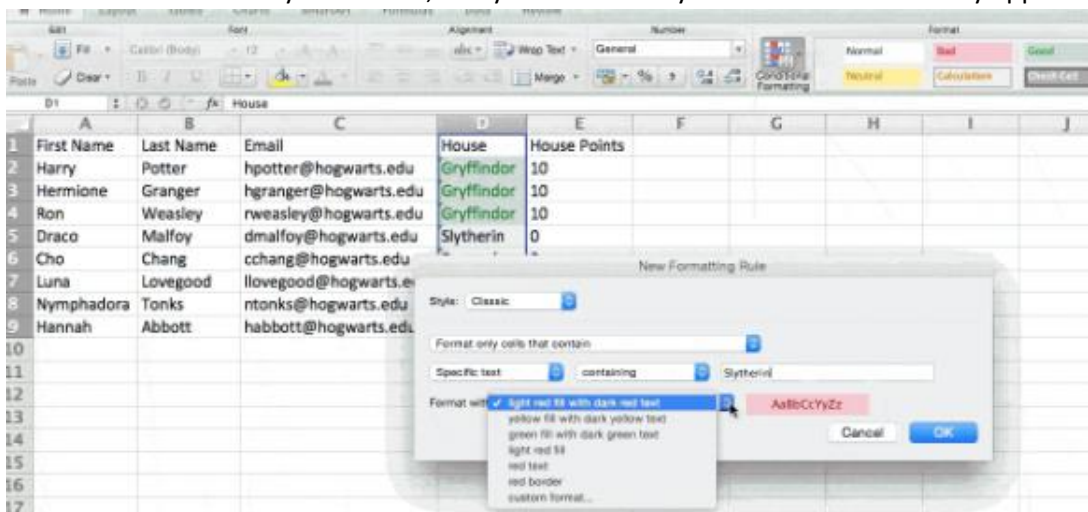
### Get the average of numbers in your cells.

If you want the average of a set of numbers, you can use the formula =AVERAGE(Cell1:Cell2). If you want to sum up a column of numbers, you can use the formula =SUM(Cell1:Cell2).

### Use conditional formatting to make cells automatically change color based on data.

Conditional formatting allows you to change a cell's color based on the information within the cell. For example, if you want to flag certain numbers that are above average or in the top 10% of the data in your spreadsheet, you can do that. If you want to color code commonalities between different rows in Excel, you can do that. This will help you quickly see information that is important to you.

To get started, highlight the group of cells you want to use conditional formatting on. Then, choose "Conditional Formatting" from the Home menu and select your logic from the dropdown. (You can also create your own rule if you want something different.) A window will pop up that prompts you to provide more information about your formatting rule. Select "OK" when you're done, and you should see your results automatically appear.



### Use the IF Excel formula to automate certain Excel functions.

Sometimes, we don't want to count the number of times a value appears. Instead, we want to input different information into a cell if there is a corresponding cell with that information.

For example, in the situation below, I want to award ten points to everyone who belongs in the Gryffindor house. Instead of manually typing in 10's next to each Gryffindor student's name, I can use the IF Excel formula to say that if the student is in Gryffindor, then they should get ten points.

The formula is: `IF(logical_test, value_if_true, [value_if_false])`

Example Shown Below: `=IF(D2="Gryffindor","10","0")`

In general terms, the formula would be `IF(Logical Test, value of true, value of false)`. Let's dig into each of these variables.

- **Logical\_Test:** The logical test is the "IF" part of the statement. In this case, the logic is `D2="Gryffindor"` because we want to make sure that the cell corresponding with the student says "Gryffindor." Make sure to put Gryffindor in quotation marks here.
- **Value\_if\_True:** This is what we want the cell to show if the value is true. In this case, we want the cell to show "10" to indicate that the student was awarded 10 points. Only use quotation marks if you want the result to be text instead of a number.
- **Value\_if\_False:** This is what we want the cell to show if the value is false. In this case, for any student not in Gryffindor, we want the cell to show "0". Only use quotation marks if you want the result to be text instead of a number.

Excel IF formula in action

Note: In the example above, I awarded 10 points to everyone in Gryffindor. If I later wanted to sum the total number of points, I wouldn't be able to because the 10's are in quotes, thus making them text and not a number that Excel can sum.

The real power of the IF function comes when you string multiple IF statements together or nest them. This allows you to set multiple conditions, get more specific results, and ultimately organize your data into more manageable chunks.

Ranges are one way to segment your data for better analysis. For example, you can categorize data into values that are less than 10, 11 to 50, or 51 to 100. Here's how that looks in practice:

`=IF(B3<11,"10 or less",IF(B3<51,"11 to 50",IF(B3<100,"51 to 100")))`

It can take some trial-and-error, but once you have the hang of it IF formulas will become your new Excel best friend.

### **Use dollar signs to keep one cell's formula the same regardless of where it moves.**

Have you ever seen a dollar sign in an Excel formula? When used in a formula, it isn't representing an American dollar; instead, it makes sure that the exact column and row are held the same even if you copy the same formula in adjacent rows.

You see, a cell reference — when you refer to cell A5 from cell C5, for example — is relative by default. In that case, you're referring to a cell that's five columns to the left (C minus A) and in the same row (5). This is called a relative formula. When you copy a relative formula from one cell to another, it'll adjust the values in the formula based on where it's moved. But sometimes, we want those values to stay the same no matter whether they're moved around or not — and we can do that by turning the formula into an absolute formula.



To change the relative formula ( $=A5+C5$ ) into an absolute formula, we'd precede the row and column values by dollar signs, like this: ( $=\$A\$5+\$C\$5$ ). (Learn more on Microsoft Office's support page [here](#).)

### **Use the VLOOKUP function to pull data from one area of a sheet to another.**

Have you ever had two sets of data on two different spreadsheets that you want to combine into a single spreadsheet?

For example, you might have a list of people's names next to their email addresses in one spreadsheet, and a list of those same people's email addresses next to their company names in the other — but you want the names, email addresses, and company names of those people to appear in one place.

I have to combine data sets like this a lot — and when I do, the VLOOKUP is my go-to formula.

Before you use the formula, though, be sure that you have at least one column that appears identically in both places. Scour your data sets to make sure the column of data you're using to combine your information is the same, including no extra spaces.

The formula: `=VLOOKUP(lookup value, table array, column number, Approximate match (TRUE) or Exact match (FALSE))`

The formula with variables from our example below: `=VLOOKUP(C2, Sheet2!A: B,2, FALSE)`

In this formula, there are several variables. The following is true when you want to combine the information in Sheet 1 and Sheet 2 onto Sheet 1.

- **Lookup Value:** This is the identical value you have in both spreadsheets. Choose the first value in your first spreadsheet. In the example that follows, this means the first email address on the list, or cell 2 (C2).
- **Table Array:** The table array is the range of columns on Sheet 2 you're going to pull your data from, including the column of data identical to your lookup value (in our example, email addresses) in Sheet 1 as well as the column of data you're trying to copy to Sheet 1. In our example, this is "Sheet2!A: B." "A" means Column A in Sheet 2, which is the column in Sheet 2 where the data identical to our lookup value (email) in Sheet 1 is listed. The "B" means Column B, which contains the information that's only available in Sheet 2 that you want to translate to Sheet 1.
- **Column Number:** This tells Excel which column the new data you want to copy to Sheet 1 is located in. In our example, this would be the column that "House" is located in. "House" is the second column in our range of columns (table array), so our column number is 2. [Note: Your range can be more than two columns. For example, if there are three columns on Sheet 2 — Email, Age, and House — and you still want to bring House onto Sheet 1, you can still use a VLOOKUP. You just need to change the "2" to a "3" so it pulls back the value in the third column: `=VLOOKUP(C2:Sheet2!A: C,3, false).`]
- **Approximate Match (TRUE) or Exact Match (FALSE):** Use FALSE to ensure you pull in only exact value matches. If you use TRUE, the function will pull in approximate matches.

In the example below, Sheet 1 and Sheet 2 contain lists describing different information about the same people, and the common thread between the two is their email addresses. Let's say we want to combine both datasets so that all the house information from Sheet 2 translates over to Sheet 1.

	A	B	C	D	E	F	G
1	First Name	Last Name	Email	House			
2	Harry	Potter	hpotter@hogwarts.edu	Gryffindor			
3	Hermione	Granger	hgranger@hogwarts.edu				
4	Ron	Weasley	rweasley@hogwarts.edu				
5	Draco	Malfoy	dmalfoy@hogwarts.edu				
6	Cho	Chang	cchang@hogwarts.edu				
7	Luna	Lovegood	llovegood@hogwarts.edu				
8	Nymphador	Tonks	ntonks@hogwarts.edu				
9	Hannah	Abbott	habbott@hogwarts.edu				
10							

So when we type in the formula `=VLOOKUP(C2, Sheet2!A: B,2, FALSE)`, we bring all the house data into Sheet 1.

Keep in mind that VLOOKUP will only pull back values from the second sheet that are to the right of the column containing your identical data. This can lead to some limitations, which is why some people prefer to use the INDEX and MATCH functions instead.

### Use INDEX and MATCH formulas to pull data from horizontal columns.

Like VLOOKUP, the INDEX and MATCH functions pull data from another dataset into one central location. Here are the main differences:

- VLOOKUP is a much simpler formula. If you're working with large data sets that would require thousands of lookups, using the INDEX and MATCH function will significantly decrease load time in Excel.
- The INDEX and MATCH formulas work right-to-left, whereas VLOOKUP formulas only work as a left-to-right lookup. In other words, if you need to do a lookup that has a lookup column to the right of the results column, then you'd have to rearrange those columns to do a VLOOKUP. This can be tedious with large datasets and/or lead to errors.

So if I want to combine the information in Sheet 1 and Sheet 2 onto Sheet 1, but the column values in Sheets 1 and 2 aren't the same, then to do a VLOOKUP, I would need to switch around my columns. In this case, I'd choose to do an INDEX and MATCH instead.

Let's look at an example. Let's say Sheet 1 contains a list of people's names and their Hogwarts email addresses, and Sheet 2 contains a list of people's email addresses and the Patronus that each student has. (For the non-Harry Potter fans out there, every witch or wizard has an animal guardian called a "Patronus" associated with him or her.) The information that lives in both sheets is the column containing email addresses, but this email address column is in different column numbers on each sheet. I'd use the INDEX and MATCH formulas instead of VLOOKUP so I wouldn't have to switch any columns around.

So what's the formula, then? The formula is the MATCH formula nested inside the INDEX formula. You'll see I differentiated the MATCH formula using a different color here.

The formula: `=INDEX(table array, MATCH formula)`

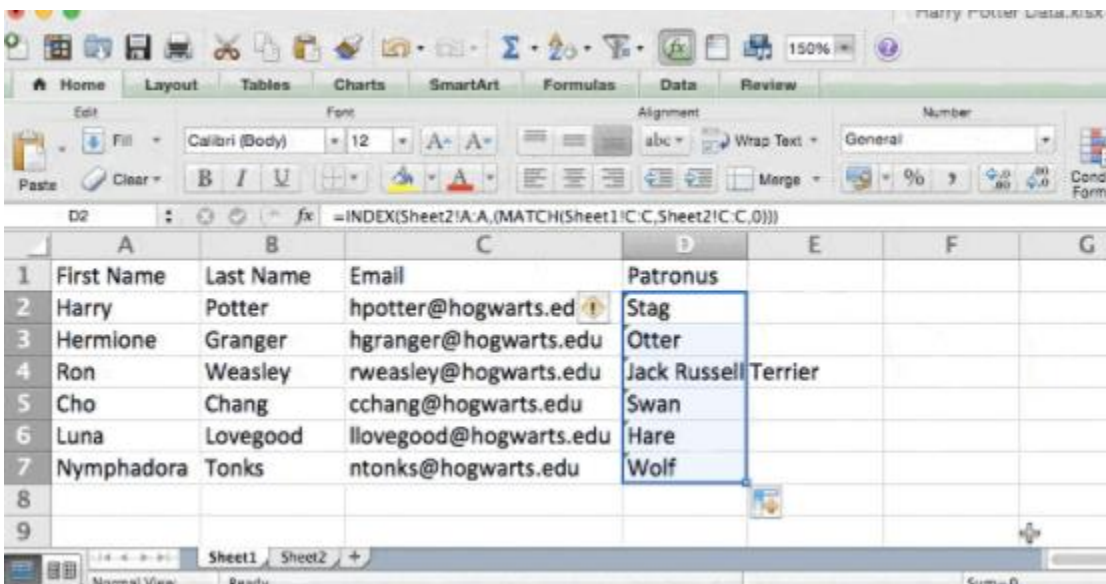
This becomes: =INDEX(table array, MATCH (lookup\_value, lookup\_array))

The formula with variables from our example below: =INDEX(Sheet2!A:A,(MATCH(Sheet1!C:C,Sheet2!C:C,0)))

Here are the variables:

- Table Array: The range of columns on Sheet 2 containing the new data you want to bring over to Sheet 1. In our example, "A" means Column A, which contains the "Patronus" information for each person.
- Lookup Value: This is the column in Sheet 1 that contains identical values in both spreadsheets. In the example that follows, this means the "email" column on Sheet 1, which is Column C. So: Sheet1!C: C.
- Lookup Array: This is the column in Sheet 2 that contains identical values in both spreadsheets. In the example that follows, this refers to the "email" column on Sheet 2, which happens to also be Column C. So: Sheet2!C: C.

Once you have your variables straight, type in the INDEX and MATCH formulas in the top-most cell of the blank Patronus column on Sheet 1, where you want the combined information to live.



The screenshot shows an Excel spreadsheet with two sheets, Sheet1 and Sheet2. Sheet1 contains columns A (First Name), B (Last Name), and C (Email). Sheet2 contains columns A (Patronus) and C (Email). The formula bar shows the formula =INDEX(Sheet2!A:A,(MATCH(Sheet1!C:C,Sheet2!C:C,0))). The Patronus column in Sheet1 is being populated with data from Sheet2.

	A	B	C	D	E	F	G
1	First Name	Last Name	Email	Patronus			
2	Harry	Potter	hpotter@hogwarts.edu	Stag			
3	Hermione	Granger	hgranger@hogwarts.edu	Otter			
4	Ron	Weasley	rweasley@hogwarts.edu	Jack Russell Terrier			
5	Cho	Chang	cchang@hogwarts.edu	Swan			
6	Luna	Lovegood	llovegood@hogwarts.edu	Hare			
7	Nymphadora	Tonks	ntonks@hogwarts.edu	Wolf			
8							
9							

### Use the COUNTIF function to make Excel count words or numbers in any range of cells.

Instead of manually counting how often a certain value or number appears, let Excel do the work for you. With the COUNTIF function, Excel can count the number of times a word or number appears in any range of cells.

For example, let's say I want to count the number of times the word "Gryffindor" appears in my data set.

The formula: =COUNTIF(range, criteria)

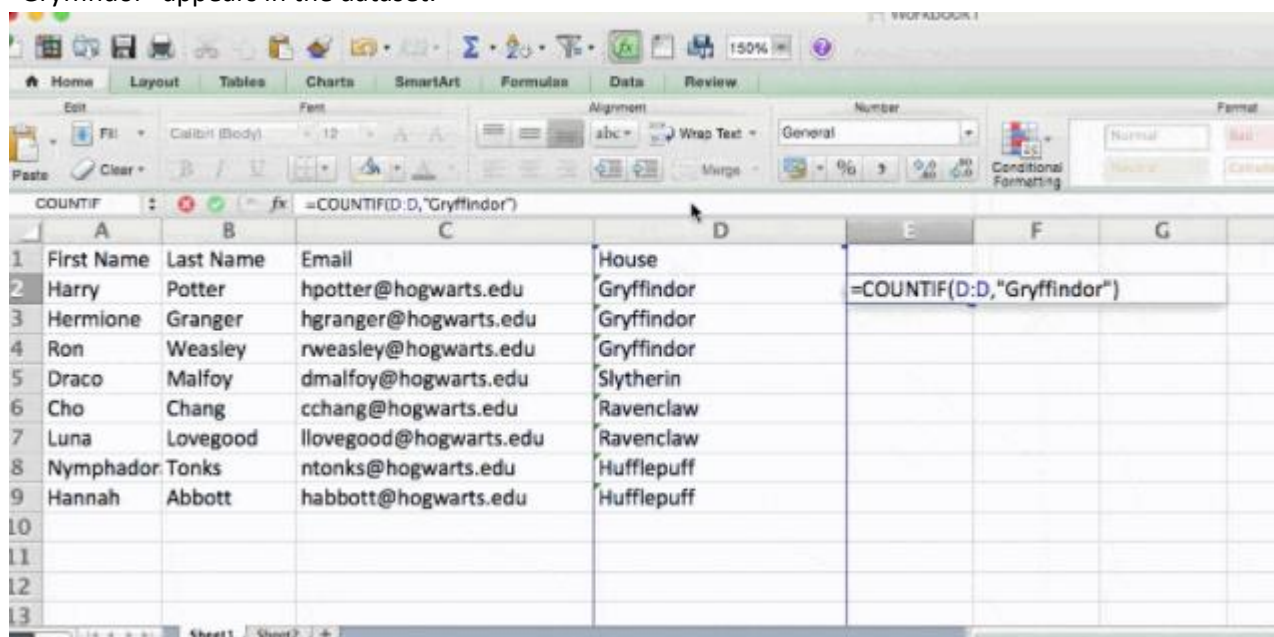
The formula with variables from our example below: =COUNTIF(D:D, "Gryffindor")

In this formula, there are several variables:

Range: The range that we want the formula to cover. In this case, since we're only focusing on one column, we use "D:D" to indicate that the first and last column are both D. If I were looking at columns C and D, I would use "C:D."

Criteria: Whatever number or piece of text you want Excel to count. Only use quotation marks if you want the result to be text instead of a number. In our example, the criteria are "Gryffindor."

Simply typing in the COUNTIF formula in any cell and pressing "Enter" will show me how many times the word "Gryffindor" appears in the dataset.



### Combine cells using &.

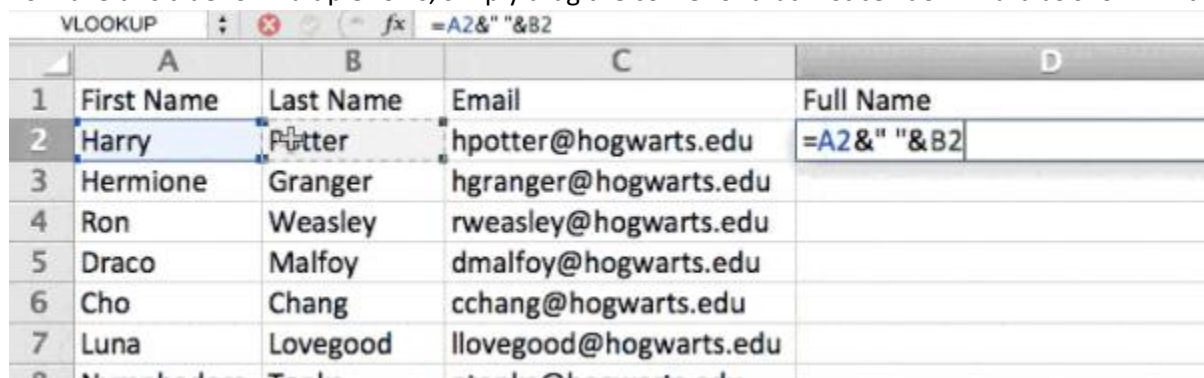
Databases tend to split out data to make it as exact as possible. For example, instead of having a column that shows a person's full name, a database might have the data as a first name and then the last name in separate columns. Or, it may have a person's location separated by city, state, and zip code. In Excel, you can combine cells with different data into one cell by using the "&" sign in your function.

The formula with variables from our example below: =A2&" "&B2

Let's go through the formula together using an example. Pretend we want to combine first names and last names into full names in a single column. To do this, we'd first put our cursor in the blank cell where we want the full name to appear. Next, we'd highlight one cell that contains a first name, type in an "&" sign, and then highlight a cell with the corresponding the last name.

But you're not finished — if all you type in is =A2&B2, then there will not be a space between the person's first name and last name. To add that necessary space, use the function =A2&" "&B2. The quotation marks around the space tell Excel to put a space in between the first and last name.

To make this true for multiple rows, simply drag the corner of that first cell downward as shown in the example.



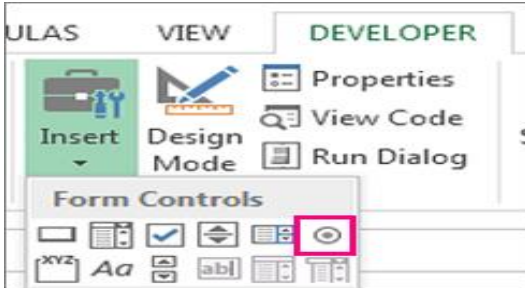
### Add checkboxes.

If you're using an Excel sheet to track customer data and want to oversee something that isn't quantifiable, you could insert checkboxes into a column.

For example, if you're using an Excel sheet to manage your sales prospects and want to track whether you called them in the last quarter, you could have a "Called this quarter?" column and check off the cells in it when you've called the respective client.

Here's how to do it.

Highlight a cell you'd like to add checkboxes to in your spreadsheet. Then, click DEVELOPER. Then, under FORM CONTROLS, click the checkbox or the selection circle highlighted in the image below.



Once the box appears in the cell, copy it, highlight the cells you also want it to appear in, and then paste it.

### Hyperlink a cell to a website.

If you're using your sheet to track social media or website metrics, it can be helpful to have a reference column with the links each row is tracking. If you add a URL directly into Excel, it should automatically be clickable. But, if you have to hyperlink words, such as a page title or the headline of a post you're tracking, here's how.

Highlight the words you want to hyperlink, then press Shift K. From there a box will pop up allowing you to place the hyperlink URL. Copy and paste the URL into this box and hit or click Enter.

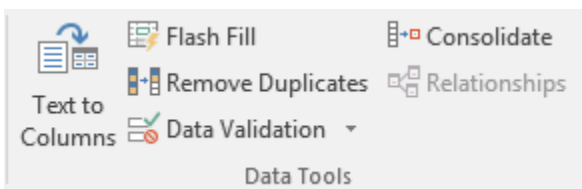
If the key shortcut isn't working for any reason, you can also do this manually by highlighting the cell and clicking Insert > Hyperlink.

### Add drop-down menus.

Sometimes, you'll be using your spreadsheet to track processes or other qualitative things. Rather than writing words into your sheet repetitively, such as "Yes", "No", "Customer Stage", "Sales Lead", or "Prospect", you can use dropdown menus to quickly mark descriptive things about your contacts or whatever you're tracking.

Here's how to add drop-downs to your cells.

Highlight the cells you want the drop-downs to be in, then click the Data menu in the top navigation and press Validation.



From there, you'll see a Data Validation Settings box open. Look at the Allow options, then click Lists and select Drop-down List. Check the In-Cell dropdown button, then press OK.

### **Use the format painter.**

As you've probably noticed, Excel has a lot of features to make crunching numbers and analyzing your data quick and easy. But if you ever spent some time formatting a sheet to your liking, you know it can get a bit tedious.

Don't waste time repeating the same formatting commands over and over again. Use the format painter to easily copy the formatting from one area of the worksheet to another. To do so, choose the cell you'd like to replicate, then select the format painter option (paintbrush icon) from the top toolbar.

### **Excel Keyboard Shortcuts**

Creating reports in Excel is time-consuming enough. How can we spend less time navigating, formatting, and selecting items in our spreadsheet? Glad you asked. There are a ton of Excel shortcuts out there, including some of our favorites listed below.

- Create a New Workbook  
PC: Ctrl-N | Mac: Command-N
- Select Entire Row  
PC: Shift-Space | Mac: Shift-Space
- Select Entire Column  
PC: Ctrl-Space | Mac: Control-Space
- Select Rest of Column  
PC: Ctrl-Shift-Down/Up | Mac: Command-Shift-Down/Up
- Select Rest of Row  
PC: Ctrl-Shift-Right/Left | Mac: Command-Shift-Right/Left
- Add Hyperlink  
PC: Ctrl-K | Mac: Command-K
- Open Format Cells Window  
PC: Ctrl-1 | Mac: Command-1
- Autosum Selected Cells  
PC: Alt-= | Mac: Command-Shift-T



# 3

## PROBABILITY DISTRIBUTION AND APPLICATIONS IN R

### Basic Concepts of Probability

A probability is a number that reflects the chance or likelihood that a particular event will occur. Probabilities can be expressed as proportions that range from 0 to 1, and they can also be expressed as percentages ranging from 0% to 100%. A probability of 0 indicates that there is no chance that a particular event will occur, whereas a probability of 1 indicates that an event is certain to occur. A probability of 0.45 (45%) indicates that there are 45 chances out of 100 of the event occurring.

The concept of probability can be illustrated in the context of a study of obesity in children 5-10 years of age who are seeking medical care at a particular pediatric practice. The population (sampling frame) includes all children who were seen in the practice in the past 12 months and is summarized below.

	Age (years)						
	5	6	7	8	9	10	Total
Boys	432	379	501	410	420	418	2,560
Girls	408	513	412	436	461	500	2,730
Totals	840	892	913	846	881	918	5,290

### Unconditional Probability

If we select a child at random (by simple random sampling), then each child has the same probability (equal chance) of being selected, and the probability is  $1/N$ , where  $N$ =the population size. Thus, the probability that any child is selected is  $1/5,290 = 0.0002$ . In most sampling situations we are generally not concerned with sampling a specific individual but instead, we concern ourselves with the probability of sampling certain types of individuals. For example, what is the probability of selecting a boy or a child 7 years of age? The following formula can be used to compute probabilities of selecting individuals with specific attributes or characteristics.

$$P(\text{characteristic}) = \# \text{ persons with characteristic} / N$$

Try to figure these out before looking at the answers:

- What is the probability of selecting a boy?
  - If we select a child at random, the probability that we select a boy is computed as follows  $P(\text{boy}) = 2,560/5,290 = 0.484$  or 48.4%.
- What is the probability of selecting a 7-year-old?
  - The probability of selecting a child who is 7 years of age is  $P(7 \text{ years of age}) = 913/5,290 = 0.173$ .
- What is the probability of selecting a boy who is 10 years of age?
  - $P(\text{boy who is 10 years of age}) = 418/5,290 = 0.079$ .

4. What is the probability of selecting a child (boy or girl) who is at least 8 years of age?

a.  $P(\text{at least 8 years of age}) = (846 + 881 + 918) / 5,290 = 2,645 / 5,290 = 0.500.$

### Conditional Probability

Each of the probabilities computed in the previous section (e.g.,  $P(\text{boy})$ ,  $P(7 \text{ years of age})$ ) is an unconditional probability because the denominator for each is the total population size ( $N=5,290$ ) reflecting the fact that everyone in the entire population is eligible to be selected. However, sometimes it is of interest to focus on a particular subset of the population (e.g., a sub-population). For example, suppose we are interested just in the girls and ask the question, what is the probability of selecting a 9-year-old from the sub-population of girls? There is a total of  $N_G=2,730$  girls (here  $N_G$  refers to the population of girls), and the probability of selecting a 9-year-old from the sub-population of girls is written as follows:

$$P(9 \text{ year old} \mid \text{girls}) = \# \text{ persons with characteristic} / N$$

where  $\mid \text{girls}$  indicate that we are conditioning the question to a specific subgroup, i.e., the subgroup specified to the right of the vertical line.

The conditional probability is computed using the same approach we used to compute unconditional probabilities. In this case:

$$P(9 \text{ year old} \mid \text{girls}) = 461 / 2,730 = 0.169.$$

This also means that 16.9% of the girls are 9 years of age. Note that this is not the same as the probability of selecting a 9-year old girl from the overall population, which is  $P(\text{girls who are 9 years of age}) = 461 / 5,290 = 0.087.$

What is the probability of selecting a boy from among the 6-year-olds?

$P(\text{boy} \mid 6 \text{ years of age}) = 379 / 892 = 0.425.$  Thus 42.5% of the 6-year-olds are boys (57.5% of the 6 year olds are girls).

### What is a Discrete Probability Distribution?

In statistics, you'll come across dozens of different types of probability distributions, like the binomial distribution, normal distribution, and Poisson distribution. All of these distributions can be classified as either a continuous or a discrete probability distribution.

A discrete probability distribution is made up of discrete variables. Specifically, if a random variable is discrete, then it will have a discrete probability distribution.

### Discrete Probability Distribution Examples

For example, let's say you had the choice of playing two games of chance at a fair.

Game 1: Roll a die. If you roll a six, you win a prize.

Game 2: Guess the weight of the man. If you guess within 10 pounds, you win a prize.

One of these games is a discrete probability distribution and one is a continuous probability distribution. Which is which?

For game 1, you could roll a 1,2,3,4,5 or 6. All of the die rolls have an equal chance of being rolled (one out of six, or  $1/6$ ). This gives you a discrete probability distribution of:



Roll	1	2	3	4	5	6
Odds	1/6	1/6	1/6	1/6	1/6	1/6

For the guess the weight game, you could guess that the mean weighs 150 lbs. Or 210 pounds. Or 185.5 pounds. Or any fraction of a pound (172.566 pounds). Even if you stick to, say, between 150 and 200 pounds, the possibilities are endless:

160.1 lbs.

160.11 lbs.

160.111 lbs.

160.1111 lbs.

160.111111 lbs.

In reality, you probably wouldn't guess 160.111111 lbs...that seems a little ridiculous. But it doesn't change the fact that you could (if you wanted to), so that's why it's a continuous probability distribution.

### **What is a Factorial?**

Factorials (!) are products of every whole number from 1 to n. In other words, take the number and multiply through by 1.

For example:

If n is 3, then 3! is  $3 \times 2 \times 1 = 6$ .

If n is 5, then 5! is  $5 \times 4 \times 3 \times 2 \times 1 = 120$ .

It's a shorthand way of writing numbers. For example, instead of writing 479001600, you could write 12! instead (which is  $12 \times 11 \times 10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$ ).

### **What is a factorial used for in stats?**

In algebra, you probably encountered ugly-looking factorials like  $(x - 10!)/(x + 9!)$ . Don't worry; You won't be seeing any of these in your beginning stats class. Phew! The only time you'll see them is for permutation and combination problems.

The equations look like this:

$$\frac{100}{5!(100-5)!}$$

Permutations and Combinations are the various ways in which objects from a set may be selected, generally without replacement, to form subsets. This selection of subsets is called a permutation when the order of selection is a factor, a combination when order is not a factor. By considering the ratio of the number of desired subsets to the number of all possible subsets for many games of chance in the 17th century, the French mathematicians Blaise Pascal and Pierre de Fermat gave impetus to the development of combinatorics and probability theory.

The concepts of and differences between permutations and combinations can be illustrated by an examination of all the different ways in which a pair of objects can be selected from five distinguishable objects—such as the letters A, B, C, D, and E. If both the letters selected and the order of selection are considered, then the following 20 outcomes are possible:

AB	BA	AC	CA	AD
DA	AE	EA	BC	CB
BD	DB	BE	EB	CD
DC	CE	EC	DE	ED

## **R Applications – Real-world Use Cases of R programming**

### **What is R used for R ? R Applications across these sectors**

Data Science and Big Data have proved themselves useful and even necessary in many different fields and industries today. It helps them to keep up with the trends and capitalize on every opportunity.

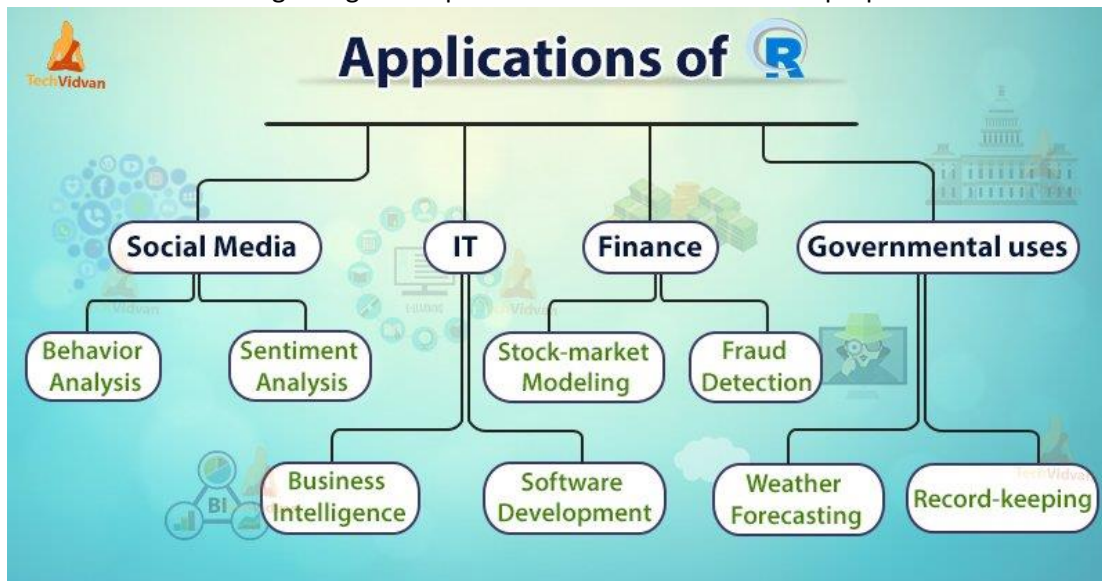
R acts as a tool, used to make sense of big data and to gain use from it. R has also proved itself useful in research by processing large amounts of data in less time.

Let's take a look at an interesting fact. "According to O'Reilly, R is the most-used data science language after SQL."

In this article, we will explore the various R applications in the real world.

### **Applications of R Programming**

Let's start from the beginning and explore the uses of R for research purposes:



### **R in Research and Academics**

R is a statistical research tool. It is still used by statisticians and students to perform various statistical computations and analyses. Statistical techniques like linear and non-linear modeling, time-series analysis, classification, classical statistic tests, clustering, and others are all implemented by R and its libraries.

R is also used for machine learning research and deep learning as well. With libraries that facilitate monitored and unmonitored learning, R is one of the most commonly used languages for machine learning.

Other research involving large data sets like big data, finding genetic anomalies and patterns, various drug compositions, all use R to sift through a large collection of relevant data and to draw meaningful conclusions from it.

## **R Use Cases in Research & Academics**

- Cornell University: Cornell recommends their researchers and students use R for all their research involving statistical computing.
- UCLA: The University of California, Los Angeles uses R to teach statistics and data analysis to its students.

Apart from research, R also has its applications in IT companies.

## **R in IT Sector**

IT companies not only use R for their business intelligence but offer such services to other small, medium, and large scale businesses as well. They use it for their machine learning products too.

They use R to build statistical computing tools and data handling products and to create other data manipulation services.

Some big IT companies that use R:

- Accenture
- IBM
- Infosys
- Paytm
- Tata Consultancy Services
- Wipro

## **R Use Cases in IT Sector**

- Mozilla: Mozilla uses R to visualize web activity for their browser firefox.
- Microsoft: Microsoft uses R as a statistical engine within the Azure Machine Learning framework. They also use it for the Xbox matchmaking service.
- Foursquare: R works behind the scenes on Foursquare's recommendation engine.
- Google: Google uses R to improve its search results, to provide better search suggestions, to calculate the ROI of its advertising campaigns, to increase the efficiency of online advertising, and to predict their economic activity.

## **R in Finance**

Other than the finance sector which industry will be dealing more with statistics as R is a statistical programming language.

R and data science find widespread use in the finance sector. R provides an advanced statistical suite for all the financial tasks and computations. Moving averages, auto-regression, time-series analysis, stock-market modeling, financial data mining, downside risk assessment are all easily done through R and its libraries.

R is also used to support the business decision-making process. R's data visualization powers can represent the findings of data analysis in multiple graphical formats like candlestick charts, density plots, and drawdown plots of high quality.

This helps the business minds to connect with the technical aspect of data analyses and their results. Companies like American Express, Bajaj Allianz Insurance, JP Morgan, and Standard Chartered use R.

## **R Use Cases in Finance**

- Lloyds of London: Lloyds of London use R for risk analysis.
- Bajaj Allianz Insurance: Bajaj Allianz uses R to make their upsell propensity models and recommendation engines. They also use it to mine data and generate actionable insights to improve customer experience.

The digital revolution has changed the world drastically. One of the most prominent changes is the fact that marketplaces have moved to the internet. The E-commerce industry makes heavy usage of R for varying purposes.

### **R in E-commerce**

In the field of finance and retail, analytics is useful for risk assessment and for devising a marketing strategy. E-commerce goes beyond that in its usage of data science. E-commerce companies use R to improve the user's experience on their site as well as for marketing and finance purposes. They use R to improve cross-product selling. When a customer is buying a product the site suggests additional products that complement their original purchase. These suggestions also work for products purchased by the customer in the past. Internet-based companies like various e-commerce sites gather and process structured and unstructured data from varying sources. R proves to be highly useful for this.

Apart from this, R is also used to help with marketing strategy, targeted advertising, sales modeling, and financial data processing.

### **R Use Cases in E-commerce**

Amazon: Amazon uses R and data analysis to improve its cross-product suggestions.

Flipkart: Flipkart uses R for predictive analysis which helps them with targetted advertisements.

Social media is the most common generator of big data today. Therefore, the most advanced and cutting-edge uses of data science can be found in the social media industry.

If you have any queries in R applications till now, mention them in the comment section.

### **R in Social Media**

Social media companies like Facebook use R for behavior analysis and sentiment analysis. They can alter and improve their suggestions to users based on the user's history, and the mood and tone of their recent posts and viewed content. The advertisements shown to the user are also adjusted according to user sentiment and history. R is also used to analyze traffic, user sessions, and content, all to improve user experience. R Use Cases in Social Media

Facebook: Facebook uses R to predict colleague interactions and update its social network graph.

Twitter: Twitter uses R for semantic clustering. They also use it for data visualization.

Another sector making great use of R's statistical computation abilities is the banking sector.

### **R in Banking**

Banking firms use R for credit risk modeling and other forms of risk analytics. Banks often use R along with other proprietary software like SAS. It is also used for fraud detection, mortgage haircut modeling, stat modeling, volatility modeling, loan stress test simulation, client assessment, and much more. Apart from statistics, banks also use R for business intelligence and data visualization.

Another use of R is in the calculation of customer segmentation, customer quality, and customer retention.

### **R Use Cases in Banking**

- ANZ: ANZ bank uses R for credit risk modeling and also in models for mortgage loss.
- Bank of America: Bank of America uses R for financial reporting and to calculate financial losses.

The healthcare industry is not one to be left behind when it comes to cutting-edge technologies:

### **R in Healthcare**

With R, you can crunch data and process information, providing an essential backdrop for further analysis and data processing. Genetics, drug discovery, bioinformatics, epidemiology, etc. are some fields in the healthcare industry that

use R heavily. It is used to analyze and predict the spreading of various diseases, for analyzing genetic sequences, to analyze drug-safety data, and to analyze various permutations and combinations of drugs and chemicals as well. R's Bioconductor package provides facilities for analyzing genomic data. Lastly, R is a god-send for pre-clinical trials of all new drugs and medical techniques.

### **R Use Case in Healthcare**

Merck: Merck & co. use the R programming language for clinical trials and drug testing.

Manufacturing companies also use R to make use of big data and to be ahead of the curve.

### **R in Manufacturing**

Various manufacturing companies use R to complement their marketing and business strategies. They analyze customer feedback to help streamline and improve their products. They also use the data to support their marketing strategies. Predicting demand and market trends to adjust their manufacturing practices is yet another use of R and data analytics.

### **R Use Cases in Manufacturing**

Ford Motor Company: Ford uses R for statistical analyses to support its business strategy and to analyze customer sentiment about its product which helps them in improving their future designs.

John Deere: John Deere uses R to forecast demand for their products and spare parts. They also use it to forecast crop yield and use that data for their business strategy and to meet market demand and downturns.

Every government has to handle a large amount of data. A country's worth of data! Many governmental departments across the world use R as well.

### **R in Governmental Use**

Many governmental departments use R for record-keeping and processing their censuses. This helps them in effective law-making and governance. They also use it for essential services like drug regulation, weather forecasting, disaster-impact analysis, and much more.

### **R Use Cases in Governmental Activities**

Food and Drug Administration: FDA uses R for drug evaluation and to perform pre-clinical trials. It also uses R to predict possible reactions and medical issues caused by various food products.

National Weather Service: The National Weather Service uses R for weather forecasts and disaster prediction. They also use it to visualize their forecasts and predictions to analyze the areas affected.

# 4

## STATISTICAL SAMPLING

### Sample

A sample is a subset of individuals from a larger population. Sampling means selecting the group that you will collect data from in your research. For example, if you are researching the opinions of students in your university, you could survey a sample of 100 students. Probability sampling means that every member of the target population has a known chance of being included in the sample. Probability sampling methods include simple random sampling, systematic sampling, stratified sampling, and cluster sampling. In statistics, sampling allows you to test a hypothesis about the characteristics of a population.

When you research a group of people, it's rarely possible to collect data from every person in that group. Instead, you select a sample. The sample is the group of individuals who will participate in the research. To draw valid conclusions from your results, you have to carefully decide how you will select a sample that is representative of the group as a whole. There are two types of sampling methods:

- Probability sampling involves random selection, allowing you to make strong statistical inferences about the whole group.
- Non-probability sampling involves non-random selection based on convenience or other criteria, allowing you to easily collect data.

You should clearly explain how you selected your sample in the methodology section of your paper or thesis.

### Population vs sample

First, you need to understand the difference between a population and a sample and identify the target population of your research.

- The population is the entire group that you want to conclude about.
- The sample is the specific group of individuals that you will collect data from.

The population can be defined in terms of geographical location, age, income, and many other characteristics.



It can be very broad or quite narrow: maybe you want to make inferences about the whole adult population of your country; maybe your research focuses on customers of a certain company, patients with a specific health condition, or students in a single school.

It is important to carefully define your target population according to the purpose and practicalities of your project.

If the population is very large, demographically mixed, and geographically dispersed, it might be difficult to gain access to a representative sample.

### **Sampling frame**

The sampling frame is the actual list of individuals that whom the sample will be drawn. Ideally, it should include the entire target population (and nobody who is not part of that population).

#### **Example**

You are researching working conditions at Company X. Your population is all 1000 employees of the company. Your sampling frame is the company's HR database which lists the names and contact details of every employee.

### **Sample size**

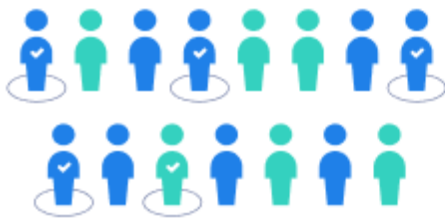
The number of individuals you should include in your sample depends on various factors, including the size and variability of the population and your research design. There are different sample size calculators and formulas depending on what you want to achieve with statistical analysis.

### **Probability sampling methods**

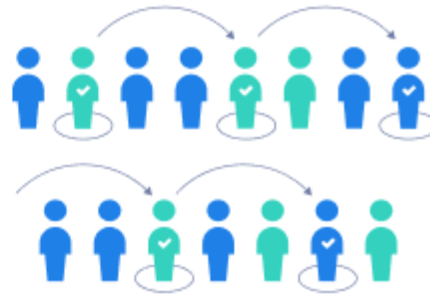
Probability sampling means that every member of the population has a chance of being selected. It is mainly used in quantitative research. If you want to produce results that are representative of the whole population, probability sampling techniques are the most valid choice.

There are four main types of a probability samples.

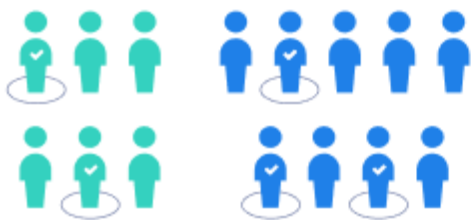
### Simple random sample



### Systematic sample



### Stratified sample



### Cluster sample



#### 1. Simple random sampling

In a simple random sample, every member of the population has an equal chance of being selected. Your sampling frame should include the whole population.

To conduct this type of sampling, you can use tools like random number generators or other techniques that are based entirely on chance.

##### Example

You want to select a simple random sample of 100 employees of Company X. You assign a number to every employee in the company database from 1 to 1000, and use a random number generator to select 100 numbers.

#### 2. Systematic sampling

Systematic sampling is similar to simple random sampling, but it is usually slightly easier to conduct. Every member of the population is listed with a number, but instead of randomly generating numbers, individuals are chosen at regular intervals.

##### Example

All employees of the company are listed in alphabetical order. From the first 10 numbers, you randomly select a starting point: number 6. From number 6 onwards, every 10th person on the list is selected (6, 16, 26, 36, and so on), and you end up with a sample of 100 people.



If you use this technique, it is important to make sure that there is no hidden pattern in the list that might skew the sample. For example, if the HR database groups employees by a team, and team members are listed in order of seniority, there is a risk that your interval might skip over people in junior roles, resulting in a sample that is skewed towards senior employees.

### **3. Stratified sampling**

Stratified sampling involves dividing the population into subpopulations that may differ in important ways. It allows you to draw more precise conclusions by ensuring that every subgroup is properly represented in the sample.

To use this sampling method, you divide the population into subgroups (called strata) based on the relevant characteristic (e.g. gender, age range, income bracket, job role).

Based on the overall proportions of the population, you calculate how many people should be sampled from each subgroup. Then you use random or systematic sampling to select a sample from each subgroup.

#### **Example**

The company has 800 female employees and 200 male employees. You want to ensure that the sample reflects the gender balance of the company, so you sort the population into two strata based on gender. Then you use random sampling on each group, selecting 80 women and 20 men, which gives you a representative sample of 100 people.

### **4. Cluster sampling**

Cluster sampling also involves dividing the population into subgroups, but each subgroup should have similar characteristics to the whole sample. Instead of sampling individuals from each subgroup, you randomly select entire subgroups.

If it is practically possible, you might include every individual from each sampled cluster. If the clusters themselves are large, you can also sample individuals from within each cluster using one of the techniques above. This is called multistage sampling.

This method is good for dealing with large and dispersed populations, but there is more risk of error in the sample, as there could be substantial differences between clusters. It's difficult to guarantee that the sampled clusters are representative of the whole population.

#### **Example**

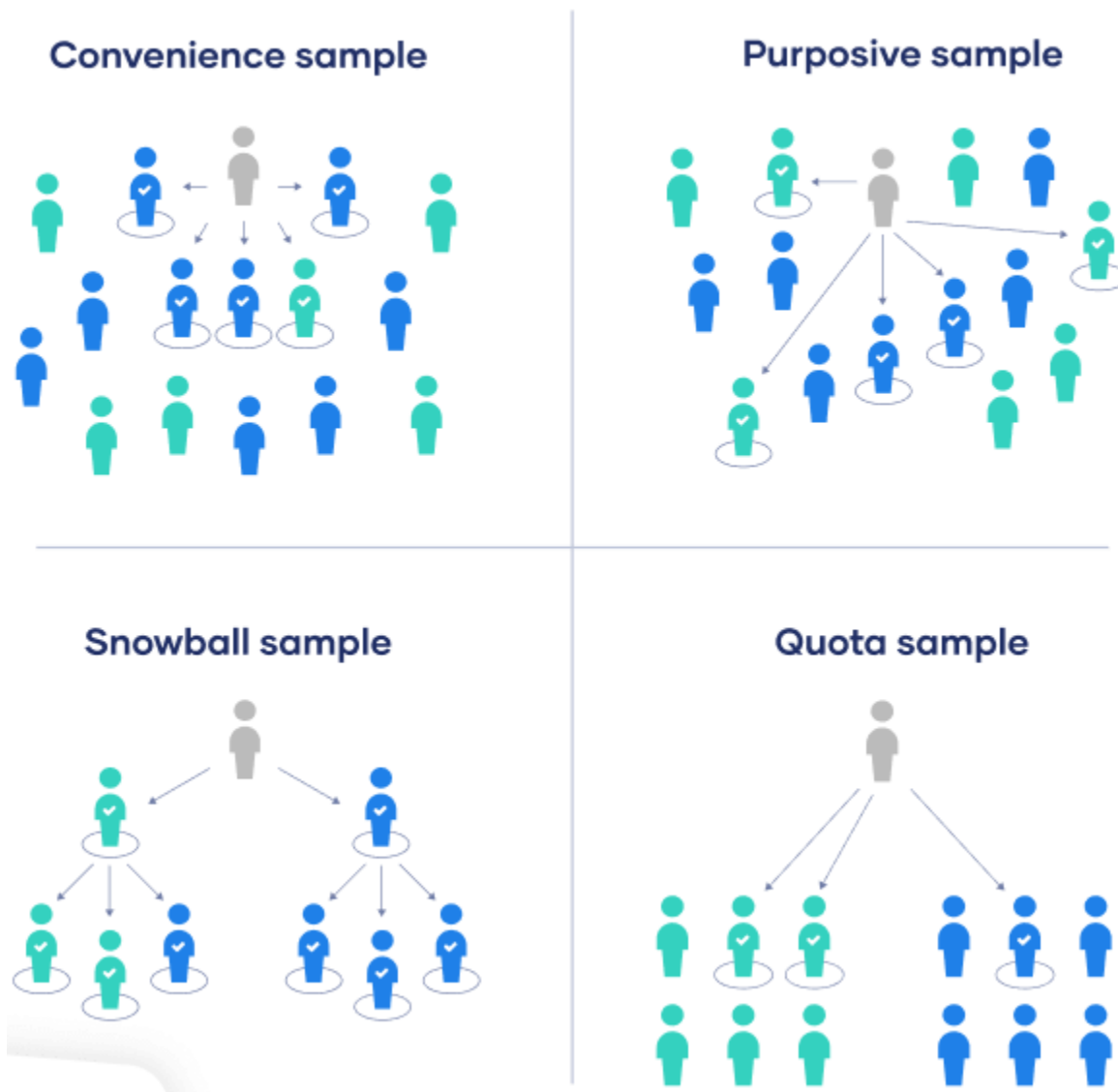
The company has offices in 10 cities across the country (all with roughly the same number of employees in similar roles). You can't travel to every office to collect your data, so you use random sampling to select 3 offices – these are your clusters.

### **Non-probability sampling methods**

In a non-probability sample, individuals are selected based on non-random criteria, and not every individual has a chance of being included.

This type of sample is easier and cheaper to access, but it has a higher risk of sampling bias. That means the inferences you can make about the population are weaker than with probability samples, and your conclusions may be more limited. If you use a non-probability sample, you should still aim to make it as representative of the population as possible.

Non-probability sampling techniques are often used in exploratory and qualitative research. In these types of research, the aim is not to test a hypothesis about a broad population, but to develop an initial understanding of a small or under-researched population.



### 1. Convenience sampling

A convenience sample simply includes the individuals who happen to be most accessible to the researcher.

This is an easy and inexpensive way to gather initial data, but there is no way to tell if the sample is representative of the population, so it can't produce generalizable results.

#### Example

You are researching opinions about student support services in your university, so after each of your classes, you ask your fellow students to complete a survey on the topic. This is a convenient way to gather data, but as you only surveyed students taking the same classes as you at the same level, the sample is not representative of all the students at your university.

### 2. Voluntary response sampling

Similar to a convenience sample, a voluntary response sample is mainly based on ease of access. Instead of the researcher choosing participants and directly contacting them, people volunteer themselves (e.g. by responding to a public online survey).

Voluntary response samples are always at least somewhat biased, as some people will inherently be more likely to volunteer than others.

#### Example

You send out the survey to all students at your university and a lot of students decide to complete it. This can certainly give you some insight into the topic, but the people who responded are more likely to be those who have strong opinions about the student support services, so you can't be sure that their opinions are representative of all students.

### **3. Purposive sampling**

This type of sampling, also known as judgment sampling, involves the researcher using their expertise to select a sample that is most useful to the purposes of the research.

It is often used in qualitative research, where the researcher wants to gain detailed knowledge about a specific phenomenon rather than make statistical inferences, or where the population is very small and specific. An effective purposive sample must have clear criteria and rationale for inclusion.

#### Example

You want to know more about the opinions and experiences of disabled students at your university, so you purposefully select many students with different support needs to gather a varied range of data on their experiences with student services.

### **4. Snowball sampling**

If the population is hard to access, snowball sampling can be used to recruit participants via other participants. The number of people you have access to "snowballs" as you get in contact with more people.

#### Example

You are researching experiences of homelessness in your city. Since there is no list of all homeless people in the city, probability sampling isn't possible. You meet one person who agrees to participate in the research, and she puts you in contact with other homeless people that she knows in the area.

# 5

## STATISTICAL INFERENCE

### Introduction

The purpose of this introduction is to review how we got here and how the previous units fit together to allow us to make reliable inferences. Also, we will introduce the various forms of statistical inference that will be discussed in this unit, and give a general outline of how this unit is organized.

In the Exploratory Data Analysis unit, we learned to display and summarize data that were obtained from a sample. Regardless of whether we had one variable and we examined its distribution, or whether we had two variables and we examined the relationship between them, it was always understood that these summaries applied only to the data at hand; we did not attempt to make claims about the larger population from which the data were obtained.

Such generalizations were, however, a long-term goal from the very beginning of the course. For this reason, in the unit on Producing Data, we took care to establish principles of sampling and study design that would be essential for us to claim that, to some extent, what is true for the sample should be also true for the larger population from which the sample originated.

These principles should be kept in mind throughout this unit on statistical inference since the results that we will obtain will not hold if there was bias in the sampling process or flaws in the study design under which variables' values were measured.

Perhaps the most important principle stressed in the Producing Data unit was that of randomization. Randomization is essential, not only because it prevents bias, but also because it permits us to rely on the laws of probability, which is the scientific study of random behavior.

In the Probability unit, we established basic laws for the behavior of random variables. We ultimately focused on two random variables of particular relevance: the sample mean ( $\bar{x}$ ) and the sample proportion ( $\hat{p}$ ), and the last section of the Probability unit was devoted to exploring their sampling distributions.

We learned what probability theory tells us to expect from the values of the sample mean and the sample proportion, given that the corresponding population parameters — the population mean ( $\mu$ ) and the population proportion ( $p$ ) — are known.

As we mentioned in that section, the value of such results is more theoretical than practical, since in real-life situations we seldom know what is true for the entire population. All we know is what we see in the sample, and we want to use this information to say something concrete about the larger population.

Probability theory has set the stage to accomplish this: learning what to expect from the value of the sample mean, given that the population mean takes a certain value, teaches us (as we'll soon learn) what to expect from the value of the unknown population mean, given that a particular value of the sample mean has been observed.

Similarly, since we have established how the sample proportion behaves relative to population proportion, we will now be able to turn this around and say something about the value of the population proportion, based on an observed sample proportion. This process — inferring something about the population based on what is measured in the sample — is (as you know) called statistical inference.

## **Types of Inference**

We will introduce three forms of statistical inference in this unit, each one representing a different way of using the information obtained in the sample to conclude the population. These forms are:

- Point Estimation
- Interval Estimation
- Hypothesis Testing

Each one of these forms of inference will be discussed at length in this section, but it would be useful to get at least an intuitive sense of the nature of each of these inference forms, and the difference between them in terms of the types of conclusions they draw about the population based on the sample results.

### **Point Estimation**

In point estimation, we estimate an unknown parameter using a single number that is calculated from the sample data.

EXAMPLE:

Based on sample results, we estimate that  $p$ , the proportion of all U.S. adults who are in favor of stricter gun control, is 0.6.

### **Interval Estimation**

In interval estimation, we estimate an unknown parameter using an interval of values that is likely to contain the true value of that parameter (and state how confident we are that this interval indeed captures the true value of the parameter).

EXAMPLE:

Based on sample results, we are 95% confident that  $p$ , the proportion of all U.S. adults who are in favor of stricter gun control, is between 0.57 and 0.63.

### **Hypothesis Testing**

In hypothesis testing, we begin with a claim about the population (we will call the null hypothesis), and we check whether or not the data obtained from the sample provide evidence AGAINST this claim.

EXAMPLE:

It was claimed that among all U.S. adults, about half are in favor of stricter gun control and about half are against it. In a recent poll of a random sample of 1,200 U.S. adults, 60% were in favor of stricter gun control. This data, therefore, provides some evidence against the claim.

Soon we will determine the probability that we could have seen such a result (60% in favor) or more extreme IF the true proportion of all U.S. adults who favor stricter gun control is 0.5 (the value in the claim the data attempts to refute).

EXAMPLE:

It is claimed that among drivers 18-23 years of age (our population) there is no relationship between drunk driving and gender.

A roadside survey collected data from a random sample of 5,000 drivers and recorded their gender and whether they were drunk.

The collected data showed roughly the same percentage of drunk drivers among males and females. These data, therefore, do not give us any reason to reject the claim that there is no relationship between drunk driving and gender. In terms of organization, the Inference unit consists of two main parts: Inference for One Variable and Inference for Relationships between Two Variables. The organization of each of these parts will be discussed further as we proceed through the unit.

## **Inference for One Variable**

The next two topics in the inference unit will deal with inference for one variable. Recall that in the Exploratory Data Analysis (EDA) unit, when we learned about summarizing the data obtained from one variable where we learned about examining distributions, we distinguished between two cases; categorical data and quantitative data.

We will make a similar distinction here in the inference unit. In the EDA unit, the type of variable determined the displays and numerical measures we used to summarize the data. In Inference, the type of variable of interest (categorical or quantitative) will determine what population parameter is of interest.

- When the variable of interest is categorical, the population parameter that we will infer is the population proportion ( $p$ ) associated with that variable. For example, if we are interested in studying opinions about the death penalty among U.S. adults, and thus our variable of interest is “death penalty (in favor/against),” we’ll choose a sample of U.S. adults and use the collected data to make an inference about  $p$ , the proportion of U.S. adults who support the death penalty.
- When the variable of interest is quantitative, the population parameter that we infer is the population mean ( $\mu$ ,  $\mu$ ) associated with that variable. For example, if we are interested in studying the annual salaries in the population of teachers in a certain state, we’ll choose a sample from that population and use the collected salary data to make an inference about  $\mu$ , the mean annual salary of all teachers in that state.

The following outlines describe some of the important points about the process of inferential statistics as well as compare and contrast how researchers and statisticians approach this process.

### **Outline of Process of Inference**

Here is another restatement of the big picture of statistical inference as it pertains to the two simple examples we will discuss first.

- A simple random sample is taken from a population of interest.
- To estimate a population parameter, a statistic is calculated from the sample. For example:
  - Sample mean ( $\bar{x}$ )
  - Sample proportion ( $\hat{p}$ )
- We then learn about the DISTRIBUTION of this statistic in repeated sampling (theoretically). We now know these are called sampling distributions!
- Using THIS sampling distribution we can make inferences about our population parameter based on our sample statistics.

It is this last step of statistical inference that we are interested in discussing now.

### **Applied Steps (What do researchers do?)**

One issue for students is that the theoretical process of statistical inference is only a small part of the applied steps in a research project. Previously, in our discussion of the role of biostatistics, we defined these steps to be:

1. Planning/design of the study
2. Data collection
3. Data analysis
4. Presentation
5. Interpretation

You can see that:

- Both exploratory data analysis and inferential methods will fall into the category of “Data Analysis” in our previous list.
- Probability is hidden in the applied steps in the form of probability sampling plans, estimation of desired probabilities, and sampling distributions.

Among researchers, the following represent some of the important questions to address when conducting a study.

- What is the population of interest?
- What is the question or statistical problem?
- How to sample to best address the question given the available resources?
- How do analyze the data?
- How to report the results?

AFTER you know what you are going to do, then you can begin collecting data!

### **Theoretical Steps (What do statisticians do?)**

Statisticians, on the other hand, need to ask questions like these:

- What assumptions can be reasonably made about the population?
- What parameter(s) in the population do we need to estimate to address the research question?
- What statistic(s) from our sample data can be used to estimate the unknown parameter(s)?
- How does each statistic behave?
  - Is it unbiased?
  - How variable will it be for the planned sample size?
  - What is the distribution of this statistic? (Sampling Distribution)

Then, we will see that we can use the sampling distribution of a statistic to:

- Provide confidence interval estimates for the corresponding parameter.
- Conduct hypothesis tests about the corresponding parameter.

### **Standard Error of a Statistic**

In our discussion of **sampling distributions**, we discussed the variability of sample statistics; here is a quick review of this general concept and a formal definition of the standard error of a statistic.

- All statistics calculated from samples are random variables.
- The distribution of a statistic (from a sample of given sample size) is called the sampling distribution of the statistic.
- The standard deviation of the sampling distribution of a particular statistic is called the standard error of the statistic and measures the variability of the statistic for a particular sample size.

The standard error of a statistic is the standard deviation of the sampling distribution of that statistic, where the sampling distribution is defined as the distribution of a particular statistic in repeated sampling.

- The standard error is an extremely common measure of the variability of a sample statistic.

**EXAMPLE:**

In our discussion of sampling distributions, we looked at a situation involving a random sample of 100 students taken from the population of all part-time students in the United States, for which the overall proportion of females is 0.6. Here we have a categorical variable of interest, gender.

We determined that the distribution of all possible values of p-hat (that we could obtain for repeated simple random samples of this size from this population) has mean  $p = 0.6$  and a standard deviation

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.6(1-0.6)}{100}} = 0.05$$

which we have now learned is more formally called the standard error of p-hat. **In this case, the true standard error of p-hat will be 0.05.**

We also showed how we can use this information along with information about the center (mean or expected value) to calculate probabilities associated with particular values of p-hat. For example, what is the probability that the sample proportion p-hat is less than or equal to 0.56? After verifying the sample size requirements are reasonable, we can use a normal distribution to approximate

$$P(\hat{p} \leq 0.56) = P\left(Z \leq \frac{0.56 - 0.6}{0.05}\right) = P(Z \leq -0.80) = 0.2119$$

**EXAMPLE:**

Similarly, for a quantitative variable, we looked at an example of household size in the United States which has a mean of 2.6 people and a standard deviation of 1.4 people.

If we consider taking a simple random sample of 100 households, we found that the distribution of sample means (x-bar) is approximately normal for a large sample size such as  $n = 100$ .

The sampling distribution of the x-bar has a mean which is the same as the population mean, 2.6, and its standard deviation is the population standard deviation divided by the square root of the sample size:

$$\frac{\sigma}{\sqrt{n}} = \frac{1.4}{\sqrt{100}} = 0.14$$

Again, this standard deviation of the sampling distribution of x-bar is more commonly called the **standard error of x-bar**, in this case, 0.14. And we can use this information (the center and spread of the sampling distribution) to find probabilities involving particular values of x-bar.

$$P(\bar{x} > 3) = P\left(Z > \frac{3 - 2.6}{\frac{1.4}{\sqrt{100}}}\right) = P(Z > 2.86) = 0.0021$$



# 6

## REGRESSION ANALYSIS

### An Introduction to Regression Analysis

Regression analysis is a statistical tool for the investigation of relationships between variables. Usually, the investigator seeks to ascertain the causal effect of one variable upon another—the effect of a price increase upon demand, for example, or the effect of changes in the money supply upon the inflation rate. To explore such issues, the investigator assembles data on the underlying variables of interest and employs regression to estimate the quantitative effect of the causal variables upon the variable that they influence. The investigator also typically assesses the “statistical significance” of the estimated relationships, that is, the degree of confidence that the true relationship is close to the estimated relationship.

Regression techniques have long been central to the field of economic statistics (“econometrics”). Increasingly, they have become important to lawyers and legal policymakers as well. Regression has been offered as evidence of liability under Title VII of the Civil Rights Act of 1964 as evidence of racial bias in death penalty litigation, as evidence of damages in contract actions, as evidence of violations under the Voting Rights Act,<sup>2</sup> and as evidence of damage in antitrust litigation, among other things.

In this course notes, we will have an overview of the most basic techniques of regression analysis—how they work, what they assume, and how they may go awry when key assumptions do not hold. To make the discussion concrete, I will employ a series of illustrations involving a hypothetical analysis of the factors that determine individual earnings in the labor market. The illustrations will have a legal flavor in the latter part of the lecture, where they will incorporate the possibility that earnings are impermissibly influenced by gender in violation of the federal civil rights laws. Also, of necessity, there are many important topics that I omit, including simultaneous equation models and generalized least squares. The lecture is limited to the assumptions, mechanics, and common difficulties with a single equation, ordinary least squares regression.

### **What is Regression?**

For purposes of illustration, suppose that we wish to identify and quantify the factors that determine earnings in the labor market. A moment’s reflection suggests a myriad of factors that are associated with variations in earnings across individuals—occupation, age, experience, educational attainment, motivation, and innate ability come to mind, perhaps along with factors such as race and gender that can be of particular concern to lawyers. For the time being, let us restrict attention to a single factor—call it education. Regression analysis with a single explanatory variable is termed “simple regression.”

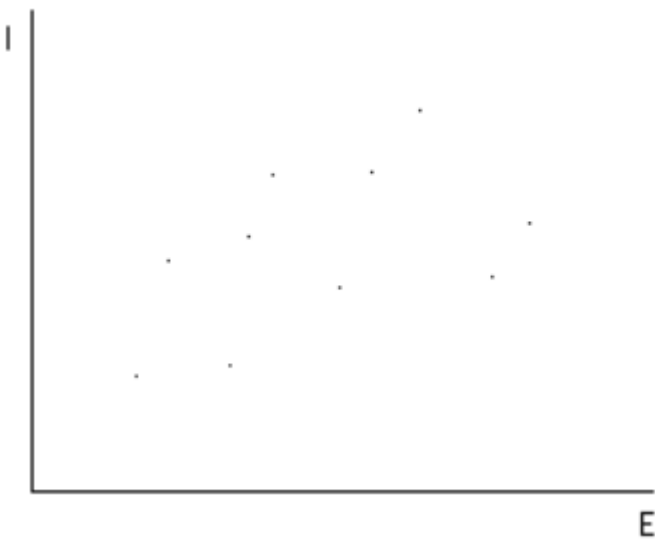
### **Simple Regression**

In reality, any effort to quantify the effects of education upon earnings without careful attention to the other factors that affect earnings could create serious statistical difficulties (termed “omitted variables bias”), which I will discuss later. But for now, let us assume away this problem. We also assume, again quite unrealistically, that “education” can be measured by a single attribute—years of schooling. We thus suppress the fact that a given number of years in school may represent widely varying academic programs.

At the outset of any regression study, one formulates some hypothesis about the relationship between the variables of interest, here, education and earnings. Common experience suggests that better-educated people tend to make more money. It further suggests that the causal relation likely runs from education to earnings rather than the other way around. Thus, the tentative hypothesis is that higher levels of education cause higher levels of earnings, other things being equal.

To investigate this hypothesis, imagine that we gather data on education and earnings for various individuals. Let  $E$  denote education in years of schooling for each individual, and let  $I$  denote that individual's earnings in dollars per year. We can plot this information for all of the individuals in the sample using a two-dimensional diagram, conventionally termed a "scatter" diagram. Each point in the diagram represents an individual in the sample.

### Chicago Working Paper in Law & Economics



The diagram indeed suggests that higher values of  $E$  tend to yield higher values of  $I$ , but the relationship is not perfect it seems that knowledge of  $E$  does not suffice for an entirely accurate prediction about  $I$ . We can then deduce either that the effect of education on earnings differs across individuals, or that factors other than education influence earnings. Regression analysis ordinarily embraces the latter explanation. Thus, pending discussion below of omitted variables bias, we now hypothesize that earnings for each individual are determined by education and by an aggregation of omitted factors that we term "noise."

To refine the hypothesis further, it is natural to suppose that people in the labor force with no education nevertheless make some positive amount of money and that education increases earnings above this baseline. We might also suppose that education affects income in a "linear" fashion—that is, each additional year of schooling adds the same amount to income. This linearity assumption is common in regression studies but is by no means essential to the application of the technique, and can be relaxed where the investigator has reason to suppose a priori that the relationship in question is nonlinear.

Then, the hypothesized relationship between education and earnings may be written

$$I = \alpha + \beta E + \varepsilon$$

where

$\alpha$  = a constant amount (what one earns with zero education);

$\beta$  = the effect in dollars of an additional year of schooling on income hypothesized to be positive; and

$\varepsilon$  = the "noise" term reflecting other factors that influence earnings.

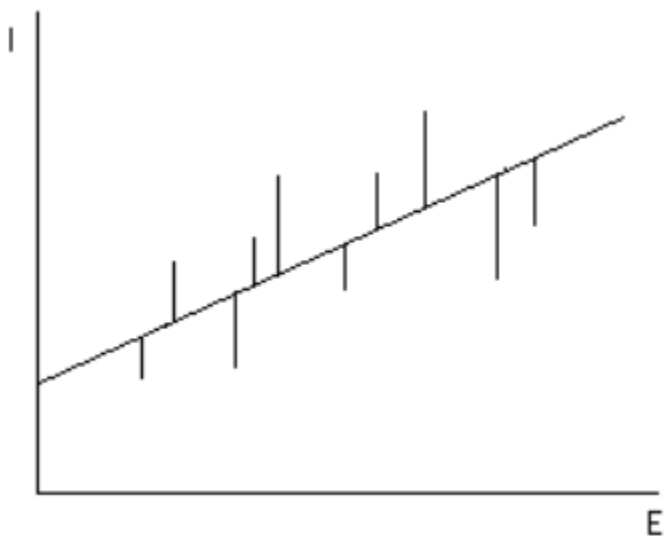
The variable  $I$  is termed the “dependent” or “endogenous” variable;  $E$  is termed the “independent,” “explanatory,” or “exogenous” variable;  $\alpha$  is the “constant term” and  $\beta$  is the “coefficient” of the variable  $E$ .

Remember what is observable and what is not. The data set contains observations for  $I$  and  $E$ . The noise component  $\epsilon$  is comprised of unobservable factors, or at least unobserved. The parameters  $\alpha$  and  $\beta$  are also unobservable. The task of regression analysis is to produce an estimate of these two parameters, based upon the information contained in the data set and, as shall be seen, upon some assumptions about the characteristics of  $\epsilon$ .

To understand how the parameter estimates are generated, note that if we ignore the noise term  $\epsilon$ , the equation above for the relationship between  $I$  and  $E$  is the equation for a line—a line with an “intercept” of  $\alpha$  on the vertical axis and a “slope” of  $\beta$ . Returning to the scatter diagram, the hypothesized relationship thus implies that somewhere on the diagram may be found a line with the equation  $I = \alpha + \beta E$ . The task of estimating  $\alpha$  and  $\beta$  is equivalent to the task of estimating where this line is located.

What is the best estimate regarding the location of this line? The answer depends in part upon what we think about the nature of the noise term  $\epsilon$ . If we believed that  $\epsilon$  was usually a large negative number, for example, we would want to pick a line lying above most or all of our data points—the logic is that if  $\epsilon$  is negative, the true value of  $I$  (which we observe), given by  $I = \alpha + \beta E + \epsilon$ , will be less than the value of  $I$  on line  $I = \alpha + \beta E$ . Likewise, if we believed that  $\epsilon$  was systematically positive, a line lying below the majority of data points would be appropriate. Regression analysis assumes, however, that the noise term has no such systematic property, but is on average equal to zero—I will make the assumptions about the noise term more precise in a moment. The assumption that the noise term is usually zero suggests an estimate of the line that lies roughly amid the data, some observations below and some observations above.

But there are many such lines, and it remains to pick one line in particular. Regression analysis does so by embracing a criterion that relates to the estimated noise term or “error” for each observation. To be precise, define the “estimated error” for each observation as the vertical distance between the value of  $I$  along with the estimated line  $I = \alpha + \beta E$  (generated by plugging the actual value of  $E$  into this equation) and the true value of  $I$  for the same observation. Superimposing a candidate line on the scatter diagram, the estimated errors for each observation may be seen as follows:



With each possible line that might be superimposed upon the data, a different set of estimated errors will result. Regression analysis then chooses among all possible lines by selecting the one for which the sum of the squares of the estimated errors is at a minimum. This is termed the minimum sum of squared errors (minimum SSE) criterion. The intercept of the line chosen by this criterion provides the estimate of  $\alpha$ , and its slope provides the estimate of  $\beta$ .

It is hardly obvious why we should choose our line using the minimum SSE criterion. We can readily imagine other criteria that might be utilized (minimizing the sum of errors in absolute value, for example). One virtue of the SSE criterion is that it is very easy to employ computationally. When one expresses the sum of squared errors mathematically and employs calculus techniques to ascertain the values of  $\alpha$  and  $\beta$  that minimize it, one obtains expressions for  $\alpha$  and  $\beta$  that are easy to evaluate with a computer using only the observed values of  $E$  and  $I$  in the data sample. But computational convenience is not the only virtue of the minimum SSE criterion—it also has some attractive statistical properties under plausible assumptions about the noise term. These properties will be discussed in a moment after we introduce the concept of multiple regression.

## Multiple Regression

Earnings are affected by a variety of factors in addition to years of schooling, factors that were aggregated into the noise term in the simple regression model above. “Multiple regression” is a technique that allows additional factors to enter the analysis separately so that the effect of each can be estimated. It is valuable for quantifying the impact of various simultaneous influences upon a single dependent variable. Further, because of omitted variables bias with simple regression, multiple regression is often essential even when the investigator is only interested in the effects of one of the independent variables.

For purposes of illustration, consider the introduction into the earnings analysis of a second independent variable called “experience.” Holding constant the level of education, we would expect someone who has been working for a longer time to earn more. Let  $X$  denote years of experience in the labor force and, as in the case of education, we will assume that it has a linear effect upon earnings that is stable across individuals. The modified model may be written:

$$I = \alpha + \beta E + \gamma X + \varepsilon$$

where  $\gamma$  is expected to be positive.

The task of estimating the parameters  $\alpha$ ,  $\beta$ , and  $\gamma$  is conceptually identical to the earlier task of estimating only  $\alpha$  and  $\beta$ . The difference is that we can no longer think of regression as choosing a line in a two-dimensional diagram—with two explanatory variables we need three dimensions, and instead of estimating a line, we are estimating a plane. Multiple regression analysis will select a plane so that the sum of squared errors—the error here is the vertical distance between the actual value of  $I$  and the estimated plane—is at a minimum. The intercept of that plane with the  $I$ -axis (where  $E$  and  $X$  are zero) implies the constant term  $\alpha$ , its slope in the education dimension implies the coefficient  $\beta$ , and its slope in the experience dimension implies the coefficient  $\gamma$ .

Multiple regression analysis is capable of dealing with an arbitrarily large number of explanatory variables. Though people cannot visualize in more than three dimensions, mathematics does not. With  $n$  explanatory variables, multiple regression analysis will estimate the equation of a “hyperplane” in  $n$ -space such that the sum of squared errors has been minimized. Its intercept implies the constant term, and its slope in each dimension implies one of the regression coefficients. As in the case of simple regression, the SSE criterion is quite convenient computationally. Formulae for the parameters  $\alpha$ ,  $\beta$ ,  $\gamma$  . . . can be derived readily and evaluated easily on a computer, again using only the observed values of the dependent and independent variables.

The interpretation of the coefficient estimates in multiple regression warrants brief comment. In the model  $I = \alpha + \beta E + \gamma X + \varepsilon$ ,  $\alpha$  captures what an individual earns with no education or experience,  $\beta$  captures the effect on the income of a year of education, and  $\gamma$  captures the effect on the income of a year of experience. To put it slightly differently,  $\beta$  is an estimate of the effect of a year of education on income, holding experience constant. Likewise,  $\gamma$  is the estimated effect of a year of experience on income, holding education constant.